



Competition, subjective feedback, and gender gaps in performance

Anna Lovász^{a,b,*}, Boldmaa Bat-Erdene^c, Ewa Cukrowska-Torzewska^d, Mariann Rigó^e,
 Ágnes Szabó-Morvai^{a,f}

^a Centre for Economic and Regional Studies, Tóth Kálmán u. 4. Budapest 1097, Hungary

^b University of Washington Tacoma, 1900 Commerce Street, Tacoma, WA 98402-3100, USA

^c Eötvös Lóránd University, Pázmány Péter sétány 1/a, Budapest 1117, Hungary

^d University of Warsaw, Faculty of Economic Sciences, Długa 44/50, Warsaw 00-241, Poland

^e Institute of Medical Sociology, Medical Faculty and University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf, Moorenstr. 5, Düsseldorf 40225, Germany

^f Economics Department, University of Debrecen, Böszörményi út 132, Debrecen 4032, Hungary

ARTICLE INFO

JEL codes:

I24
 J16
 J24
 M54

Keywords:

Gender Gaps
 Performance
 Competition
 Feedback

ABSTRACT

We use an online game with randomized treatments to study gender differences in the impacts of competition and subjective feedback. 5191 participants were randomly selected into 8 groups: players either saw a Top 10 leaderboard or not (*competition*), and within these, they received no subjective feedback, supportive feedback, rewarding feedback, or "trash talk" (*feedback type*). Seeing a leaderboard increases the persistence (number of games played) of all players, but only increases the performance (score) of male players. When the leaderboard is combined with supportive feedback, the performance of female players increases as well. This points to important heterogeneities by feedback type and individual characteristics and suggests that personalized feedback may be key for decreasing gender gaps, particularly in competitive settings such as STEM fields.

1. Introduction

An increasing strand of literature on gender inequalities highlights the impact of gender differences in psychological traits and preferences on individual educational and career outcomes (Niederle, 2016; Eckel & Grossman, 2008; Croson & Gneezy, 2009). One important factor appears to be that women tend to choose to compete less often (Booth & Nolen, 2012; Gneezy et al., 2009; Healy & Pate, 2011; Niederle & Vesterlund, 2007, 2011; Wozniak et al., 2014) and to perform worse in competitive situations (Cai et al., 2019; Cotton et al., 2013; Gneezy et al., 2003; Gneezy & Rustichini, 2004). These competition-related gender differences have been shown to contribute to gaps in educational (Buser et al., 2014, 2020; Ors et al., 2013; Reuben et al., 2017) and labor market outcomes (Azmat, Calsamiglia, & Iriberry, 2015; Bertrand, 2011; Joensen & Nielsen, 2009). Differences in attitudes towards competition can impact outcomes through key decisions, such as field of study (Buser et al., 2014; Kirkeboen et al., 2016; Osborne et al., 2003), which contribute significantly to the gender gaps in earnings that we still observe today (Bertrand, 2020; Macis, 2017). Competitive attitudes also impact choices within a field or occupation, for example, women may

choose to participate less often in challenging tasks and striving for promotions (Bertrand, 2011; Kauhanen & Napari, 2015), and perform worse in high-stakes competitive settings (Jurajda & Münich, 2011; Ors et al., 2013).

There is much debate about what can be done to mitigate disadvantages due to gender differences in preferences and traits, and thereby decrease gender gaps in educational and career outcomes. One approach, termed "fix institutions," emphasizes the idea that certain institutional elements can be altered to "achieve outcomes that better reflect underlying abilities" (Niederle, 2016). Certain elements of the current institutional design – such as feedback policies, or assessment methods – may favor males due to gender differences in individual traits. For example, competition may hurt the performance of those with lower confidence due to increased stress (Azmat, Calsamiglia, & Iriberry, 2015). Women tend to have lower confidence even given equal abilities, especially in traditionally male tasks (Lloyd et al., 2005; McCarty, 1986), and may therefore suffer a relative disadvantage. Previous evidence shows that altering certain elements of the institutional design can decrease or even eliminate gender gaps in competitive settings: for example, single-sex tournaments (Datta Gupta, Poulsen, & Villeval,

* Corresponding author at: Centre for Economic and Regional Studies, Tóth Kálmán u. 4., Budapest 1097, Hungary.

E-mail address: lovasz.anna@rtk.hu (A. Lovász).

2005; Gneezy et al., 2003), quota-style affirmative policies (Balafoutas & Sutter, 2012; Niederle et al 2013), simple advice regarding the gender gap in willingness to compete and its consequences for earnings (Kessel et al., 2021), feedback about relative performance (Ertac & Szentes, 2011; Wozniak et al., 2014; Wozniak et al., 2016), and performance feedback followed by sequential choice in entering a tournament (Niederle & Yestrumskas, 2008). We use a pre-registered (Lovasz (2022)) large-scale randomized online experiment to test whether the provision of certain types of positive feedback, such as encouragement and praise, can also be a tool used to counteract the disadvantage of women in competitive settings. We test whether their provision can motivate women similarly to men, and lead to their higher performance in a competitive setting.

We analyze data from a sample of 5191 individuals who played a simple online game of visual perception that was advertised on social media. During the game, players received randomized treatment in the form of simple texts and graphics, which appeared as pop-up messages on the screen before, during, and after the game. Treatment was randomized among a total of eight groups, along two dimensions: competitiveness and subjective feedback type. Players either saw a Top 10 leaderboard or did not (*competitive element*), and, within each of these categories, they received either no subjective feedback, supportive feedback, rewarding feedback, or trash talk (*feedback type*). This experimental design allows us to test the impacts of competition and the three feedback types when they are given separately, as well as their joint impacts compared to the control group who saw no leaderboard and received no feedback. We estimate the effects of the treatments on outcome measures capturing persistence (number of games played) and performance (accuracy, mean score, best score). We test two main hypotheses. First, we replicate previous results regarding the relatively unfavorable impact of competition on female players. In our case, competition refers to the presence of social incentives, rather than financial incentives, similarly to some previous studies (Gérxhani, 2020; Schram et al., 2019). Second, we assess how players are impacted by competition when there is no subjective feedback compared to when competition is coupled with subjective feedback. We hypothesize that female players will also increase their performance when they are given positive feedback, since such feedback may lower the pressures faced in more competitive, and therefore, higher stakes environments.

While objective performance feedback has been studied as a potential mitigator of gender gaps in outcomes in the economic literature (Azmat & Iriberry, 2010; Bandiera et al., 2015; Ertac & Szentes, 2011; David Wozniak et al., 2016), subjective feedback has received significant attention in the psychology, educational psychology, and human resource management literature (e.g., Deci & Ryan, 1985; Dweck, 2007; Johnson, 2013; Khan et al., 2014; Locke, 1996; Posner & Kouzes, 1999; Wong, 2015), but has not been evaluated in the economics literature. Praise – a positive valuation of performance or effort – has been studied in its role as a verbal performance incentive, with the finding that its impact can be greater than that of a financial reward (Ariely, 2016b). Differences have been highlighted in the perception and impact of feedback on motivation by confidence and gender (Chang et al., 2012; Healy & Pate, 2011; Wozniak et al., 2016). Our study combines these strands of research and tests the interactions of competition and subjective feedback. In a previous study using the same online game (Lovász et al., 2022), we showed that supportive feedback (encouragement) had a positive impact on female players when no competition (leaderboard) was present. Here, we extend the methodology to test whether such feedback could be especially beneficial in a more competitive environment. We focus on three common types of subjective feedback that have received attention in previous empirical work: supportive feedback (encouragement), rewarding feedback (praise), and "trash talk" (competitive incivility). To our knowledge, there is no previous empirical evidence on the interactions of these feedback types and a competitive environment. Yet in real-life settings, these elements are often present simultaneously and their impacts may be interdependent,

which can affect gender gaps in outcomes.

The results provide evidence of significant heterogeneities by gender, and interdependencies in the impacts of competition and subjective feedback that can contribute to gender gaps in outcomes. While competition increases the persistence of both genders, it only improves the performance of male players. However, when competition is coupled with supportive feedback, the performance of female players increases as well. Praise has a similar, though slightly less significant impact, while trash talk (at least our version of it) counteracts the positive impact of seeing a leaderboard for both genders. The use of an online game in a randomized experiment is a somewhat novel approach. It provides access to a larger and potentially more diverse pool of candidates, and allows for the observation of real-life behavior in a natural context, representing a sort of lab in the field method as discussed in Gneezy and Imas (2017). However, it is in the context of a game with low stakes and a short-term interaction, where no financial incentives are present, and individuals are intrinsically motivated to play. The literature on selection in online (versus lab) experiments and the validity of non-incentivized experiments (versus those with monetary incentives) is still emerging. Some recent studies have shown that tests of individual time and risk preferences and performance on cognitive reflection tests provide similar results whether participants are paid or not (Brañas-Garza, Estepa-Mohedano, et al., 2021; Brañas-Garza, et al., 2019; Brañas-Garza, et al., 2021; Taylor, 2013), and similar selection processes and results for online and lab experiments (Arechar et al., 2018; Dandurand et al., 2008; Jorrat, 2021). However, the use of an online game setting has not been directly compared to other settings, and may impact the relevance of the results, as we discuss in Section 2.4.

Our findings support studies on school effectiveness that call attention to the importance of students' social emotional learning needs (Rutledge et al., 2015) and adds to the literature evaluating the implications for gender equity-minded policies.¹ The results point to the importance of recognizing individual heterogeneities when giving feedback and designing assessment methods, particularly in competitive environments. Rather than suggesting the provision of gender-based feedback, the main implication is that personalized feedback can be an important tool for decreasing gender gaps in educational and workplace outcomes. Although we explore some heterogeneities within genders (Section 3.3), we do not answer the question of what the targeting of feedback should be based on. The impact of feedback likely differs by personality traits we do not observe in this experiment, as well as by tasks, environmental factors, and over time. Our evidence provides support for the importance of efforts made by teachers and managers to tailor communication to the individual needs of their students and workers. We show that the provision of uniform feedback that does not take individual needs into account leads to efficiency losses through the lower performance of certain individuals and can contribute to group-level inequalities.

A further direct application of these findings pertains to the increasingly widespread use of adaptive educational software, which was accelerated by the COVID-19 pandemic. The use of such technologies provides a new opportunity, since they allow for the provision of more frequent personalized feedback compared to traditional in-person supervision, and can free up supervisors' time for more targeted in-depth interactions (Luckin et al., 2016; Perrotta & Selwyn, 2019). Our study thus adds to the growing literature on low-cost behavioral interventions in educational settings (e.g. Aronson et al., 2002; Bettinger et al., 2018, Damgaard & Nielsen, 2018). Our results point to the potential positive impact of such technologies on equity and diversity. Significant research is being carried out in the field of adaptive software development to improve targeting algorithms based on the available highly detailed observable data on learners' characteristics (e.g. Narciss

¹ For example, Solanki & Xu (2018) show that having a female instructor can decrease gender gaps in motivation in STEM fields.

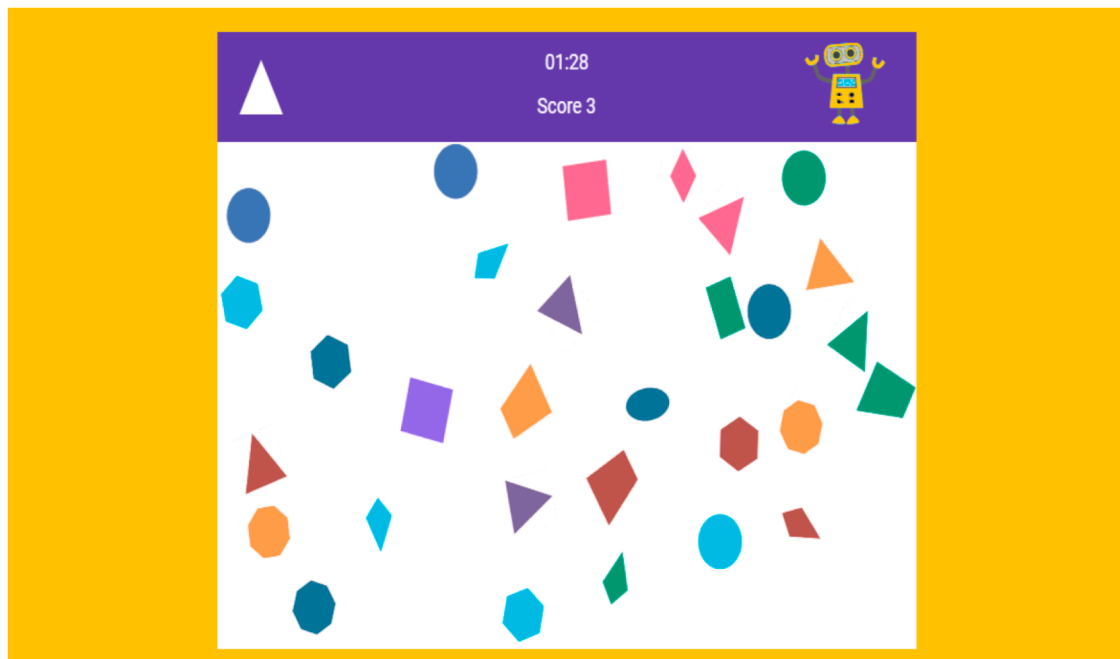


Fig. 1. Screenshot of the Shape Game.

Table 1
Summary of Treatment Groups.

Group	Leaderboard	Feedback
1 (Baseline)	No	None
2	No	Supportive feedback (encouragement)
3	No	Rewarding feedback (praise)
4	No	Trash talk
5	Yes	None
6	Yes	Supportive feedback (encouragement)
7	Yes	Rewarding feedback (praise)
8	Yes	Trash talk

et al., 2014; Young et al., 2021). As these data-driven feedback mechanisms improve, we may see a positive impact on gender equality.

2. Methodology

2.1. Experimental design



























We utilize a simple online game developed for this research, during which players receive randomized treatment. The Shape Game (Fig. 1) is a simple game of visual perception that requires both concentration and effort. The task is to click on a given geometric shape that is displayed in the top left corner of the screen (target shape), from the set of shapes that appear on the screen. The shapes move around the screen, and players must find and click on all shapes that match the target shape shown. The target shape then changes to a new shape. The game takes two minutes, and the goal is to score as many points as possible. Players see the remaining game time and their cumulative score in the upper corners of the screen during the entire game. The Shape Game was developed specifically for our research purposes, and was previously used in an experiment focused on measuring the impact of subjective feedback types (Lovász et al., 2022). We chose this task for several reasons. It is simple, the task is easy to understand, and independent of language skills. It can be entertaining, but requires focus, effort, and skill if a player aims to achieve a high score. It is novel but resembles some popular games in the online market. Overall, our goal was to create a game that allows us to capture real-life behavior in a setting in which individuals would naturally participate.

