DEPARTMENT OF ECONOMICS
JOHANNES KEPLER UNIVERSITY OF LINZ

Gender differences in overconfidence and decision-making in high-stakes competitions: evidence from freediving contests

by

Mario LACKNER

Hendrik SONNABEND

Working Paper No. 2016

September 2020

Johannes Kepler University of Linz
Department of Economics
Altenberger Strasse 69
A-4040 Linz - Auhof, Austria
www.econ.jku.at

Corresponding author: mario.lackner@jku.at

Gender differences in overconfidence and decision-making in high-stakes competitions: evidence from freediving contests

Mario Lackner[*] Hendrik Sonnabend[†]

14th September 2020

Abstract

This study examines gender differences in overconfidence and decision-making in a high-stakes environment. Using data on more than 40,000 individual attempts from international freediving competitions, we provide evidence that women, on average, are less likely than men to overestimate their ability. This result is robust to different measures of overconfidence and can be partly explained by experience. There are no substantial gender differences on the intensive margin of overconfidence. In terms of performance, results suggest that women suffer more from overconfidence than men.

JEL-Code: D03, D81, J16, Z2

Keywords: overconfidence; gender; decision-making; competition; freediving

[*] Mario.Lackner@jku.at, Johannes Kepler University Linz (JKU), Department of Economics

[†] Hendrik.Sonnabend@fernuni-hagen.de, University of Hagen, Department of Economics, Hagen, Germany (*corresponding author*).

Acknowledgements

We would like to thank Christian Deutscher, Ulf Rinne, Sascha Serwe, W. David Allen, Joachim Grosser, and Michael Möcker for their helpful comments. In addition, we are grateful to Houda Ben-Said for her excellent assistance in preparing the data.

1 Introduction

It is a well-established result that people find it difficult to correctly estimate their abilities. Flawed self-assessment has profound consequences in many everyday fields of economic activity such as entrepreneurship (e.g. Astebro et al., 2014), finance (e.g. Glaser and Weber, 2010), company management (e.g. Malmendier and Tate, 2015), contract design (e.g. Grubb, 2015), education (e.g. Reuben et al., 2017), and labour markets (e.g. Spinnewijn, 2015; Hoffman and Burks, 2020; Cooper and Kuhn, 2020).

Furthermore, there is evidence that men and women are likely to differ in self-confidence. For instance, in their seminal paper, Niederle and Vesterlund (2007) show that their ground-breaking result that "women shy away from competition and men embrace it" can partly be explained by differences in participants' beliefs about their relative performance. Generally, a bias in self-assessment may lead to sub-optimal decision-making. This is exactly the case in Niederle and Vesterlund (2007), where low-ability male participants opt into the tournament scheme too often (in terms of payoff maximisation), whereas the opposite is true for female participants.[1]

In this paper we seek to investigate gender differences related to a type of overconfidence commonly referred to as (ex ante) *overestimation* (e.g. Moore and Healy, 2008). Overestimation refers to situations where the individual's estimation of their own abilities is systematically biased upwards.[2] In a typical laboratory setting, participants would be asked to predict their performance before carrying out a task and/or to assess their performance afterwards (e.g. Clark and Friesen, 2009; Kamas and Preston, 2012). In one of the rare real-world studies, Bengtsson et al. (2005) document gender gaps in (ex post) self-assessment among students who have to decide whether their performance in a written exam is good enough to aim for the highest grade. Krawczyk and Wilamowski (2017) document that male marathon participants (ex ante) overestimate their target time more than females. They also tend to slow down more in the second half of the

[1] Further evidence is given by Dohmen and Falk (2011) and Niederle and Vesterlund (2011).

[2] Moore and Healy (2008) also define two other types of overconfidence. The first is overplacement (denoting an individual's tendency to believe that their performance relative to others is better than it actually is, also known as the 'better-than-average effect'). The second is overprecision (meaning an excessive certainty regarding the accuracy of one's beliefs).

race, suggesting that the gender gap in overestimation translates into performance.[3]

This study provides clear evidence of gender differences in overestimation among professionals performing a gender-neutral task in a high-stakes environment. For this we leveraged a unique rule in international freediving contests whereby athletes have to officially announce their intended performance (depth, distance, or duration) the day before the event. This gives us a continuous measure of overconfidence. Falling below that value is penalised whereas if athletes go too far, a loss of consciousness not only results in disqualification, but also poses a serious health threat.[4] There is some similarity to the weightlifting scenario in Genakos and Pagliero (2012). Yet, those authors do not consider potential gender differences. Weightlifting contests also do not provide an opportunity to measure the intensive margin of overconfidence as competitors either fail or succeed. That is, there is no information about how close they were to success.

We find that females, on average, are less likely to show overconfidence and to take too much risk.[5] This result is robust to different measures of overconfidence and can be partly explained by experience. On the intensive margin, our results do not show substantial gender differences. Still, we find that women suffer more from overconfidence than men in terms of competition outcomes, and we show that the overall negative effect of overconfidence on relative performance is more pronounced for better athletes. Finally, with reference to prior research on how peer information affects the decision-making process in strategic situations (e.g. Beshears et al., 2015; Dechenaux et al., 2015; Brookins et al., 2016; Gamba et al., 2017), we investigate the causal effect of confidence (proxied by the announced performance) on (realised) performance. Taking an instrumental variable approach, we find a negative effect of competition-induced (over)confidence on relative performance which is again more pronounced for female

[3] In another study taken from sport, Anbarci et al. (2016) document that male tennis players are more likely to make ‘embarrassing’ (i.e. noticeably wrong) line-call challenges. This, however, could also be explained by strategic behaviour like catching their breath or disrupting their opponent’s rhythm.

[4] Note that Frick (2020) also uses data from freediving events, among other sources, in his study of sensation-seeking behaviour.

[5] Overconfidence and risk-taking are two closely related concepts. Since “the overconfidence literature has occasionally branched out from the domain of judgment to that of decision-making” (Campbell et al., 2004, p.299), a decision which involves too much risk can be broadly seen as a consequence of overconfidence. According to Bertrand (2011), “a gender gap in overconfidence is often offered as an explanation for the gender gap in risk attitudes” (p.1550).

divers.

The remainder of the article is structured as follows: Section 2 gives some background information on freediving and describes the data set. Next, Section 3 presents the empirical strategy and the results. Finally, Section 4 concludes with a discussion.

2 Institutional background and data

Basically, freediving means breath-holding without equipment until resurfacing. While there are early references to freediving from the beginning of written history, modern competitive freediving has its origins in the 1950s. A milestone was the founding of the International Association for the Development of Apnea (AIDA) in 1992 which organises events and verifies record attempts (Lahtinen et al., 2015).

In contests, points are awarded according to the duration of immersion (STA), the depth of the dive (CWT, CWTB, CNF, FIN), or the distance covered (DYN, DYNB, DNF). See Table 1 for an overview of standard disciplines. The conversion is defined as follows: 1 second = 0.2 points (duration), 1 meter = 0.5 points (distance), 1 meter = 1 point (depth). Men and women compete for their own titles in mixed- and same-sex groups.

Moreover, contestants have to officially announce their intended performance (depth, distance, or duration) the day before the event. For depth disciplines, the announced performance (AP) is the maximum value which can be achieved. When the realised performance (RP) is *below* AP, a penalty occurs which is a linear function of AP - RP: $\alpha\,(AP - RP)$ with $\alpha \in (0, 1)$.[6] Since the difference between AP and RP also works as a tie-breaker, it can be interpreted as substantial sanction for overconfidence. Announcements remain private information until the event starts.

Obviously, the inherent risks caused by hypoxia pose serious health threats, either directly by underwater blackouts[7] or indirectly (for possible long-term sequelae see e.g. Ridgway and McFarland, 2006; Dujic and Breskovic, 2012; Pearn et al., 2015). For this

[6] In detail, $\alpha = 0.2$ for time competitions, $\alpha = 0.5$ for distance competitions, and $\alpha = 0.1$ for depth competitions.

[7] Two of the most prominent cases are the deaths of former American record holder Nicholas Mevoli in 2013 and multiple world-record holder Natalia Molchanova in 2015.

reason, all AIDA freediving events include extensive safety procedures. For instance, an athlete's level of hypoxia is tested via the Surface Protocol (SP). Violation of the SP leads to disqualification.[8]

Despite the real risks involved, freediving is far from being an 'adrenaline' or 'action sport' for athletes who have an appetite for risk. The reason is that adrenaline consumes oxygen. Consequently, unlike other sports, athletes in their forties or fifties have comparative advantages because their metabolism has slowed down. Freediving is often regarded as a 'pure' sport which demands great focus and where complacency is scorned.[9] It is sometimes also referred to as a 'spiritual experience'.[10] We thus can observe a comparably high proportion of female divers. This is an important point as prior research has shown that gender gaps in confidence are the greatest for tasks commonly perceived as masculine (e.g. Lundeberg et al., 1994; Beyer and Bowden, 1997; Shurchkov, 2012; Coffman, 2014; Dreber et al., 2014; Bordalo et al., 2019).

Table 1: Freediving disciplines.

Abbr.	Name	Description
CWT	Constant weight apnea	Depth competition with fin and a small amount of weight.
CWTB	Constant weight apnea with bifins	Same as CWT without monofins.
CNF	Constant weight apnea without fins	Same as CWT without fins.
FIM	Free immersion apnea	Same as CNF, only guide rope is allowed for propulsion.
DYN	Dynamic apnea with fins	Pool disciplines (horizontal distance) with fin.
DYNB	Dynamic apnea with bifins	Same as DYN without monofins.
DNF	Dynamic apnea without fins	Same as DYN without fins.
STA	Static apnea	Duration competition.

The data we use in our empirical analysis was taken from the official AIDA list of international freediving tournaments between 2006 and 2019. It covers $41,947$ dives from $1,378$ different events. Figure B.1 in the Appendix illustrates the number of events by year. An event typically lasts one or more days and consists of multiple competitions for male and female divers in various disciplines. We define a competition as the sum

[8] The official rules and safety standards can be found at https://www.aidainternational.org/Documents.

[9] Note that Alkan and Akış (2013) found that freediving athletes perform better in tests for situational psychological factors (e.g. stress level, anxiety, and affectivity) and stable psychological factors (e.g. locus of control and coping with stress) compared to non-athletes in their sample.

