Gender and Performance Under Competitive Pressure: A Meta-Analysis of Experimental Studies

Christopher W. Gardiner[*] Eva Markowsky[†]

January 2025

Abstract

This paper analyses a crucial aspect of the gender gap in competitive behaviour: performance under competitive pressure. We rely on existing experimental evidence to test the prevalent hypothesis that women 'choke' under pressure while men increase their performance in high-pressure environments. To this aim, we conduct a comprehensive quantitative review and synthesis of 70 experimental studies reporting 237 effect sizes that compare gender differences in performance in various real-effort tasks in non-competitive and competitive settings. Summarising across effect sizes, irrespective of competition, results in a gender gap of 4.5 per cent in absolute performance in favour of men. The magnitude of the gap is sensitive to the subject pool, the setting, and the task of the experiment. University students performing mathematical tasks in a computer laboratory exhibit the largest gender difference. Overall, gendered performance differences are qualitatively small, measured by standard instruments to evaluate effect size magnitude. Contrary to the prevalent belief, the gender gap in performance does not increase under competitive pressure.

Keywords: Competitiveness, Performance, Experiments, Gender, Meta-analysis

JEL Codes: J16, D91, C9

[*] Universität Hamburg

[†] Universität Potsdam. Corresponding author. Email: eva.markowsky@posteo.de. This work was supported by the Joachim-Herz-Stiftung's Add-on Fellowship for Interdisciplinary Economics. We thank Florian Baumgartinger and Marie Schäfer for excellent research assistance, and seminar participants at Maastricht University's Interdisciplinary Brownbag Seminar of Behavioral Sciences and the 16th Maastricht Behavioral and Experimental Economics Symposium for helpful comments. All remaining errors are our own.

1 Introduction

In 2018, Forbes Magazine compiled a list of who were considered the 75 most powerful people in the world. The ranking was based on the following criteria: It was determined how many other people an individual has influence over, the amount of financial resources they control, and whether their influence encompasses more than one sphere of society. The resulting list presents a selection of politicians, academics, media moguls, business people, and investors. What every individual on the list has in common is that their careers were accompanied by fierce competition - be it in the form of electoral campaigns, fighting over market shares, taking over rivalling firms, competing for a promotion within an organisation, or striving for recognition as a leading researcher in one's field. Among the 75 named individuals, only five are women (Forbes, 2018). The Forbes list makes no claim to accuracy, but it does illustrate the following two circumstances very clearly: Firstly, societal influence is gained in competitive environments. And secondly, women are vastly underrepresented in the most powerful positions across various spheres of society.

Besides institutional barriers and discrimination, part of the gender gap in leadership positions might be caused by behavioural differences, for example, through self-selection in high- or low-pressure jobs (e.g., Alan et al., 2020). One particularly essential trait in this context is competitiveness. The broad and latent trait competitiveness comprises numerous aspects: risk and feedback aversion, confidence, the willingness to enter competitions, the ability to perform under pressure, and other related traits, like altruism (Shurchkov & Eckel 2018). Together, these behavioural aspects determine how individuals fare in competitive settings. Each of these aspects gives rise to potential gender gaps that might help explain gender differences in behaviour in competitive settings and thus the persistent under-representation of women in influential positions.

Experimental economics and neighbouring sciences like psychology have produced large and fascinating bodies of literature on gender gaps in risk attitudes (see, e.g., the meta-analyses by Nelson 2015 and Filippin & Crosetto 2016), confidence (e.g., Zuckerman et al., 2016; Gentile et al., 2009; Bandiera et al., 2022), and the willingness to enter competition (e.g., Niederle, 2017; Markowsky & Beblo,2022).[1] However, one essential element of competitiveness has, to date, received much less attention: The performance reaction to competitive pressure.

This paper fills this gap by systematically exploring the relationship between gender, performance, and competitive pressure. While many experimental papers on competitiveness report figures relating to this question, the respective answers are highly specific to the experimental

[1] See also Croson & Gneezy (2009), Niederle et al. (2015) and Shurchkov & Eckel (2018) for general reviews on gender differences in economic preferences.

context at hand. By combining evidence from a large number of studies, our paper allows a more thorough conclusion on the question of whether women 'choke' under pressure to a higher degree than men do.

In their review of economic research relating to "Gender Differences in Behavioral Traits and Labor Market Outcomes", Shurchkov & Eckel (2018) state that "women [...] often underperform relative to men in tournaments, especially when under pressure or when the task is stereotyped to be male oriented." (p. 488). At the same time, they point out that women's and men's performances in competitive settings vary with the setting and subject pool (ibid., footnote 20). In fact, the existent literature is ambivalent on whether and under what circumstances there are systematic differences in women's and men's performances in competitive situations. In their influential experiment, Gneezy et al. (2003) find a gender performance gap of 13.4 per cent (albeit insignificant) in favour of men in a maze-solving task, which increases to 28 per cent when introducing competitive pressure. The experiment by Shurchkov (2012) leads to similar results for a math task but finds no gender difference in performance in either non-competitive or tournament settings for a verbal task.[2] Several other studies confirm the importance of task differences and task stereotypicality in moderating gendered performances in competitions (e.g., Günther et al., 2010, Iriberri & Rey-Biel, 2017). However, there are also experiments that produce no gender differences in competitive performances in stereotypical male tasks (e.g., Dreber et al., 2014; Geraldes et al., 2021).

In field experiments and studies using observational data, results are also inconclusive. Ors et al. (2013) demonstrate that women perform worse than men in a highly competitive entrance test for a renowned French Business School, but not in *baccalaureát* exams taken by the same students, which are less competitive because they are not associated with a fixed number of successful candidates (i.e., passing the high school exam is not a tournament). In the study by Jurajda & Münich (2011), men perform better than women in entrance exams to more competitive higher education institutions but not in less prestigious ones. Similarly, Morin (2015) shows that male university students perform better than female ones when inter-student competition increases. On the other hand, Lavy (2013) finds no gender difference in performance in a tournament among Israeli school teachers based on class performance. Finally, Paserman (2023) reports no performance difference (in the number of forced errors) in highly competitive tennis matches, and Bedard & Fischer (2019) find no gender differences in incentivised classroom quizzes.

Our paper extends the current understanding of gender differences in performance reactions to

[2] Additionally, Shurchkov (2012) demonstrates that the gender performance gap in math tournaments disappears when the task is carried out with low time pressure. In the verbal task, women outperform men when time pressure is decreased in the tournament.

competitive pressure by utilising a large body of existing experimental literature. Combining evidence from a large number of studies allows comparisons across multiple relevant dimensions, like tasks, participant groups, and experimental contexts.

In this paper, therefore, we synthesise and systematise the existing evidence on gender differences in competitive performance with meta-analytic tools. The total number of experimental studies on this question that ensure sufficient comparability is limited. However, studies investigating another element of competitiveness, namely the willingness to enter a competition, regularly report gender-specific performance scores under different payment schemes as a side result and are conducted under highly standardised conditions.

Departing from and extending the experiment in Gneezy et al. (2003), Niederle & Vesterlund (2007) conducted a laboratory experiment to elicit preferences regarding willingness to enter competitions, which proved to be highly influential. Since its publication, numerous studies have replicated or refined their study design. In these experiments, participants are asked to perform a real-effort task in three stages. In the first stage, subjects are rewarded solely based on their own performance. This is called the "non-competitive" stage. In the second stage, the tournament, they are rewarded a higher payoff if they perform better than a (group of) randomly selected opponent(s) - but get nothing otherwise. The measure of interest, namely the preference for competition, is elicited in the third stage: Subjects are asked whether they prefer to solve the task in a tournament or by themselves in the final performance round. While performance is not the main focus of this body of experimental literature, many papers report gender-specific performance scores for the non-competitive and competitive stages of the experiment. We use this information to investigate whether women's relative performance suffers under competitive pressure in a meta-analytic setting.

To this end, we combine the results of 70 experimental studies on gender differences in competitive preferences and performance with a total number of about 32,000 participants. The effect size of choice is the gender difference in performance, measured as the percentage difference between women's and men's average performances to ensure comparability across different experimental tasks. First, we calculate a weighted mean gender gap in overall performance and perform a subgroup analysis thereof to investigate whether relative performance varies with tasks, participant groups, types of experiments, or other features of the study design. Second, we test for publication bias to ensure that our meta-analytic results are unbiased by researchers' and editors' tendency to only report results that are statistically significant or verify existing expectations regarding the underlying relationship. Third, we perform a meta-regression analysis to estimate the influence of competitive pressure on women's relative performance.

We find a precision-weighted overall gender difference in performance of 4.5 per cent in favour of men. This is equivalent to a standardised mean difference following Cohen (*Cohen's d*, 1988) of 0.13. This difference is driven by the large proportion of performances (circa 40 per cent) by students solving calculus problems in a laboratory, where gender gaps in performances are most pronounced. The magnitude of the gender gap in performance is low compared to gender gaps in related behavioural traits. Furthermore, we find no evidence of a disadvantageous reaction to competition in terms of performance for women compared to men.

The rest of this paper is structured as follows: First, we present our empirical set-up and the meta-data set. The following sections summarise the findings of our research. We discuss limitations in terms of generalisability and validity in light of effect size dependencies that call for a cautious interpretation of our results and conclude with implications for organisational practices and future research.

2 Data

In this section, we describe the process of constructing the meta-data set. We first exemplify our inclusion criteria and research strategy, then move on to calculating effect sizes and operationalising other variables of interest.