Players initially land on the game homepage (Fig. A1.a). Individuals are informed of the experimental purpose of the game and the details of data collection (Fig. A1.b). Players are shown a description and a brief demo video of the game (Fig. A1.c). The game is preceded by a simple survey (Fig. A1.d). This asks for basic demographic information: gender, age, country, and level of education. It includes two further questions related to the individual's own experience with games (plays often, sometimes, never), and to their task-related confidence. Players are asked how good they consider themselves to be at computer games: excellent, pretty good, ok, pretty bad, or very bad. Since the question is asked after the game description and demo, the players' responses likely reflect beliefs regarding how well they will play this particular game. The survey was designed to be quick and easy to fill out, asking only for anonymous information similar to what is often requested on gaming sites. Overall, our goal was for players to focus on the game itself, and to observe real-life behavior in a natural game setting. The survey also asks players to give a nickname, which is shown if the player achieves a high score in the Top 10 leaderboard. Data is also collected automatically to account for whether the device the game is played on is a touchscreen or not, as well as screen size, both of which can impact performance.

When players click to start playing the game, they are randomly selected to be in one of the treatment groups, as summarized in Table 1. Treatment is varied along two dimensions: whether players see a Top 10 leaderboard or not (*competitiveness*), and in terms of the subjective feedback they receive (*feedback type*). This gives us a total of 8 groups. Seeing a leaderboard or no leaderboard is interacted with four types of subjective feedback (including a control with no feedback, supportive feedback, praise of performance, and trash talk). This setup allows us to estimate the effects of the leaderboard and each of the feedback types individually, as well as their joint effects. We consider the no leaderboard, no feedback group to be the baseline.

Showing a leaderboard before and after the game provides players with relative performance information, as well as the potential for social recognition. While most of the previous economic literature on gender differences in competitive attitudes has focused on the behavior of participants competing for resources (physical or financial incentives tied to performance rankings), competition also entails a performance and social ranking dimension (Schram et al., 2019). This may also lead to performance differences, and previous experimental evidence

Table 2
Treatment specifications and timing.

Group number	Leaderboard	Subjective Feedback type	Graphic 1	Graphic 2	Graphic 3	Graphic 4	Graphic 5
Timing	Beginning, end		Beginning	2nd shape change	5th shape change	8th shape change	10th shape change
1	No	None					
2	No	Supportive feedback	 GIVE IT A TRY!	 YOU CAN DO IT!	 KEEP IT UP!	 GREAT EFFORT!	 YOU'RE GETTING REALLY GOOD!
3	No	Rewarding feedback		 GREAT START!	 WELL DONE!	 GREAT JOB!	 YOU ROCK!
4	No	Trash talk		 IS THAT THE BEST YOU CAN DO?!	 MY HAWSTER CAN DO BETTER THAN THAT!!	 ARE YOU ASLEEP?!	 IS THAT THE BEST YOU CAN DO?!
5	Yes	None					
6	Yes	Supportive feedback	 GIVE IT A TRY!	 YOU CAN DO IT!	 KEEP IT UP!	 GREAT EFFORT!	 YOU'RE GETTING REALLY GOOD!
7	Yes	Rewarding feedback		 GREAT START!	 WELL DONE!	 GREAT JOB!	 YOU ROCK!
8	Yes	Trash talk		 IS THAT THE BEST YOU CAN DO?!	 MY HAWSTER CAN DO BETTER THAN THAT!!	 ARE YOU ASLEEP?!	 IS THAT THE BEST YOU CAN DO?!

suggests that women respond more negatively to social rankings compared to men. Women’s performance was shown to decrease when they were told prior to a task that their performance would be ranked upon the completion of the task, while that of men increased (Gérxhani, 2020; Schram et al., 2019). Public performance rankings are typically a main motivator of participation in online games, and as such, can lead to competitive pressures, with participants competing for a scarce resource, social status. We therefore refer to the leaderboard treatment as “competition” throughout our discussions.

Our choice of subjective feedback types was motivated by previous evidence on their impact. The phrases in the supportive feedback treatment referred to expressions of support regarding the player’s expected future performance, i.e., encouragement (“You can do it!”), and acknowledgments of the player’s effort (“Great effort!”), without any reference to their actual performance. Encouragement has been shown to positively impact female students’ participation in competitive settings, such as economics or accounting majors and STEM fields (Bedard et al., 2021; Khan et al., 2014; Unkovic et al., 2016). Based on these previous results, we expect encouragement to have a positive impact on female players, which could counteract the negative relative impacts due to competition.

The phrases in the rewarding feedback treatment included frequently used positive valuations of performance, i.e., praise (“Good job!”). The empirical evidence regarding the impact of praise is mixed.

On the one hand, teachers are often encouraged to use praise as a reinforcer of a desired behavior, as some previous studies found that it enhances motivation and leads to improvement of individuals’ performance (Cameron & Pierce, 1994; Deci & Ryan, 1985; Dev, 1997). Such feedback has been shown to motivate individuals and improve performance in a workplace setting, acting as a verbal reward (Ariely, 2016a). On the other hand, others argue that these estimated effects often rely on methodologically flawed studies (Henderlong & Lepper, 2002). Baumeister et al. (1990) show that the effect can be positive or negative: it improved students’ performance on a pure effort task but decreased their performance in skilled tasks. In terms of gender differences, praise has been shown to have a more negative impact on female students, and to be less beneficial compared to effort-based feedback (Roberts & Nolen-Hoeksema, 1989; Zeldin & Pajares, 2000).

Finally, we included a particular form of negative feedback, a “trash talk” treatment. This treatment included phrases such as “Is that the best you can do?” and “Are you asleep?!” as well as humorous graphics that were meant to capture the spirit of such feedback. The inclusion of this form of feedback was motivated by previous evidence on its impact as a motivator in a workplace human resources context, termed “competitive incivility” (Yip et al., 2017), the widespread use of such incivilities in the gaming industry, as well as the feedback we received from participants during the small-sample testing of the experiment.

Table 2 summarizes the categories and gives details regarding the

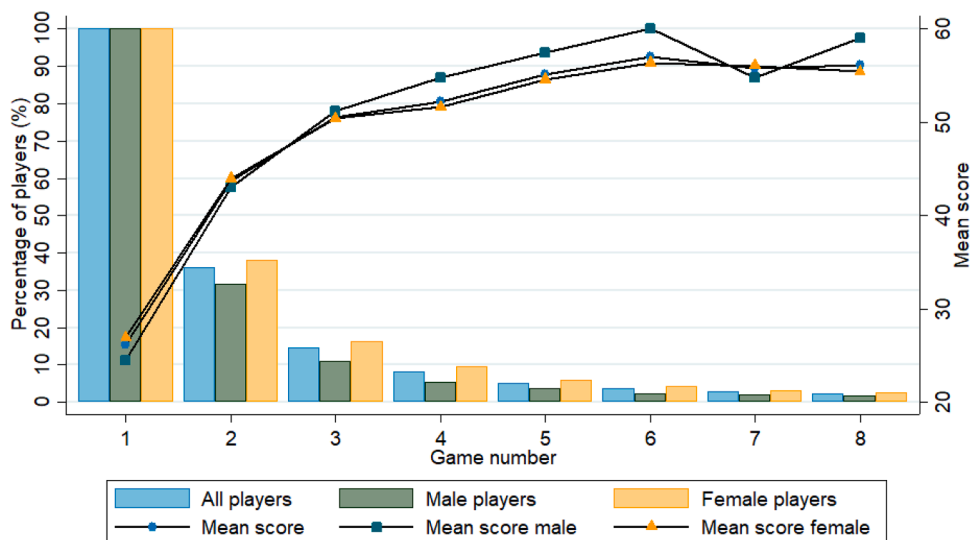


Fig. 2. Number of games played and mean game score.

Table 3
Selection statistics of participants.

	Reached (number who saw ad)	Ad link clicked	Click Rate (link clicks / reached)	Sample (played the game)	Play Rate (sample / link clicks)
Female	1,497,174	57,450	0.038	3,635	0.063
Female (%)	57.1%	63.8%		70.0	
Male	1,122,880	32,558	0.029	1,556	0.048
Male (%)	42.9%	36.2%		30.0	
Female-Male difference (pp)	14.2	27.6		40	
Total	2,620,054	90,008	0.034	5,191	0.058

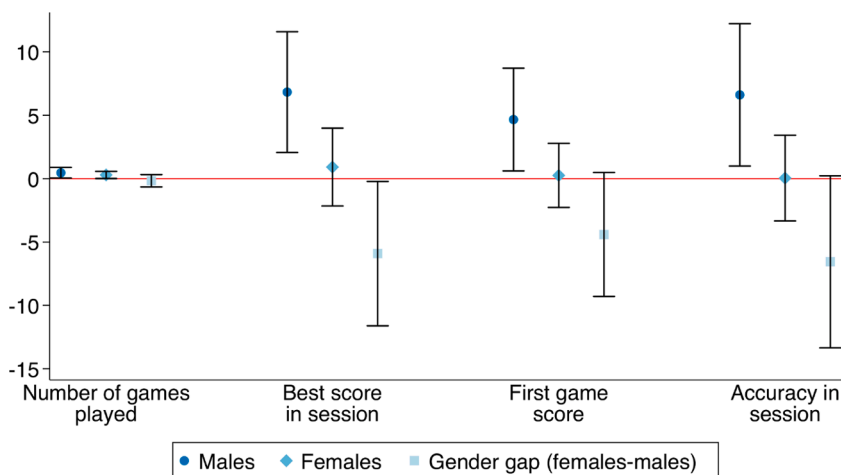


Fig. 3. Treatment Effects of Competition (Leaderboard) with no Subjective Feedback. Notes: Y axis shows number of games played or score. Treatment effects and gaps are calculated based on the linear combinations of coefficient estimates from the OLS estimation of equation (1). 95% confidence intervals are shown for the treatment effect estimates and the gender gaps in the treatment effects. Full OLS results are given in Table A3.a, treatment effect calculations and p-values are summarized in Table A3.b.

specifications and the exact timing and feedback content. The leaderboard was shown at the beginning and at the end of each game, while feedback was given at various stages during the game, as shown in the second row of Table 2. Subjective feedback was given in the form of text and graphics. As our goal was to collect data internationally, we used commonly known English phrases and simple, culturally neutral emoticons and pictures in the treatments.

2.2. Data

The Shape Game is freely available on a website. Participants were recruited using paid social media advertisements, over the course of

targeted at the age group of 18–45-year-olds, from Central and Eastern European countries. Data was collected between April 2019 and February 2020. Participants did not receive any financial compensation, and there were no financial rewards paid to top performers. The resulting data sample is comprised of 5191 individuals, who played a total of 9557 games. It is important to note that different players played a different number of games, and as we discuss later on, this itself may be impacted by the treatments. During a single gaming session – defined as all the games played in a single web browser session – players received the same treatment in every game. This allows us to study slightly longer-run (session long) impacts on persistence and performance. Although some players returned for further sessions, we limited our

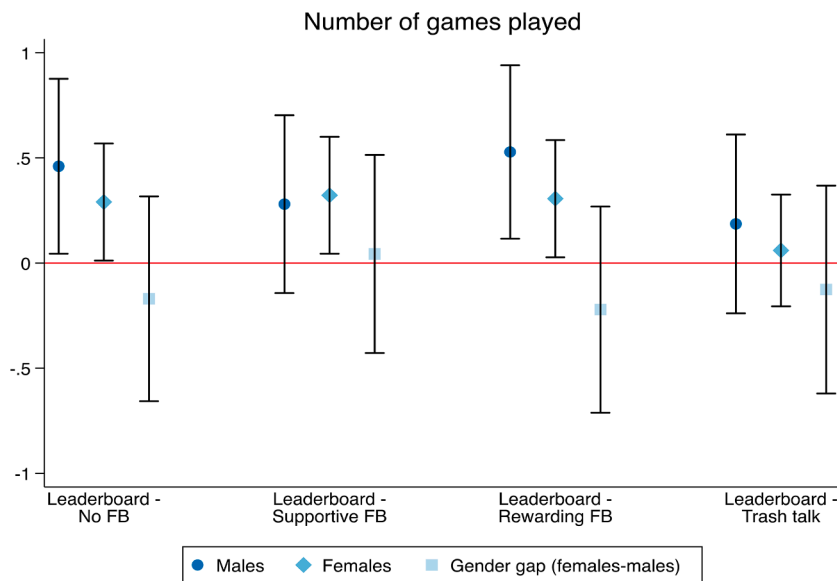


Fig. 4. Treatment effects on persistence by gender. Notes: Persistence is measured as the number of games played by the player in the session. “FB” refers to feedback. Treatment effects and gaps are calculated based on the linear combinations of coefficient estimates from the OLS estimation of equation (1). 95% confidence intervals are shown for the treatment effect estimates and the gender gaps in the treatment effects. Full OLS results are given in Table A3.a, treatment effect calculations and p-values are summarized in Table A3.b.

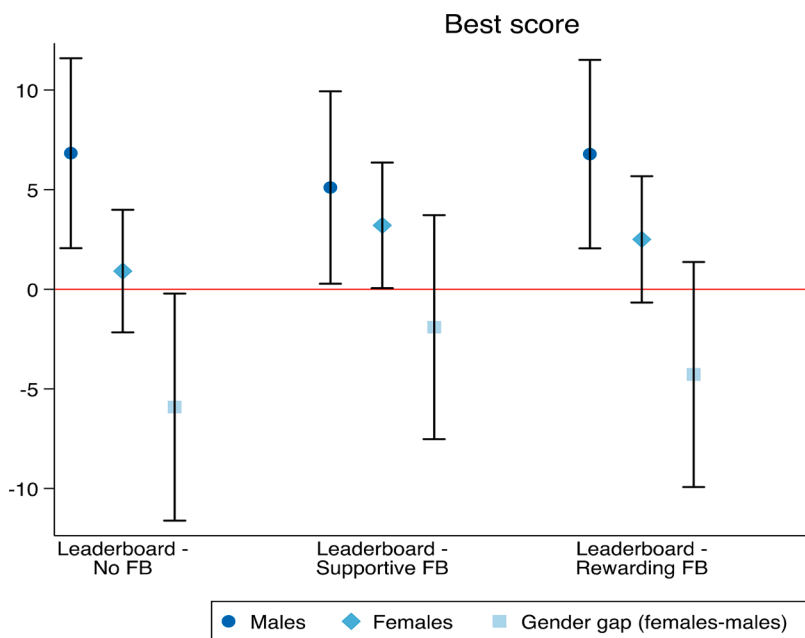


Fig. 5. Treatment effects on performance (best score) by gender. Notes: Performance is measured by the best score achieved by the player in the session. “FB” refers to feedback. Treatment effects and gaps are calculated based on the linear combinations of coefficient estimates from the OLS estimation of equation (1). 95% confidence intervals are shown for the treatment effect estimates and the gender gaps in the treatment effects. Full OLS results are given in Table A3.a, treatment effect calculations and p-values are summarized in Table A3.b.

sample to the first session of each player based on automatically collected tracking token information that identified individual devices, so players in our sample only received one type of treatment.² Players who stopped playing (at any time) before the two-minute time limit of their first game are also included in our sample. Their game score is the score they had achieved at the time they stopped clicking. This means that our analysis does not suffer from a potential drop-out problem often seen in experiments, as we observe all players who start playing the game, and any impacts realized through treated players being more

² We further verified the first session of each player based on their answer to the question “Have you played this game before?” When players return for a new browser session, the survey had to be filled out again and resubmitted. Neither of these could ensure that players were not included who played a second session, however, since they could play on a different device, and respond falsely to the question in the survey.

likely to drop out earlier in the game are also picked up in our estimated treatment effects.