[10] See, for instance, the website of diver Sara Campbell, who is part of our sample: http://www.discoveryourdepths.com/freediving-with-us.

of attempts in one of the disciplines presented in Table 1 by gender at an event. Our data covers a total of $5,375$ competitions. Each observation is associated with a (valid or invalid) diving attempt. $3,463$ (8.3%) attempts ended in a disqualification.

For all years and competitions, we can identify a total of $5,078$ individual divers being involved in our data. The proportion of observed attempts relating to female divers equals 31.96%. For each diver, we observe an average of 32.80 dives. Since our analysis relies on information on past performances, data that involves single observations and first events by diver were discarded.

Figures 1 and 2 present kernel density estimates for performances and the difference between announced and realised performance in distance and depth contests (on the left) and time contests (on the right). It shows that while gender gaps in divergence are discernible, gender gaps in performance are not.[11]

Furthermore, descriptive statistics for the main variables used in the analysis (by gender and discipline) can be found in Tables 2 and 3. Table 2 shows that the realised distances and times exceed the announcements whereas the opposite is true for depth competitions. The reason is that announcements can work like a strategic element to determine the running order in (horizontal) distance and duration competitions. For instance, athletes aiming for a low start number will declare a value like one metre/minute since exceeding the pre-announced performance is not penalised. On the contrary, in depth competitions, this is not possible as the ‘tag’, i.e. the marker at the bottom plate that must be retrieved, is placed at the pre-announced depth. In other words, the announced performance is a binding constraint. We therefore mainly focus on depth competitions in the empirical analysis.

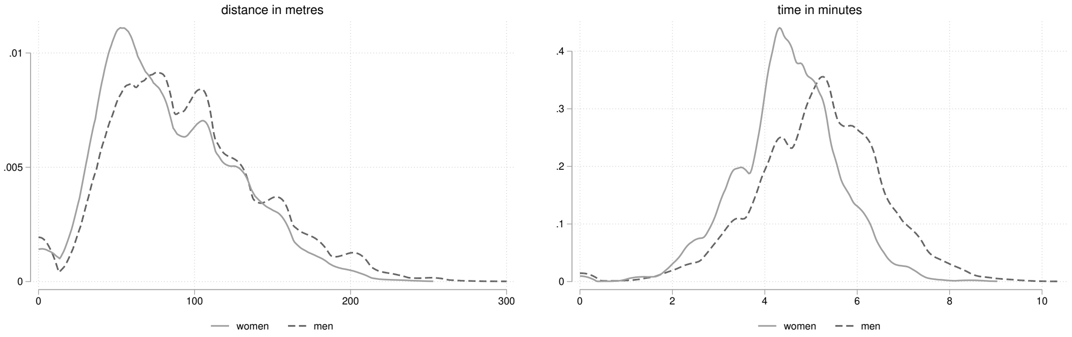
Table 3 presents summary statistics for our outcome variables. Throughout the paper, we define overconfidence as the difference between the announced performance and the realised performance (AP and RP). In the same manner, an attempt is categorised as ‘overconfident’ when the athlete falls below the prior announcement (RP $<$ AP).

[11] Note that despite advantages of male subjects with regard to aerobic capacity, lung volume, and haematological index, Jay and White (2006) as well as Cherouveim et al. (2013) find no gender differences in breath-hold time.

This applies to 920 dives (7.34%) assigned to female athletes and $2,426$ dives (9.35%) assigned to male athletes. As explained above, we expect depth competitions to provide the most reliable measure of overconfidence.

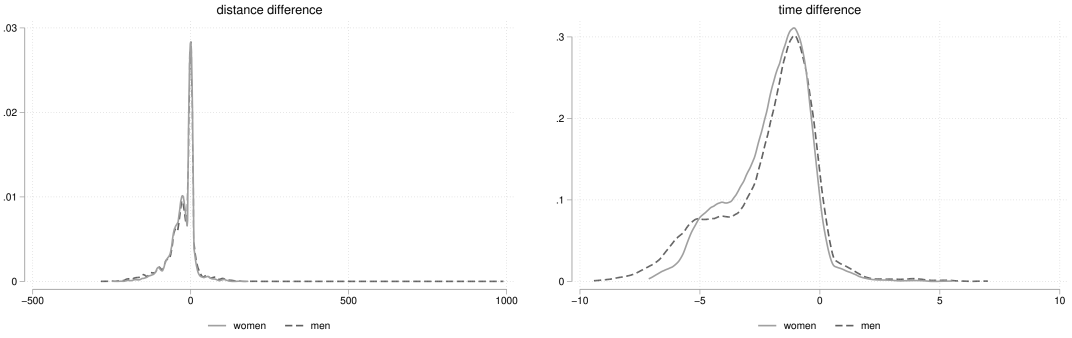
The score is a function of the performance (e.g., one metre in depth gives one point) and the potential penalties imposed for RP $<$ AP and infractions of the rules (like late starts or illegal techniques, cf. AIDA (2020)). The fact that the average difference between the score and the highest score in that contest is considerably higher in men's competitions than in women's competitions hints at differences in the competitive balance.

Figure 1: Realised performances by gender (kernel density estimates)



Notes: Realised performances for depth and distance competitions in metres (left), as well as time competitions (right) in minutes.

Figure 2: Difference between performance announcements and realisations by gender (kernel density estimates)



Notes: Differences in announced performance minus realised performance for depth and distance competitions in metres (left), as well as time competitions (right) in minutes.

Table 2: Summary statistics of main variables by competition type and gender.

	depth		*distance*		*time*	
	women	men	women	men	women	men
number of competitors	10.69	14.97	12.95	21.21	11.96	21.61
	(9.55)	(13.01)	(10.48)	(14.17)	(9.78)	(13.93)
realised depth/distance[a]	50.00	58.77	104.54	113.43		
	(19.36)	(23.74)	(39.80)	(46.61)		
announced depth/distance[b]	53.35	63.39	63.34	71.21		
	(17.58)	(21.77)	(39.46)	(44.43)		
personal best[c]	55.08	63.67	107.89	119.72		
(depth or distance)	(20.73)	(25.74)	(43.12)	(50.27)		
realised time					4.49	5.16
					(1.11)	(1.37)
announced time					2.41	3.01
					(1.56)	(1.86)
personal best[c]					4.78	5.45
(time)					(1.16)	(1.40)
N	3,888	7,744	6,056	13,598	3,461	7,200

Notes: Means for all outcome variables by discipline type and gender. Standard deviations are reported in parentheses. [a] The realised depth or distance measured in metres for all competition formats categorised as depth or distance competitions. For 827 observed dives, no valid performance was recorded due to disqualification or other reasons. [b] Announced performances in metres for all competition format categorised as depth or distance competitions. [c] Personal bests in terms of distance and time are recorded for each diver at past events in the same competition format as the observed event. Since we can only rely on dives recorded in our data, we expect this variable to be 'noisy' and that the accuracy depends on the number of observations and on the (unobserved) previous diver experience.

Table 3: Descriptives: outcome variables by competition type and gender

	depth		*distance*		*time*	
	women	men	women	men	women	men
overconfident	0.20	0.25	0.05	0.08	0.03	0.05
(1 = yes, 0 = no)	(0.40)	(0.43)	(0.22)	(0.27)	(0.17)	(0.22)
disqualified[a]	0.07	0.10	0.07	0.09	0.05	0.08
(1 = yes, 0 = no)	(0.26)	(0.30)	(0.25)	(0.29)	(0.22)	(0.27)
blackout[b]	0.02	0.03	0.03	0.03	0.01	0.02
(1 = yes, 0 = no)	(0.15)	(0.17)	(0.16)	(0.18)	(0.11)	(0.14)
score	48.54	56.18	52.26	57.33	53.21	61.14
	(19.76)	(24.09)	(19.35)	(22.11)	(14.18)	(17.50)
score difference[c]	21.30	28.95	22.02	31.98	16.22	23.71
	(18.94)	(24.30)	(18.21)	(22.58)	(14.75)	(17.85)
N	3,888	7,744	6,056	13,598	3,461	7,200

Notes: Means for all outcome variables by discipline type and gender. Standard deviations are reported in parentheses. [a] $disqualified$ is a binary variable equal to one if the observed dive is recorded as a disqualification, zero otherwise. [b] $blackout$ is a binary variable equal to one if the observed attempt results in the blackout of the diver, zero otherwise. [c] $scoredifference$ is defined as the difference between the score and the highest score in the contest.

3 Empirical strategy and results

The main objective of the empirical analysis is to estimate gender differences in overconfidence. To identify overconfidence, we take the difference between the announced and the realised performance, AP and RP. In our setting, athletes face the following trade-off: choosing a low level of AP decreases the chance of winning, but also mitigates the risk of penalties, disqualification, and health impairments. If men are more overconfident than females, we expect them to end up with RP $<$ AP more often. As a second aspect of overconfidence, we also investigate the intensive margin (measured in depth, distance, or time) of the over-announcement for overconfident athletes.

Furthermore, we examine the consequences of overconfidence. If male divers are more overconfident and therefore overestimate their ability, we expect them to have a higher rate of disqualification because of hypoxia issues across disciplines. Additionally, the outcome is qualified in terms of points and final rankings. This allows checking for sub-optimal decision-making.

Specifically, we estimate the following fixed effects model:

$$Y_{i,d,e,k} = \beta_0 + \beta_1 female_d + \delta' \mathbf{X_{i,d,c}} + \xi_e + \phi_k + \varepsilon_{i,d,e,k} \ , \tag{1}$$

where $Y_{i,d,e,k}$ is the outcome of interest (overconfidence and its consequences) for diver d in attempt i at event e competing in discipline k. Each event covers one or more competitions of various disciplines. The coefficient β_1 then captures the gender gaps in overconfidence. Furthermore, the vector $\mathbf{X_{i,d,c}}$ consists of diver and competition-level control variables which account for the number of total competitors and relative ability differences. The number of competitors controls for the size of the competition. For instance, a higher number of competitors may evoke a higher level of uncertainty about the 'right' announcement. And, as we cannot rely on individual fixed effects (FEs) to control for heterogeneity in terms of (relative) ability, we use data on past performances instead. More specifically, *relative ability* is defined as $PB_{d,c} - \overline{PB}_{-d,c}$, where $PB_{d,c}$ is the (observed) personal best of diver d before competition c and $\overline{PB}_{-d,c} =$

$\frac{1}{n_c}\sum_{-d=1}^{n_c} personal\ best_{-d,c}$ is the average personal best of all other (n_c) participants in the competition c.[12] We estimate discipline (as defined in Table 1) fixed effects (ϕ_k) to account for the discipline-specific strategic concerns of all competing divers. Finally, the model also includes event fixed effects (ξ_e) to control for time trends and unobserved event-specific characteristics (like mixed- or same-sex groups of competitors) that may impact on announcements and performances. Coefficients are estimated in an ordinary least squares (OLS) regression framework.