2.1 Data collection

Our literature search is focused on experimental studies on competitiveness, in which participants perform real-effort tasks. We aim to include all studies reporting gender-specific performance values, which allow the calculation of a gender gap in performance. For example, a common measure of performance in a mathematics task is the "number of solved problems within five minutes" (e.g., Niederle & Vesterlund, 2007). We use the study pool of the meta-analysis in tournament entry by Markowsky & Beblo (2022), along with their list of excluded studies, as a starting point for our literature search. The next step involves a search on Google Scholar, as well Web of Science, Web of Knowledge, the Social Science Research Network, IDEAS, and JSTOR using the keywords 'gender', 'performance', 'competition', 'competitive', and 'pressure'. The cutoff for the inclusion of studies was May 2022. Effect sizes are excluded from the data set whenever no gender-specific sample size is mentioned or missing for either gender. During screening, we further identified effect sizes based on samples split by multiple dimensions (e.g., split by age and cultural background). Inclusion of all such effect sizes would constitute "double counting" of observations, leading to biased results. Therefore, after deciding which dimension would be the most informative for the present research, it was taken care that every observation

Table 1: List of included studies

Almas et al. 2016
Alnamlah and Gravert 2020
Apicella et al. 2017
Apicella et al. 2020
Balafoutas and Sutter 2012
Balafoutas and Sutter 2019
Balafoutas et al. 2012
Bedard and Fischer 2019
Berlin and Dargnies 2016
Bjorvatn et al. 2016
Booth and Nolen 2012
Booth and Nolen 2021
Booth et al. 2019
Brandts et al. 2015
Buehren et al. 2016
Burow et al. 2017
Buser 2016
Buser et al. 2014
Buser et al. 2017
Buser et al. 2017a
Buser et al. 2018
Buser et al. 2021
Buser et al. 2021a
Bönte et al 2018
Cassar and Rigdon 2021
Cassar and Zhang 2022
Cahlikova et al. 2020
Charness et al. 2017
Charness et al. 2022
Comeig et al. 2016
Czibor and Martinez 2019
Dasgupta et al. 2019
Fu and Zhong 2019
Geraldes et al. 2021
Gill and Prowse 2014
Grosse et al 2014
Halko and Sääksvuori 2017
Hallady and Landsman 2022
Hauge et al. 2020
He et al. 2021
Healy and Pate 2011
Hoyer et al. 2016
Hässler and Schneider 2020
Ifcher and Zarghamee 2016
Iriberri and Rey-Biel 2017
Ivanova-Stenzel and Kübler 2011
Jung and Vranceanu 2019
Kamas and Preston 2012
Kessel et. al 2021
Klinwoski 2019
Kuhn and Villeval 2013
Kuhnen and Tymula 2012
Lee et al. 2014
Masclet et al. 2015
Mayr et al. 2012
Meier et al. 2017
Müller and Schwieren 2012
Niederle and Vesterlund 2007
Niederle et al. 2013
Price 2012
Price 2020
Reuben et al. 2017
Reuben et al. 2019
Saccardo et al. 2018
Shastry et al. 2020
Shurchkov 2012
Tungodden and Willén 2022
van Veldhuizen 2018
Yagasaki 2022
Zhong et al. 2018

appears only once in the data set.

Another prerequisite for inclusion is the reporting of within-group standard deviations based on gender or the availability of statistics that allow the calculation of missing standard deviations. As proposed by Debray et al. (2018), we calculate missing standard deviations using test statistics from two-sample t-tests, if available. Where significance levels are reported instead of detailed test statistics, we code the upper limit of the reported significance range (e.g., 0.1 for "$0.05 < p < 0.1$"). This procedure is not feasible for differences labelled "non-significant" (Higgins et al., 2022); therefore, such values are dropped from the data set. Non-parametric significance tests (e.g., Mann-Whitney U test) do not allow for a reliable estimation of the standard deviation. Hence, these effect sizes are also excluded. Only performance values from mandatory stages of the experiments are included to rule out selection issues.

Having started with 108 studies and 431 effect sizes, we end up with 70 studies and 237 effect sizes after screening according to these rules. Table 1 shows the list of included studies.

Among the remaining effect sizes, we distinguish between two levels of reliability, depending on whether and how the standard deviation was calculated: We classify as highly reliable any effect sizes based on studies that explicitly report within-gender group standard deviations or that clearly state that gender differences were evaluated based on t-tests. In cases where it is unclear from which test the reported test statistics are taken, we assume the default procedure to

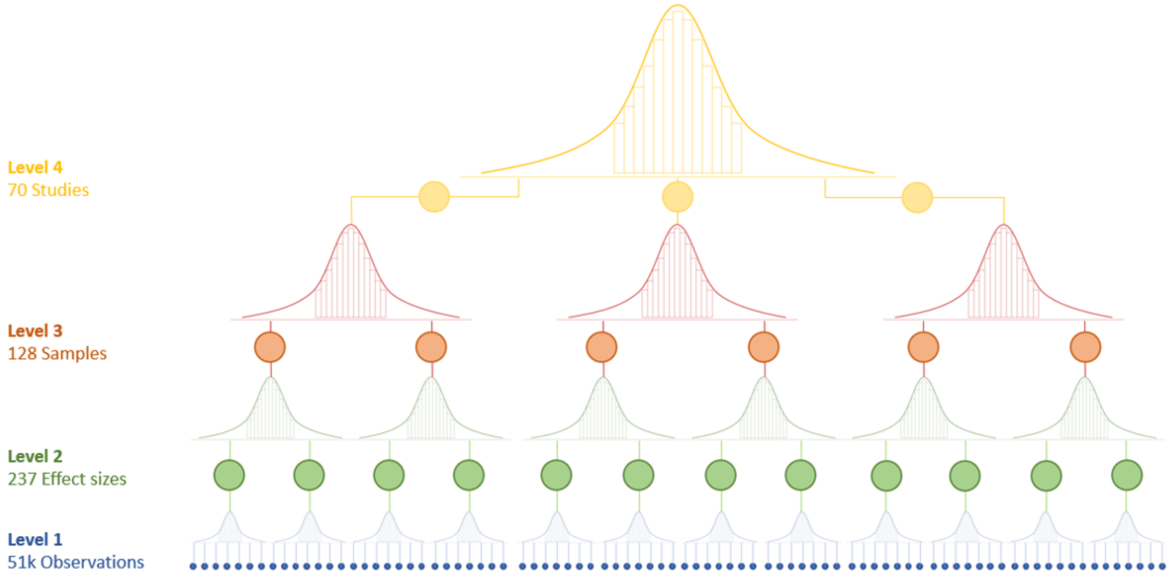

Figure 1: Hierarchical structure of the meta-data (illustration derived from Harrer et al., 2021)

be t-testing but label the affected effect sizes as "less reliable". Standard deviation estimations based on the upper limit of the significance range are also considered "less reliable".[3]

The experimental studies in the meta-data set include varying numbers of samples (e.g., multiple treatment groups), and each sample can produce multiple effect sizes (within-group treatments or different tasks). This causes a multi-levelled structure of effect size dependency. The data set has the following four-level hierarchy, as illustrated in Figure 1: The 237 effect sizes (Level 2) are nested within 128 unique samples (Level 3). These are, in turn, nested within the 70 studies (Level 4). Adding up the unique samples yields a total number of 31,912 participants in the meta-data set, which account for 50,597 observations at the individual level (Level 1). [4]

2.2 Effect size construction

We construct an intuitive measurement of the gender gap in performance as our main outcome, namely the standardised performance difference (SPD). It is calculated by subtracting the average male performance from the average female performance and using the male average as the base value for standardisation (Equation 1).

[3] Any imputation of standard deviations by this method is only an estimation because it assumes that the compared groups have identical standard deviations. However, those imputed standard deviations yield the correct standard error of the group differences (in which lies our main interest), because they are derived from the correct test statistic regarding the group differences. Therefore, the procedure as such does not impair the validity of our analysis.

[4] Table 9 in the Appendix provides more details about the meta-data set and its structure.

$$SPD = (Perf_F - Perf_M)/Perf_M \tag{1}$$

The main advantage of this measure is that it allows a comparison of performance values across experimental tasks measured on different scales. Another feature is its interpretability: Multiplied by 100, this effect size can be interpreted as the percentage amount by which women's group performance differs from men's. A positive sign indicates a performance advantage for women. The within-group standard deviations (SD) were adjusted accordingly (Equations 2 and 3; denoted SSD after adjusting).

$$SSD_F = SD_F/Perf_M \tag{2}$$

$$SSD_M = SD_M/Perf_M \tag{3}$$

We also calculate *Cohen's d* as an alternative effect size. Here, it is computed by subtracting the mean female performance from the mean male performance and dividing the difference by the pooled within-group standard deviation. In the meta-analysis literature, it is a commonly used standardised measure for group differences (e.g., Lindberg et al., 2010; Hyde et al., 1990), making it an appropriate choice for comparing our results with the related literature.[5]

2.3 Effect size dependence

As illustrated in Figure 1, the data set includes repeated measures. In many experiments, a group of participants solves a task at a piece rate, and shortly after, the same group solves the same task under a competitive incentive scheme.[6] Naturally, many effect sizes are correlated at the sample level. Meta-analyses in which effect sizes are based on mean differences presuppose that all effect sizes come from independent groups as opposed to paired groups. For the present data, one could only obtain independent groups for meta-analysis via the within-subject changes, which would then be synthesised across unique samples by gender. However, since the overwhelming proportion of studies in our pool do not report this information on within-subject changes along with the necessary statistics, we have to resort to comparing absolute performance means, which often constitute paired units. This violation of the assumption of independent groups may lead

[5] Note that the Cohen's d effect size has the opposite sign of our SPD measure by construction, in line with the meta-analysis literature on gender differences that uses Cohen's d.

[6] Technically, most underlying primary experiments elicit within-subject performance differences. However, since we only have access to the data reported in the papers, we cannot observe performances at the individual level, but rely on reported summary statistics of performance by gender, thus treating the data uniformly as between-subject measures of performance changes.

to distorted standard errors of the group differences, resulting in incorrect weights (Dunlap et al., 1996).

One possible solution would be disregarding any within-sample variation and including only one effect size per sample (e.g., use the authors' "preferred" effect size measure; Stanley & Doucouliagos, 2012: 32). In our case, this is not feasible since our analysis crucially depends on comparing performances in different settings. Furthermore, any sub-selection of observations would be arbitrary and might introduce bias in itself. It also considerably lowers statistical power compared to including all observations.[7] Instead, we attempt to mitigate the issue with the help of meta-regression analysis tools. Since we are confronted with several levels of hierarchy, the usual meta-analytic tools to deal with effect size dependency, like cluster-robust variance estimation (Stanley & Doucouliagos, 2012), might not be sufficient since they allow to take into account only one level of dependence. Instead, we model the multi-layered structure of the data via a multilevel random effects meta-regression (Harrer et al., 2021) and use cluster-robust variance estimation as a robustness check, testing different levels of clustering.