The data was collected at the event level, meaning we observed every click made by players, as well as target shape changes and feedback shown. This information on player behavior (clicks) and performance (score) was linked to the demographic and other individual information given in the survey, and the automatically collected technical data on device type (as well as screen size). The event-level data was aggregated to the game level, and then to the individual player level. We focus our analysis on player level outcomes: the number of games played as a measure of persistence, first game score, mean game score, best game score, as well as player accuracy (score/clicks) as alternative measures of performance. Fig. 2 shows the ratio of players by game number, as well as the mean score by game number. About one third of players chose to play a second game, and the ratio of players decreased sharply in subsequent games. We also see a sharp increase in mean score for those who play further after the first game, which is in line with

Table A7

Within-Gender Heterogeneities by game playing frequency and performance level, OLS regressions by gender.

	Women			Men			Women			Men		
	Game playing frequency						Performance					
	Rarely	Sometimes	Often	Rarely	Sometimes	Often	Low	Medium	High	Low	Medium	High
Leaderboard	2.076 (3.590)	2.483 (2.135)	-4.954 (3.616)	20.03*** (7.240)	4.725 (3.264)	6.492* (3.878)	-0.482 (1.121)	0.566 (1.016)	-1.730 (1.364)	2.186 (1.392)	-0.587 (1.587)	0.542 (3.074)
Supportive FB	6.892* (3.531)	1.822 (2.077)	0.0743 (3.649)	8.422 (6.672)	5.414 (3.326)	0.153 (3.889)	1.051 (1.128)	1.979** (0.950)	-1.042 (1.382)	-0.169 (1.375)	-0.744 (1.536)	-0.522 (3.185)
Rewarding FB	5.997* (3.639)	1.917 (2.124)	2.956 (3.587)	11.81* (6.804)	2.533 (3.453)	3.137 (3.805)	0.617 (1.166)	1.912** (0.958)	-2.628* (1.383)	2.042 (1.365)	-1.135 (1.578)	-2.074 (3.213)
Trash talk FB	1.453 (3.487)	2.118 (2.175)	-1.659 (3.585)	4.318 (7.395)	1.029 (3.227)	0.0457 (4.279)	0.00956 (1.146)	0.900 (0.975)	-0.586 (1.413)	0.375 (1.386)	0.765 (1.559)	-0.623 (3.799)
Leaderboard x supportive FB	-2.623 (4.979)	-1.450 (2.984)	6.180 (5.078)	-13.85 (9.286)	-7.651 (4.743)	0.464 (5.555)	0.0112 (1.608)	-2.536* (1.398)	2.845 (1.874)	-1.430 (2.020)	3.273 (2.217)	0.750 (4.292)
Leaderboard x rewarding FB	-1.327 (5.207)	-2.102 (2.985)	0.172 (5.102)	-18.99* (9.721)	0.214 (4.723)	-4.541 (5.499)	0.264 (1.636)	-2.877** (1.384)	4.937** (1.918)	-3.148 (1.992)	1.235 (2.198)	3.697 (4.269)
Leaderboard x trash talk FB	-1.721 (4.940)	-3.968 (3.055)	2.260 (5.036)	-13.81 (10.07)	-6.166 (4.672)	-8.186 (5.973)	1.152 (1.585)	-2.629* (1.427)	1.532 (1.935)	-3.286 (2.024)	-0.347 (2.247)	-0.523 (5.023)
Observations	743	2,115	777	216	721	618	1,434	1,274	927	721	505	329
R-squared	0.075	0.119	0.173	0.262	0.230	0.258	0.024	0.056	0.053	0.097	0.107	0.086

Notes: Estimates based on OLS regressions run on separate samples by gender and by game playing frequency or performance level (lowest, medium, or top thirds), with interaction terms of leaderboard and feedback type dummy variables. Standard errors are shown in parentheses. Controls included: age, education, region, touchscreen, pixel ratio.

Table A8

Treatment effects by game playing frequency and performance level.

	Game playing frequency						Performance					
	Rarely		Sometimes		Often		Low		Medium		High	
	estimate	p	estimate	p	estimate	p	estimate	p	estimate	p	estimate	p
LB women	1.746	0.626	2.549	0.223	-5.189	0.152	-0.433	0.688	0.583	0.566	-1.808	0.235
LB men	21.321	0.003	5.040	0.146	6.016	0.117	2.042	0.178	-0.587	0.709	0.280	0.909
Gender Gap LB	-19.575	0.016	-2.491	0.538	-11.205	0.034	-2.476	0.184	1.170	0.531	-2.087	0.468
LB+suppFB women	5.921	0.085	2.859	0.176	1.556	0.662	0.585	0.598	-0.016	0.987	0.192	0.897
LB+suppFB men	14.989	0.031	2.069	0.562	6.849	0.088	0.370	0.811	1.870	0.234	0.180	0.944
Gender Gap LB+suppFB	-9.068	0.243	0.790	0.849	-5.294	0.323	0.215	0.910	-1.887	0.309	0.012	0.997
LB+rewFB women	6.550	0.071	2.333	0.262	-1.540	0.672	0.390	0.727	-0.387	0.687	0.566	0.714
LB+rewFB men	11.191	0.119	7.670	0.022	5.239	0.185	0.923	0.545	-0.597	0.693	1.182	0.630
Gender Gap LB+rewFB	-4.640	0.564	-5.337	0.175	-6.779	0.207	-0.533	0.778	0.210	0.907	-0.616	0.832
LB+trFB women	1.380	0.685	0.640	0.764	-3.789	0.289	0.756	0.478	-1.104	0.277	-0.759	0.624
LB+trFB men	11.704	0.108	-0.891	0.800	-1.694	0.684	-1.041	0.498	-0.338	0.831	-0.392	0.888
Gender Gap LB+trFB	-10.324	0.199	1.531	0.710	-2.095	0.703	1.797	0.338	-0.766	0.684	-0.366	0.908

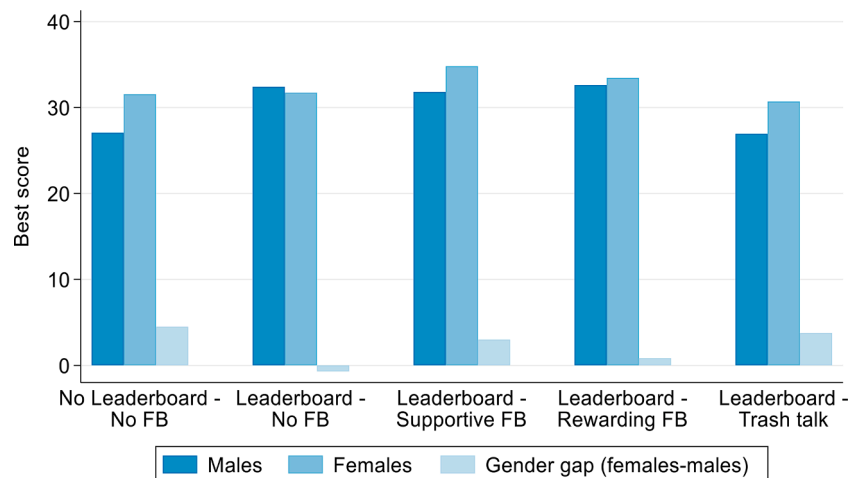


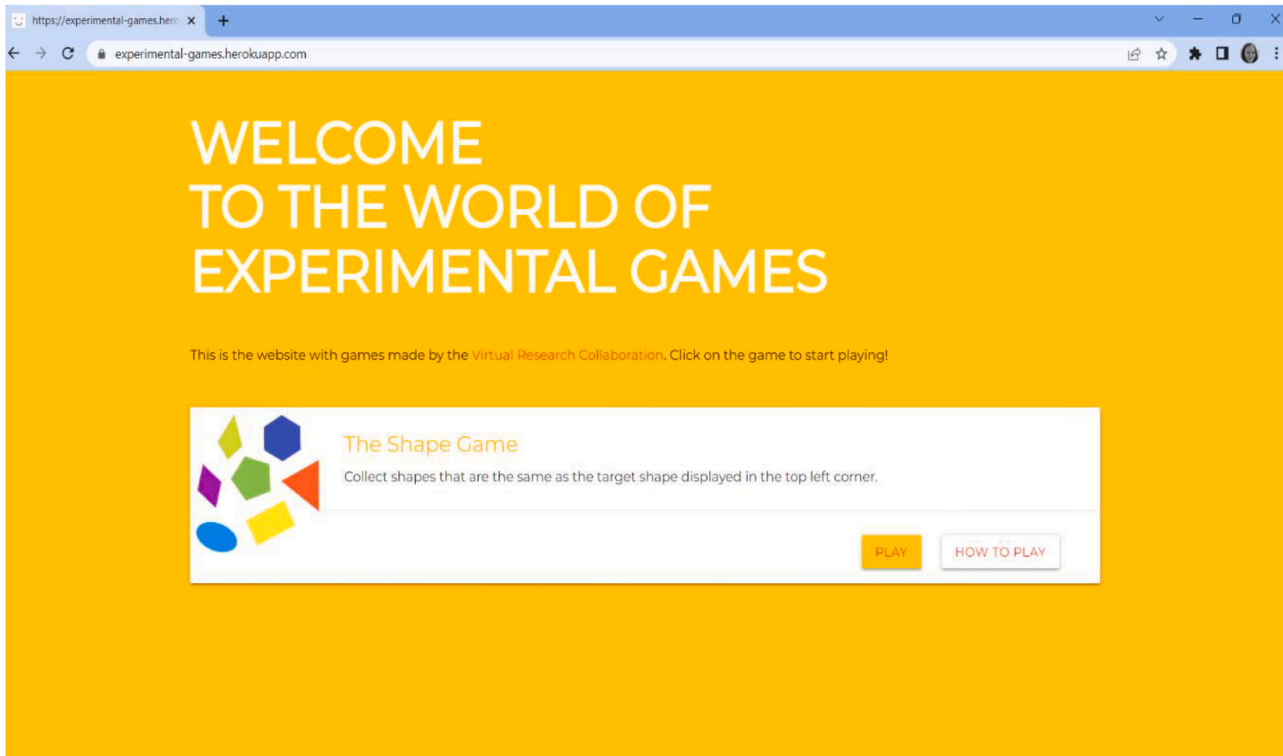
Fig. 6. Performance levels (best score) by gender and treatment group. Notes: The figure shows mean scores by gender and treatment group, and the gender gap in mean scores. "FB" refers to feedback.

learning.

The descriptive statistics of the estimation sample are presented in Table A1. These show that the sample is generally skewed towards younger individuals with higher education. About 55 percent of players are aged 24 or below, and about 78-80 percent have secondary or higher

level of education. The sample consists of only around 27 percent frequent game players, and the majority of players consider themselves to be "okay" at playing games (55 percent), only 7 percent consider themselves to be excellent. Table A1 also shows that our sample is skewed towards female players, who comprise about two thirds of the

a. Game website



b. Terms and Conditions page

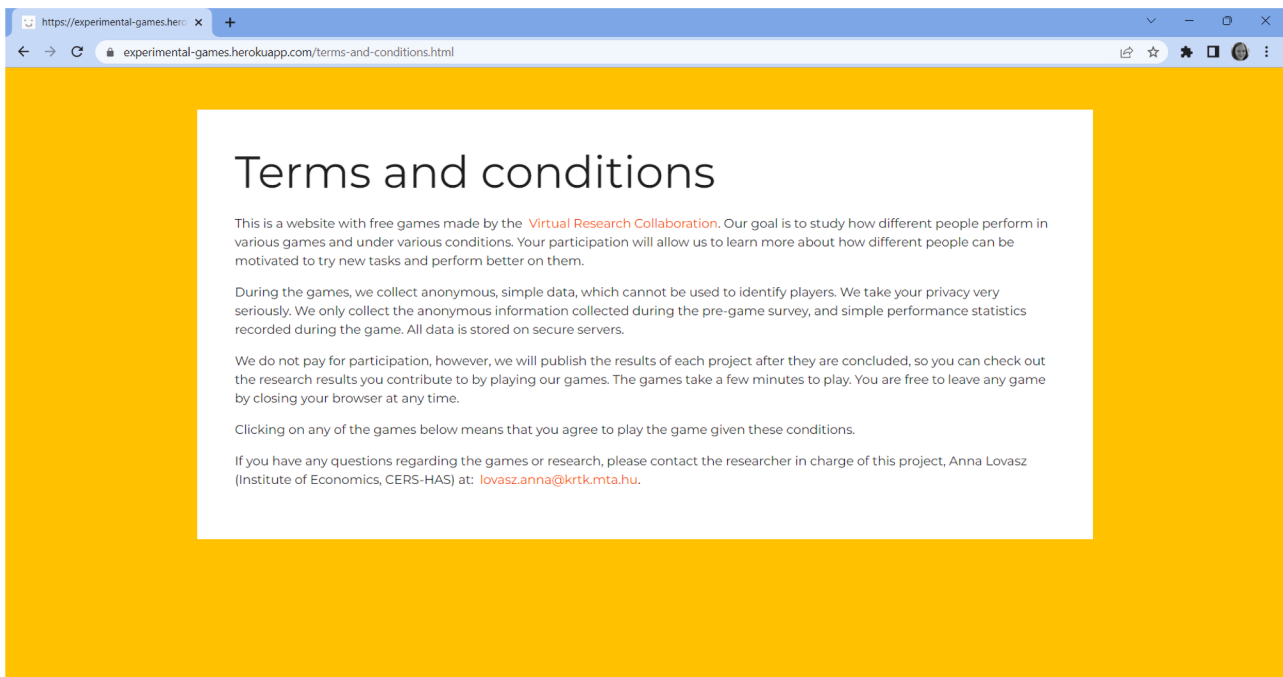


Fig. A1. Screenshots of The Shape Game. a. Game website, b. Terms and Conditions page, c. Instructions page, d. Pre-game Survey.