3.1 Main results

Table 4 presents the results for estimating model (1) with a binary dependent variable which equals 1 if the attempt is classified as overconfident (RP $<$ AP), and zero otherwise. We split the sample into the three main types of competitions: depth competitions, distance competitions, and time competitions.[13] In addition to plain models including only fixed effects, we present results for a full specification including competition-level control variables. The results from the full model with all control variables provide evidence in favour of a significant gender gap in overconfidence of 3.8 percentage points (ppts) for depth competitions. The estimated coefficients are considerably lower for distance competitions and time competitions (1.1 ppts and 1.7 ppts, respectively), but are less reliable due to the strategic use of announcements as explained above.[14] Evaluated at the sample mean, female competitors are about 17.8% less overconfident than men in depth competitions (distance competitions: 28%, time competitions: 45%). The Table also shows that neither the size of the competitor field nor the relative ability is associated with overconfident announcements for depth and distance competitions. In time contests, higher relative ability decreases the probability of observing overconfid-

[12] Peer effects on performance (see, e.g., Falk and Ichino, 2006; Mas and Moretti, 2009; Carrell et al., 2009; Jane, 2015) and risk-taking (e.g. Yechiam et al., 2008; Cooper and Rege, 2011; Lahno and Serra-Garcia, 2015) are well documented in the literature.

[13] Note that we exclude all observations where a disqualification was issued. The reason is that we do not have clear information on the actual performance for these observations and thus cannot unambiguously measure overconfidence.

[14] Since we use a binary dependent variable to measure gender difference in overconfidence, the estimates presented in Table 4–the differences in overconfidence between men and women in percentage points–are scale free and independent of sample size. Hence, there is no need for methods to quantify the effect size using a standardized measure like $Cohen's\ d$ as suggested by Nelson (2015).

ence. The average estimated marginal effects from probit regressions are very similar, and these results are reported in Table A.1 of Appendix A. Taken together, we conclude that, on balance, men show more overconfidence than women.

Table 4: Gender and overconfidence (RP $<$ AP)

comp. type:	**depth**		**distance**		**time**	
	(1)	(2)	(3)	(4)	(5)	(6)
female	-0.044*** (0.008)	-0.038*** (0.011)	-0.012*** (0.003)	-0.011*** (0.004)	-0.014*** (0.003)	-0.017*** (0.005)
number of competitors		0.001 (0.001)		0.000 (0.000)		-0.000 (0.000)
relative ability[a] (in distance)		-0.000 (0.000)		0.000 (0.000)		
relative ability[a] (in time)						-0.004*** (0.001)
event FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
discipline FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
mean dep. var.	0.214		0.039		0.038	
N	10,589		17,997		9,898	
R^2	0.086	0.086	0.076	0.076	0.132	0.133

Notes: Robust standard errors clustered on the competition level in round parentheses. *, ** and *** indicate statistical significance at the 10%, 5%, and 1% level. All specifications include competition event-effects. The dependent variable is 1 if the announce performance exceeds the realised performance ('overconfident'), 0 otherwise.
[a] *relative ability* is the athlete's observed relative ability compared to the other contestants in a competition. It is defined as the athlete's best prior performance relative to the mean of the competitors' prior best performances. An increase in this measure is equivalent to a relative improvement compared to all opponents.

Disqualifications and blackouts One potential outcome of overconfidence is an invalid attempt. In most cases, this is due to a hypoxic loss of consciousness (blackout) or a failure to complete surface protocol: that is, when athletes drastically overestimate their abilities.

To investigate this alternative binary measure of overconfidence (and its consequences), we estimate model (1) using a binary measure which equals 1 if an athlete was disqualified (zero otherwise) as the dependent variable. The results are presented in the first three columns of Table 5. We estimate a significant lower disqualification probability for women of 2.5 (depth), 1.4 (distance), and 2.2 (time) ppts.

The most severe consequence of overconfidence (and a particular type of disqualification) is to black out during the competition, because it not only leads to a disqualification but also carries the risk of lasting medical consequences. To examine the gender differences in overconfidence manifested in losing conscious during competition, we use an additional dependent variable equal to 1 if the athlete passed out during the dive (zero otherwise).

Results are presented in columns 4 to 6 of Table 5. We find a significant gender gap in the probability of a blackout of 0.9 and 0.7 ppts in distance and time competitions (significant at the 5% level). For depth competitions, the estimated β_1 is smaller and not significantly different from zero. This finding might be best explained by the life-threatening risks of a loss of consciousness at a depth of say 60 metres. Finally, there is also some evidence that the number of competitors and our proxy for (relative) ability slightly increase the probability of a disqualification and a blackout (for depth and time competitions). This could be interpreted as peer effects.

Table 5: Gender and overconfidence: probability of disqualification and blackout

	disqualification			*blackout*		
	depth	**distance**	**time**	**depth**	**distance**	**time**
female	-0.025***	-0.014**	-0.022***	-0.000	-0.009**	-0.007**
	(0.007)	(0.006)	(0.006)	(0.004)	(0.004)	(0.003)
number of competitors	0.000 (0.001)	0.001*** (0.000)	0.000 (0.000)	0.001 (0.001)	-0.000 (0.000)	-0.000 (0.000)
relative ability^{a} (in distance)	0.001*** (0.000)	-0.000 (0.000)		0.000*** (0.000)	0.000 (0.000)	
relative ability^{a} (in time)			0.004** (0.002)			0.002** (0.001)
event FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
discipline FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
mean dep. var.	0.090	0.084	0.072	0.027	0.030	0.018
N	11,632	19,654	10,661	11,632	19,654	10,661
R^2	0.088	0.082	0.119	0.067	0.069	0.099

Notes: Standard errors (clustered for competition ID) in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All specifications include competition fixed effects. The dependent variable for columns (1) to (3) is 1 if the observed competitor's attempt results in a disqualification, 0 if the attempt is valid. The dependent variable for columns (4) to (6) is 1 if the observed competitor's attempt results in the diver blacking out during any stage of the attempt, 0 otherwise.

[a] *relative ability* is the athlete's observed relative ability compared to the other contestants in a competition. It is defined as the athlete's best prior performance relative to the mean of competitors' previous best performances. An increase in this measure is equivalent to a relative improvement compared to all opponents.

Gender and overconfidence: intensive margin. Next, we aim to quantify the level of overconfidence. That is, we estimate its (gender-specific) intensive margin. Therefore, two additional covariates were added to model (1): (i) *overconfident* is a binary variable which equals 1 if the announced performance exceeds the realised performance ($AP > RP$) (zero otherwise), and (ii) *female x overconfident* is the interaction term between *overconfident* and the gender dummy.[15]

Table 6 presents the results. It shows that overconfident athletes, on balance, announce about 3.1 metres more in terms of depth, 35.4 metres more in terms of distance, and 1.3 additional minutes in terms of time compared to non-overconfident athletes (columns 1 to 3). In terms of the difference between announced and realised performance, we estimate a negative and significant coefficient for *female x overconfident* for depth competitions (column 4). Here, the difference between AP and RP is about 1.2 metres less in absolute terms for overconfident women than for overconfident men. To qualify this difference, we follow Nelson (2015) and find that $Cohen's\ d = -0.162$ ($SE = 0.040$), which is a comparatively low value. Moreover, we do not find any gender gaps when we standardise the extend of overconfidence. We thus conclude that gender differences on the intensive margin are negligible.

Table 6: Gender and overconfidence: intensive margin

	announcement			*announcement - realisation*		
	depth	**distance**	**time**	**depth**	**distance**	**time**
female	-12.058***	-6.229***	-0.398***	0.245**	4.827***	0.110***
	(0.489)	(0.886)	(0.038)	(0.112)	(1.043)	(0.039)
overconfident	3.142***	35.395***	1.278***	12.773***	70.712***	2.709***
	(0.329)	(2.463)	(0.079)	(0.342)	(2.604)	(0.092)
female×	-0.001	-6.788**	0.129	-1.221**	-5.024	0.067
overconfident	(0.553)	(3.407)	(0.143)	(0.538)	(3.988)	(0.179)
number of	0.051	-0.130**	0.009***	0.044***	0.106	0.005*
competitors	(0.043)	(0.066)	(0.003)	(0.013)	(0.072)	(0.003)
relative ability^a	0.544***	0.273***		0.011***	-0.160***	
(in distance)	(0.010)	(0.008)		(0.003)	(0.007)	
relative ability^a			0.293***			-0.247***
(in time)			(0.017)			(0.015)
event FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
discipline FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
mean dep. var.	59.190	68.249	2.797	2.677	-45.436	-2.158
N	10,589	17,997	9,898	10,589	17,997	9,898
R^2	0.760	0.302	0.330	0.504	0.302	0.340

Notes: Standard errors (clustered for competition ID) in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All specifications include event and discipline fixed effects. ^{a} *relative ability* is the athlete's observed relative ability compared to the other contestants in a competition. It is defined as the athlete's best previous performance relative to the mean of the competitors' previous best performances. An increase in this measure is equivalent to a relative improvement compared to all opponents.

The role of experience Intuitively, one may expect that learning in the sense of acquiring self-knowledge should improve the accuracy of self-estimation (e.g., Clark and

[15] Note that we refrain from using a relative measure of performance announcements or divergences, because the gender dummy already controls for gender differences in performance levels.

Friesen, 2009).[16] Thus, in the case of freediving, more experienced divers should be less overconfident. However, there is also theoretical and empirical work on the dynamics of overconfidence which suggests that overconfidence may also persist or even increase with experience by virtue of ego-preserving biases and the 'good-news, bad-news' effect for example (e.g. Gervais and Odean, 2001; Mobius et al., 2011; Eil and Rao, 2011; Grossman and Owens, 2012).

