STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's Thesis in Economics
Academic Year 2022-2023

Do Football Players Give Female Coaches the Red Card?

Investing gender bias in football coaching through an experimental approach

Erica Froste Myrin (24025) and Sigrid Holmgren (24177)

Abstract: In Swedish elite football, the skewed gender distribution between coaches is conspicuous, and this thesis aims to investigate one channel as to why there are so few female coaches in this environment. Previous research has found that students evaluate female teachers more critically than male teachers, even in components that they cannot control; might this be the case in Swedish elite football as well? We have conducted a framed field experiment on the teams in the highest football division for players 19 years or under in Sweden. 505 participants watched a video with instructions of a football skill and evaluated it, half of the players were instructed by a female coach and the other half by a male coach. We find no evidence of a gender bias in this sample at the statistical significance level of 5 %. Neither did we find any significant results when controlling for participants being a member of a male or female team, nor when controlling for currently or previously having exposure to a female coach. A series of robustness checks were conducted and we find our results to be robust.

Keywords: gender bias, sport, leadership, football, evaluation

JEL: J160, J710, Z130, Z220

Supervisor:	Anna Dreber Almenberg
Date submitted:	December 5, 2022
Date examined:	December 16, 2022
Discussant:	Loise Hedberg
Examiner:	Kelly Ragan

Acknowledgements

We would like to express our deepest appreciation to Professor Anna Dreber Almenberg, for her feedback and support throughout the process of planning, executing and writing this thesis.

We would also like to extend our thanks to Linn Eriksson at Elite Football Women (EFD) for her invaluable help and engagement in the project and for reaching out to the teams and encouraging them to participate in our study.

Furthermore, we would like to extend thanks to Kristoffer Östlin and Maja Wangerheim for volunteering as football players in our experiment video, and to Jehnny Johansson and Lucas Riedel for the manuscript and voice-overs. We want to thank Jens Grönlund and Felix Fors for letting us run our pilot study on their football teams, and to the Media Committee at the Student Association at the Stockholm School of Economics for lending out their video camera to us for the purpose of filming the video used in the experiment.

We would also like to mention Anton Persson, Mika Lindgren and Sigurd Log Roeran, as co-authors to the research proposal this master thesis built upon. Without them, this master thesis would look a lot different.

Lastly, we would like to recognize Thomas Hasselgren at Swedish Elite Football (SEF) for his help with the conduction of the experiment, to Pontus Granström and Robert Arreman for their football expertise and advice when designing the experiment, and - to all sports managers, academy managers, coaches and players that participated in the research and thus, made this thesis possible.

Contents

1 Introduction **4**

2 Literature Review **8**

2.1 Women in Leadership Positions 8

2.2 Gender and Leadership in Football 10

2.2.1 Macro level 10

2.2.2 Meso level 10

2.2.3 Micro level 12

2.3 Prejudice in Teaching Evaluations 13

2.4 Our Contribution to the Literature 15

3 Method **17**

3.1 Setup and Conditions 17

3.1.1 Experimental design 17

3.1.2 Pilot study 17

3.1.3 Population and sample size 18

3.1.4 Procedure 19

3.1.5 Design of video 20

3.1.6 Design of survey 22

3.1.7 Classification of experiment 23

3.2 Data 23

3.2.1 Dependent variables 23

3.2.2 Independent variables 24

3.2.3 Outliers and exclusion 25

3.3 Statistical Methods 25

3.3.1 Main regression 25

3.3.2 Subanalyses 26

3.3.3 Robustness checks 28

3.3.4 Further statistical considerations 29

3.4 Hypotheses 30

4 Results **32**
4.1 Descriptive Data . . . 32
4.2 Main Regression . . . 33
4.2.1 Main regression . . . 33
4.3 Subanalyses . . . 35
4.3.1 Subanalysis one . . . 35
4.3.2 Subanalysis two . . . 36
4.4 Robustness Checks . . . 36
4.4.1 On each question separately . . . 36
4.4.2 High and low dispersion . . . 38
4.4.3 Treatment question . . . 38
4.5 Summary of Results . . . 38