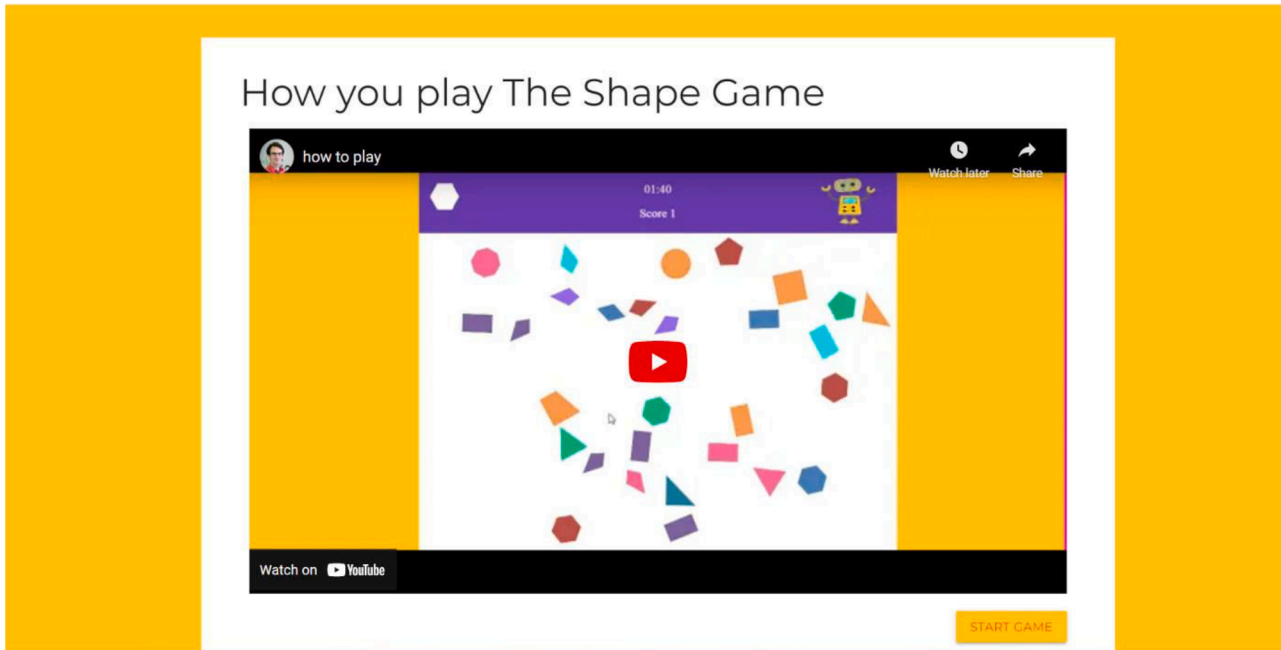
observations. We discuss this and sample representativeness in general in Section 2.4. Appendix Table A2 shows the descriptive statistics of the sample by gender, which show that randomization was well balanced within genders. As shown in Fig. A2, in our sample, male players tend to play computer games more often than female players, and they tend to have higher confidence in their game-playing ability. As we will see

later, the lower confidence of female players is not in line with their performance, suggesting that they tend to undervalue their ability.

2.3. Empirical analysis

The empirical analysis relies on ordinary least squares (OLS)

c. Instructions page



d. Pre-game Survey

https://experimental-games.herokuapp.com/questionnaire.html

Please answer the following questions!

Whats your name / nick name?

Gender

Please select

How often do you play computer games?

Please select

Have you played this game before?

Please select

How old are you?

Where are you from?

Are you good at playing computer games?

Please select

What is the highest level of education you have completed or are pursuing?

Please select

By clicking SUBMIT you agree on [terms and conditions](#) of the game

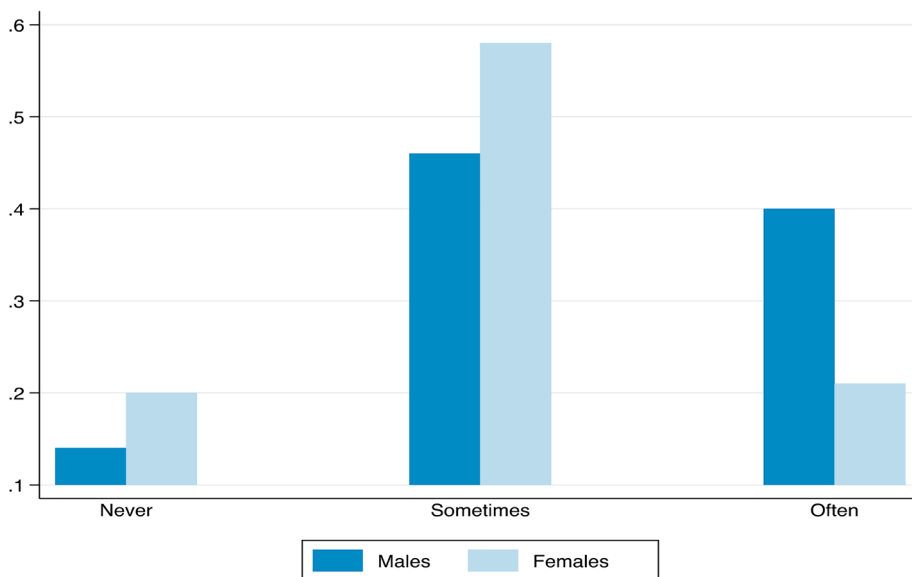
Fig. A1. (continued).

regressions for the estimation of the various treatment effects. We estimate the impact of each treatment compared to the control group, who saw no leaderboard and did not receive any subjective feedback. The estimated equations also control for the observable characteristics in our data seen in Table A1: the age, country, and education level of the individual, whether they are playing on a touchscreen device, and their screen size. Controlling for these characteristics should only impact estimates if the sample size is not large enough to guarantee the randomness among groups in terms of individual characteristics, or if there is some problem with the randomization. The results shown do not

differ significantly from treatment effects estimated without the control variables, supporting the validity of our randomization method (Table A6).

We estimate the effects of the various treatments on two main player level outcomes: the number of games the player played in the session, or players' persistence, and the best score they achieved in the session, or the players' performance. We also estimate impacts on further outcomes to learn more about the mechanisms: the score of players in the first game in order to see the immediate short-run impacts, and players' accuracy to see the role this played in shaping their scores. We use the

a. How often do you play computer games?



b. How good are you at playing computer games?

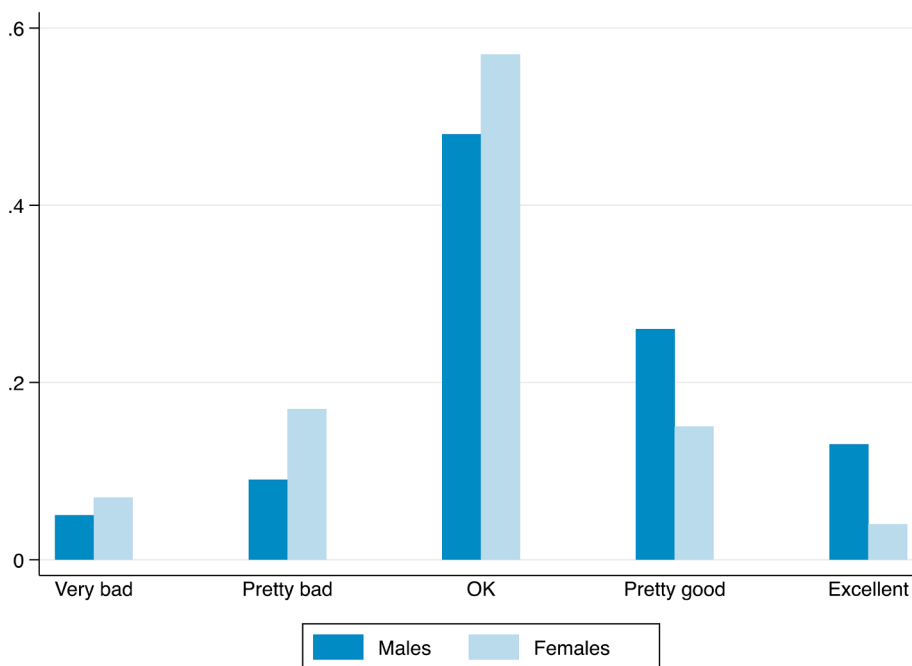


Fig. A2. Frequency of game playing and self-reported confidence by gender. a. How often do you play computer games? How good are you at playing computer games?.

pooled sample of all treatment groups in our estimation. We include dummy variables indicating whether a player saw a leaderboard, whether they received one of the subjective feedback types, the player’s gender, and the interactions of these as explanatory variables. Including the interaction terms makes it possible to estimate the combined effects of competition and feedback and the differential impacts by gender. The estimated regressions are of the form:

$$\begin{aligned}
 Y_i = & \alpha_0 + \alpha_1 LB_i + \alpha_2 Fem_i + \sum_{k=1}^3 \beta_k FB_i^k + \alpha'_3 X_i + \\
 & \pi LB_i \times Fem_i + \sum_{k=1}^3 \gamma_k FB_i^k \times LB_i + \sum_{k=1}^3 \delta_k FB_i^k \times Fem_i +
 \end{aligned} \tag{1}$$

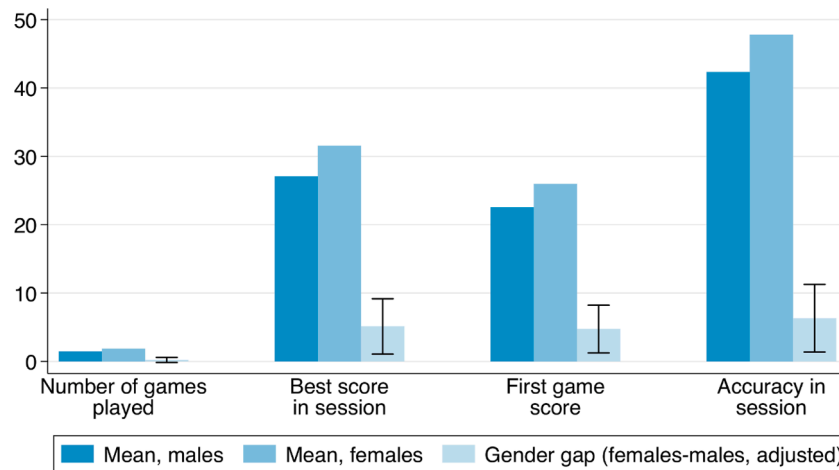


Fig. A3. Mean outcomes by gender, baseline group. Notes: 95% confidence intervals are shown for the gender gaps. Detailed information on the coefficients can be found in Tables A3 and A4.

$$\sum_{k=1}^3 \varphi_k FB_i^k \times Fem_i \times LB_i + \delta_i$$

where Y_i represents the various player-session level outcome variables for individual i , X_i is a matrix of control variables, such as age group, education level, region, touchscreen and screen size, LB_i represents seeing a top ten leaderboard, FB_i^k represents subjective feedback type k , and Fem_i represents the gender of individual i . The coefficient estimates are used to calculate the different treatment effects (as linear combinations of the relevant coefficients, see Table A3.b. for specifications), gender differences in the treatment effects, and the relevant p-values.

Our focus is on the coefficient π , which measures the gender difference in the impact of competition, and the linear combination of coefficients $\pi + \delta_k + \varphi_k$, which captures the gender difference in the combined impact of competition and subjective feedback type k . Based on our first hypothesis, we expect $\pi < 0$, as female players are expected to respond less positively to competition compared to male players. Based on our second hypothesis, we expect $\pi + \delta_k + \varphi_k = 0$ in the case of supportive or rewarding feedback, if the addition of positive feedback counteracts this relative disadvantage and leads to the higher performance of female players as well.

We performed several checks to assess the robustness of our estimates, these are summarized in Appendix Table A6. We analyzed OLS equations without controls (column 1), and with additional controls for self-reported game-playing frequency (column 9). We estimated the impacts of competition and the feedback types on subsamples of two or four groups, rather than pooling all groups (columns 4 and 5). We estimated the effects on subsamples by gender (columns 2 and 3). The estimates had lower significance levels due to the smaller sample sizes but were also comparable to the main results. We checked the sensitivity of the results to certain sample restrictions: dropping players who clicked less than 3 (or 5) times in the game (columns 6 and 7), and dropping outlier observations with extremely high scores, as two players achieved scores above 150 (column 8). We estimated the impact of treatments using the game level dataset, on game level performance (game end score). We looked at performance in the first game each player played, as well as based on the sample of pooled games (Appendix Table A4). The latter specification is complicated by the fact that different players play different number of games, which is a potential channel through which treatment impacts performance through learning. In these specifications, we calculated standard errors clustered at the player level. The game level analysis yielded similar overall results, with higher significance in the pooled game specifications due to the large sample size. Finally, we address concerns regarding multiple testing using the Bayesian multilevel modeling method proposed by

Gelman et al. (2012). The results (Table A5 and Fig. A6) support the main conclusion that for female players, the treatment combining the leaderboard and supportive feedback is the most advantageous compared to the baseline of no leaderboard and no feedback, while the treatment with leaderboard and no feedback is less advantageous.

2.4. External relevance

Given the novel nature of the online game setting, it is important to discuss the external relevance of our estimates. One question that arises is that of selection. Compared to lab experiments - which often recruit university students as participants - larger samples can be achieved at lower cost, and the target population is potentially more diverse (Anwyl-Irvine et al., 2020; Dandurand et al., 2008). Recently, several studies have provided evidence of the generalizability of the results of lab experiments, for example, by comparing measured preferences of participants to those in the population (Cleave et al., 2013), or by using various recruitment and incentive schemes (Abeler & Nosenzo, 2015; Brañas-Garza, Estepa-Mohedano, et al., 2021; Brañas-Garza, Jorrat, et al., 2021). We have no reason to believe that selection is a bigger problem in our online setting. In fact, participation in our game has very low costs compared to participating in a lab, and the task is one that many people take part in in everyday life. Although selection issues of lab experiments have been studied more extensively, available comparisons of outcomes using lab and online samples support the reliability of online experiments, particularly in the case of simple and relatively short experiments such as ours (Arechar et al., 2018; Dandurand et al., 2008; Jorrat, 2021).