For our analysis, experience is proxied by the number of (observed) dives before attempt i. Then, a median split between those with one to four attempts versus those with five or more attempts results in two groups: *low* and *high* (experience). In the same way, $experienced$ is binary variable, such that $experienced = 1$ indicates an athlete with five or more attempts before attempt i, and $experienced = 0$ otherwise. We then reestimate model 1 for both groups and for the whole sample.

Table 7 shows the results. More experienced divers are 1.2 ppts less likely to blackout or to be disqualified with no gender difference (column (6)). Since the strategic use of announcements in time and distance competitions (see Section 2) does not affect the probability of disqualification and blackout, we use data from all disciplines.[17]

On the contrary, the sample is restricted to depth competitions when we estimate the effect of experience on the probability of being overconfident (columns (1) to (3)). We find that experienced divers are 8.7 ppts *more* likely to be overconfident. However, the significant and negative coefficient for the *female x experienced* interaction indicates that this effect is driven by male divers (column (3)). The estimates presented in columns (1) and (2) reflect this finding. For experienced divers, we estimate a gender gap of 6.5 ppts in the probability to be overconfident, whereas the estimate for the low-experience group is -0.4 ppts and insignificant.

In summary, our main results indicate that female divers are less likely to be overconfident. There is a significant and sizeable gender gap in the probability of overconfidence and different indicators for overconfidence, and experience appears to be an

[16] In psychology, this is called the Dunning–Kruger effect (Kruger and Dunning, 1999).

[17] Athletes with records from 2006 (i.e., the first year in our sample) are excluded, because we lack of data on their careers before the observation period. Note also that due to different sample size, the median value of our proxy for experience also differs.

important driver. For athletes showing overconfidence, we do not find a substantial gender gap in the extend of overconfidence.

Table 7: Gender and overconfidence - the role of experience

	overconfident[a]			*blackout or disqualified*[b]		
	(1)	(2)	(3)	(4)	(5)	(6)
experience:	low	high	pooled	low	high	pooled
female	-0.004	-0.065***	0.001	-0.024***	-0.025***	-0.027***
	(0.016)	(0.018)	(0.015)	(0.007)	(0.008)	(0.006)
number of competitors	0.001 (0.002)	0.003 (0.002)	0.002 (0.001)	0.001 (0.001)	0.000 (0.001)	0.000 (0.000)
relative ability[c]	-0.000	-0.002***	-0.001***	0.000**	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
experienced			0.087***			-0.012**
			(0.016)			(0.006)
female× experienced			-0.068*** (0.020)			0.005 (0.008)
event FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
discipline FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
mean exp.	1.45	16.26	8.42	1.07	13.07	7.59
max exp.	4	116	116	3	118	118
R^2	4,437	3,936	8,373	13,216	12,409	25,625
N	0.140	0.143	0.099	0.121	0.119	0.085

Notes: Robust standard errors clustered on the competition level in round parentheses. *, ** and *** indicate statistical significance at the 10 % level, 5 % level, and 1 % level. All specifications include event- and discipline-fixed effects. The dependent variable is 1 if the announce performance exceeds the realised performance ('overconfident'), 0 otherwise. [a] The dependent variable is equal to 1 if the observed attempt is characterized by $AP > RP$, 0 otherwise. Invalid attempts are excluded. [b] The dependent variable is equal to 1 if the observed attempt is invalid and thus results in a disqualification, 0 otherwise. All competition types (depth, distance and time) are used. [c] *relative ability* is the athlete's observed relative ability compared to the other contestants in a competition. It is defined as the athlete's best previous performance relative to the mean of the competitors' previous best performances. An increase in this measure is equivalent to a relative improvement compared to all opponents.

3.2 Overconfidence and relative performance

So far, the analysis has focused on the consequences of overconfidence in terms of disqualifications and blackouts. In this section we examine how overconfidence affects (relative) performance. This is important because it makes a difference in understanding biases in the decision-making process whether overestimation causes only moderate or severe losses. As measures of performance, we use diver *i*'s (i) final rank and (ii) the difference in final scores compared to the competition's best attempt. Since announcements are more reliable in depth competitions (see Section 2), only data drawn from these kind of freediving disciplines is used in the regressions.[18]

An obvious issue when comparing men and women is the need to control for unobserved heterogeneity in diver characteristics, such as age and differences in abilities. We therefore split the overall sample into women's and men's competitions and estimate the following fixed effects model:

$$Y_{i,d,c} = \gamma_0 + \gamma_1 \, overconfident_{i,d,c} + \gamma_2 \, ability_{i,d,c} + \psi_c + \lambda_d + \mu_{i,d,c} \ , \tag{2}$$

where $Y_{i,d,c}$ is the outcome of interest (relative performance) for diver d in attempt i at competition c. Again, $overconfident_{i,d,c}$ equals one if the observed athlete falls below the previous announcement (RP $<$ AP) at attempt i, zero otherwise. Then, γ_1 measures the effect of overconfidence (RP $<$ AP) on relative performance. Since we focus on the average ability of all other competitors as a key factor for the effect of overconfidence on relative performance, we control for diver i personal ability, measured as their personal best recorded in our data before competition c ($ability_{i,d,c}$). We use diver fixed effects (λ_d) to control for unobserved heterogeneity (e.g. in competitiveness) of competing divers. Competition fixed effects (ψ_c) account for time trends, unobserved heterogeneity in competition characteristics, and different types of competition formats.

Since we expect men's and women's contests to differ in the depths achieved and in

[18] For full transparency and as a robustness check, we replicated our analysis for distance and time competitions. The results can be found in the Appendix.

the number of athletes, relative performance measures are defined as

$$\textit{rel. rank}_{i,d,c} = \frac{rank_{i,d,c}}{max.\ rank_c} * 100 \tag{3}$$

and

$$rel.\ scoredifference_{i,d,c} = \frac{max.\ score_c - score_{i,d,c}}{max.\ score_c} * 100\ , \tag{4}$$

where $max.\ rank_c$ is the total number of valid attempts and $max.\ score_c$ is the score for the best attempt in competition c. The rank is calculated based on single attempts.

Figure 3 illustrates the distributions of both absolute (*rank* and *score difference*) and relative measures of performance (*rel. rank* and *rel. score difference*) by gender. Note that $rel.\ scoredifference$ has a maximum distinctively exceeding 100 due to the fact that both positive and negative scores are possible. The Figure shows that the distribution of score differentials in female contests is more skewed to the right compared to the distribution for males, whereas the distributions are close to each other for *rel. score difference*.

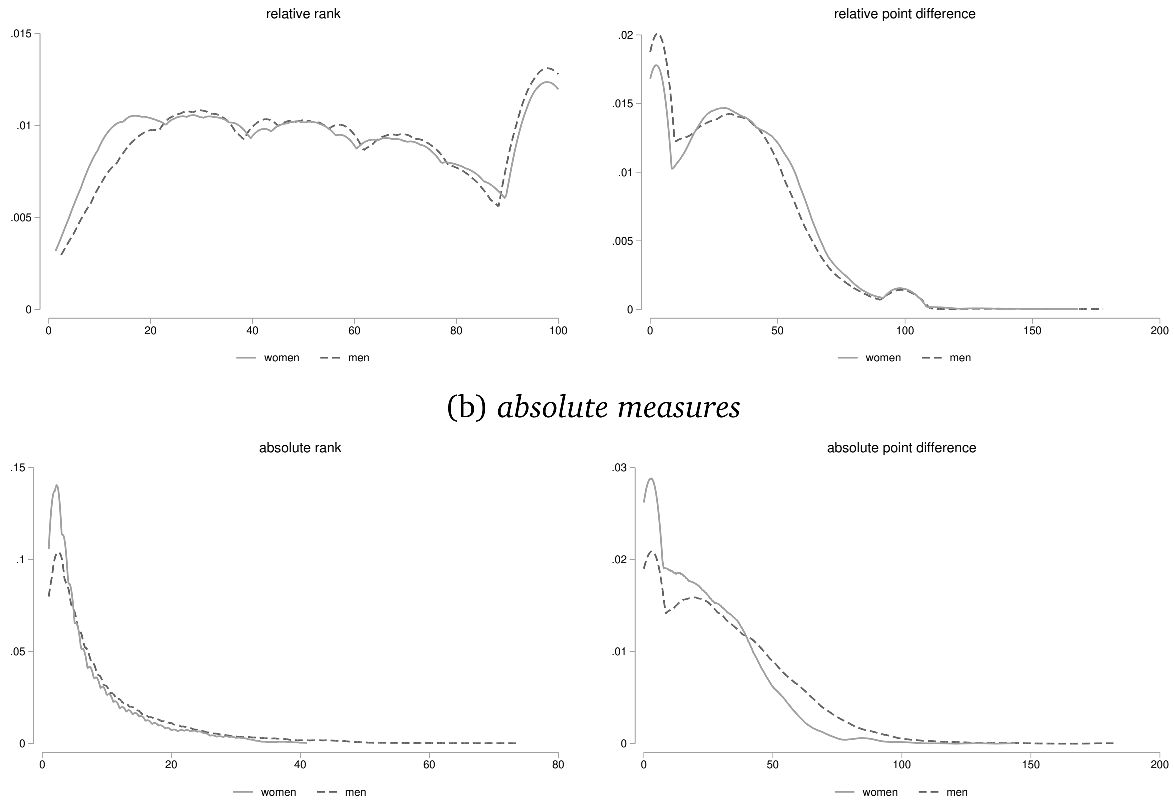
Table 8 presents the results from regression analyses.[19] First, Panel B indicates that the estimated γ_1 is positive for *rank* and *score difference*, meaning that overconfidence affects performance negatively.[20] Moreover, the estimated coefficients suggest a gender gap to the disadvantage of men.

However, things change when we use relative performance measures. Specifically, Panel A shows that in depth competitions overconfident women lose about 32.3 ppts relative to the best ranking whereas overconfident men lose 31.1 ppts (with the difference being significant at the 5% level). In terms of the score differential (which is the more sensitive measure), overconfident women lose about 28.1 ppts compared to non-overconfident competitors whereas overconfidence has a smaller negative impact

[19] Again, the data used for estimations presented in Table 8 does not include invalid attempts resulting in disqualifications, blackouts or implausible scores like negative depths or miscoded results.