2.4 Explanatory variables

The explanatory variable of interest for analysing the relationship between performance and competitive pressure, *Competition*, is a dummy variable indicating whether the participants perform a task under competitive pressure or not. We define competitive pressure as having one's performance evaluated in relation to a certain benchmark. Usually, successful completion of a competitive task involves a reward, whereas failure to reach the benchmark yields a zero payoff. Most experiments invoke competitive pressure by forming tournaments with two or more players. The variable *Subject pool* categorises the participants by adolescents, students, and non-student adults. Children account for only three effect sizes, so we refrain from including them in the analysis. *Type of task* categorises the experimental tasks the participants engage in: *Math* signifies mathematical calculations. *Visual* describes tasks that involve (spatial) visual thinking (e.g., solving mazes or mental rotation tasks). *Attention* refers to tasks that require a particularly high level of concentration (e.g., counting zeros in a matrix). The final category, *physical*, indicates exercises that measure physical abilities (e.g., running, throwing). *Other* includes categories that do not offer any explanatory power due to insufficient observations (e.g., tasks that involve searching for or forming words fall into this category, logical reasoning exercises, or creative tasks). *Type of experiment* describes the setting of the experiment. We distinguish between online-, laboratory-, and field experiments.

[7] As the procedure can still provide some insight into whether effect size dependencies might distort our results, it will function as a complementary robustness check for the main analysis.

First performance measures the timing of performances whenever participants are asked to engage in a series of tasks. This variable is intended to capture learning effects. *First performance* is dummy coded as 1 for the first incident where the subjects perform a task, i.e., they have no prior experience in this specific assignment, and 0 for any later performance. We assume a diminishing learning effect over time. That is, most of the learning happens during the initial performance round, and any learning happening afterwards is negligible. A more detailed distinction of the time dimension is not feasible since the duration of rounds varies between experiments, and learning rates differ across tasks.

Publication year refers to the year in which a particular article was published. *Share of male authors* indicates the percentage of authors who are male in a given study.*Peer reviewed* signifies whether a study or article has undergone an evaluation process by independent experts in the same field before publication. *Performance main outcome* indicates whether a given study principally investigates performance differences as its outcome. *High reliability*indicates whether an effect size meets the reliability criteria described in Section 2.1. *Group size in competition* describes how many participants a subject's performance is compared to in the competitive scheme. *Winning chances* depicts the percentage of winning performances in a group. For instance, if the goal is to outperform a single opponent, that gives an average winning chance of 50 per cent; if scoring among the top two in a group of six is required to win, that gives an average winning chance of one-third.

Table 2 depicts descriptive statistics of the effect size and moderators. The average gender gap lies at - 6.2 per cent, ranging from − 62 per cent to + 40 per cent. The negative sign indicates higher mean performances among men. Keep in mind that our sample is somewhat selected because we had to omit effect sizes with missing standard deviations, which are mostly the gender gaps described as "insignificant", i.e., small or imprecisely measured effect sizes. We further explore this matter in the following sections.

45 per cent of the tasks were performed under competitive pressure. The highest proportion of effect sizes, 69 per cent, stems from students, followed by adults, with about 19 per cent. Adolescents make up 11 per cent. Most tasks (64 per cent) are of a mathematical nature. 76 per cent of effect sizes come from laboratory experiments. The imbalances regarding the subject pool, type of task, and type of experiment dimensions point out one particularly influential type of experiment: the subsample of students performing mathematical tasks in a laboratory (henceforth referred to as “SML”). They account for 103 effect sizes (44 per cent). 56 per cent of the effect sizes come from initial performances. The average publication year of studies is 2017, ranging from 2007 to 2022. The gender of authors is balanced; within study, the mean

Table 2: Summary statistics

	Mean	Min	Max	N
Gender Gap (SPD)	-0.062	-0.621	0.404	237
Competition	0.451	0.000	1.000	235
Subject pool				
Adolescents	0.114	0.000	1.000	237
Adults	0.194	0.000	1.000	237
Students	0.692	0.000	1.000	237
Type of task				
Math	0.641	0.000	1.000	237
Attention	0.156	0.000	1.000	237
Visual	0.101	0.000	1.000	237
Other	0.101	0.000	1.000	237
Type of experiment				
Lab	0.760	0.000	1.000	233
Online	0.077	0.000	1.000	233
Field	0.163	0.000	1.000	233
SML subsample	0.435	0.000	1.000	237
First performance	0.559	0.000	1.000	211
Publication year	2016.9	2007	2022	237
Share of male authors	0.478	0.000	1.000	227
Peer reviewed	0.814	0.000	1.000	237
Performance main outcome	0.224	0.000	1.000	237
High reliability	0.789	0.000	1.000	237
Comp. group size				
1	0.028	0.000	1.000	107
2	0.495	0.000	1.000	107
4	0.374	0.000	1.000	107
6	0.065	0.000	1.000	107
8	0.009	0.000	1.000	107
23	0.019	0.000	1.000	107
50	0.009	0.000	1.000	107

Notes: The table reports descriptive statistics on effect size and moderators. SPD is the standardised performance difference, as defined above. SLM stands for students performing math tasks in the laboratory.

share of male authors is 49 per cent. Most studies are peer-reviewed (81 per cent of effect sizes). Only a minority of effect sizes stem from studies with performance as their main outcome of interest (22 per cent). In the remaining 78 per cent of studies, performance scores are reported as side results. Nearly 80 per cent of effect sizes included in the main analysis reach a high level of reliability. Most frequently, tournaments are held between two people or in groups of four. Competitive tasks involving a comparison with one's own prior performance (2.8 per cent) and large competition groups with more than 4 participants are rare (10 per cent).

3 Meta-summary

In the following, we investigate the relationship between gender and performance by estimating a weighted mean (standardised) gender difference in task scores. As a first insight into sources of heterogeneity, we also conduct a subgroup analysis by grouping the data according to the main moderators introduced in Section 4. The analysis is carried out in Stata 16. We use a random effects model, specifically a Restricted Maximum Likelihood (REML) estimator, as recommended for continuous data (Veroniki et al., 2016). This model assumes the presence of between-effect size heterogeneity that goes beyond sampling error, while the alternative common effects model presupposes that all effect sizes are drawn from the same population and heterogeneity stems solely from sampling error (Stanley & Doucouliagos, 2012). As established by previous meta-analyses on the topic of gender differences in competition and performance (see Section 2), differences in methodology cause heterogeneous results. Therefore, a random effects estimator is the appropriate choice.

For computing the overall effect size, the individual standardised differences are weighted by a multiplicative combination of the following two factors: firstly, by the inverse of their standard errors (to account for effect size precision), and secondly, by a random element that models the between effect size heterogeneity (Konstantopoulos, 2011).

As can be seen at the top of Table 3, the overall effect size, as measured by the SPD (Equation 1), is - 0.045. It is significant at the 0.1 per cent level (p $<$ 0.001). According to this figure, women perform 4.5 per cent lower than men, on average. Unsurprisingly, and as reflected by the Q value of 758, we find highly significant heterogeneity, reinforcing our choice of meta-analytic model. Cohen's d lies at 0.13 and is highly significant, too (p $<$ 0.001). Although these overall effect sizes differ statistically significantly from zero, judging their magnitude is a different matter: Cohen (1988) suggests measures of $d < 0.2$ to be "small". For comparison, we consider effect sizes calculated by earlier meta-analyses on gender gaps in related economic preferences. For risk attitudes, Cohen's d lies between 0.17 for the Holt & Laury (2002) task and 0.55 in the

Table 3: Meta summary results

	Group	SPD	p-value	Cohen's d	p-value	N
Overall		-.045	0.000	.134	0.000	237
Competition						
	0	-.041	0.000	.126	0.000	129
	1	-.049	0.000	.141	0.000	106
Subject Pool						
	Students	-.055	0.000	.174	0.000	164
	Adolescents	-.029	0.196	.090	0.135	27
	Adults	-.013	0.556	.042	0.269	46
Type of task						
	Math	-.057	0.000	.145	0.000	152
	Attention	.018	0.021	-.086	0.001	37
	Visual	-.120	0.000	.373	0.000	24
	Other	-.049	0.020	.160	0.029	24
Type of experiment						
	Lab	-.053	0.000	.164	0.000	177
	Field	-.036	0.092	.072	0.106	38
	Online	.012	0.480	.011	0.767	18
SML subsample						
	0	-.032	0.001	.101	0.000	134
	1	-.065	0.000	.178	0.000	103
First performance						
	0	-.038	0.000	.117	0.000	93
	1	-.053	0.000	.154	0.000	118
Larger group						
	0	-.059	0.000	.180	0.000	53
	1	-.039	0.004	.106	0.005	47

Notes: The table reports the results of meta-summary analyses across the whole sample (top panel) and by subgroups (bottom panels). Results are reported for two alternative effect size measures: the standardised gender difference in performance (SPD) and *Cohen's d*, with corresponding p-values. SML refers to the experiments where students perform math tasks in the laboratory.

measure introduced by Eckel & Grossman (2002) (Filippin & Crosetto 2016). For self-esteem, Gentile et al. (2009) report effect sizes (measured again as Cohen's *d*) between -0.38 (women are more confident in the moral-ethical domain) and 0.41 (for athletic confidence). In their meta-analysis of experimental studies on competition *entry*, Markowsky & Beblo (2022) report an overall Cohen's *d* of 0.34.

Expressed in terms of the "Common Language Effect Size Indicator" by McGraw & Wong (1992), our calculated overall effect size means that when drawing a female and a male performance at random from the pooled data set, there is only a 54 per cent probability that the male performance is higher than the female one. In other words, knowing the sex of two randomly chosen participants adds almost no information compared to the 50 per cent chance of simply guessing who scored higher (see also Nelson 2016 for interpretations of effect size measures in terms of difference versus similarity.)

The results of the subgroup analysis show how the gap varies across different dimensions. Note that all results are qualitatively the same between the SPD and Cohen's d, with highly similar p-values. Recall that the signs of the SPD and Cohen's d are in opposite directions by construction. The following descriptions refer to the SPD measure unless stated otherwise.