As shown in Table A1, about 70 percent of our sample is female, however, this is not necessarily indicative of a selection issue. Table 3 shows some available statistics on the selection of participants into our sample by gender. We can see that more women saw the ad, about 57 percent of those reached were female. Out of the target population who saw the ad, about 3.4 percent clicked on the link to the game website, and about 5.8 percent of these individuals ended up playing the game. Women were somewhat more likely both to click on the link in the ad (3.8 vs 2.9 percent), and to play the game once they clicked on the link (6.3 vs 4.8 percent). This suggests that the gender difference in our sample is due to both the initial ad targeting and differences in participation rates. Historically, more males played online games, although the ratio of female players has steadily increased (Leonhardt & Overå, 2021), and the types of games available have also adapted, from traditionally male-oriented themes (cars, monsters, weapons, explosions, etc.) to increasingly diverse themes. Given our game's simplicity and theme, it may be less attractive to frequent game players, who make up a

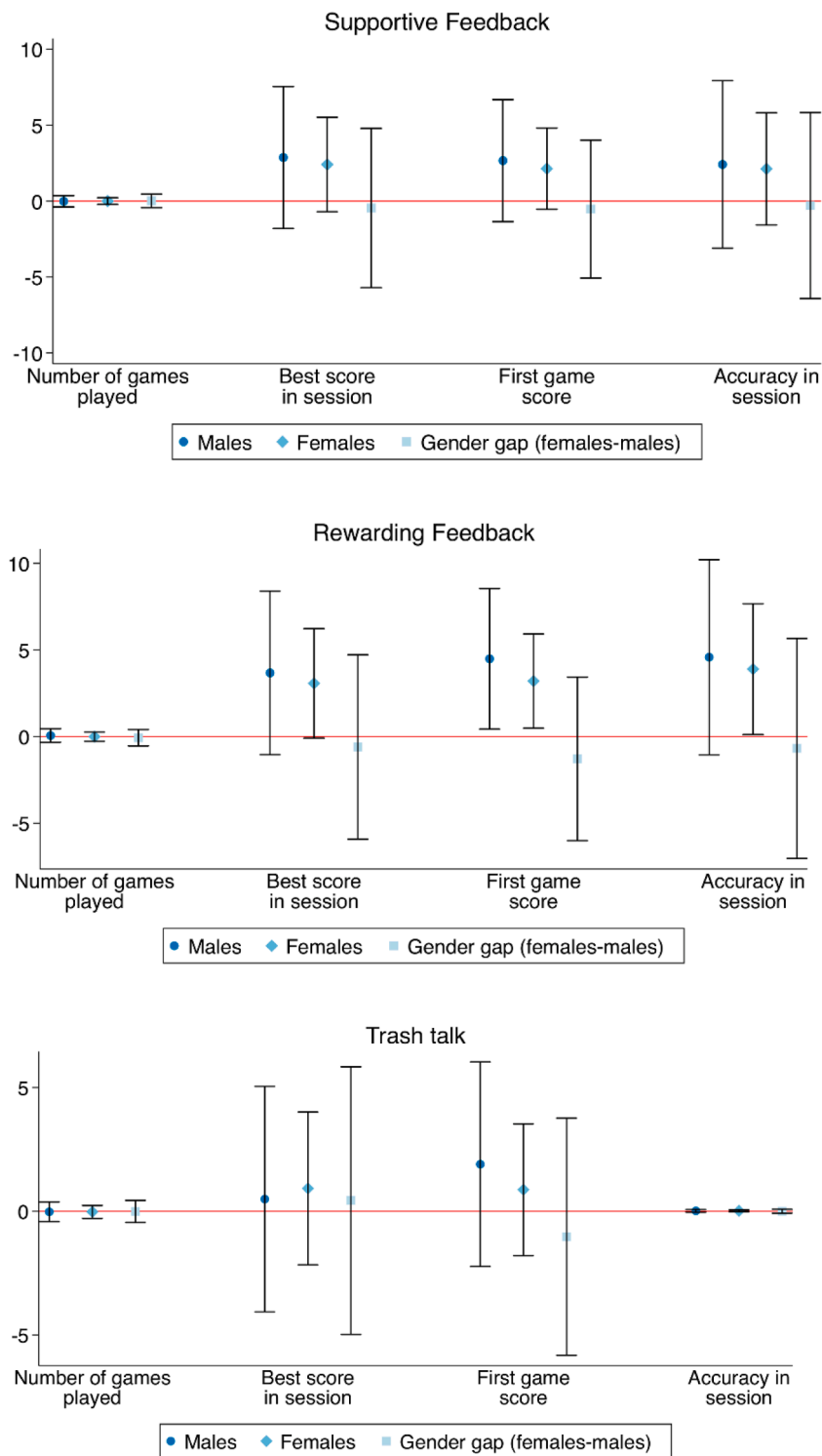


Fig. A4. . Treatment Effects of the Three Subjective Feedback Types. Notes: 95% confidence intervals are shown for the treatment effect estimates and the gender gaps in the treatment effects. Detailed information on the coefficients can be found in Table A3.

higher proportion of men. We expect that more serious gamers would have a higher preference for competition – as this is typical in more sophisticated games – and less preference for friendly, positive feedback. However, we do not have information that could help us confirm this.

The second issue relates to the behavior of participants. The use of an online game represents a sort of lab in the field method – as discussed in Gneezy and Imas (2017)- in the sense that it allows us to maintain experimental control while observing real-life behavior in a natural

setting. The anonymous online setting does not lead to behavioral biases such as experimenter demand effects, which pose a problem in lab settings. Our results likely reflect real-life behavior more closely. On the other hand, it is important to note that this is not a work-related or educational setting. The task may be considered entertaining rather than a tedious task, and there are no monetary incentives present, only an intrinsic motivation to play. As noted earlier, there is some evidence emerging that behavior may be similar whether participants are paid or

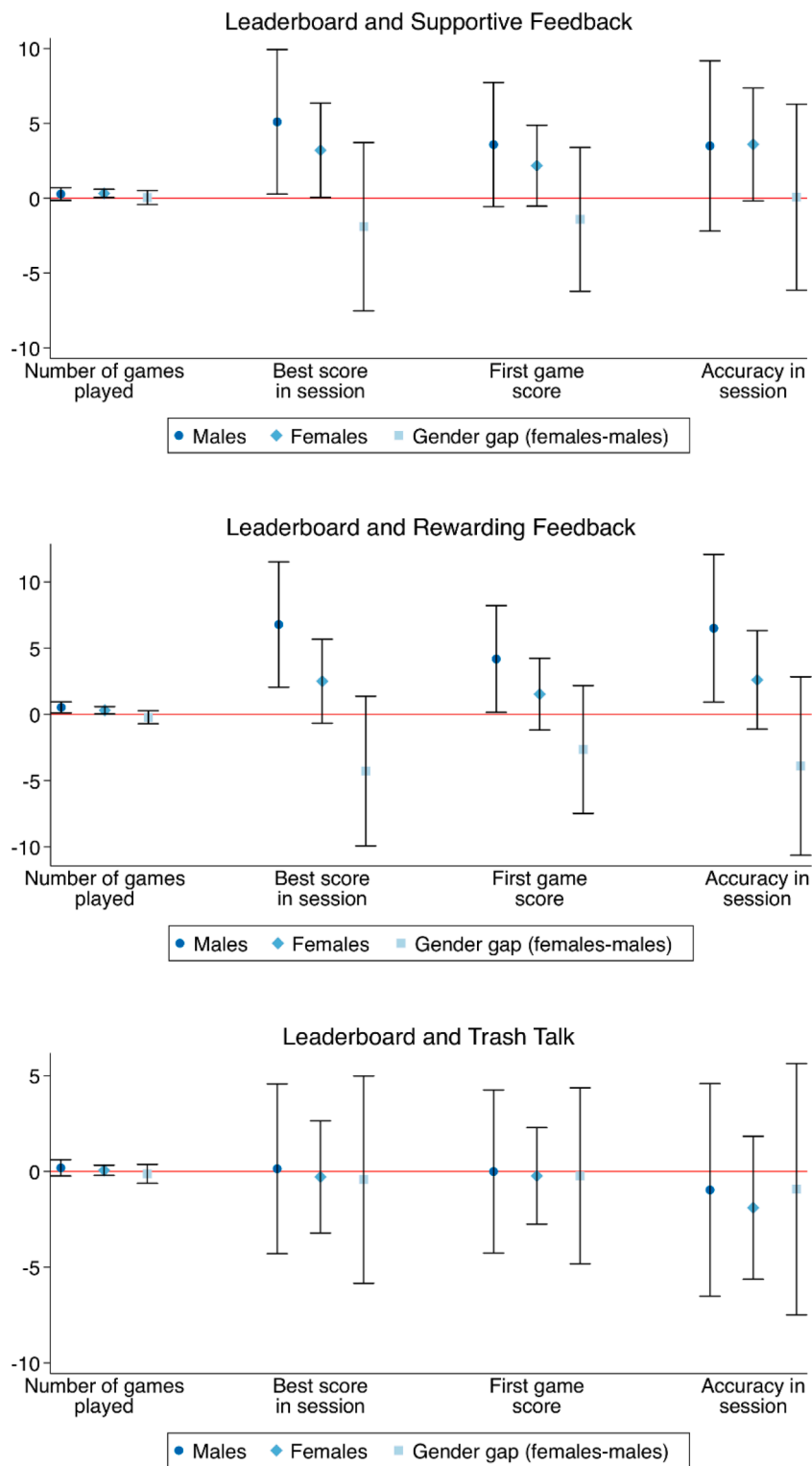


Fig. A5. Combined treatment effects of a leaderboard and subjective feedback. Notes: 95% confidence intervals are shown for the treatment effect estimates and the gender gaps in the treatment effects. “FB” refers to feedback. Detailed information on the coefficients can be found in Table A3.

not (Brañas-Garza, Estepa-Mohedano, et al., 2021; Brañas-Garza et al., 2019, 2021), and in online and lab experiments (Arechar et al., 2018; Dandurand et al., 2008; Jorrat, 2021). Some previous studies have confirmed the validity of hypothetical choices in real-life settings (Taylor, 2013), especially under low-incentive conditions (Holt & Laury, 2002). Similarly, behavior we observe during a game may be indicative of behavior in a school or workplace context. However, there

are no previous empirical results directly comparing experimental results from such a game-based to a workplace-based task environment that we could rely on to support the relevance of our results. Relying on the psychology literature, we can argue that the results reflect key lifelong character traits related to individual behavior when facing a new task. Attitudes towards new tasks and challenges are key determinants of educational achievement (Henderson & Dweck, 1991;

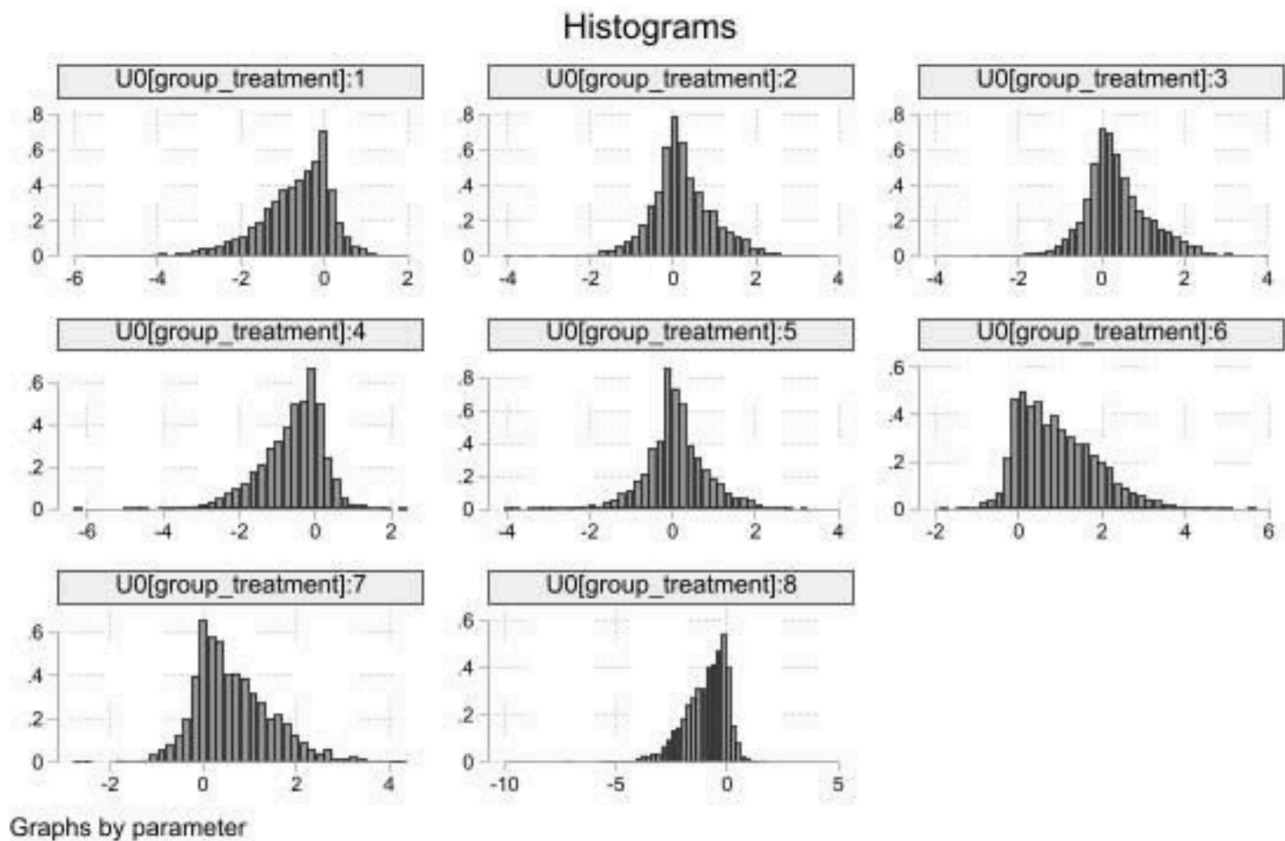


Fig. A6. Bayesian multilevel model results: posterior distributions of gender gap by treatment type. Notes: Bayesian multilevel model with Gibbs sampling, 12,500 MCMC iterations, the burn-in is set to 2500. We use default priors. The dependent variable is best session score, and the right-hand side variable is gender dummy, which is 1 if female. The overall average female coefficient is 3.2, as reported in Table A5. The histograms report the difference from the average in case of the eight treatment types, where Treatment 1: no LB no FB (control group); Treatment 2: no LB SuppFB; Treatment 3: no LB RewFB; Treatment 4: no LB TrashFB; Treatment 5: with LB no FB; Treatment 6: with LB SuppFB; Treatment 7: with LB RewFB; Treatment 8: with LB TrashFB. Treatment 6 is relatively the most advantageous for female players. The coefficients in the treatment groups 1 and 6 are significantly different at the 5% level (Table A5).

Hong et al., 1999), and impact gender differences in choices and outcomes (Dweck, 2006; Lloyd et al., 2005). Furthermore, our findings have direct relevance for some intrinsically motivated settings that are important for labor market success, such as everyday human capital investment or self-improvement activities.

Several further aspects of the experimental design are key to the interpretation of the results. First, we observe behavior in the short run, based on short-term interactions. The results may not generalize to longer-term interactions. For example, it is possible that in the longer run, the supportive feedback given becomes routine, loses its credibility, and becomes less effective. On the other hand, it could be that in the longer run, such feedback may help build a stronger relationship with the supervisor, and thus become more important and effective. Second, the source of the subjective feedback in the game is clearly pre-programmed, not a real-life supervisor. Individual reactions could differ in the case of more personal feedback received from a real-life supervisor, as the stakes would be higher, although some studies have shown that responses are not necessarily sensitive to the perceived source of the feedback (Lipnevich & Smith, 2009). Third, as in the case of any experiment, the results may be task specific. Online gaming in general, and visual perception in particular, are often considered to be stereotypically male tasks, which may exacerbate gender differences in confidence and competitive attitudes. Finally, our results are specific to the particular subjective feedback content (phrases and graphics). However, the goal of this study is not to provide specific suggestions for content, but rather to highlight the heterogeneity in its impact, and the importance of this heterogeneity as a potential contributor to gender inequalities.

3. Results

Fig. A3 shows that in the baseline group who did not see a leaderboard or receive subjective feedback, female players have similar persistence (number of games played), but better performance (score, accuracy). The performance advantage of female players is likely related to selection into our sample, as discussed earlier. If a higher ratio of males in the population are serious (high ability) gamers, and serious gamers are less likely to play such a game, this could lead to the higher mean scores among female players observed in our sample. It is also possible that the male players in our sample are simply not motivated to play as well when a leaderboard is not shown, while female players are more intrinsically motivated. As there is no previous evidence suggesting that males have lower ability in this type of task (Shaqiri et al., 2016), the gender gap observed in favor of female players is likely reflective of one of these two mechanisms.