[20] This is because final ranks are assigned to performances in ascending order.

Figure 3: Distribution of performance measures by gender



(a) *relative measures*

(b) *absolute measures*

Notes: Kernel density estimates for absolute and relative performance measures.

of about 24.9 ppts for men. The difference between both estimates is significant at the 1% level.

From this we conclude that female divers, on average, suffer more from being overconfident than male divers. It would be misleading to take absolute measures of performance because of the different absolute levels of performance across gender.

Looking more closely at what drives the result on score differentials, we estimate in auxiliary regressions that overconfident female divers increase their announcement by about 2.9 metres whereas overconfident male divers increase announcements by 3.4 metres. Yet, the relative deviation from the target score (i.e., the score which corresponds to the announced performance)[21] is 40.5 for women and 37.5 for men in cases of overconfidence. That is, missing the announced performance by say three metres comes at a higher cost (in terms of lost points) for female athletes than for male athletes. Detailed results are provided Table A.2 in the Appendix.

[21] The exact definition is $\frac{expected\ score\ from\ AP - actual\ score}{actual\ score} \cdot 100$.

Heterogeneity: underdogs vs. favourites Next, we examine the role that the heterogeneity in abilities among divers plays in the relationship between overconfidence and relative performance. For instance, top athletes may show more pronounced overconfidence because of media attention and sponsors' expectations. Alternatively, the pressure from a strong field of opponents could translate into higher levels of overconfidence.

Like in the beginning of this section, let $PB_{d,c}$ be the personal best of diver d before event c and $\overline{PB}_{-d,c}$ the average personal best of all other participants. Then, in order to investigate the effect of relative ability on the relationship between overconfidence and relative performance, we split the sample along the median of $rel.\ ability = PB_d - \overline{PB}_{-d}$. Consequently, the above-the-median subsample consists of ex ante favourites, which means athletes do better than their competitors, whereas the below-the-median subsample includes the ex ante underdogs.

Figure 4 illustrates the results from estimating model (2) by gender, relative performance measure, and relative ability of the observed diver. Panel (a) illustrates the estimated effect of overconfidence on our relative performance measures, as defined at the beginning of this section. Overconfident female and male divers who are ex ante favourites suffer more from being overconfident compared to underdogs. That is, they lose more in terms of relative ranking and the relative point difference when they are overconfident. Furthermore, the difference between the estimated γ_1 for men and women is statically not different from zero at conventional levels of statistical significance. Results are confirmed for the absolute performance measures (panel (b)).

Why do favourites suffer more from being overconfident than underdogs? Auxiliary regressions with the announced depth, as well as the difference between announced and realised depth as the dependent variable, shed some light on the underlying mechanism. It shows that the difference in announcements between overconfident and non-overconfident athletes is smaller for favourites (females: 2.2 metres, men: 2.8 metres) than for underdogs (women: 3.4 metres, men: 3.5 metres). In contrast, overconfident favourites (females: 13.1 metres, men: 13.8 metres) miss their targets (RP vs. AP) to a larger extent than overconfident underdogs (females: 11.2 metres, men: 12.3 metres). A possible explanation is that the 'choking under pressure' phenomenon is more prevalent among favourites (see, for instance, Harb-Wu and Krumer, 2019).

As a robustness check, we repeat this analysis for the two other types of competitions. Corresponding estimates for distance and time competitions are illustrated in Figures B.2 and B.3 in Appendix B. These estimates largely confirm the previous finding from depth competitions: favourites suffer more from overconfidence than underdogs.

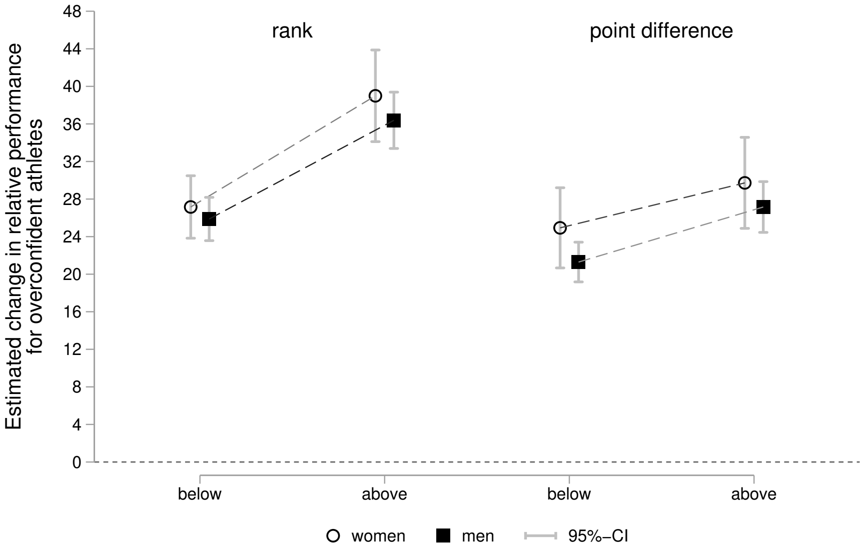
Table 8: Absolute and relative performance measures (depth competitions only)

	rank		*score difference*	
	(1)	(2)	(3)	(4)
	female	**male**	**female**	**male**
Panel A. Relative performance measures [a]				
overconfident[c]	32.304***	31.121***	28.093***	24.874***
	(1.139)	(0.820)	(1.229)	(0.775)
ability[d]	-0.416***	-0.277***	-0.299***	-0.192***
	(0.066)	(0.040)	(0.044)	(0.029)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
difference male-female	-1.183**		-3.219***	
mean dep. var.	57.054	54.490	28.962	31.979
N	3,609	6,980	3,609	6,980
Panel B. Absolute performance measure [b]				
overconfident[c]	5.232***	6.016***	20.194***	21.591***
	(0.402)	(0.443)	(0.821)	(0.621)
ability[d]	-0.067***	-0.062***	-0.270***	-0.213***
	(0.013)	(0.011)	(0.033)	(0.025)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
difference male-female	0.783***		1.397***	
mean dep. var.	6.904	9.126	21.168	28.610
N	3,609	6,980	3,609	6,980

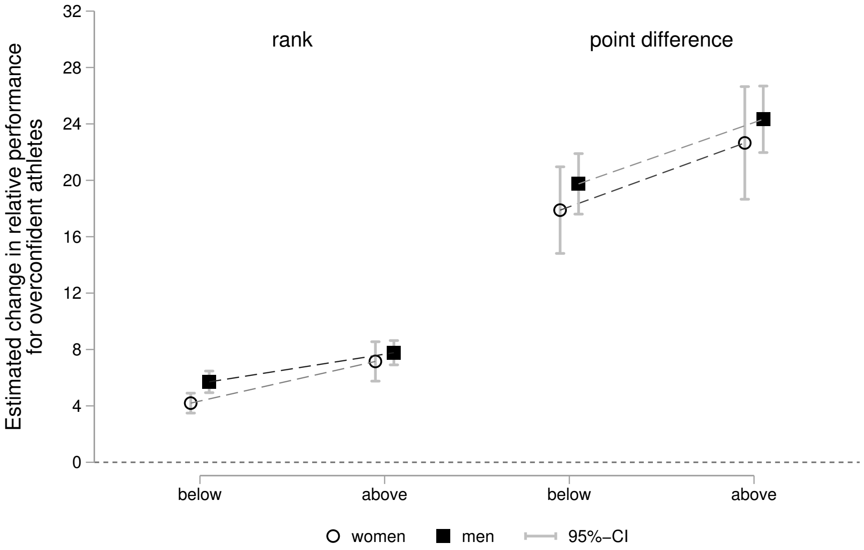
Notes: Robust standard errors clustered on the diver level in round parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level. [a] The dependent variables are *rel. rank* (columns (1) and (2)) and *rel. score difference* (columns (3) and (4)). [b] The dependent variables are the ranking (columns (1) and (2)) and the (absolute) score difference to the best recorded performance in the actual competition (columns (3) and (4)). [c] $overconfident$ equals one if the announced performance is below the recorded performance, zero otherwise. [d] $ability$ is the personal best performance so far for the observed diver in depth competitions.

Figure 4: Effect heterogeneity: underdogs vs favourites

(a) *relative measure*



(b) *absolute measure*



Notes: All illustrated coefficients are estimated for different samples. Full estimation reports are presented in the Appendix. *below* and *above* indicate the past-performance level of the diver who performs the observed attempt, relative to all other competitors.

3.3 Peer information, decision-making, and performance

In the earlier sections, we analysed how the nexus between overconfidence and (relative) performance varies across gender. Additionally, it has been found that underdogs suffer less from overconfidence than favourites. We now take an additional step by investigating how an increase in competition affects self-estimation and performance. That is, we want to know how peer information affects the decision-making process with regard to the announced performance. For instance, Schoenberg and Haruvy (2012) document larger laboratory asset-market bubbles when subjects possess information about the performance of top traders in their group. In our setting, contestants may react to a strong field of competitors by choosing higher values for AP so that peer pressure causes athletes to take more risks.

Yet, the drawback of using plain announcements is that the announcement depends on the (general) level of performance, and the level of performance is different for male and female divers. Although we can control for ability in our empirical model (past best performance, diver fixed effects), a measure of confidence which already incorporates ability would be superior. Therefore, we propose a relative measure of what we call (self-)*confidence* which relates diver d's personal best prior to event c to the actual announcement:

$$confidence_{i,d,t,k} = \frac{PB_{d,c} - AP_{i,d,t,k}}{AP_{i,d,t,k}} * 100. \tag{5}$$

In other words, $confidence$ is a measure of how close the announcement is to the athlete's previous best result.[22] We interpret this as (self-)confidence since the decision to go for the limit typically takes a certain level of self-assurance in your personal ability.

Our aim is to measure the causal effect of (self-)confidence on performance. An obvious issue is that despite the use of individual fixed effects, it cannot be ruled out that unobserved variables will bias our results. For example, short-term fluctuations in performance potentials (i.e., form on the day) would influence both the announcement

[22] Figure B.4 in Appendix B illustrates the distribution of the variable.

and the performance. OLS estimates would therefore be biased.