The gap for competitive settings (-4.9 per cent, $p < 0.001$) is slightly larger than in non-competitive settings (-4.1 per cent, $p < 0.001$). However, the difference between both groups is insignificant ($p = 0.58$).

Being the most prevalent group in the data set, unsurprisingly, the gap among university students roughly corresponds with the overall gap in sign and magnitude (-5.5 per cent, $p < 0.001$). Among adolescents and non-student adults, the gap is negative too, albeit smaller and insignificant at common significance levels. These group differences are statistically significant at the 10 per cent level ($p = 0.08$). In mathematical tasks, women perform significantly worse than men by an estimated 5.3 per cent ($p < 0.001$). The largest difference between genders is found for visual tasks, where women's average performance is 12 per cent lower than men's ($p < 0.001$). For tasks that require attention, however, the relationship is reversed: We find that women, on average, perform better than men, the gap reaching significance on the 5 per cent level (+ 1.8 per cent, $p = 0.02$). The p-value of heterogeneity between task groups is less than 0.001, indicating substantial group differences. The SPD for experiments conducted in computer laboratories corresponds to -5.0 per cent ($p < 0.001$). Among field experiments, the mean performance of women lies an estimated 3.6 per cent below the men's average ($p = 0.092$), while online experiments produce no significant gender differences in performance. Again, group differences are highly significant. The mean effect size for the SML subsample is twice as large as for the remaining sample (-0.65

and –0.32, respectively). For both subsamples, the gender difference is highly statistically significant. A test of group differences using the Q-statistic reveals a significant difference between them (at the 1-per cent level). As reflected by the test for residual heterogeneity, heterogeneity is considerably lower among the SML-subsample ($Q = 160$) than among the remaining observations ($Q = 567$). This implies that the dimensions subject pool, type of task, and type of experiment capture most of the heterogeneity in our sample. However, the remaining heterogeneity is still considered significant ($p < 0.001$).

An analysis of the timing dimension reveals a negative gap that narrows slightly when tasks are repeated (among first performances, -5.3 per cent, $p < 0.001$; among later performances, -3.8 per cent, $p < 0.001$). However, this group difference is not significant ($p = 0.26$). After categorising *Group size in competition* into "2-person tournaments" and "groups of 4 and above", we find smaller gender gaps in larger groups.[8] However, the margins at which men outperform women do not differ significantly by group size ($p = 0.286$).

Like the overall mean gender difference, most of our subgroup analysis splits imply a significant but qualitatively small performance advantage for men. Only few exceptions indicate the opposite. Additionally, the subgroup analysis of first and later performances supports the existence of an overall gap in light of the interdependence issues elaborated in Section 4.3: Among first performances, there are no repeated measures and the effect size is about the same magnitude as the overall gap. Therefore, the finding that men (only) slightly outperform women on average is not due to any distortion from interdependence.

The analysis also hints at considerable heterogeneity between most groups, which motivates further investigation using regression analysis. However, the subgroup differences identified so far offer only limited explanatory value. They merely represent correlations, with any causal effect possibly masked or distorted by unobserved confounding factors.

4 Selection bias

Before exploring the influence of competition on the performance of women and men conditionally on potential confounders in a meta-regression setting, we conduct standard meta-analytic checks for publication bias to ensure the validity of our summary and regression results.

The issue of publication bias comprises two phenomena: The first occurs when authors or publishers refrain from publishing a study because the identified effect is insignificant, especially when

[8] Given that there are only three cases of a self-competition scheme which constitutes its own category, we refrain from including these observations in the subgroup analysis.

the result contradicts a theory or existing evidence. Research has shown that in economics, as well as in other sciences, more extensive and statistically significant results are more likely to be published in peer-reviewed journals (e.g., Card & Krueger, 1995; Brodeur et al., 2016). This results in a bias in the set of all published studies and consequently threatens the validity of meta-analytic conclusions.

The second phenomenon appears at the study level. Authors of studies that report multiple effect sizes might "hand-select" these by the same logic as mentioned above. The present meta-analysis could be especially susceptible to such under-reporting: It contains a large proportion of studies whose primary outcome of interest is the decision to compete. In many cases, performance differences only constitute a side result and might only be reported in detail if deemed particularly striking. Furthermore, as explained in Section 4, we had to exclude effect sizes that were reported as insignificant without indicating precise standard errors.

As a first approach to assessing whether our compiled data set suffers from bias by excluding "insignificant" gender gaps in performance, we calculate a rudimentary mean gender gap in performance for the effect sizes we had to drop from the data set. To improve comparability with the result from our main analysis, we weight the simple mean of the excluded effect sizes by the respective sample sizes as an approximate measure of precision. This way, among excluded sample sizes, we obtain a mean of -0.40. The relatively small difference to our main result of -0.45 provides some reassurance that excluding these effect sizes does not substantially bias our analysis. [9]

As a subsequent step, we present a funnel plot that visualises individual effect sizes in relation to their precision in Figure 2. The ordinate of this plot shows effect sizes' standard errors as measures of precision. The most precisely estimated effect sizes cluster around the overall effect size at the top of the graph (Note that the ordinate is reversed). In the absence of publication bias, the less precisely estimated effect sizes should scatter symmetrically around the overall effect size towards the bottom of the graph (Stanley & Doucouliagos, 2012). Visual inspection suggests our data exhibit some asymmetry, with fewer effect sizes located to the right of the mean. That is, fewer effect sizes report performance gaps in favour of women than expected if the probability of retrieving a given study was independent of its results. We visually distinguish between effect sizes, which are the main outcome (as indicated by the green dots), and those that are a secondary outcome (blue dots). Of the two kinds, effect sizes that are the main outcome exhibit greater asymmetry as a larger number of imprecise effect sizes is located to the left of the mean. These findings indicate that the overall effect size could be biased by imprecise estimates

[9] After weighting the included effect sizes by sample size for comparison (instead of using the reverse standard error, as we do in our baseline calculation), we find an even smaller gender gap of -0.36.

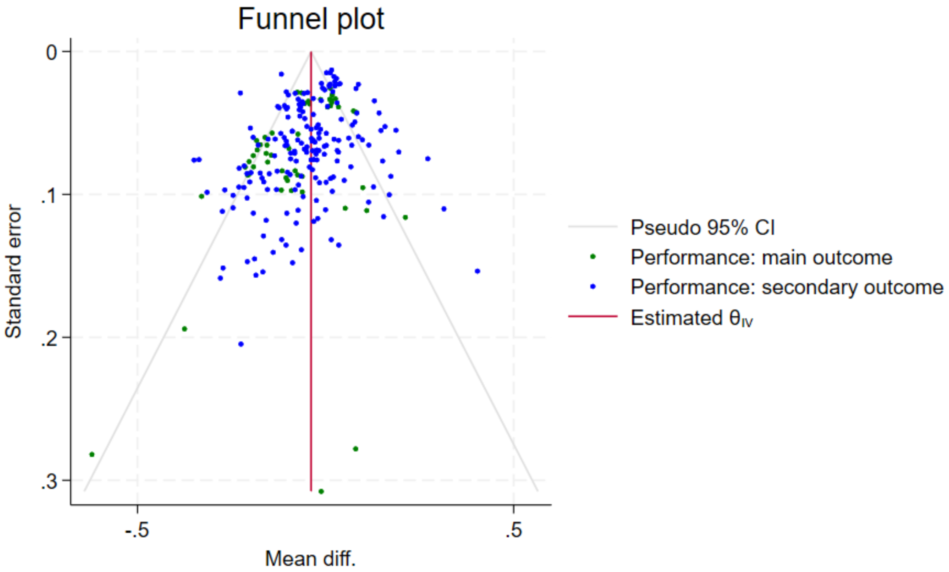

Figure 2: Funnelplot of the SPD by focus of the experiment

that indicate rather large negative gender performance gaps. This circumstance might be driven by the fact that studies are more likely to be published when their main outcome is significant and/ or has the expected sign.

As a more formalised test of publication bias, we also calculate a regression-based test on funnel-plot asymmetry following Egger et al. (1997). The null hypothesis, the absence of asymmetry, is rejected at the highest significance level ($p < 0.0001$). The relationship between effect size precision and magnitude weakens if we include an indicator for the gender gap in performance being the main outcome, along with controls for subject pool, task, and type of experiment, even though the relationship is still significant at the 5 per cent level ($p = 0.046$).

Doleman et al. (2020) demonstrate that Egger's test might indicate the presence of publication bias, even when there is none when applied to continuous outcomes. We, therefore, additionally employ their recommended novel funnel plot-based method to estimate publication bias for continuous outcomes. The results indicate asymmetry that is only marginally significant without controls ($p = 0.095$). However, the authors recommend further research before applying this method to standardised mean differences.

In any case, we do not rule out the potential presence of publication bias. Our analysis illustrates that taking into account the focus of the respective experiment as well as the fundamental dimensions of heterogeneity can mitigate the possible effects of selection bias on our results. Consequently, we incorporate these factors into our subsequent meta-regression analysis.

5 Meta-regression analysis

So far, we have established that aggregating the existing literature suggests that men slightly outperform women in most tasks that are used in economic experiments on competitiveness. The overall gender difference in performance seems to be slightly inflated by the selective nature of our sample, but this bias can be mitigated by taking into account the focus of the experiment and the basic dimensions of heterogeneous research designs.

In the following, we want to examine how the gender gap in performance differs between competitive and non-competitive payment options. We use meta-regression analysis with the (standardised) gender gap in performance as the dependent variable and *Competition* as the main explanatory variable of interest. All observable dimensions of heterogeneity, the *moderators* introduced in Section 2.4, are included as controls. Before reporting results, we briefly exemplify how we deal with effect size dependency in this setting. After discussing the results of the principal analysis, we conduct additional robustness checks to secure our main results.