3.1. The impact of competition

We first examine our first hypothesis regarding the impact of competition (a leaderboard) when there is no subjective feedback given. The estimated treatment effects are shown in Fig. 3 and are based on the comparison of the competition only treatment (leaderboard, no feedback) and control (no leaderboard, no feedback) groups. The results suggest that both genders increase the number of games they play - their persistence - when they see a leaderboard at the beginning and end of each game, by 0.3 games for female players, and 0.46 for male players (with p-values of 0.04 and 0.03, respectively). Though the increase is

Table A1
Descriptive statistics of the sample

	Total	1. No LB + no FB (control)	2. No LB + supportive FB	3. No LB + rewarding FB	4. No LB + trash talk FB	5. LB + no FB	6. LB + supportive FB	7. LB + rewarding FB	8. LB + trash talk FB
N (individuals)	5191	644	688	655	620	652	644	659	629
N (games)	9557	1110	1167	1114	1037	1310	1312	1372	1135
Number of games played	1.84	1.72	1.70	1.70	1.67	2.01	2.04	2.08	1.80
Female	0.70	0.69	0.70	0.69	0.71	0.69	0.71	0.68	0.72
Age									
>17	0.25	0.24	0.26	0.26	0.26	0.27	0.22	0.26	0.25
18-24	0.30	0.30	0.29	0.30	0.30	0.34	0.30	0.27	0.30
25-34	0.18	0.18	0.17	0.18	0.18	0.15	0.21	0.19	0.18
35-44	0.13	0.15	0.14	0.13	0.12	0.12	0.14	0.14	0.14
45-64	0.11	0.10	0.12	0.11	0.10	0.11	0.11	0.11	0.11
<65	0.02	0.02	0.02	0.03	0.04	0.02	0.02	0.03	0.02
Education									
Elementary	0.18	0.17	0.18	0.17	0.21	0.17	0.18	0.17	0.20
Secondary	0.36	0.32	0.36	0.36	0.33	0.41	0.34	0.36	0.37
College or university	0.46	0.51	0.46	0.46	0.46	0.42	0.48	0.47	0.43
Plays games often									
Never	0.18	0.18	0.19	0.18	0.19	0.17	0.21	0.16	0.21
Sometimes	0.55	0.55	0.56	0.53	0.55	0.55	0.52	0.58	0.53
Often	0.27	0.27	0.26	0.29	0.25	0.28	0.27	0.26	0.26
Confidence in game playing									
Very bad	0.07	0.06	0.08	0.07	0.08	0.06	0.08	0.05	0.06
Pretty bad	0.14	0.15	0.15	0.12	0.16	0.16	0.14	0.14	0.15
Ok	0.54	0.55	0.51	0.56	0.53	0.52	0.55	0.54	0.56
Pretty good	0.18	0.17	0.21	0.17	0.17	0.19	0.17	0.19	0.17
Excellent	0.07	0.07	0.06	0.08	0.06	0.07	0.06	0.08	0.07
Region									
Hungary	0.46	0.46	0.44	0.46	0.49	0.47	0.45	0.47	0.49
North America	0.06	0.06	0.06	0.06	0.07	0.06	0.05	0.05	0.05
Other	0.24	0.24	0.26	0.25	0.23	0.21	0.24	0.24	0.23
Poland, Czech Republic, Slovakia	0.18	0.18	0.17	0.17	0.17	0.18	0.19	0.18	0.17
Western Europe	0.07	0.07	0.07	0.06	0.05	0.08	0.07	0.07	0.06
Touchscreen	0.69	0.68	0.67	0.71	0.70	0.71	0.68	0.71	0.68

higher for male players, there is no significant difference in the impact by gender. This means that in terms of persistence, we do not find evidence confirming the hypothesis that female players suffer a relative disadvantage.

The estimated effect for best score, on the other hand, indicate clear gender differences in the response to competition. Male players' performance increases in the session, their best scores improve by 6.8 points. This is a significant magnitude: compared to the baseline mean score of around 26, it represents a positive impact of around 26 percent. Their performance improves already in the first game score and is supported by their increased accuracy. Female players' performance, on the other hand, does not increase as a result of being shown a leaderboard. Even though they play more games, their scores and accuracy do not improve as a result of seeing a leaderboard. The gender gap in the impact of competition on performance is significant. These results confirm our first hypothesis ($\pi < 0$ based on Eq. 1), and are in line with previous evidence pointing to the relative disadvantage of women in competitive settings, in particular, the evidence regarding the impact of competitive pressures due to performance rankings and related social status (Gérxhani, 2020; Schram et al., 2019). Overall, we see that even though both male and female players are motivated to play more games, for the latter accuracy and performance do not improve as a result.

3.2. The impact of competition and subjective feedback

Appendix Fig. A4 shows that in the absence of a leaderboard, supportive and rewarding feedback have a positive impact on the performance (scores and accuracy) of both male and female players, though the impact is only significant in the case of rewarding feedback on

female players and the first game score of male players. The trash talk treatment provides a useful contrast, showing no impact on accuracy and smaller, insignificant impacts on performance. We next turn our attention to our main question and the second hypothesis. Given the relative disadvantage of female compared to male players in the competitive setting, can better outcomes for female players be achieved if competition is paired with some form of positive feedback? Appendix Fig. A5 shows the combined impacts of seeing a leaderboard and receiving subjective feedback on the full set of outcome measures, separately for the three subjective feedback types. In Figs. 4 and 5 we focus on the two main outcome measures representing persistence and performance and compare the effects of the various combined treatments to the treatment with only a leaderboard.

In Fig. 4, we can see that seeing a leaderboard generally has a positive impact on persistence. For male players, the positive impact is the highest and most significant when rewarding feedback is given in addition to the leaderboard, however, this does not differ significantly from the impacts of supportive feedback and leaderboard or no feedback and leaderboard. For female players, the impact is somewhat smaller in magnitude but significant and stable across the two positive feedback types. Adding trash talk to the leaderboard, however, decreases the beneficial impact on persistence for both genders. Overall, the gender gaps in the treatment effects are not significant for any treatment.

Fig. 5 illustrates the treatment effects on performance (best score) separately for male and female players. We focus on this performance measure as it represents a session-level performance measure and is the typical self-set goal of players of such games. It is important to note that the estimated impact of treatment on the players' best scores includes any impact that occurs through increased persistence and learning.

Table A2
Descriptive statistics of the sample by gender.

	Total	Treatment Group							
		1. No LB, no FB	2. No LB, supp. FB	3. No LB, rew. FB	4. No LB, trash FB	5. LB, no FB	6. LB, supp. FB	7. LB, rew. FB	8. LB, trash FB
N (individuals)	5191	644	688	655	620	652	644	659	629
N (games)	9557	1110	1167	1114	1037	1310	1312	1372	1135
Female	0.70	0.69	0.70	0.69	0.71	0.69	0.71	0.68	0.72
Males									
N (individuals)	1556	201	206	201	178	202	185	210	173
N (games)	2544	293	296	300	251	380	322	415	287
Number of games/player	1.635	1.46	1.44	1.49	1.41	1.88	1.74	1.98	1.66
Age									
>17	0.31	0.29	0.30	0.33	0.31	0.35	0.28	0.32	0.31
18-24	0.30	0.35	0.31	0.30	0.28	0.31	0.34	0.27	0.27
25-34	0.16	0.15	0.17	0.12	0.16	0.13	0.16	0.17	0.17
35-44	0.10	0.08	0.11	0.12	0.09	0.08	0.11	0.09	0.10
45-64	0.09	0.07	0.09	0.08	0.08	0.10	0.08	0.10	0.12
Education									
Elementary	0.20	0.18	0.23	0.18	0.26	0.19	0.20	0.17	0.19
Secondary	0.36	0.32	0.32	0.38	0.31	0.43	0.36	0.39	0.41
College or university	0.43	0.50	0.45	0.44	0.43	0.38	0.44	0.44	0.40
Plays games									
Never	0.14	0.10	0.17	0.17	0.12	0.12	0.16	0.12	0.14
Sometimes	0.46	0.49	0.41	0.37	0.54	0.45	0.44	0.51	0.50
Often	0.40	0.41	0.41	0.46	0.33	0.43	0.39	0.36	0.36
Confidence									
Very bad	0.05	0.04	0.07	0.06	0.06	0.04	0.06	0.05	0.02
Pretty bad	0.09	0.08	0.10	0.07	0.11	0.09	0.10	0.05	0.13
Ok	0.47	0.54	0.40	0.44	0.48	0.48	0.51	0.48	0.46
Pretty good	0.25	0.21	0.29	0.26	0.24	0.27	0.22	0.28	0.26
Excellent	0.13	0.12	0.14	0.17	0.11	0.11	0.11	0.14	0.12
Region									
Hungary	0.37	0.39	0.32	0.34	0.39	0.42	0.36	0.38	0.40
North America	0.06	0.06	0.08	0.07	0.08	0.05	0.05	0.06	0.06
Other	0.28	0.25	0.33	0.33	0.29	0.22	0.30	0.29	0.26
Poland, Czech Republic, Slovakia	0.21	0.22	0.19	0.21	0.19	0.24	0.23	0.22	0.21
Western Europe	0.06	0.08	0.07	0.05	0.04	0.07	0.06	0.05	0.07
Touchscreen	0.62	0.61	0.61	0.64	0.66	0.64	0.62	0.62	0.57
Females									
N (individuals)	3635	443	482	454	442	450	459	449	456
N (games)	7013	817	871	814	786	930	990	957	848
Number of games/player	1.929	1.84	1.81	1.79	1.78	2.07	2.16	2.13	1.86
Age									
>17	0.23	0.22	0.24	0.23	0.24	0.23	0.20	0.22	0.22
18-24	0.30	0.28	0.28	0.29	0.31	0.35	0.28	0.27	0.31
25-34	0.19	0.20	0.17	0.21	0.19	0.16	0.23	0.20	0.19
35-44	0.15	0.18	0.16	0.13	0.13	0.13	0.15	0.17	0.15
45-64	0.12	0.12	0.13	0.12	0.11	0.12	0.12	0.12	0.11
<65	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Education									
Elementary	0.17	0.16	0.16	0.17	0.19	0.16	0.17	0.16	0.21
Secondary	0.35	0.33	0.37	0.36	0.34	0.40	0.34	0.35	0.36
College or university	0.47	0.52	0.46	0.47	0.47	0.44	0.49	0.49	0.44
Plays games									
Never	0.20	0.22	0.19	0.18	0.22	0.19	0.22	0.18	0.23
Sometimes	0.58	0.58	0.62	0.60	0.56	0.59	0.56	0.61	0.54
Often	0.21	0.21	0.19	0.22	0.22	0.21	0.22	0.21	0.23
Confidence in game playing									
Very bad	0.07	0.07	0.08	0.07	0.08	0.07	0.09	0.06	0.07
Pretty bad	0.17	0.17	0.17	0.14	0.18	0.19	0.15	0.18	0.15
Ok	0.56	0.55	0.56	0.61	0.55	0.53	0.57	0.56	0.59
Pretty good	0.15	0.16	0.17	0.13	0.14	0.16	0.15	0.15	0.14
Excellent	0.04	0.05	0.02	0.04	0.05	0.05	0.04	0.05	0.05
Region									
Hungary	0.50	0.49	0.49	0.50	0.52	0.50	0.48	0.51	0.52
North America	0.05	0.06	0.05	0.06	0.06	0.06	0.05	0.05	0.04
Other	0.22	0.23	0.23	0.22	0.21	0.21	0.22	0.21	0.22
Poland, Czech Republic, Slovakia	0.16	0.16	0.16	0.15	0.16	0.15	0.17	0.16	0.15
Western Europe	0.07	0.06	0.07	0.06	0.05	0.08	0.07	0.07	0.06
Touchscreen	0.72	0.71	0.70	0.74	0.72	0.75	0.71	0.74	0.73

Table A3
Main estimation results.

a. Full OLS results, pooled groups				
VARIABLES	(1) First game score	(2) Number of games played	(3) Best score in session	(4) Accuracy in session
Female	4.225 (1.772)	0.284 (0.181)	4.89 (2.062)	0.0667 (0.0246)
leaderboard	4.659 (2.068)	0.46 (0.212)	6.826 (2.407)	0.0661 (0.0287)
supportive FB	2.666 (2.059)	-0.0160 (0.211)	2.872 (2.396)	0.0241 (0.0286)
rewarding FB	4.488 (2.071)	0.0603 (0.212)	3.674 (2.410)	0.0458 (0.0287)
trash talk	1.903 (2.137)	-0.0188 (0.219)	0.493 (2.487)	0.0173 (0.0297)
leaderboard x supportive FB	-3.742 (2.948)	-0.165 (0.302)	-4.592 (3.431)	-0.0550 (0.0409)
leaderboard x rewarding FB	-4.972 (2.912)	0.00693 (0.298)	-3.720 (3.389)	-0.0467 (0.0404)
leaderboard x trash talk FB	-6.568 (3.031)	-0.255 (0.310)	-7.182 (3.528)	-0.0931 (0.0421)
female x leaderboard	-4.406 (2.491)	-0.170 (0.255)	-5.916 (2.899)	-0.0656 (0.0346)
female x supportive FB	-0.531 (2.471)	0.0183 (0.253)	-0.461 (2.876)	-0.00291 (0.0343)
female x rewarding FB	-1.289 (2.493)	-0.0688 (0.255)	-0.608 (2.901)	-0.00687 (0.0346)
female x trash talk FB	-1.029 (2.552)	-0.00613 (0.261)	0.431 (2.969)	0.00159 (0.0354)
female x leaderboard x supportive FB	3.529 (3.529)	0.195 (0.361)	4.474 (4.107)	0.0693 (0.0490)
female x leaderboard x rewarding FB	3.043 (3.510)	0.0172 (0.359)	2.245 (4.085)	0.0337 (0.0487)
female x leaderboard x trash talk FB	5.207 (3.610)	0.0503 (0.369)	5.058 (4.201)	0.0547 (0.0501)
Constant	8.468 (3.138)	1.119 (0.321)	11.58 (3.652)	0.216 (0.0436)
Observations	5,190	5,190	5,190	5,190
R-squared	0.143	0.044	0.142	0.273

b. Treatment effects, best score in session							
Treatment effect	Linear combination	Supportive feedback estimate	p	Rewarding feedback estimate	p	Trash talk estimate	P
LB with no FB females	LB + fe_LB	0.911	0.573				
LB with no FB males	LB	6.826	0.005				
gender gap LB with no FB	fe_LB	-5.916	0.041				
FB with no LB females	LB + fe_LB + FB + fe_FB + LB_FB + fe_LB_FB	2.412	0.129	3.066	0.057	0.924	0.569
FB with no LB males	LB + FB + LB_FB	2.872	0.231	3.674	0.127	0.493	0.843
gender gap FB with no LB	fe_LB + fe_FB + fe_LB_FB	-0.461	0.873	-0.608	0.834	0.431	0.885
LB+FB females	FB + fe_FB	3.204	0.046	2.501	0.122	-0.289	0.857
LB+FB males	FB	5.106	0.038	6.780	0.005	0.137	0.956
gender gap LB+FB	fe_FB	-1.902	0.518	-4.279	0.138	-0.427	0.886

Notes: Estimates based on OLS regressions of the form Eq. (1). Standard errors are shown in parentheses. Controls include: age, education, region, touchscreen, pixel ratio.