To establish a causal link between $confidence$ and (relative) performance, we follow Böheim and Lackner (2015) and use information on the previous best performances of competitors as an exogenous shock on the diver's announcement. This information is publicly available and it is highly plausible that athletes have (implicit) knowledge about the past performance of their competitors. First, we define the average of the previous personal bests of all other (n_c) participants in depth competition c excluding diver d as

$$\overline{PB}_{-d,c} = \frac{1}{n_c}\sum_{-d=1}^{n_c} personal\ best_{-d,c}. \tag{6}$$

Then, our instrument $Z_{i,d,c}$ is equal to 1 if $\overline{PB}_{-d,c}$ for a given competition c is in the fourth quartile of the overall distribution of $\overline{PB}_{-d,c}$ for all competitions in our sample, 0 otherwise. In other words, $Z_{i,d,c} = 1$ indicates a strong field of athletes.

Finally, our model can be written as

$$\begin{aligned} rel.\ performance_{i,d,t,k} &= \alpha + \gamma_1\ \overline{confidence}_{i,d,t,k} + \gamma_2\ ability_{i,d,t} + \\ &+ \gamma_3\ no.\ of\ competitors_c + \pi_t + \lambda_d + \omega_k + \epsilon_{i,d,t,k}, \end{aligned} \tag{7}$$

where the first stage is

$$\begin{aligned} confidence_{i,d,t,k} &= \pi_0 + \pi_1 Z_{i,d,c} + \alpha_1\ ability_{i,d,t} + \alpha_2\ no.\ of\ competitors_c + \\ &+ \pi_t + \lambda_d + \omega_k + \nu_{i,d,t,k}. \end{aligned} \tag{8}$$

The coefficient γ_1 measures the causal effect of a competition-induced increase in *confidence* on realised relative or absolute performance. A negative coefficient would indicate a positive effect on the outcome variable (lower final rank, lower difference to the winner's score).

Our instrument is valid if $cov(Z_{i,d,c}, \epsilon_{i,d,t,k}) = 0$. Hence, a potential concern for our identification strategy is that the competitors' abilities may affect diver d's performance. We therefore control for the diver d's ability $ability_{i,d,t}$ (proxied by diver d's personal best in previous depth competitions). The number of competitors, $no.\ of\ competitors_c$, controls for the size of the competition. We also use diver (λ_i), discipline (ω_k), and year (π_t) fixed effects.

The exclusion restriction holds as the average personal bests of all the other participants in a contest have no direct effect on the dependent variable when we control for the difference between diver d's personal best and the average personal best of all other contestants. Furthermore, we expect our instrument to be valid, because a stronger field of competitors will put pressure on athletes to increase their announcement in order to retain their chances of winning.

The LATE estimates of model (7) are presented in Table 9. We find that an increase in our measure of (self-)confidence by 1 causes an increase in the relative final rank of 2.4 percentage points for female divers and 1.5 percentage points for male divers. In addition, the score ratio to the contest winner increases by 1.3 percentage points for women and 0.81 percentage points for men. Hence, a competition-induced boost in *confidence* decreases relative performance. The gender gaps are statistically significant at the 5% level. Furthermore, the F-statistics from the first-stage regressions imply that we do not have to worry about weak instruments.

The first stage results also indicate that men in our sample react more strongly to an increase in competition than women: $\hat{\pi}_1 = 9.852$ for male divers and $\hat{\pi}_1 = 6.717$ for female divers. This is in line with previous findings, such as Gneezy and Rustichini (2004). Yet, despite the stronger reaction, male divers suffer less from competition-induced variations in *confidence*.

A comparison of the instrumental variable results with those obtained using simple OLS regression reveals that the OLS results indicate a positive association between pre-announcements (i.e., *confidence*) and relative performance. This result is not surprising, given that better athletes not only announce larger target depths but also reach greater depths in contests. Thus, the OLS estimates suffer from a classic ability bias. Nevertheless, the OLS results confirm a stronger effect for men than for women.[23]

[23] There is evidence that peer information in a contest environment may affect favourites and underdogs in different ways (e.g. Brookins et al., 2016). However, we refrain from splitting the overall sample of

Table 9: Causal effect of announcements on relative performance

	final rank		*points difference*	
	female	**male**	**female**	**male**
confidence^a	2.366*** (0.507)	1.516*** (0.190)	1.322*** (0.295)	0.791*** (0.118)
difference (male - female)^b	-0.849 **		-0.530**	
ability (in metres depth)	2.851*** (0.703)	1.776*** (0.261)	1.546*** (0.402)	0.876*** (0.156)
number of competitors	-0.999*** (0.179)	-0.518*** (0.084)	0.378*** (0.107)	0.433*** (0.053)
Diver FEs	*Yes*	*Yes*	*Yes*	*Yes*
Discipline FEs	*Yes*	*Yes*	*Yes*	*Yes*
Year FEs	*Yes*	*Yes*	*Yes*	*Yes*
first stage coefficient	6.717*** (1.204)	9.852*** (0.947)	6.717*** (1.204)	9.852*** (0.947)
F stat.	31.116	108.172	31.116	108.172
OLS^c	-0.269*** (0.040)	-0.231*** (0.021)	-0.175*** (0.018)	-0.152*** (0.014)
N	3,609	6,980	3,609	6,980

Notes: Robust standard errors clustered on the diver level in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level. ^{a} *confidence* is a measure of how close the announcement is to the diver's previous best result (see equation (5)). ^{b} A permutation test (5, 000 iterations) was used to calculate the statistical significance of the difference between the estimated coefficients of *confidence*. *, **, and *** indicate a rejection of the of the null hypothesis of equal coefficients at the 10%, 5%, and 1% level, respectively. ^{c} Estimates of γ_1 with simple OLS regressions.

our IV estimates into (relative) favourites and underdogs for each competition since this split would be endogenous.

4 Conclusion and discussion

Using data from a large number of professional freediving competitions, we investigate gender differences in a facet of overconfidence called overestimation. The specifics of freediving contests provide an ideal opportunity to assess the extensive and intensive margin of overestimation. In particular, we compare actual to announced performances of female and male athletes in different types of competitions (chiefly, depth competitions).

As a result, we find evidence that women are, on balance, less likely to show overconfidence. Specifically, female divers are about 18% less overconfident than male divers. This finding is robust to different measures of overconfidence like being disqualified or blacking out. The fact that we can confirm prior experimental findings of men having a higher tendency towards overconfidence in a real-world contest situation with high stakes where a bias in self-assessment can pose serious health threats is notable. This is because studies like Azmat et al. (2016) suggest that raising the stakes decreases gender gaps in performance.

Furthermore, our data allows us to calculate the intensive margin of overestimation. To the best of our knowledge, this has not been done before in a non-experimental setting. We find a negligible gender gap in the extent of overestimation (i.e., the difference between announced and actual performance) of about 1.4 metres in depth competitions which completely disappears when we use standardised measures.

Regarding the consequences of overconfidence, we provide evidence that female athletes suffer more from overconfidence than male athletes in terms of their relative performances in contests. This result is robust to employing a fixed-effect strategy to account for unobserved ability differences. Furthermore, we find that better (i.e., above-the-median) athletes lose more in terms of relative ranking and the relative point difference when they are overconfident. Finally, we use an instrumental variable approach to investigate the effects competition-induced variations in *confidence* (which is a measure of how close the announcement is to the athlete's prior best result) on relative performance. Results show that whereas male athletes react more strongly to an increase in competition, female athletes suffer more from the adjustment of *confidence*. Against this background, the fact that women in our sample are (consciously or unconsciously) less vulnerable to overconfidence on the extensive and intensive margin appears as rational behaviour.

Our analysis also indicates that experience improves the accuracy of self-estimation in terms of blackouts and disqualifications. In other words, athletes learn to decrease the risk of a zero score and, above all, the risk for health. When it comes to overconfidence in terms of missing the announced performance, we find that these gender gaps can be attributed to more experienced men. That is, it appears that male divers 'learn' to be overconfident. While there are some explanations for overconfidence that grows with experience like ego-preserving biases (Gervais and Odean, 2001) and herding (Avery and Chevalier, 1999), it might also be the case that, in a competitive setting like freediving, announcements may also have a goal-setting role. For instance, Clark et al. (2016) document that goal setting can motivate college students to work harder and achieve better outcomes. The authors also highlight gender differences in the effectiveness of task-based goals in favor of male students. In Dalton et al. (2016), self-chosen goal contracts in a laboratory experiment have a positive impact on performance for men but not for women (compared to piece rate contracts). The fact that we find negative effects of overconfidence on relative performance does not necessarily contradict this hypothesis as there is no way to measure the performance-enhancing effect of ambitious announcements.

Our results have considerable implications for financial markets and other markets prone to bubble formation. As suggested by Eckel and Füllbrunn (2015), a higher share of female investors whose decisions are less biased by overconfidence could have a dampening effect on magnitudes and likelihood of speculative bubbles.

References

Alkan, N. and Akış, T. (2013). Psychological characteristics of free diving athletes: A comparative study. *International Journal of Humanities and Social Science*, 3(15):150–157.

Anbarci, N., Lee, J., and Ulker, A. (2016). Win at all costs or lose gracefully in high-stakes competition? gender differences in professional tennis. *Journal of Sports Economics*, 17(4):323–353.

Association Internationale pour le Développement de l'Apnée (AIDA) (2020). Competition rules and guidelines. `https://www.aidainternational.org/Documents`. Accessed July 2020.

Astebro, T., Herz, H., Nanda, R., and Weber, R. A. (2014). Seeking the roots of entrepreneurship: Insights from behavioral economics. *Journal of Economic Perspectives*, 28(3):49–70.

Avery, C. N. and Chevalier, J. A. (1999). Herding over the career. *Economics Letters*, 63(3):327–333.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6):1372–1400.

Bengtsson, C., Persson, M., and Willenhag, P. (2005). Gender and overconfidence. *Economics Letters*, 86(2):199–203.

Bertrand, M. (2011). New perspectives on gender. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, volume 4B, chapter 17, pages 1543–1590. Elsevier.

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., and Milkman, K. L. (2015). The effect of providing peer information on retirement savings decisions. *The Journal of Finance*, 70(3):1161–1201.

Beyer, S. and Bowden, E. M. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, 23(2):157–172.