5.1 Multilevel meta-regression

Like in most meta-analyses in economics and social sciences, we acknowledge that variation between the studies in our pool is not solely attributed to sampling error. Rather, studies differ in their "true" effect due to heterogeneous research designs, leading to a diverse distribution of effect sizes. The yellow bell-shaped curve in Figure 1 serves as a stylised representation of the distribution of studies' "true" effect sizes (Level 4). For this reason, meta-regression analyses in economics commonly account for the differences between heterogeneous effect sizes (i.e., "random elements") of studies. Considering our data's hierarchical structure of dependencies, we also model heterogeneous effect sizes *within* studies by introducing effect size dependence on the sample level into our meta-analytic model. Remember, studies can report gender differences in performance for several unique samples, e.g., groups that receive different treatments. The gender gaps in performance, i.e., the effect sizes in our data, are therefore nested within samples that are nested within studies. Where conventional meta-analysis estimates one random component on the study level, we, therefore, estimate two heterogeneity variance parameters: one on the study level and one on the sample level (see: Harrer et al., 2021).[10] Therefore, we employ a multilevel random effects meta-regression (Multilevel MRA), which allows us to model

[10] In theory, there is another level of dependence within samples, as some groups of participants complete multiple performance rounds *within* a particular treatment. In practice, however, this variance is fully accounted for by our control variables, and, consequently, the heterogeneity variance parameter within samples is estimated to be zero. We, therefore, disregard this level of dependence in the following.

Table 4: Main Multilevel MRA

	(1) Mean diff.	(2) Mean diff.	(3) Mean diff.	(4) Mean diff.	(5) Mean diff.
Competition	−0.0013	−0.0006	−0.0009	−0.0010	0.0005
	(0.0072)	(0.0072)	(0.0072)	(0.0072)	(0.0072)
Subject pool	no	yes	yes	yes	yes
Type of task	no	no	yes	yes	yes
Type of experiment	no	no	no	yes	yes
Study moderators	no	no	no	no	yes
Constant	-0.0414^{***}	-0.0539^{***}	-0.0610^{***}	-0.0612^{***}	−0.6118
	(0.0108)	(0.0128)	(0.0135)	(0.0140)	(7.0331)
Observations	221	221	221	221	221
Qres	691.4	668.8	560.3	535.5	516.6

Notes: Dependent variable: Standardised gender difference in performance (SPD). Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Study moderators are listed in Table 9.

random components between units at each of these levels (Konstantopoulos, 2011). To this end, we employ the *metafor* package in R (Viechtbauer, 2010).[11]

In addition to the moderators at the effect size level described in Section 2.4 (competitive pressure, timing, subject group, type of task, type of experiment),[12] we control for the moderators at the study level depicted in Table 9 - *publication year*, *share of male authors*, *peer-reviewed*, *Performance main outcome.*

Table 4 shows the regression results. Each column adds a set of moderators, with the fifth column including the full set of moderators. Remember that the dependent variable is relative female performance, measured on a percentage scale with a negative baseline value (i.e., averaging across effect sizes, women perform worse than men). Negative coefficients thus mean that the gender gap in performance widens.

The coefficient on the *Competition* variable reflects our main result. Throughout all specifications, this coefficient is small and insignificant. This therefore provides no evidence that competitive pressure widens or closes the gender gap in performance.

[11] For comparison, see appendix table 10 for a conventional random effects meta-regression not accounting for a multi-level structure.

[12] *Group size in competition* is not part of the main specification as it restricts the sample to competitive tasks and therefore does not offer variation for *Competition.*

The difference between the coefficient displayed in Column 1 and the raw difference between the gender performance gap in competitive and non-competitive settings from above is explained by the somewhat smaller sample size in the regression framework due to incomplete information on moderators in a handful of studies. If we estimate the specification in Column 1 on the full sample, the coefficient remains statistically insignificant.

Regarding *Subject pool*, *Type of task*, and *Type of experiment*, we find no significant influences on the mean gender gap in performance using our main specification in column (5). The one exception is tasks that require attention, for which we find that women's relative performance is about ten percentage points higher compared to the reference category of mathematical tasks (significant on the one-per cent level).[13]

Performance main outcome is negative and significant at the 5 per cent level, indicating studies that focus on the analysis of performance report women's relative performance as 7.7 percentage points worse compared to studies where performance is reported as a secondary result. None of the remaining study-level moderators is significant.

Every set of moderators has considerable explanatory value for the outcome. This is illustrated by the Q value, reflecting residual heterogeneity, decreasing upon the inclusion of each set. The test for residual heterogeneity is still highly significant ($p < 0.001$) and yields a Q value of 516.6 for the full specification. This circumstance implies that unobserved factors still partly determine gender differences in performance.

Summing up, our main results imply no relationship between women's relative performance and competitive pressure.

[13] Full results in appendix table 11.

Table 5: Multilevel-MRA - Cohen's d

	(1) Mean diff.	(2) Mean diff.	(3) Mean diff.	(4) Mean diff.	(5) Mean diff.
Competition	0.0024	0.0003	0.0023	0.0028	0.0004
	(0.0216)	(0.0216)	(0.0215)	(0.0215)	(0.0216)
Subject pool	no	yes	yes	yes	yes
Type of task	no	no	yes	yes	yes
Type of experiment	no	no	no	yes	yes
Study moderators	no	no	no	no	yes
Constant	0.1373***	0.1691***	0.1697***	0.1737***	6.8292
	(0.0284)	(0.0335)	(0.0332)	(0.0339)	(16.7335)
Observations	221	221	221	221	221
Qres	646.6	600.0	473.8	464.0	457.2

Notes: Dependent variable: Standardised gender difference in performance (SPD). Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Study moderators are listed in Table 9.

5.2 Robustness checks

We perform four robustness checks on our primary regression analysis. First, Table 5 shows the results with *Cohen's d* as an alternative effect size. As judged by the signs of the coefficients and p-values, the results are qualitatively the same as when using the SPD.

Second, Stanley & Doucouliagos (2012) raise doubts about the reliability of multilevel random effects models when conducting meta-research in economics, especially in the presence of publication or selection bias (pp. 84, 150). We, therefore, employ their preferred "unrestricted" weighted least squares model (unrestricted WLS, see Stanley & Doucouliagos, 2017) as a second robustness check. Unlike the multilevel MRA, unrestricted WLS does not allow to account for effect size dependencies on multiple levels. Instead, we estimate an unrestricted WLS regression with standard errors clustered at the sample level. [14] The results are depicted in Table 6. Like in our main estimation, the coefficient on *Competition* is small and insignificant. The differential constant compared to the main results is due to unrestricted WLS effectively estimating a weighted common-effects model instead of random effects.

Most experiments on competitive behaviour follow a typical sequence of incentive schemes: initially, participants perform a task without competitive pressure, and in the second phase, they

[14] Clustering standard errors at the study level yields qualitatively identical results.

Table 6: Unrestricted WLS

	(1) Mean diff.	(2) Mean diff.	(3) Mean diff.	(4) Mean diff.	(5) Mean diff.
Competition	-0.0028	-0.0016	-0.0008	-0.0026	-0.0011
	(0.0087)	(0.0095)	(0.0092)	(0.0081)	(0.0079)
Subject pool	no	yes	yes	yes	yes
Type of task	no	no	yes	yes	yes
Type of experiment	no	no	no	yes	yes
Study moderators	no	no	no	no	yes
Constant	-0.0175*	-0.0299***	-0.0484***	-0.0481***	-1.5715
	(0.0091)	(0.0107)	(0.0135)	(0.0133)	(4.0112)
Observations	221	221	221	221	221

Notes: Dependent variable: Standardised gender difference in performance (SPD). Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Study moderators are listed in Table 9.

participate in a tournament. If there are systematic differences between how women and men enhance their performance over time, the coefficient of *Competition* also captures these relative effects. In the analysis presented in Table 7, we investigate the possibility that gender-specific learning rates might confound the effect indicated by the *Competition* variable. We introduce the variable *First performance* to account for learning effects. Across all five columns, the coefficients on *Competition* remain qualitatively unchanged compared to the main analysis. Simultaneously, the negative coefficients on *First performance* suggest that the gender performance gap narrows with repeated tasks. However, these coefficients are insignificant (p > 0.35 in the full model), supporting our main conclusion that competitive pressure does not influence the gender performance gap.

Table 8 shows the results of four robustness checks involving restricted samples: The specification in Column (1) is based on the SML subsample. We find an insignificant effect of competitive pressure, as for the whole sample. We, therefore, conclude that the relationship between the gender gap in performance and competitive pressure does not deviate significantly between students performing math tasks in the lab and other experimental settings.

To eliminate repeated performances as a possible cause of effect size dependence, we also conduct a regression on a subsample consisting solely of first performances (Column 2). Despite the coefficient showing a larger magnitude than the main result (-0.03), it is statistically insignificant. This suggests that within-sample correlations arising from repeated measures do not affect the main result. It is important to note that there is limited variation in the competitive dimension

Table 7: Multilevel-MRA - Controlling for order effects

	(1) Mean diff.	(2) Mean diff.	(3) Mean diff.	(4) Mean diff.	(5) Mean diff.
Competition	-0.0208	-0.0197	-0.0183	-0.0171	-0.0113
	(0.0151)	(0.0151)	(0.0152)	(0.0153)	(0.0156)
First performance	-0.0221	-0.0218	-0.0201	-0.0188	-0.0138
	(0.0146)	(0.0146)	(0.0146)	(0.0147)	(0.0150)
Subject pool	no	yes	yes	yes	yes
Type of task	no	no	yes	yes	yes
Type of experiment	no	no	no	yes	yes
Study moderators	no	no	no	no	yes
Constant	-0.0230	-0.0384**	-0.0449**	-0.0467**	1.8354
	(0.0182)	(0.0195)	(0.0198)	(0.0200)	(7.1039)
Observations	205	205	205	205	205
Qres	589.7	561.5	463.9	421.1	410.5

Notes: Dependent variable: Standardised gender difference in performance (SPD). Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Study moderators are listed in Table 9.

when considering only initial performances. As a consequence, the statistical precision of this particular specification is constrained.

For the subsample of peer-reviewed studies (supposedly of higher average quality), the main result still holds: the gap does not change in response to competitive pressure, the measured effect being small and insignificant (Column 3). Similarly, when restricting the sample to highly reliable effect sizes, we find no significant effect (Column 4).