5 Discussion **40**
5.1 Analysis of Results . . . 40
5.2 Internal Validity . . . 43
5.3 External Validity . . . 44
5.4 Future Research . . . 45
5.5 Going Forward . . . 46

6 Conclusion **48**

References **49**

A Appendix: The Experiment **53**

B Appendix: Robustness Check - Separated by Question **60**

C Appendix: Robustness Check - High and Low Dispersion **63**

D Appendix: Robustness Check - Treatment Question **64**

1 Introduction

Over the last decades, there has been a substantial growth in the number of women in leadership positions across society (Grant Thornton 2022; Economic co-operation and Development 2018). Growth has been especially strong in the Nordic welfare countries, which are seen as forerunners in gender equality (World Economic Forum 2022). Despite this positive trend at the aggregate societal level, the number of female leaders in sports remains low, especially among top management positions (Acosta and Carpenter 2012).

In Sweden, the sports movement is one of the largest movements within civil society. The Swedish Sports Confederation has more than 3,3 million members, which means that about one third of all Swedes are organized members (Riksidrottsförbundet 2022). Of all affiliated sports, football is by far the most popular in terms of active participants (ibid.) and the largest among both women and men (Sportstatistik 2022) - in 2021, more than one million Swedes were active in a football club. Of these, men comprised 68 % of the participants (ibid.). Although not gender equal, yet, in terms of number of participants; the appreciation of women's football, the number of people who consume it and participate in it, has steadily increased (FIFA 2019). During the Summer Olympics in Tokyo 2020, the Women's Final in Football between Sweden and Canada was the most watched broadcast Olympic event in Sweden, with 1,6 million viewers (MMS 2021). Nonetheless, the increasing numbers of female participants and consumers has not been matched with the number of women in top management positions in football. In Sweden's top divisions for men and women (OBOS Damallsvenskan, Allsvenskan, Elitettan, and Superettan), only 6 % of the head coaches were women in the 2022 season. How come this is the case?

The Swedish Sports Confederation has set goals to achieve gender equality. One of the goals is that the share of male/female coaches should be at least 40 % by 2025 within each sport at the youth level as well as at the national team level (Riksidrottsförbundet 2022). To achieve this ambitious goal, more knowledge is needed about the drivers of the gender gap in sports coaching. As football plays an integral role in societies all across the world - not least in Sweden - inequality in football organizations likely reflects other parts of society as well. Research

aiming to reveal gender differences within the football community is therefore not only important for understanding the gender gap in football, but also in order to understand gender inequality in society at large. This is in turn critical, as the gender of an individual should not prevent him/her from reaching decision-making positions.

One explanation for the low number of female coaches in sports is the existence of gender-segregated barriers to promotions and career advancements (LaVoi and Baeth 2018). Despite the lack of available information on how the recruitment process for football coaches in Sweden is structured, it is not unrealistic to assume that men are more likely to be hired for coaching positions given the historical male dominance of the profession. However, there could be other reasons for the low number of female football coaches on the elite level. For instance, could it be the case that football players evaluate female and male coaches differently? Does there exist a gender bias discouraging female coaches? To our surprise and knowledge, there is little to no research on gender bias in the perception of football coaches. Therefore, in this study, we investigate whether gender has an effect on how players assess the quality of coaching. To test for this, we exploit an experimental design on a sample of Swedish elite football players aged 14-20. Gender bias exists if female and male coaches receive different evaluations, which cannot be explained by objective differences in coaching quality. To identify the effect of gender on how the players perceive the coach's ability to instruct a technical skill through video, two versions of a video are randomly assigned to both female and male players. To hold instructor quality constant, the two instruction videos are identical apart from the gender of the coach, who instructs through a voice-over. In addition to the player's subjective evaluation of the coach's ability to instruct, additional data is collected on various control variables.