Notes: LB refers to leaderboard, and FB refers to subjective feedback. Estimated treatment effects are presented, based on OLS estimates of Eq. (1) (shown in Table A3. a., column 3), calculated as linear combinations of coefficients along with p-values. Dependent variable is the player's best score in the session.

Players may play more games due to a treatment and achieve a higher best score because they improve as they play more. Best score therefore measures the overall impact of the treatments on performance in the gaming session, including direct impacts on accuracy, effort within games, as well as longer term impacts through persistence and learning.

The results indicate significant gender differences in treatment effects by gender. Male players generally respond positively to seeing a leaderboard. They achieve higher scores by about 5-7 points in all treatments regardless of whether feedback is given and the type of feedback, except for the treatment combining the leaderboard with trash talk. Adding positive feedback (supportive or rewarding) does not alter the positive impact of competition, however, adding trash talk counteracts it. For female players, as we saw earlier, a leaderboard does not

in itself significantly increase performance. However, the combined impact of a leaderboard and supportive feedback is positive and significant. Compared to the baseline case of no leaderboard or subjective feedback, female players' scores increase by about 3.5 points, an increase of 13.5 percent. The effect when the leaderboard is paired with rewarding feedback is similar, though slightly less statistically significant. The impacts of supportive or rewarding feedback combined with a leaderboard do not differ significantly. Fig. A5 shows that this performance increase among female players from the combined leaderboard and supportive or rewarding feedback treatments can already be seen during the first game played and reflects an increase in female players' accuracy. Similarly to the case of male players, the combined treatment of leaderboard and trash talk has no impact. In terms of gender

Table A4
Game level results: summary of treatment effects.

Treatment effect	Linear combination	First game score		Trash talk estimate	Trash talk p	Game end score, pooled games							
		Supportive FB estimate	p			Supportive FB estimate	p	Rewarding FB estimate	p	Trash talk estimate	p		
LB with no FB females	LB + fe_LB	0.401	0.773			0.337	0.758						
LB with no FB males	LB	4.900	0.018			5.742	0.000						
gender gap LB with no FB	fe_LB	-4.499	0.071			-5.405	0.006						
FB with no LB females	LB + fe_LB + FB + fe_FB + LB_FB + fe_LB_FB	2.197	0.107	0.577	0.609	-0.334	0.772	2.275	0.034	2.606	0.016	0.499	0.650
FB with no LB males	LB + FB + LB_FB	2.789	0.175	4.022	0.030	-0.430	0.828	2.677	0.099	4.006	0.013	0.535	0.751
gender gap FB with no LB	fe_LB + fe_FB + fe_LB_FB	-0.592	0.811	-3.445	0.112	0.096	0.967	-0.402	0.836	-1.400	0.473	-0.036	0.986
LB+FB females	FB + fe_FB	2.199	0.111	1.601	0.138	0.473	0.677	2.364	0.030	1.486	0.171	-0.716	0.512
LB+FB males	FB	3.652	0.084	7.627	0.000	3.062	0.108	4.045	0.015	4.586	0.004	-0.351	0.836
gender gap LB+FB	fe_FB	-1.453	0.565	-6.026	0.003	-2.589	0.243	-1.682	0.398	-3.099	0.110	-0.366	0.856

Notes: LB refers to leaderboard, and FB refers to subjective feedback. Estimated treatment effects are presented, based on OLS estimates of Eq. (1), calculated as linear combinations of coefficients along with p-values. First game score results estimated using game level data from only the first game played by each individual, with the end of game score as the outcome measure. Pooled games results estimated on game level data consisting of all the games played by each individual, with the end of game score as the outcome measure, with estimated standard errors clustered at the player level.

Table A5
Bayesian multilevel model regression results.

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred. interval]
Female	3.204		0.785	3.198	1.648
Constant	29.532		0.828	29.524	27.925
U0[Treatment]					
1	-0.707	0.898	0.048	-0.532	-2.918
2	0.204	0.779	0.029	0.130	-1.356
3	0.380	0.797	0.031	0.261	-1.075
4	-0.607	0.848	0.048	-0.448	-2.597
5	0.074	0.764	0.026	0.036	-1.542
6	0.943	0.963	0.051	0.786	-0.421
7	0.644	0.852	0.043	0.498	-0.735
8	-0.964	0.966	0.063	-0.773	-3.223
Treatment					
U0:sigma2	1.523	2.100	0.117	0.892	0.013
Best score in session					
sigma2	671.789	13.372	0.135	671.456	646.272

Notes: Bayesian multilevel model with Gibbs sampling, 12,500 MCMC iterations, the burn-in is set to 2500. We use default priors. The dependent variable is best session score, and the right-hand side variable is gender dummy, which is 1 if female. Treatment 1: no LB no FB (control group); Treatment 2: no LB SuppFB; Treatment 3: no LB RewFB; Treatment 4: no LB TrashFB; Treatment 5: with LB no FB; Treatment 6: with LB SuppFB; Treatment 7: with LB RewFB; Treatment 8: with LB TrashFB. Full distributions of the estimates shown in Fig. A6.

differences in the impacts, Fig. 5 shows that only the treatment with a leaderboard alone has a significantly different impact by gender, favoring males, while the gender gaps in the other treatment effects are insignificant. This is in line with our second hypothesis, $\pi + \delta_k + \varphi_k = 0$ based on Eq. (1). It shows that when positive feedback is provided in addition to the leaderboard, the performance of female players increases similarly to that of male players.

3.3. Within-gender heterogeneities/heterogeneity analysis

We next consider within-gender heterogeneities in the impacts. We focus on two aspects for which we have data available: game playing frequency (or experience), which players self-reported in the survey (rarely, sometime, or often), and performance level, defined as being in

the lowest, middle, or top third of the distribution of scores in the game. Overall, we find some evidence of within-gender heterogeneities, however, the estimates are generally less significant due to the smaller sample sizes within these subgroups.

Table A7 reports OLS coefficient estimates separately for women and men. The left-hand side panel shows that the mean results by gender are most strongly driven by those who rarely play online games. Among men, we see a positive response to seeing a leaderboard in each category, however, it is particularly large among those who play rarely. Among women, we see insignificant impacts of seeing a leaderboard in all subgroups. Those who play rarely show a significant positive response to supportive and rewarding feedback even in the absence of a leaderboard, and positive combined impacts of seeing a leaderboard and receiving supportive feedback. This is shown directly in Table A8, which

Table A6
Robustness checks.

	1	2	3	4	5	6	7	8	9
	No controls	Men	Women	Sample: groups 1 & 5	Sample: groups 1,2,5,6	Sample: total clicks>3	Sample: total clicks>5	Sample: score<150	Controls: play often
Female	4.488** (0.042)			5.031** (0.016)	5.068** (0.015)	5.530** (0.008)	5.054** (0.017)	4.858** (0.018)	5.134** (0.013)
Leaderboard	5.351** (0.038)	6.898*** (0.003)	0.819 (0.616)	7.279*** (0.003)	7.094*** (0.003)	6.281*** (0.009)	5.257** (0.031)	6.833*** (0.004)	6.898*** (0.004)
Supportive FB	2.547 (0.321)	3.023 (0.195)	2.438 (0.129)		2.949 (0.222)	3.873 (0.114)	3.443 (0.163)	2.881 (0.227)	3.067 (0.201)
Rewarding FB	2.149 (0.406)	3.754 (0.109)	2.945* (0.070)			3.740 (0.126)	3.154 (0.201)	3.674 (0.125)	3.865 (0.109)
Trash talk	-1.183 (0.657)	0.784 (0.746)	0.939 (0.567)			0.237 (0.925)	0.823 (0.749)	0.484 (0.845)	0.532 (0.830)
Leaderboard x supportive FB	-3.152 (0.392)	-4.686 (0.160)	-0.141 (0.951)		-4.943 (0.153)	-2.635 (0.453)	-1.273 (0.718)	-5.355 (0.117)	-4.692 (0.171)
Leaderboard x rewarding FB	-1.946 (0.592)	-3.809 (0.247)	-1.420 (0.538)			-3.992 (0.243)	-2.561 (0.455)	-4.355 (0.196)	-3.919 (0.248)
Leaderboard x trash talk FB	-4.297 (0.256)	-7.702** (0.025)	-2.104 (0.361)			-6.602* (0.065)	-6.687* (0.063)	-7.166** (0.041)	-7.188** (0.042)
Female x leaderboard	-5.178* (0.096)			-6.217** (0.032)	-6.076** (0.038)	-6.470** (0.027)	-4.902* (0.095)	-5.908** (0.040)	-6.042** (0.037)
Female x supportive FB	-0.826 (0.789)				-0.482 (0.868)	-1.251 (0.671)	-0.609 (0.837)	-0.712 (0.871)	-0.712 (0.805)
Female x rewarding FB	0.390 (0.900)					-1.455 (0.621)	-0.884 (0.766)	-0.606 (0.834)	-0.890 (0.759)
Female x trash talk FB	1.826 (0.566)					-0.746 (0.805)	-1.181 (0.699)	0.446 (0.880)	0.402 (0.892)
Female x leaderboard x supp FB	4.506 (0.306)				4.625 (0.264)	3.467 (0.409)	1.824 (0.665)	5.219 (0.201)	4.702 (0.252)
Female x leaderboard x rew FB	1.136 (0.795)					3.847 (0.351)	1.807 (0.662)	2.853 (0.483)	2.496 (0.541)
female x leaderboard x tr FB	2.621 (0.561)					7.064* (0.0970)	6.442 (0.132)	5.022 (0.229)	5.144 (0.221)
Observations	5,191	1,555	3,635	1,296	2,628	4,848	4,722	5,188	5,190
R-squared	0.008	0.228	0.109	0.166	0.157	0.143	0.145	0.141	0.143

Notes: Estimates based on OLS regressions of the form equation (1). P values are shown in parentheses. Controls included: age, education, region, touchscreen, pixel ratio.

summarizes the treatment effects and the gender gaps in these based on pooled gender regressions. The gender gap in the effect of the leaderboard treatment is highly significant among those who play games rarely, and smaller but still significant among those who play often. There is no gender gap in the effects of the treatments that combine the leaderboard with supportive or rewarding feedback among any of the game playing frequency categories.

Table A7 also reveals some heterogeneity in the impacts by performance level (right-hand side panel). The impact of seeing a leaderboard is not significant in any subgroup, though the positive impact for men appears to be the largest among the lowest performers. The impact of positive feedback does show some differences among the three performance levels, which may be linked to its perception. A previous study showed gender differences in the response to unexpected negative feedback: men only attributed such feedback to their ability when it confirmed their prior beliefs, while for women, unexpected negative feedback reduced the likelihood of choosing to enter a tournament (Shastry et al., 2020). In our case, positive feedback could be perceived as not matching the player’s own perception of their performance. We find some evidence of a perceived mismatch lowering the effectiveness of such feedback among female players. Among those in the middle performance category, supportive and rewarding feedback both have a positive impact when there is no leaderboard shown, however, this impact disappears when they see a leaderboard. This suggests that when they can compare their scores to top performers, the positive feedback is no longer perceived as genuine or well deserved, which has a

discouraging effect. Among high performing female players, on the other hand, rewarding feedback has a negative effect when no leaderboard is shown, which turns positive when players see the leaderboard. This may suggest that when high performing female players realize they are among the top performers, they perceive the feedback to be well deserved and therefore respond more positively to it.

3.4. Discussion

To assess the main results and their implications, we summarize the levels of performance by gender for each treatment group in Fig. 6, depicting the mean levels of best scores for each treatment group and the adjusted gender gap in these. We can see the previously documented performance advantage of female players in the baseline group who did not see a leaderboard or receive subjective feedback. This advantage disappears when competition is added, due to the fact that the mean score of male players increases significantly while that of female players does not. The combined treatment of a leaderboard and supportive feedback, on the other hand, increases the mean scores of both male and female players compared to the control group. We again see a gender performance gap in favor of female players as a result. When competition is combined with rewarding feedback, female players’ performance increases less compared to the baseline, therefore, their performance advantage is again closer to zero. The combined treatment with trash talk preserves the baseline performance advantage of female players, however, it decreases mean scores for both genders.

The results imply that there are important heterogeneities in the impacts of competition, subjective feedback types and content, and the combined impacts of these by gender. The magnitudes of the treatment effects and their gender differences are far from negligible. For example, the treatments with the most beneficial impacts by gender can improve scores by 6.5 for male players (leaderboard alone or leaderboard combined with rewarding feedback) and 3.5 for female players (leaderboard combined with supportive feedback), representing increases of about 24 percent and 13.5 percent over the baseline scores, respectively. The best performance outcomes are achieved with different treatment schemes for male and for female players. Male players appear to be sufficiently motivated by competition alone, and this motivation translates to higher persistence and performance. Female players are motivated by competition as well, increasing their persistence. However, they only increase their performance when positive feedback – especially supportive feedback – is provided at the same time. This is due to an increase in female players' accuracy when playing (Fig. A5), which does not occur as a result of adding competition alone. This finding may be suggestive of female players being more at ease in a competitive setting when positive feedback is given at the same time, leading to fewer mistakes made while clicking. The provision of positive subjective feedback can counteract some of the relative disadvantage of female players in competitive settings.

4. Conclusion

In this experiment, we vary the level of competition (no leaderboard or leaderboard shown) and the type of feedback given (none, supportive, rewarding, or trash talk) during an online game. We then assess whether there are any mean differences by gender in players' response (number of games played, score). Overall, the results show that (1) the subjective content of supervisory feedback is a factor that affects individuals' performance during a task, (2) the impacts of competition and subjective feedback elements are interrelated, and (3) there are significant heterogeneities in the impacts of competition and feedback, in particular, we show evidence of mean differences by gender. On average, male players improve their performance significantly when a leaderboard is shown, while for female players, this improvement is realized when positive (supportive or rewarding) feedback is given in addition to the leaderboard being shown. We find some evidence of heterogeneities within genders, in particular, players who have less task-related experience respond most strongly, and positive feedback is the most effective when it is perceived as matching the player's performance.

The main implication of our findings is that personalized feedback, rather than uniform feedback, can achieve higher efficiency and decrease gender inequalities. Although the results we present reflect mean gender differences, this does not suggest that feedback should be targeted by gender. Rather, the results provide evidence of the importance of personalized feedback as a potential tool for decreasing gender gaps, particularly in competitive settings. It is important to note that the beneficial impact of personalized feedback can only be realized if receiving feedback is not a choice, or if men and women are equally receptive to receiving feedback, as suggested by some recent evidence (Coffman and David, 2022). The impact of feedback likely differs by individual characteristics (for example, confidence level, stress resilience, previous experience with the task, etc.), by task (for example, its perception as a typically male or female task), by task environment (for example, competitive or cooperative, high or low stakes, etc.), and over time. Our findings therefore support (1) the importance of personal attention to students' and workers' needs and supervisory communication that is tailored to those needs, and (2) further research on the targeting of feedback based on individual characteristics, including the growing data-driven research on feedback algorithms in educational and HR management software.

Declarations of Competing Interest

None

Data Availability

Data will be made available on request.

Acknowledgements

The project leading to this application received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 691676. The research was supported by grants FK 124658 and 121267-PD of the National Research, Development and Innovation Office of Hungary. The experiment is registered under AEA RCT registry ID #AEARCTR-0003984. We would like to thank Andor Zöldeši for developing the game website, as well as Dániel Horn, János Hubert Kiss, Andrea Kiss, Barbara Pertold-Gebicka, and commenters at conference and seminar presentations for their valuable feedback.