Table 8: Multilevel MRA - Restricted samples

	(1) SML	(2) First perf.	(3) Peer reviewed	(4) High reliability
Competition	0.0026	-0.0307	-0.0015	0.0015
	(0.0138)	(0.0304)	(0.0095)	(0.0083)
Subject pool	/	yes	yes	yes
Type of task	/	yes	yes	yes
Type of experiment	/	yes	yes	yes
Study mod.	yes	yes	yes	yes
Constant	-2.4698	2.5595	-3.0981	2.1238
	(7.1708)	(6.5792)	(6.7617)	(7.2689)
Observations	102	114	178	171
Qres	150.3	215.5	373.3	371.1

Notes: Dependent variable is the standardised gender difference in performance. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Column 1 restricts the sample to experiments with students performing math tasks in the laboratory. Column 2 restricts the sample to first performances and, therefore, uses only one effect size per study. Column 3 restricts the sample to peer-reviewed studies. Column 4 shows the regression results when restricting to the effect sizes of high reliability, i.e., those where group-specific standard deviations were reported in the paper or it clearly stated that group differences were tested with t-tests.

6 Discussion and conclusion

This meta-study analyses gender differences in performance in competitive environments, a crucial aspect of competitive behaviour that has been understudied. If women perform poorly in competitive situations relative to men, this might help explain persistent gender gaps in positions of societal influence, which are mainly obtained through competitive processes. To contribute to understanding gender differences in competitive behaviour, we synthesise experimental evidence from 70 studies that conduct experiments where subjects perform real-effort tasks under competitive and non-competitive payment options.

While we find a statistically significant gender gap in performance of 4.5 per cent in favour of men, this difference is qualitatively small compared to effect sizes found in previous meta-analyses on gender gaps in economic preference traits, like risk attitudes, confidence, and willingness to enter competitions. Expressed in terms of an experiment where the performances of one woman and one man are chosen at random, our overall effect size suggests that the man would outperform the woman with a probability of 54 per cent. Thus, knowing the sex of two randomly chosen participants adds almost no information compared to the 50 per cent probability associated with naively guessing who performs better. This gender performance gap is insensitive to competitive pressure. That is, we find no evidence of an adverse effect of competition on women's relative performances.

Our findings suggest that competitive pressure does not inherently disadvantage women's performances, which challenges long-standing assumptions about gender and competition. Besides extending the current knowledge on gender differences in competitive behaviour, our research is informative for practitioners as well, for example, for organisations aiming to design evidence-based diversity interventions.

While it significantly contributes to understanding gender gaps in experimental performances by allowing comparisons across a wide range of dimensions, our meta-analysis suffers from some limitations: The fact that the meta-data set contains disproportionately many students solving mathematical problems in a computer laboratory does not speak for the generalisability of our results. However, by the same token, it allows stronger inferences about this particular group: The gender gap in performance of -6.5 per cent for a group exhibiting little residual heterogeneity in relation to the overall sample constitutes a very robust finding.

Additionally, based on the existing experimental evidence, we cannot satisfactorily distinguish between the performance reaction to competitive pressure and learning since the underlying experiments contain only limited variation in the order of competitive and non-competitive experimental stages. If women have higher learning rates than men, on average, this might mask gender gaps in competitive performances since 90 per cent of experiments have the competitive follow the non-competitive stage. However, we do not see reason to expect women to exhibit significantly higher learning rates across a relatively large number of different experimental tasks.

We also identify essential aspects of gender, performance, and competition that our study must remain unanswered. The first aspect regards further possible determinants of performance differences between men and women. Moderators, such as the presence of stereotype threat or the preferential treatment of one gender, have been subject to investigation regarding tournament entry as the outcome (see, Markowsky & Beblo, 2022). We cannot identify their potential impact on performance (and the dynamic relationship with competition) because such treatments are introduced after the performance elicitation in the vast majority of studies (e.g., Leibbrandt et al., 2018). By adjusting the experimental design accordingly, that is, shifting the timing of the intervention back, it would be possible to conduct such an analysis in the future.

Another aspect concerns the *causal* relationship between competitive pressure and *absolute* performance. To investigate this relationship, one would have to compare performance measures on identical scales while accounting for learning effects. Among experiments that use the same scale, we did not find sufficient variation regarding competitive pressure to do so. The alternative, standardising performances across scales, is only possible with individual-level performance data.

All in all, our findings challenge simplistic assumptions about gender differences in competitive performance, highlighting the importance of empirical investigation of all aspects of economically relevant behavioural traits. Our study shows how careful empirical analyses of existing evidence can help us test seemingly intuitive beliefs about gender gaps in economic preference traits.

Appendix

Table 9: Key information

	Mean	Min	Max	N
Panel A: Studies				
Publication year	2016.9	2007	2022	70
Share of male authors	.48	0	1	69
Peer reviewed	.74	0	1	70
No. of samples	1.83	1	8	70
No. of effect sizes	3.39	1	30	70
Panel B: Samples				
No. of effect sizes	1.85	1	4	128
No. of participants	249.3	12	2304	128
Panel C: Effect sizes				
No. of individual obs.	213.5	7	2304	237
High reliability	.79	0	1	237

Notes: The table lists characteristics of all studies, samples, and effect sizes. Each study can contain multiple samples, each of which can result in multiple effect sizes. *Share of male authors* has one observation less because we could not conclusively identify each author's gender from their names or personal websites for one paper. *High reliability* refers to the effect sizes where the paper clearly reports within-gender standard deviations or explicitly states that group differences are evaluated with t-tests.

Table 10: Random effects MRA

	(1) Mean diff.	(2) Mean diff.	(3) Mean diff.	(4) Mean diff.	(5) Mean diff.
Competition	-0.0057	-0.0031	-0.0030	-0.0030	-0.0004
	(0.0135)	(0.0134)	(0.0122)	(0.0122)	(0.0122)
Subject pool	no	yes	yes	yes	yes
Type of task	no	no	yes	yes	yes
Type of experiment	no	no	no	yes	yes
Study moderators	no	no	no	no	yes
Constant	-0.0400***	-0.0540***	-0.0622***	-0.0624***	-2.0793
	(0.0091)	(0.0102)	(0.0109)	(0.0110)	(4.3781)
Observations	221	221	221	221	221

Notes: Dependent variable: Standardised gender difference in performance (SPD). Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Main multilevel MRA

	(1)	(2)	(3)	(4)	(5)
	Mean diff.	Mean diff.	Mean diff.	Mean diff.	Mean diff.
Competition	-0.0013	-0.0006	-0.0009	-0.0010	0.0005
	(0.0072)	(0.0072)	(0.0072)	(0.0072)	(0.0072)
Adolescents		0.0214	0.0309	0.0400	0.0324
		(0.0278)	(0.0268)	(0.0365)	(0.0398)
Adults		0.0510**	0.0343	0.0533	0.0652
		(0.0254)	(0.0256)	(0.0452)	(0.0519)
Attention			0.0798***	0.0824***	0.0966***
			(0.0237)	(0.0267)	(0.0296)
Visual			-0.0674**	-0.0655**	-0.0539
			(0.0279)	(0.0305)	(0.0332)
Other task			-0.0091	-0.0073	0.0145
			(0.0285)	(0.0295)	(0.0321)
Online				-0.0266	-0.0583
				(0.0555)	(0.0607)
Field				-0.0167	-0.0388
				(0.0398)	(0.0436)
Performance main outcome					-0.0769**
					(0.0321)
Publication Year					0.0003
					(0.0035)
Share of male authors					0.0095
					(0.0314)
Peer reviewed					-0.0292
					(0.0278)
Constant	-0.0414***	-0.0539***	-0.0610***	-0.0612***	-0.6118
	(0.0108)	(0.0127)	(0.0135)	(0.0140)	(7.0331)
Observations	221	221	221	221	221
Qres	691.4	668.8	560.3	535.5	516.6

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

Alan, S. & Ertac, S. (2019), 'Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment', *Journal of the European Economic Association* **17**(4), 1147–1185.

Alan, S., Ertac, S., Kubilay, E. & Loranth, G. (2020), 'Understanding gender differences in leadership', *The Economic Journal* **130**(626), 263–289.

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø. & Tungodden, B. (2016), 'Willingness to compete: Family matters', *Management Science* **62**(8), 2149–2162.

Alnamlah, M. & Gravert, C. A. (2020), 'She could not agree more: The role of failure attribution in shaping the gender gap in competition persistence', *CEBI Working Paper 25/20, Available at SSRN* .

Apicella, C. L., Demiral, E. E. & Mollerstrom, J. (2017), 'No gender difference in willingness to compete when competing against self', *American Economic Review* **107**(5), 136–140.

Apicella, C. L., Demiral, E. E. & Mollerstrom, J. (2020), 'Compete with others? no, thanks. with myself? yes, please!', *Economics Letters* **187**, 108878.

Balafoutas, L., Kerschbamer, R. & Sutter, M. (2012), 'Distributional preferences and competitive behavior', *Journal of Economic Behavior and Organization* **83-334**(1), 125–135.

Balafoutas, L. & Sutter, M. (2012), 'Affirmative action policies promote women and do not harm efficiency in the laboratory', *Science (New York, N.Y.)* **335**(6068), 579–582.

Balafoutas, L. & Sutter, M. (2019), 'How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior', *European Economic Review* **118**, 1–13.

Bandiera, O., Parekh, N., Petrongolo, B. & Rao, M. (2022), 'Men are from Mars, and Women Too: A Bayesian Meta-analysis of Overconfidence Experiments', *Economica* **89**(S1), S38–S70. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecca.12407.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecca.12407*

Bedard, K. & Fischer, S. (2019), 'Does the response to competition depend on perceived ability? evidence from a classroom experiment', *Journal of Economic Behavior and Organization* **159**, 146–166.

Berlin, N. & Dargnies, M.-P. (2016), 'Gender differences in reactions to feedback and willingness to compete', *Journal of Economic Behavior and Organization* **130**, 320–336.

Bjorvatn, K., Falch, R. & Hernæs, U. (2016), 'Gender, context and competition: Experimental evidence from rural and urban uganda', *Journal of Behavioral and Experimental Economics* **61**, 31–37.