Böheim, R. and Lackner, M. (2015). Gender and risk taking: evidence from jumping competitions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4):883–902.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Brookins, P., Brown, J., and Ryvkin, D. (2016). Peer information and risk-taking under competitive and non-competitive pay schemes. Working Paper 22486, National Bureau of Economic Research.

Campbell, W. K., Goodie, A. S., and Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, 17(4):297–311.

Carrell, S. E., Fullerton, R. L., and West, J. E. (2009). Does your cohort matter? measuring peer effects in college achievement. *Journal of Labor Economics*, 27(3):439–464.

Cherouveim, E. D., Botonis, P. G., Koskolou, M. D., and Geladas, N. D. (2013). Effect of gender on maximal breath-hold time. *European Journal of Applied Physiology*, 113(5):1321–1330.

Clark, D., Gill, D., Prowse, V., and Rush, M. (2016). Using goals to motivate college students: Theory and evidence from field experiments. *Review of Economics and Statistics*, pages 1–45.

Clark, J. and Friesen, L. (2009). Overconfidence in forecasts of own performance: An experimental study. *The Economic Journal*, 119(534):229–251.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660.

Cooper, D. J. and Rege, M. (2011). Misery loves company: Social regret and social interaction effects in choices under risk and uncertainty. *Games and Economic Behavior*, 73(1):91–110.

Cooper, M. and Kuhn, P. (2020). Behavioral job search. In Zimmermann, K. F., editor, *Handbook of Labor, Human Resources and Population Economics*, pages 1–22. Springer International Publishing, London (UK).

Dalton, P. S., Gonzalez, V., and Noussair, C. N. (2016). Self-chosen goals: Incentives and gender differences. *CentER Discussion Paper Series No. 2016-036*.

Dechenaux, E., Kovenock, D., and Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101(2):556–590.

Dreber, A., von Essen, E., and Ranehill, E. (2014). Gender and competition in adolescence: task matters. *Experimental Economics*, 17(1):154–172.

Dujic, Z. and Breskovic, T. (2012). Impact of breath holding on cardiovascular respiratory and cerebrovascular health. *Sports Medicine*, 42(6):459–472.

Eckel, C. C. and Füllbrunn, S. C. (2015). Thar she blows? gender, competition, and bubbles in experimental asset markets. *American Economic Review*, 105(2):906–20.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

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

Frick, B. (2020). Gender differences in risk-taking and sensation-seeking behavior: Empirical evidence from “extreme sports”. Mimeo.

Gamba, A., Manzoni, E., and Stanca, L. (2017). Social comparison and risk taking behavior. *Theory and Decision*, 82(2):221–248.

Genakos, C. and Pagliero, M. (2012). Interim rank, risk taking, and performance in dynamic tournaments. *Journal of Political Economy*, 120(4):782–813.

Gervais, S. and Odean, T. (2001). Learning to be overconfident. *The Review of Financial Studies*, 14(1):1–27.

Glaser, M. and Weber, M. (2010). Overconfidence. In Baker, H. K. and Nofsinger, J. R., editors, *Behavioral Finance: Investors, Corporations, and Markets*, chapter 13, pages 241–258. Wiley (UK).

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Grossman, Z. and Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510–524.

Grubb, M. D. (2015). Overconfident consumers in the marketplace. *Journal of Economic Perspectives*, 29(4):9–36.

Harb-Wu, K. and Krumer, A. (2019). Choking under pressure in front of a supportive audience: Evidence from professional biathlon. *Journal of Economic Behavior & Organization*, 166:246–262.

Hoffman, M. and Burks, S. V. (2020). Worker overconfidence: Field evidence and implications for employee turnover and firm profits. *Quantitative Economics*, 11(1):315–348.

Jane, W.-J. (2015). Peer effects and individual performance: Evidence from swimming competitions. *Journal of Sports Economics*, 16(5):531–539.

Jay, O. and White, M. D. (2006). Maximum effort breath-hold times for males and females of similar pulmonary capacities during sudden face-only immersion at water temperatures from 0 to 33° c. *Applied Physiology, Nutrition, and Metabolism*, 31(5):549–556.

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

Krawczyk, M. and Wilamowski, M. (2017). Are we all overconfident in the long run? Evidence from one million marathon participants. *Journal of Behavioral Decision Making*, 30(3):719–730.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121.

Lahno, A. M. and Serra-Garcia, M. (2015). Peer effects in risk taking: Envy or conformity? *Journal of Risk and Uncertainty*, 50(1):73–95.

Lahtinen, K., Kurra, S., and Nissinen, A. (2015). *Freediving*. Deep Ideas Oy, Finland.

Lundeberg, M. A., Fox, P. W., and Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86(1):114–121.

Malmendier, U. and Tate, G. (2015). Behavioral CEOs: The role of managerial overconfidence. *Journal of Economic Perspectives*, 29(4):37–60.

Mas, A. and Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1):112–45.

Mobius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2011). Managing self-confidence: Theory and experimental evidence. Working Paper 17014, National Bureau of Economic Research.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502.

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.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Pearn, J. H., Franklin, R. C., and Peden, A. E. (2015). Hypoxic blackout: diagnosis, risks, and prevention. *International Journal of Aquatic Research and Education*, 9(3):342–347.

Reuben, E., Wiswall, M., and 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.

Ridgway, L. and McFarland, K. (2006). Apnea diving: long-term neurocognitive sequelae of repeated hypoxemia. *The Clinical Neuropsychologist*, 20(1):160–176.

Schoenberg, E. J. and Haruvy, E. (2012). Relative performance information in asset markets: An experimental approach. *Journal of Economic Psychology*, 33(6):1143–1155.

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.

Spinnewijn, J. (2015). Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association*, 13(1):130–167.

Yechiam, E., Druyan, M., and Ert, E. (2008). Observing others' behavior and risk taking in decisions from experience. *Judgment and Decision Making*, 3(7):493.

A Additional Tables

Table A.1: Gender and overconfidence: missed targets (PROBIT estimates)

	depth		distance		time	
	(1)	(2)	(3)	(4)	(5)	(6)
female	-0.041***	-0.045***	-0.013***	-0.013***	-0.015***	-0.020***
	(0.010)	(0.008)	(0.003)	(0.004)	(0.003)	(0.005)
Add. control variables[a]	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
event FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
discipline FEs	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
mean dep. var.	0.214		0.039		0.038	
N	10,589		17,997		9,898	

Notes: Reported coefficients are marginal effects computed from PROBIT estimations. Robust standard errors clustered on the competition level in round parentheses. *, ** and *** indicate statistical significance at the 10%, 5%, and 1% level. All specifications include competition fixed effects. The dependent variable is 1 if the announce performance exceeds the realised performance (‘overconfident’), 0 otherwise.

[a] control variables as reported in Table 4.

Table A.2: Announcements and relative point loss (depth competitions)

	announced depth		*relative point loss*	
	(1) **female**	(2) **male**	(3) **female**	(4) **male**
overconfident[a]	2.869***	3.410***	40.453***	37.479***
	(0.315)	(0.288)	(1.435)	(0.884)
ability[b]	0.305***	0.209***	0.025	-0.016
	(0.029)	(0.020)	(0.038)	(0.022)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
mean dep. var.	53.048	62.365	8.245	9.637
N	3,609	6,980	3,609	6,980

Notes: Robust standard errors clustered on the diver level in round parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level. [a] $overconfident$ equals one if the announced performance is below the recorded performance, zero otherwise. [b] $ability$ is the personal best performance so far for the observed diver in depth competitions.

Table A.3: Absolute and relative performance measures (distance competitions only)

	rank		*score difference*	
	(1)	(2)	(3)	(4)
	female	**male**	**female**	**male**
Panel A. Relative performance measures [a]				
overconfident[c]	40.404***	40.006***	37.891***	33.857***
	(2.531)	(1.362)	(2.909)	(1.274)
ability[d]	-0.211***	-0.088***	-0.150***	-0.069***
	(0.037)	(0.014)	(0.022)	(0.009)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
difference male-female	-0.398*		-4.034 ***	
mean dep. var.	57.141	53.197	28.160	34.327
N	5,641	12,356	5,641	12,354
Panel B. Absolute performance measure [b]				
overconfident[c]	6.452***	8.298***	28.357***	30.941***
	(0.878)	(0.670)	(2.060)	(1.424)
ability[d]	-0.014	-0.023***	-0.129***	-0.093***
	(0.010)	(0.004)	(0.017)	(0.009)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
difference male-female	1.847***		2.584***	
mean dep. var.	6.859	9.642	21.727	31.601
N	5,641	12,356	5,641	12,354

Notes: Robust standard errors clustered on the competition level in round parentheses. *, ** and *** indicate statistical significance at the 10%, 5%, and 1% level. [a] The dependent variables are *rel. rank* (columns (1) and (2)) and *rel. score difference* (columns (3) and (4)). [b] The dependent variables are the absolute ranking recorded (columns (1) and (2)) and the absolute score difference to the best recorded performance in the observed competition (columns (3) and (4)). [c] Binary variable equal to one if the announced performance is below the recorded performance, zero otherwise. [d] The personal best performance for the observed diver in depth competitions so far.

Table A.4: Absolute and relative performance measures (time competitions only)

	rank		*score difference*	
	(1)	(2)	(3)	(4)
	female	**male**	**female**	**male**
Panel A. Relative performance measures [a]				
overconfident[c]	39.086***	34.306***	36.175***	28.384***
	(4.455)	(2.021)	(4.145)	(1.893)
ability[d]	-3.613**	-3.439***	-1.831**	-2.114***
	(1.481)	(0.816)	(0.774)	(0.562)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
difference male-female	-4.780*		-7.790***	
mean dep. var.	58.989	53.393	21.449	26.698
N	3,276	6,622	3,276	6,621
Panel B. Absolute performance measure [b]				
overconfident[c]	5.457***	7.742***	23.944***	23.970***
	(1.065)	(1.169)	(2.583)	(1.658)
ability[d]	-0.129	-0.903***	-1.156**	-2.095***
	(0.282)	(0.272)	(0.547)	(0.526)
Competition fixed effects	Yes	Yes	Yes	Yes
Diver fixed effects	Yes	Yes	Yes	Yes
difference male-female	2.285*		0.026***	
mean dep. var.	6.310	9.761	15.557	23.155
N	3,276	6,622	3,276	6,621

Notes: Robust standard errors clustered on the diver level in round parentheses. *, ** and *** indicate statistical significance at the 10%, 5%, and 1% level. [a] The dependent variables are *rel. rank* (columns (1) and (2)) and *rel. score difference* (columns (3) and (4)). [b] The dependent variables are the absolute ranking recorded (columns (1) and (2)) and the absolute score difference to the best recorded performance in the observed competition (columns (3) and (4)). [c] Binary variable equal to one if the announced performance is below the recorded performance, zero otherwise. [d] The personal best performance for the observed diver in depth competitions so far.