Our thesis contributes to the literature on 1) female leadership in football, and on 2) gender bias in evaluations, by investigating whether there exists a gender bias in the evaluation of coaches in the Swedish football domain. A large part of the previous literature on female leadership in football is focused on the North American context, and especially the US intercollegiate system (Burton and Leberman 2017). This is a context that differs significantly from that of Sweden, due to the Swedish welfare model and its ideology of equality at all levels and sectors (Hau-

denhuyse, Theeboom, and Skille 2014). As the possibilities and requirements for women's advancement should differ depending on the context, we find it relevant to focus on gender (in)equality within the coaching positions in Swedish football.

We also build on to the existing literature on gender bias in evaluations. In order to look for gender differences in how subordinates rate female and male superiors, researchers have been studying the context of academia in particular, with student evaluations of teachers serving as a tool to capture gender discrepancies (Boring 2017; MacNell, Driscoll, and Hunt 2015; Mengel, Sauermann, and Zölitz 2019). These subjective measures have proved evidence for the existence of gender bias against female teachers, even in samples where teaching quality and resources are controlled for (MacNell, Driscoll, and Hunt 2015). We find it natural to build on this literature, as academia and football share many similarities. They are both hierarchical, high-performing, and pedagogical environments, where the subordinates are relatively young with extensive experience of evaluating their superiors. Although the methodology in our research does not use evaluations in which students fill out a form after completion of a course, we consider our design of letting players evaluate an instruction video to be similar. However, we use a more controlled setting to hold teaching quality fixed.

We have structured the thesis in the following manner:

i) We start by providing an overview of the previous literature and discuss some of the broader research on women in leadership positions. We continue by giving an overview of the literature on football as a gendered space, female leadership in this domain, and finishes off by narrowing down on prior research on gender bias in evaluations.

ii) We move on to explain and motivate our experimental design, data collection procedure, statistical methods, and what we hypothesize.

iii) Following the method section, we present the result of our study, our sub-analyses and robustness checks.

iv) We then discuss our results, their implications, internal and external validity, and the limitations of the study. We give suggestions on further avenues of

research.

v) The paper concludes with a brief summary of our results and the most important implications.

2 Literature Review

Over the past 30 years, researchers have studied women in leadership positions in attempts to explain the underrepresentation. Metaphors such as 'glass ceiling', 'leaking pipeline', and 'firewalls' have been frequently used, and both supply-side and demand-side perspectives of the underrepresentation have been examined.

The literature review is organized as follows: i) we begin by a short examination on the research that has been made on female leadership more generally. In this part, we put emphasis on the foundational management theories of discrimination towards female leaders that a lot of economics literature in the field builds upon and tests with observational data. We then move on to ii) zooming in on female leadership in the football context, bringing the economics literature into context, and thereafter iii) going deeper into the main component of focus in this paper - leadership evaluations. More specifically, in this part, we examine relevant research on leadership evaluation in the academic context. We wrap up the literature section by iv) outlining our contribution to the literature.

As gender equality is constantly progressing and changing in form, the majority of the work presented in this review of the literature has been published in the last 15 years, in an attempt to capture the most recent advances in this research area.

2.1 Women in Leadership Positions

Leadership comes in a variety of forms and is highly contextual (Ayman 2004). Both what type of socio-cultural norm the leader operates in, what kind of organizational culture, industry characteristics, and the type of followers it has will shape the leader's behavior. Despite the variability in appropriate characteristics, leadership has historically been described in masculine terms (Van Velsor, Taylor, and Leslie 1993). Generally, women are expected to be communal and have traits such as gentleness, kindness, and concern for others, while men are expected to be agentic, having traits such as aggressiveness, confidence, and self-direction (Powell and Butterfield 2003). Leaders are generally described to have more agentic traits than communal traits, thus showcasing typical masculine characteristics (ibid.). This has resulted in men fitting into the leadership stereotypes more easily