Bönte, W., Procher, V. & Urbig, D. (2018), 'Gender differences in selection into self-competition', *Applied Economics Letters* **25**(8), 539–543.

Booth, A., Fan, E., Meng, X. & Zhang, D. (2019), 'Gender differences in willingness to compete: The role of culture and institutions', *The Economic Journal* **129**(618), 734–764.

Booth, A. L. & Nolen, P. J. (2021), 'Gender and psychological pressure in competitive environments', *SSRN Electronic Journal* .

Booth, A. & Nolen, P. (2012), 'Choosing to compete: How different are girls and boys?', *Journal of Economic Behavior and Organization* **81**(2), 542–555.

Brandts, J., Groenert, V. & Rott, C. (2015), 'The impact of advice on women's and men's selection into competition', *Management Science* **61**(5), 1018–1035.

Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. (2016), 'Star Wars: The Empirics strike Back', *American Economic Journal: Applied Economics* **8**(1), 1–32.

Buehren, N., Goldstein, M. P., Vasilaky, K. N., Leonard, K. L., Montalvao, J. & Vasilaky, K. (2016), 'Women's empowerment, sibling rivalry, and competitiveness: Evidence from a lab experiment and a randomized control trial in uganda', *World Bank Policy Research Working Paper No. 7699, Available at SSRN* .

Buser, T. (2016), 'The impact of losing in a competition on the willingness to seek further challenges', *Management Science* **62**(12), 3439–3449.

Buser, T., Cappelen, A. W. & Tungodden, B. (2021), 'Fairness and willingness to compete', *SSRN Electronic Journal* .

Buser, T., Dreber, A. & Mollerstrom, J. (2017), 'The impact of stress on tournament entry', *Experimental economics* **20**(2), 506–530.

Buser, T., Geijtenbeek, L. & Plug, E. (2018), 'Sexual orientation, competitiveness and income', *Journal of Economic Behavior and Organization* **151**, 191–198.

Buser, T., Niederle, M. & Oosterbeek, H. (2014), 'Gender, competitiveness, and career choices *', *The Quarterly Journal of Economics* **129**(3), 1409–1447.

Buser, T., Peter, N. & Wolter, S. (2017a), 'Willingness to compete, gender and career choices along the whole ability distribution', *Working Paper* .

Buser, T., Ranehill, E. & van Veldhuizen, R. (2021a), 'Gender differences in willingness to compete: The role of public observability', *Journal of Economic Psychology* **83**, 102366.

Cadsby, C. B., Servátka, M. & Song, F. (2013), 'How competitive are female professionals? a tale of identity conflict', *Journal of Economic Behavior and Organization* **92**, 284–303.

Cahlíková, J., Cingl, L. & Levely, I. (2020), 'How stress affects performance and competitiveness across gender', *Management Science* **66**(8), 3295–3310.

Card, D. & Krueger, A. B. (1995), 'Time-Series Minimum-Wage Studies: A Meta-Analysis', *The American Economic Review* **85**(2), 238–243.

Carpenter, J. P., Frank, R. & Huet-Vaughn, E. (2017), 'Gender differences in interpersonal and intrapersonal competitive behavior', *SSRN Electronic Journal* .

Cassar, A. & Rigdon, M. L. (2021), 'Prosocial option increases womens entry into competition', *Proceedings of the National Academy of Sciences* **118**(45), e2111943118.

Cassar, A., Wordofa, F. & Zhang, Y. J. (2016), 'Competing for the benefit of offspring eliminates the gender gap in competitiveness', *Proceedings of the National Academy of Sciences of the United States of America* **113**(19), 5201–5205.

Cassar, A. & Zhang, Y. J. (2022), 'The competitive woman: Evolutionary insights and cross-cultural evidence into finding the femina economica', *Journal of Economic Behavior and Organization* **197**, 447–471.

Castilla, E. J. & Benard, S. (2010), 'The paradox of meritocracy in organizations', *Administrative Science Quarterly* **55**(4), 543–676.

Charness, G., Dao, L. & Shurchkov, O. (2022), 'Competing now and then: The effects of delay on competitiveness across gender', *Journal of Economic Behavior and Organization* **198**, 612–630.

Charness, G., Rustichini, A. & van de Ven, J. (2018), 'Self-confidence and strategic behavior', *Experimental Economics* **21**(1), 72–98.

Clot, S., Della Giusta, M. & Razzu, G. (2020), 'Gender gaps in competition: New experimental evidence from uk professionals.', *IZA Discussion Paper No. 13323, Available at SSRN* .

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Routledge, New York.

Comeig, I., Grau-Grau, A., Jaramillo-Gutiérrez, A. & Ramírez, F. (2016), 'Gender, self-confidence, sports, and preferences for competition', *Journal of Business Research* **69**(4), 1418–1422.

Croson, R. & Gneezy, U. (2009), 'Gender Differences in Preferences', *Journal of Economic Literature* **47**(2), 448–474.
URL: *http://www.jstor.org/stable/27739928*

Czibor, E. & Dominguez Martinez, S. (2019), 'Never too late: Gender quotas in the final round of a multistage tournament', *The Journal of Law, Economics, and Organization* **35**(2), 319–363.

Dasgupta, U., Mani, S., Sharma, S. & Singhal, S. (2019), 'Can gender differences in distributional preferences explain gender gaps in competition?', *Journal of Economic Psychology* **70**, 1–11.

Datta Gupta, N., Poulsen, A. & Villeval, M. C. (2013), 'Gender matching and competitiveness: Experimental evidence', *Economic Inquiry* **51**(1), 816–835.

Davoli, M. (2021), 'A, b, or c? question format and the gender gap in financial literacy', *Working Paper* .

Debray, T. P. A., Moons, K. G. M. & Riley, R. D. (2018), 'Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests', *Research synthesis methods* **9**(1), 41–50.

Doleman, B., Freeman, S. C., Lund, J. N., Williams, J. P. & Sutton, A. J. (2020), 'Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test', *Research Synthesis Methods* **11**(4), 522–534.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1414*

Dreber, A., von Essen, E. & Ranehill, E. (2014), 'Gender and Competition in Adolescence: Task Matters', *Experimental Economics* **17**(1), 154–172.
URL: *http://link.springer.com/10.1007/s10683-013-9361-0*

Dunlap, W. P., Cortina, J. M., Vaslow, J. B. & Burke, M. J. (1996), 'Meta-analysis of experiments with matched groups or repeated measures designs', *Psychological Methods* **1**, 170–177.

Egger, M., Davey Smith, G., Schneider, M. & Minder, C. (1997), 'Bias in meta-analysis detected by a simple, graphical test', *BMJ (Clinical research ed.)* **315**(7109), 629–634.

Filippin, A. & Crosetto, P. (2016), 'A Reconsideration of Gender Differences in Risk Attitudes',

Management Science **62**(11), 3138–3160. Publisher: INFORMS.
URL: *https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2015.2294*

Fletschner, D., Anderson, C. L. & Cullen, A. (2010), 'Are women as likely to take risks and compete? behavioural findings from central vietnam', *The Journal of Development Studies* **46**(8), 1459–1479.

Forbes (2018), 'The World's Most Powerful People', `https://www.forbes.com/powerful-people/list/`. Accessed: 2023-07-28.

Fu, J. & Zhong, S. (2019), 'She could not agree more: The role of failure attribution in shaping the gender gap in competition persistence', *Working Paper, Available at SSRN* .

Gentile, B., Grabe, S., Dolan-Pascoe, B., Twenge, J. M., Wells, B. E. & Maitino, A. (2009), 'Gender Differences in Domain-Specific Self-Esteem: A Meta-Analysis', *Review of General Psychology* **13**(1), 34–45. Publisher: SAGE Publications Inc.
URL: *https://doi.org/10.1037/a0013689*

Geraldes, D., Riedl, A. & Strobel, M. (2021), 'Gender differences in performance under competition: Is there a stereotype threat shadow?', *CESifo Working Paper no. 8809* .

Gill, D. & Prowse, V. (2014), 'Gender differences and dynamics in competition: The role of luck', *Quantitative Economics* **5**(2), 351–376.

Gneezy, U., L.Leonard, K. & A.List, J. (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).

Griselda, S. (2021), 'The gender gap in math: What are we measuring?', *Working Paper* .

Grosse, N., Riener, G. & Dertwinkel-Kalt, M. (2014), 'Explaining gender differences in competitiveness: Testing a theory on gender-task stereotypes', *Working Paper, Available at SSRN* .

Günther, C., Ekinci, N. A., Schwieren, C. & Strobel, M. (2010), 'Women can't jump?—An experiment on competitive attitudes and stereotype threat', *Journal of Economic Behavior & Organization* **75**(3), 395–401.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268110000855*

Halko, M.-L. & Sääksvuori, L. (2017), 'Competitive behavior, stress, and gender', *Journal of Economic Behavior and Organization* **141**, 96–109.

Halladay, B. & Landsman, R. (2022), 'Perception matters: The role of task gender stereotype on confidence and tournament selection', *Journal of Economic Behavior and Organization* **199**, 35–43.

Harrer, M., Cuijpers, P., A, F. T. & Ebert, D. D. (2021), *Doing Meta-Analysis With R: A Hands-On Guide*, 1st edn, Chapman Hall/CRC Press, Boca Raton, FL and London.

Harrison, G. W. & List, J. A. (2004), 'Field experiments', *Journal of Economic Literature* **42**(4), 1009–1055.

Hässler, P. & Schneider, S. (2020), '3 effects of gender differences in competition on creativity', *Gender Differences in Technology and Innovation Management* p. 29.

Hauge, K., Kotsadam, A. & Riege, A. (2020), 'Culture and gender differences in willingness to compete', *Working Paper* .

He, J. C., Kang, S. K. & Lacetera, N. (2021), 'Opt-out choice framing attenuates gender differences in the decision to compete in the laboratory and in the field', *Proceedings of the National Academy of Sciences of the United States of America* **118**(42).

Healy, A. & Pate, J. (2011), 'Can teams help to close the gender competition gap?', *The Economic Journal* **121**(555), 1192–1204.