B Additional Figures

Figure B.1: Number of events by year, 2006–2019

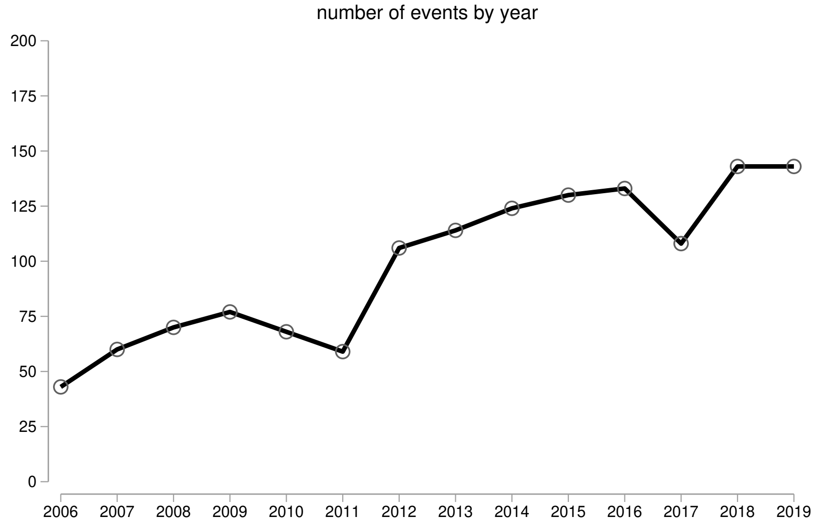
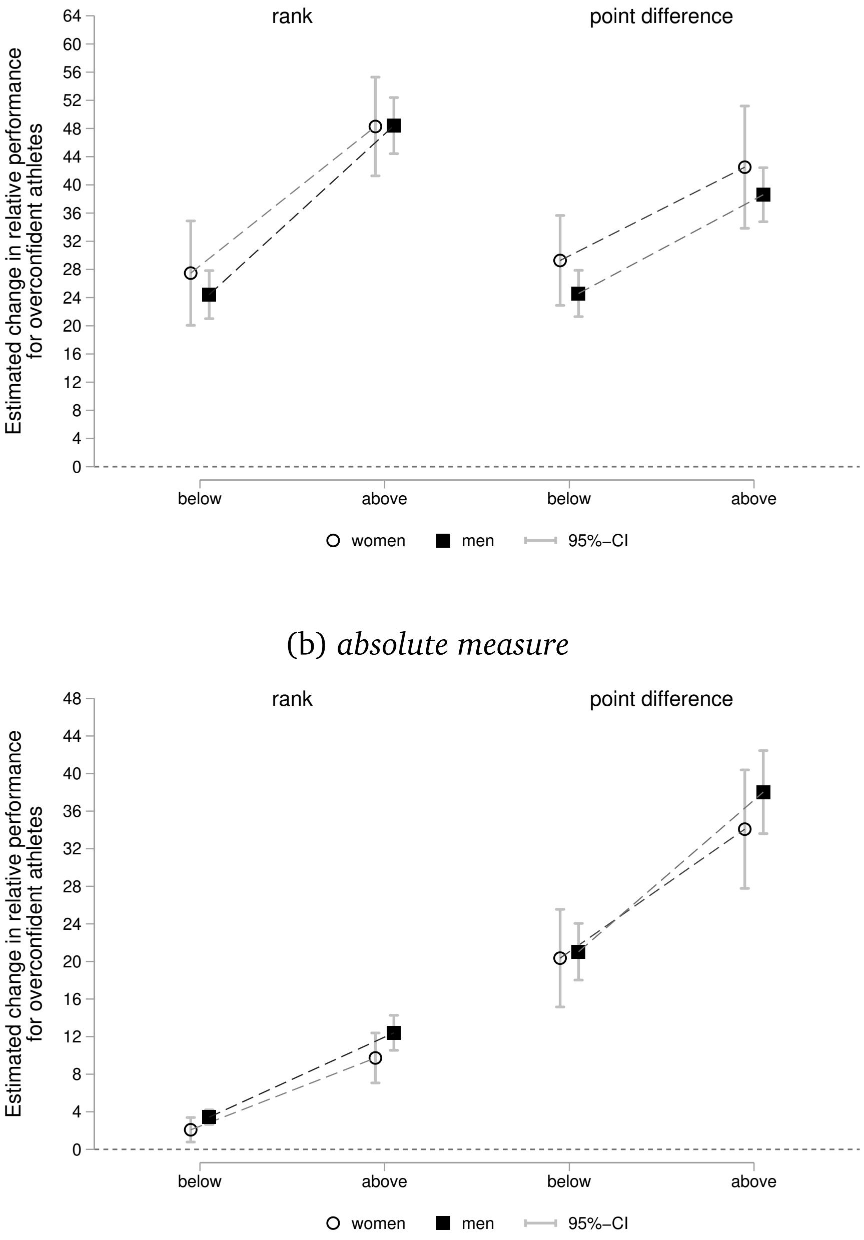


Figure B.2: Effect heterogeneity: relative ability of contestants (distance competitions)



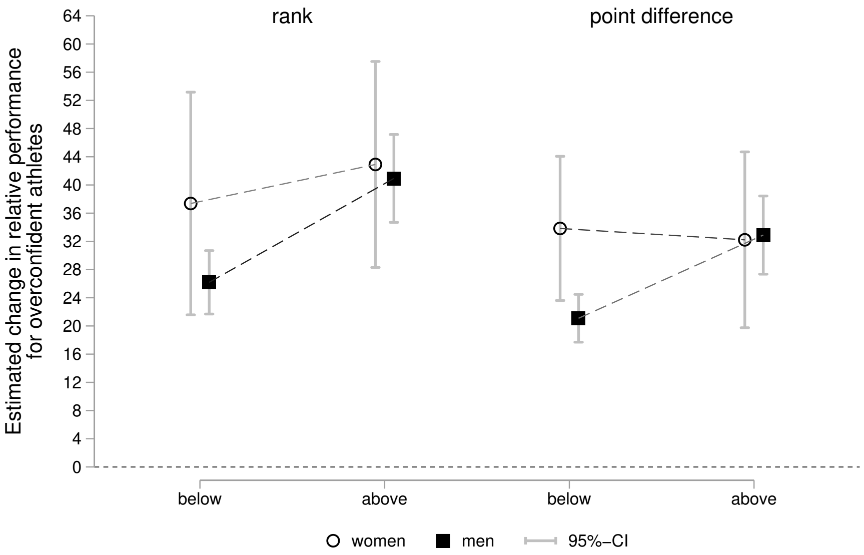
(a) *relative measure*

(b) *absolute measure*

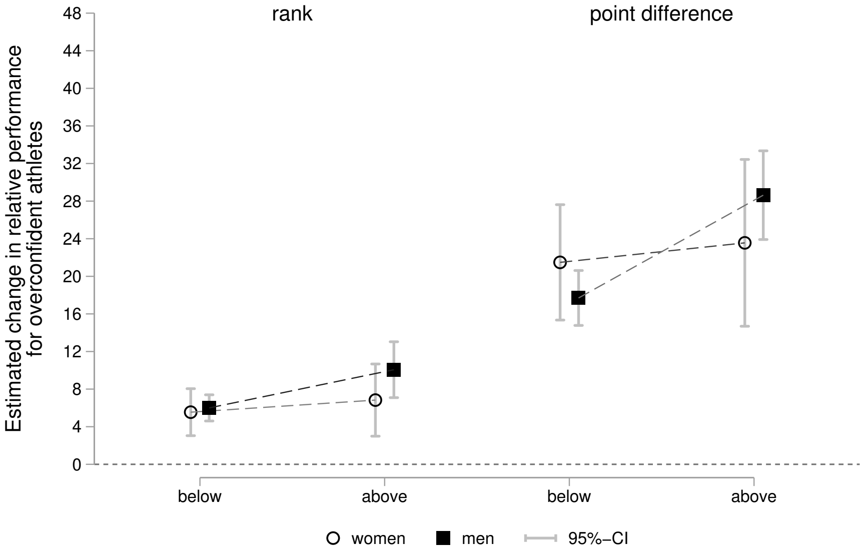
Notes: All illustrated coefficients are estimated for different samples. Only distance competitions are analysed. Detailed estimation results are available upon request. *below* and *above* indicate the past-performance level of the diver performing the observed attempt, relative to all other competitors.

Figure B.3: Effect heterogeneity: relative ability of contestants (time competitions)

(a) *relative measure*

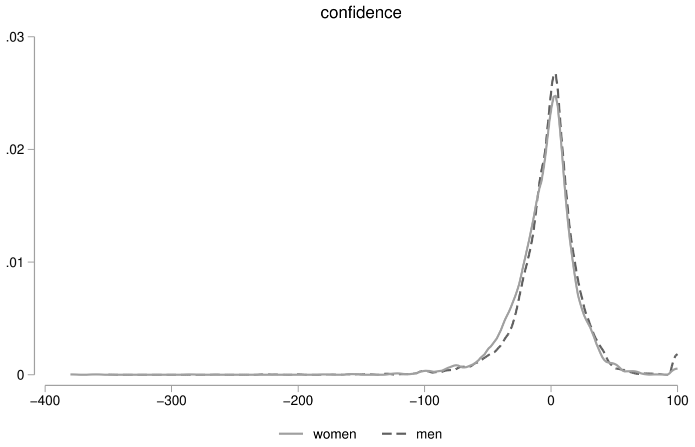


(b) *absolute measure*



Notes: All illustrated coefficients are estimated for different samples. Only distance competitions are analysed. Detailed estimation results are available upon request. *below* and *above* indicate the past-performance level of the diver performing the observed attempt, relative to all other competitors.

Figure B.4: Kernel density estimate of confidence measure.


confidence
women
men