than women, making the former appear as more natural leaders. Because of this, women face the so-called double bind (Eagly and Karau 2002). As leaders, they are expected to be agentic, demonstrating typical masculine traits such as confidence, but they are also expected to meet their female gender role, showcasing communal traits, which sometimes can appear incompatible with being agentic, thus, being a leader. This puts women in a more vulnerable position of being the target of prejudice. A woman who demonstrates agentic traits risks being labeled as 'unfeminine,' while a woman who demonstrates communal skills risks being labeled as not having the right leadership characteristics. (ibid.). These differences in gender and leader stereotypes are decreasing over time, and compared to previous research, more recent studies have shown that current views of leaders include more communal traits and less agentic traits (Duehr and Bono 2006). However, the differences have not disappeared, and the incongruity between leader prototypes and gender stereotypes can perpetuate a gender gap in leadership by driving differential evaluations of female and male leaders (Eagly and Karau 2002).

The above mentioned incongruity is commonly referred to as the role congruity theory (ibid.). This theory further suggests that the prejudice towards female leaders who face this incongruity may vary depending on the context and the characteristics of the leader's followers. The theory states that as the group composition becomes more gender diverse, the prejudice towards the female leader will weaken. This statement is built upon the sex-matching model of Kiesler (1975) which suggests that men and women are matched to different jobs depending on the sex-ratio of the people currently occupying these jobs. That is, a man is more likely to be considered a good match for a job where a majority of the people in similar jobs are men. The same holds true for women. The role congruity theory then shows that female characteristics will be perceived as more valuable in female dominated fields, such as nursing, while male characteristics will be perceived as more desirable in male dominated contexts. This provides a further explanation for why certain occupations and titles seem to be partially restricted to a specific gender (Eagly and Karau 2002).

To conclude the section, the notion that leadership traditionally has been characterized as a masculine sphere, together with the role congruity theory, forms two types of discrimination towards female leaders. First, women are exposed

to discrimination and prejudice due to incongruency between the communal gender role and the agentic leadership prototype. Second, when women engage in male-dominated contexts, they are evaluated less favorably than their male counterparts, as they, as a minority of the group, are perceived as a less good match for the occupation.

2.2 Gender and Leadership in Football

To examine the context on which we focus, namely the football industry, we have adopted the framework developed by Cunningham and Chelladurai (2015), in turn based on the Kozlowski and Klein (2000) multilevel organizational theory. This framework examines leadership in the football context by looking at both the social-cultural perspective (macro), the organizational perspective (meso), and the individual perspective (micro). To limit the scope of our research, this thesis focuses closer on one of the components on the organizational (meso) level, namely prejudice in leadership evaluation.

2.2.1 Macro level

Historically, football has been a domain dominated by masculine hegemony (Fink 2008; Whisenant 2008). Anderson (2009) argues that sport in general *"actively constructs boys and men to exhibit, value, and reproduce traditional notions of masculinity"*. As described in the section above, this may lead to women being evaluated as less capable leaders in football administrations, regardless of their traits or characteristics. Therefore, when discussing female leadership in the football domain, gender serves as a fundamental aspect of organizational and social processes (Burton and Leberman 2017). Furthermore, gender not only forms identities in these organizations, but also serves as an axis of power - influencing organizational structures and interactions of the sport organization (Shaw and Frisby 2006).

2.2.2 Meso level

At the meso level, gender is embedded in the different structural and interactional processes of an organization. This includes, but is not limited to, bias in decision making, policies, power-relations and organizational culture (Cunningham and

Chelladurai 2015). The latter aspect, culture, is something that gets passed on and is maintained over time, resulting in that it is usually taken for granted (ibid.). Therefore, structures and values that privileges men over women might be difficult to observe and critically question. According to Acker (1990) Burton and Leberman (2017), the general assumption is often that work and organizational practices are gender neutral. At the meso level, there are two main components of how gender (in)equality takes its form, namely, through i) stereotypes and ii) prejudice.