Appendix

References

- Abeler, J., & Nosenzo, D. (2015). Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*, 18(2), 195–214. <https://doi.org/10.1007/s10683-014-9397-9>
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-020-01501-5>
- Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1), 99–131. <https://doi.org/10.1007/s10683-017-9527-2>
- Ariely, D. (2016a). *Payoff: The Hidden Logic That Shapes Our Motivations*. TED Books.
- Ariely, D. (2016b). *Payoff: The Hidden Logic That Shapes Our Motivations*. TED Books. <https://www.amazon.com/Payoff-Hidden-Logic-Shapes-Motivations/dp/1501120042>.
- Aronson, J., Fried, C., & Good, C. (2002). Reducing the effects of stereotype threat on african american college students by shaping theories of intelligence. *Journal of Experimental Social Psychology - J EXP SOC PSYCHOL*, 38, 113–125.
- Azmat, G., & Iriberrí, N. (2010). *The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students* (Working Papers No. 444).
- Balafoutas, L., & Sutter, M. (2012). Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory. *Science*, 335(6068), 579–582. <https://doi.org/10.1126/science.1211180>.
- Bandiera, O., Larcinese, V., & Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics, Elsevier*, 34(C), 13–25.
- Baumeister, R. F., Hutton, D. G., & Cairns, K. J. (1990). Negative effects of praise on skilled performance. *Basic and Applied Social Psychology*, 11(2), 131–148. https://doi.org/10.1207/s15324834baspp1102_2
- Bedard, K., Dodd, J., & Lundberg, S. (2021). Can positive feedback encourage female and minority undergraduates into economics? *AEA Papers and Proceedings*, 111, 128–132. <https://doi.org/10.1126/pandp.20211025>
- Azmat, G., Calsamiglia, C., & Iriberrí, N. (2015). Gender differences in response to big stakes. *Gender Differences in Response to Big Stakes*. Journal of the European Economic Association, Volume 14, Issue 6, 1 December 2016, Pages 1372–1400, <https://doi.org/10.1111/jeea.12180>.
- Bertrand, M. (2011). New perspectives on gender. In D.E. Card & O. Ashenfelter (eds.), *Handbook of labor economics* (1st ed., Vols. 4B, 4, pp. 1543–1590). North-Holland.
- Bertrand, M. (2020). *Gender in the 21st Century*. American Economic Association.
- Bettinger, E., Ludvigsen, S., Rege, M., Solli, I. F., & Yeager, D. (2018). Increasing perseverance in math: Evidence from a field experiment in Norway. *Journal of Economic Behavior & Organization*, 146, 115.
- Booth, A., & Nolen, P. (2012). Choosing to compete: how different are girls and boys? *Journal of Economic Behavior & Organization*, 81(ue 2), 542–555.
- Brañas-Garza, P., Estepa-Mohedano, L., Jorrat, D., Orozco, V., & Rascón-Ramírez, E. (2021). To pay or not to pay: measuring risk preferences in lab and field. *Judgment and Decision Making*, 16(5), 24.
- Brañas-Garza, P., Jorrat, D., Espín, A., & Sánchez, A. (2021). Paid and hypothetical time preferences are the same: lab, field and online evidence. *Working Papers* (No. 54);

- Working Papers). *Red Nacional de Investigadores en Economía (RedNIE)* <https://ideas.repec.org/p/aoz/wpaper/54.html>.
- Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: whom, how, when. *Journal of Behavioral and Experimental Economics*, 82, Article 101455. <https://doi.org/10.1016/j.socec.2019.101455>
- Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3), 1409–1447.
- Buser, T., Niederle, M., & Oosterbeek, H. (2020). Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures. *Tinbergen Institute Discussion Papers*. Tinbergen Institute (No. 20-048/1; Tinbergen Institute Discussion Papers) <https://ideas.repec.org/p/tin/wpaper/20200048.html>.
- Cai, X., Lu, Y., Pan, J., & Zhong, S. (2019). Gender gap under pressure: evidence from China's national college entrance examination. *The Review of Economics and Statistics*, 101(2), 249–263.
- Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: a meta-analysis. *Review of Educational Research*, 64(3), 363–423. <https://doi.org/10.3102/00346543064003363>
- Chang, C., FD, L., Johnson, R. E., Rosen, C., & Tan, J. A. (2012). Core self-evaluations. *Journal of Management*, 38(1), 81–128.
- Cleave, B. L., Nikiforakis, N., & Slonim, R. (2013). Is there selection bias in laboratory experiments? The case of social and risk preferences. *Experimental Economics*, 16(3), 372–382. <https://doi.org/10.1007/s10683-012-9342-8>
- Coffman, K.B. and David K. "Gender and preferences for performance feedback." Working Paper, May 2022.
- Cotton, C., McIntyre, F., & Price, J. (2013). Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior & Organization*, 86 (February), 52–66.
- Crosno, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2), 448–474.
- Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, 64, 313–342.
- Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, 40(2), 428–434. <https://doi.org/10.3758/BRM.40.2.428>
- Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press. <https://doi.org/10.1007/978-1-4899-2271-7>
- Dev, P. C. (1997). Intrinsic motivation and academic achievement: what does their relationship imply for the classroom teacher? *Remedial and Special Education*, 18(1), 12–19. <https://doi.org/10.1177/074193259701800104>
- Dweck, C. S. (2006). Is math a gift? Beliefs that put females at risk. In S. J. Ceci, & W. Williams (Eds.), *Why Aren't More Women in Science? Top Researchers Debate the Evidence* (pp. 47–55). American Psychological Association.
- Dweck, C. S. (2007). The secret to raising smart kids. *Scientific American: Mind*. December/January, 36–43.
- Eckel, C., & Grossman, P. (2008). Men, women and risk aversion: experimental evidence. *Handbook of Experimental Economics Results*, 1.
- Ertac, S., & Szentos, B. (2011). *The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence*. Koc University-TUSIAD Economic Research Forum [Koc University-TUSIAD Economic Research Forum Working Papers].
- Gelman, A., Hill, J., & Yajima, M. (2012). Why we (Usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. <https://doi.org/10.1080/19345747.2011.618213>
- Gërçhani, K. (2020). Status ranking and gender inequality: A cross-country experimental comparison. *Research in Social Stratification and Mobility*, 65, Article 100474. <https://doi.org/10.1016/j.rssm.2020.100474>
- Datta Gupta, N., Poulsen, A., & Villeva, M. C. (2005). Male and female competitive behavior: experimental evidence. GATE Working Paper No. W.P.05-12, <https://doi.org/10.2139/ssrn.906766>.
- Gneezy, U., & Imas, A. (2017). Chapter 10 – lab in the field: measuring preferences in the wild. In A.V. Banerjee & E. Duflo (eds.), *Handbook of Economic Field Experiments* (Vol. 1, pp. 439–464). North-Holland. 10.1016/bs.hefe.2016.08.003.
- Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5), 1637–1664.
- Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3), 1049–1074.
- Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review Papers and Proceedings*, 377–381.
- Healy, A., & Pate, J. (2011). Can teams help to close the gender competition gap? *The Economic Journal*, 121, 1192–1204.
- Henderlong & Lepper. (2002). *The effects of praise on children's intrinsic motivation: A review and synthesis*. - *PsycNET*. Henderlong & Lepper. <https://doi.apa.org/doiLandIng?doi=10.1037%2F0033-2909.128.5.774>.
- Henderson, V., & Dweck, C. S. (1991). Adolescence and Achievement. In S. S. Feldman, & G. R. Elliott (Eds.), *At the Threshold: Adolescent Development* (pp. 197–216). Harvard University Press.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655. <https://doi.org/10.1257/000282802762024700>
- Hong, Y. Y., Chiu, C. Y., Dweck, C. S., Lin, D. M.-S., & Wan, W. (1999). Implicit theories, attributions, and coping: a meaning system approach. *Journal of Personality and Social Psychology*, 77(3), 588–599.
- Joensen, J. S., & Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources*, 44(1), 171–198.
- Johnson, D. A. (2013). A component analysis of the impact of evaluative and objective feedback on performance. *Journal of Organizational Behavior Management*, 33(2), 89–103.
- Jorrat, D. (2021). Recruiting experimental subjects using WhatsApp. *Journal of Behavioral and Experimental Economics*, 90, Article 101644. <https://doi.org/10.1016/j.socec.2020.101644>
- Jurajda, S., & Münich, D. (2011). Gender gap in performance under competitive pressure: admissions to Czech universities. *American Economic Review*, 101(3), 514–518.
- Kauhanen, A., & Napari. (2015). Gender differences in careers. *Annals of Economics and Statistics*, 117/118, 61.
- Kessel, D., Mollerstrom, J., & van Veldhuizen, R. (2021). Can simple advice eliminate the gender gap in willingness to compete? *European Economic Review*, (C), 138. <https://ideas.repec.org/a/eee/eecrev/v138y2021ics0014292121001306.html>.
- Khan, A. A., Hamdan, R. A. R., & Mustaffa, M. S. (2014). Educational encouragement, parenting styles, gender and ethnicity as predictors of academic achievement among special education students. *International Education Studies*, 7(2), 18.
- Kirkeboen, L. J., Leuven, E., & Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3), 1057–1111.
- Leonhardt, M., & Overå, S. (2021). Are there differences in video gaming and use of social media among boys and girls? – A mixed methods approach. *International Journal of Environmental Research and Public Health*, 18(11), 6085. <https://doi.org/10.3390/ijerph18116085>
- Lipnevich, & Smith. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, 15(4), 319–333.
- Lloyd, J. E. V., Walsh, J., & Yailagh, M. S. (2005). Sex differences in performance attributions, self-efficacy, and achievement in mathematics: If I'm so smart, why don't I know it? *Canadian Journal of Education*, 28(3), 384–408.
- Locke, E. A. (1996). Motivation through conscious goal setting. *Applied and Preventive Psychology*, 5(2), 117–124.
- Lovasz, A. (2022). *Supervisory Feedback, Effort, and Performance – A Randomized Experiment Using an Online Game [Data set]*. n.d. American Economic Association. <https://doi.org/10.1257/rct.3984>
- Lovász, A., Cukrowska-Torzewska, E., Rigó, M., & Szabó-Morvai, Á. (2022). Gender differences in the effect of subjective feedback in an online game. *Journal of Behavioral and Experimental Economics*, 98, Article 101854. <https://doi.org/10.1016/j.socec.2022.101854>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.
- Macis, M. (2017). Gender differences in wages and leadership. *IZA World of Labor*.
- McCarty, P. A. (1986). Effects of feedback on the self-confidence of men and women. *Academy of Management Journal*, 29(4), 840–847.
- Narciss, S., Sosnovsky, S., Schnaubert, L., Andres, E., Eichelmann, A., Gogudze, G., & Er, M. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, 71, 56–76.
- Niederle, M. (2016). Gender. In A. E. Roth, & J. H. Kagel (Eds.), *Handbook of Experimental Economics* (2nd ed., pp. 481–553). Princeton University Press.
- Niederle, M., Segal, C., & Vesterlund, L. (2013). How Costly Is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness. *Management Science*, 59(1), 1–16.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122(3), 1067–1101.
- Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annu. Rev. Econ.* 3(1), 601–630.
- Niederle, M., & Yestrumkas, A. H. (2008). *Gender differences in seeking challenges: The role of institutions*. National Bureau of Economic Research (Working Paper No. 13922; Working Paper Series).
- Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics*, 31(3), 443–499.
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079.
- Perrotta, C., & Selwyn, N. (2019). Deep learning goes to school: Toward a relational understanding of AI in education. *Learning, Media and Technology*, 45(3), 251–269.
- Posner, B. Z., & Kouzes, J. M. (1999). *Encouraging the heart: A leader's guide to rewarding and recognizing others*. Wiley.
- Reuben, E., Wiswall, M., & Zafar, B. (2017). Preferences and biases in educational choices and labour market expectations: shrinking the black box of gender. *The Economic Journal*, 127(604), 2153–2186. <https://doi.org/10.1111/eoj.12350>
- Roberts, T.-A., & Nolen-Hoeksema, S. (1989). Sex differences in reactions to evaluative feedback. *Sex Roles*, 21(11–12), 725–747. <https://doi.org/10.1007/BF00289805>
- Rutledge, S. A., Cohen-Vogel, L., Osborne-Lampkin, L., & Roberts, R. L. (2015). Understanding effective high schools: Evidence for personalization for academic and social emotional learning. *American Educational Research Journal*, 52(6), 1060–1092. <https://doi.org/10.3102/0002831215602328>
- Schram, A., Brandts, J., & Gërçhani, K. (2019). Social-status ranking: A hidden channel to gender inequality under competition. *Experimental Economics*, 22(2), 396–418. <https://doi.org/10.1007/s10683-018-9563-6>
- Shaqiri, A., Brand, A., Roinishvili, M., Kunchulia, M., Sierro, G., Willemin, J., Chkonia, E., Iannantuoni, L., Pilz, K., Mohr, C., & Herzog, M. (2016). Gender differences in visual perception. *Journal of Vision* September, 16, 207.
- Shastri, G. K., Shurchkov, O., & Xia, L. L. (2020). Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics*, 88, Article 101592. <https://doi.org/10.1016/j.socec.2020.101592>

- Solanki, S. M., & Xu, D. (2018). Looking beyond academic performance: The influence of instructor gender on student motivation in STEM Fields. *American Educational Research Journal*, 55(4), 801–835. <https://doi.org/10.3102/0002831218759034>
- Taylor, M. P. (2013). Bias and brains: Risk aversion and cognitive ability across real and hypothetical settings. *Journal of Risk and Uncertainty*, 46(3), 299–320. <https://doi.org/10.1007/s11166-013-9166-8>
- Unkovic, C., Sen, M., & Quinn, K. M. (2016). Does encouragement matter in improving gender imbalances in technical fields? Evidence from a randomized controlled trial. *PLOS ONE*, 11(4), Article e0151714. <https://doi.org/10.1371/journal.pone.0151714>
- Wong, Y. J. (2015). The psychology of encouragement: Theory, research, and applications Ψ . *The Counseling Psychologist*, 43(2), 178–216.
- Wozniak, D., Harbaugh, W., & Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, 32(1), 161–198.
- Wozniak, D., Harbaugh, W. T., & Mayr, U. (2016). The effect of feedback on gender differences in competitive choices. *SSRN Scholarly Paper ID 1976073*. Social Science Research Network.
- Yip, J. A., Schweitzer, M. E., & Nurmohamed, S. (2017). Trash-talking: Competitive incivility motivates rivalry, performance, and unethical behavior. *Organizational Behavior and Human Decision Processes*, 144, 125–144.
- Young, K. R., Schaffer, H. E., James, J. B., & Gallardo-Williams, M. T. (2021). Tired of failing students? Improving student learning using detailed and automated individualized feedback in a large introductory science course. *Innovative Higher Education*, 46(2), 133–151. <https://doi.org/10.1007/s10755-020-09527-5>
- Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, 37(1), 215–246. <https://doi.org/10.2307/1163477>