Heinz, M., Normann, H.-T. & Rau, H. A. (2016), 'How competitiveness may cause a gender wage gap: Experimental evidence', *European Economic Review* **90**, 336–349.

Higgins, J. P., Li, T. & Deeks, J. J. (2022), Chapter 6: Choosing effect measures and computing estimates of effect., *in* J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, V. Welch & E. Flemying, eds, 'Cochrane Handbook for Systematic Reviews of Interventions', Cochrane.

Hoyer, B., van Huizen, T., Keijzer, L., Khavas, T. R., Rosenkranz, S. & Westbrock, B. (2016), 'Do talented women shy away from competition?', *Working Papers* (16-06).

Hoyer, B., van Huizen, T., Keijzer, L., Rezaei, S., Rosenkranz, S. & Westbrock, B. (2020), 'Gender, competitiveness, and task difficulty: Evidence from the field', *Labour Economics* **64**, 101815.

Hyde, J., Fennema, E. & Lamon, S. (1990), ‘Gender differences in mathematics performance: a meta-analysis.’, *Psychol Bull.* **107**(2).

Ifcher, J. & Zarghamee, H. (2016), ‘Pricing competition: a new laboratory measure of gender differences in the willingness to compete’, *Experimental Economics* **19**(3), 642–662.

Iriberri, N. & Rey-Biel, P. (2017), ‘Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision’, *Journal of Economic Behavior and Organization* **135**, 99–111.

Ivanova-Stenzel, R. & Kübler, D. (2011), ‘Gender differences in team work and team competition’, *Journal of Economic Psychology* **32**(5), 797–808.

Jung, S. & Vranceanu, R. (2019), ‘Willingness to compete: Between- and within-gender comparisons’, *Managerial and Decision Economics* **40**(3), 321–335.

Jurajda, & Münich, D. (2011), ‘Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities’, *The American Economic Review* **101**(3,), 514–518.
URL: *http://www.jstor.org/stable/29783799*

Kamas, L. & Preston, A. (2012), ‘The importance of being confident; gender, career choice, and willingness to compete’, *Journal of Economic Behavior and Organization* **83**(1), 82–97.

Kessel, D., Mollerstrom, J. & van Veldhuizen, R. (2021), ‘Can simple advice eliminate the gender gap in willingness to compete?’, *European Economic Review* **138**, 103777.

Khachatryan, K., Dreber, A., von Essen, E. & Ranehill, E. (2015), ‘Gender and preferences at a young age: Evidence from armenia’, *Journal of Economic Behavior and Organization* **118**, 318–332.

Klinowski, D. (2019), ‘Selection into self-improvement and competition pay: Gender, stereotypes, and earnings volatility’, *Journal of Economic Behavior & Organization* **158**, 128–146.

Klonner, S., Pal, S. & Schwieren, C. (2020), ‘Equality of the Sexes and Gender Differences in Competition: Evidence from Three Traditional Societies’, *Discussion Paper* .

Konstantopoulos, S. (2011), ‘Fixed effects and variance components estimation in three-level meta-analysis’, *Research synthesis methods* **2**(1), 61–76.

Kuhn, P. & Villeval, M. C. (2015), ‘Are women more attracted to co-operation than men?’, *The Economic Journal* **125**(582), 115–140.

Kuhnen, C. M. & Tymula, A. (2012), 'Feedback, self-esteem, and performance in organizations', *Management Science* **58**(1), 94–113.

Lavy, V. (2013), 'Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-based Pay Tournaments among Teachers*', *The Economic Journal* **123**(569), 540–573. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0297.2012.02542.x.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0297.2012.02542.x*

Lee, S., Niederle, M. & Kang, N. (2014), 'Do single-sex schools make girls more competitive?', *Economics Letters* **124**(3), 474–477.

Leibbrandt, A., Wang, L. C. & Foo, C. (2018), 'Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women', *Management Science* **64**(8), 3501–3516.

Lindberg, S. M., Hyde, J. S., Petersen, J. L. & Linn, M. C. (2010), 'New trends in gender and mathematics performance: A meta analysis', *Psychological Bulletin* **136**(6), 1123–1135.

Markowsky, E. & Beblo, M. (2022), 'When do we observe a gender gap in competition entry? a meta-analysis of the experimental literature', *Journal of Economic Behavior and Organization* **198**, 139–163.

Masclet, D., Peterle, E. & Larribeau, S. (2015), 'Gender differences in tournament and flat-wage schemes: An experimental study', *Journal of Economic Psychology* **47**, 103–115.

Mayr, U., Wozniak, D., Davidson, C., Kuhns, D. & Harbaugh, W. T. (2012), 'Competitiveness across the life span: the feisty fifties', *Psychology and aging* **27**(2), 278–285.

McGraw, K. O. & Wong, S. P. (1992), 'A common language effect size statistic.', *Psychological bulletin* **111**(2), 361.

Meier, K., Niessen-Ruenzi, A. & Ruenzi, S. (2017), 'The impact of role models on women's self-selection in competitive environments', *SSRN Electronic Journal* .

Morin, L.-P. (2015), 'Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform', *Journal of Labor Economics* **33**(2), 443–491. Publisher: The University of Chicago Press.

URL: *https://www.journals.uchicago.edu/doi/abs/10.1086/678519*

Müller, J. & Schwieren, C. (2012), 'Can personality explain what is underlying women's unwillingness to compete?', *Journal of Economic Psychology* **33**(3), 448–460.

Nelson, J. A. (2015), 'Are Women Really More Risk-Averse Than Men? A Re-Analysis of the Literature Using Expanded Methods', *Journal of Economic Surveys* **29**(3), 566–585. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12069.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12069*

Nelson, J. A. (2016), 'Not-so-strong evidence for gender differences in risk taking', *Feminist Economics* **22**(2), 114–142.

Niederle, M. (2017), 'A Gender Agenda: A Progress Report on Competitiveness', *American Economic Review* **107**(5), 115–119.

URL: *https://pubs.aeaweb.org/doi/10.1257/aer.p20171066*

Niederle, M., Kagel, J. H. & Roth, A. E. (2015), Gender, *in* 'The Handbook of Experimental Economics', Vol. 2, Princeton University Press, Princeton, pp. 481–553.

URL: *https://www.degruyter.com/document/doi/10.1515/9781400883172-009/html*

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**, 1067 – 1101.

Ors, E., Palomino, F. & Peyrache, E. (2013), 'Performance Gender Gap: Does Competition Matter?', *Journal of Labor Economics* **31**(3), 443–499. Publisher: The University of Chicago Press.

URL: *https://www.journals.uchicago.edu/doi/abs/10.1086/669331*

Paserman, M. D. (2023), 'Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players', *Journal of Economic Behavior & Organization* **212**, 590–609.

URL: *https://www.sciencedirect.com/science/article/pii/S0167268123001956*

Petersen, J. (2018), 'Gender difference in verbal performance: a meta-analysis of united states state performance assessments', *Educational Psychology Review* **30**(4), 1269–1281.

Price, C. R. (2012), 'Gender, competition, and managerial decisions', *Management Science* **58**(1), 114–122.

Price, C. R. (2020), 'Do women shy away from competition? Do men compete too much? : A (failed) replication', *Economics Bulletin* **40**(2), 1538–1547.

Reuben, E., Sapienza, P. & Zingales, L. (2019), 'Taste for competition and the gender gap among young business professionals', *Working paper* .

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.

Saccardo, S., Pietrasz, A. & Gneezy, U. (2018), 'On the size of the gender difference in competitiveness', *Management Science* **64**(4), 1541–1554.

Samak, A. C. (2013), 'Is there a gender gap in preschoolers' competitiveness? an experiment in the u.s', *Journal of Economic Behavior and Organization* **92**, 22–31.

Shastry, 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**, 101592.

Shurchkov, O. (2012), 'Under pressure: Gender differences in output quality and quantity under competition and time constraints', *Journal of the European Economic Association* **10**(5), 1189–1213.

Shurchkov, O. & Eckel, C. C. (2018), Gender Differences in Behavioral Traits and Labor Market Outcomes, *in* S. L. Averett, L. M. Argys & S. D. Hoffman, eds, 'The Oxford Handbook of Women and the Economy', Oxford University Press, pp. 480–512.
URL: *http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190628963.001.0001/oxfordhb-9780190628963-e-14*

Stanley, T. D. & Doucouliagos, H. (2012), *Meta-regression analysis in economics and business*, Routledge, New York.

Stanley, T. D. & Doucouliagos, H. (2017), 'Neither fixed nor random: weighted least squares meta-regression', *Research synthesis methods* **8**(1), 19–42.

Tungodden, J. & Willén, A. (2023), 'When parents decide: Gender differences in competitiveness', *Journal of Political Economy* **131**(3), 751–801.

van Veldhuizen, R. (2018), 'Gender differences in tournament choices: Risk preferences, overconfidence or competitiveness?', *Working Paper* .

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D. & Salanti, G. (2016), 'Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis', *Research Synthesis Methods* **7**(1), 55–79.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1164*

Viechtbauer, W. (2010), 'Conducting meta-analyses in r with the metafor package', *Journal of Statistical Software* **36**, 1–48.

World Bank Group (2022), 'PPP conversion factor, GDP (LCU per international $)', `https://data.worldbank.org/indicator/PA.NUS.PPP`. Accessed: 2022-05-30.

Wozniak, D., Harbaugh, W. T. & Mayr, U. (2016), 'The effect of feedback on gender differences in competitive choices.', *Available at SSRN 1976073* .

Yagasaki, M. (2022), 'Encouraging women to compete under social image concerns', *Available at SSRN 3416380* .

Zhang, Y. J. (2019), 'Culture, institutions and the gender gap in competitive inclination: Evidence from the communist experiment in china', *The Economic Journal* **129**(617), 509–552.

Zhong, S., Shalev, I., Koh, D., Ebstein, R. P. & Chew, S. H. (2018), 'Competitiveness and stress', *International Economic Review* **59**(3), 1263–1281.

Zuckerman, M., Li, C. & Hall, J. A. (2016), 'When men and women differ in self-esteem and when they don't: A meta-analysis', *Journal of Research in Personality* **64**, 34–51.
URL: *https://www.sciencedirect.com/science/article/pii/S0092656616300873*