Stereotypes are the notion of what traits a leader should possess within a specific context. As the prototypical leader within football still is associated with masculine behavior (Burton and Leberman 2017; Grappendorf et al. 2008; Hovden 2010), this can hinder women from succeeding in the field (Cunningham and Chelladurai 2015). Even though the association of masculinity and leadership is slowly breaking down over the years, Cunningham and Chelladurai (2015) argues that it still remains embedded in the structure and culture of sports. As a result, women interested in the role as coach may be seen as less good fits to the position if the role is defined with masculine characteristics. This may in turn lead to women themselves being discouraged to apply, as they view themselves being less capable of succeeding in the role (Eagly and Karau 2002; Cunningham and Sagas 2007). To support these theories of psychology and sociology, economists Akerlof and Kranton (2000) developed an economic framework to describe the economic outcomes of stereotypes. As explained in the theories of role congruency and double-bind, an individual's actions are often affected by how they ought to behave with respect to their social category. To deviate from the prescribed form of behavior - for example, if a woman were to take on more masculine characteristics - would induce a cost for the individual. For example, Bowles, Babcock, and Lai (2007) found that women face a social cost from negotiating assertively, especially when they have a male counterpart. With the utility model

$$U_j = U_j(a_j, a_{-j}, I_j) \tag{2.1}$$

Akerlof and Kranton (2000) describes that an individual's (denoted j) utility is a function of its own actions, a_j, everyone else's actions a_{-j}, and the individuals

own identity, I_j. The identity is in turn dependent on

$$I_j = I_j(a_j, a_{-j}; C_j, \epsilon_j, P) \tag{2.2}$$

where C is the individuals assigned social category, ϵ_j is how well the individual match the ideal of its assigned category, P, and (aj, a_{-j}) is to which extent the individual's and others actions correspond to the prescribed behavior indicated by their assigned categories. Related back to the context of football, their model indicates that female football coaches may incur a loss in utility, if they feel that their actions as a coach are in conflict with the prescribed female behavior.

The second component of bias at the organizational level is **prejudice**, which occurs when one group is evaluated differently from another (Brewer 2007). This is what drives the double bind mentioned in 2.1, and it can be a driver of differential evaluations of female and male leaders (Eagly and Karau 2002). This thesis contributes to the literature by empirically investigate whether there are support for prejudice being a driver in inequality in the Swedish football context.

When an individual unintentionally attributes certain characteristics and/or stereotypes to someone else because of their gender, this forms the so-called gender bias. As earlier stated, this gender bias can hinder the entry of women into leadership positions in sports, as these activities are strongly associated with men (Akerlof and Kranton 2000; Eagly and Karau 2002). In turn, low exposure to female leaders may fuel the biased perceptions of female (in)effectiveness. In their paper, Beaman et al. (2009) ran an experiment across Indian village councils to investigate whether having a female chief councillor affected public opinion about female leaders. They found that having exposure to female leaders reduces gender bias and weakens stereotypes about gender roles in leadership positions, as well as eliminating the negative bias of females efficiency among males. It should be noted, however, that this research was conducted in India, that ranks low on the Gender Inequality Index , so while their results are interesting in them selves, its applicability to the Swedish setting are uncertain.

2.2.3 Micro level

The third level of the framework of (Cunningham and Chelladurai 2015) is the micro level. That is, the individual perspective of leaders. Research on female

leadership in the football industry has indicated that relative to men, women leave the coaching profession at an earlier age (Knoppers et al. 1991) and are less interested in becoming a head coach (Cunningham and Sagas 2002). In order to understand this discrepancy, researchers have studied the respective return to human- and social capital (greater for men) (Cunningham and Chelladurai 2015), differences in self-efficacy (Cunningham and Sagas 2007), and the anticipated outcome of being a head coach (ibid.). (Sartore and Cunningham 2007) suggested that one plausible reason for the difference in numbers of female and male leaders may be that women unconsciously produce self-limiting behavior due to the male-dominated context. This would prevent women from viewing themselves as leaders when comparing themselves to the prototypical sports leader, which in turn prevents them from acting as leaders (ibid.). Born, Ranehill, and Sandberg (2018), supported this finding with an experiment where 580 participants of both genders were tested on their willingness to lead female-dominated versus male-dominated teams. They found that there exists a gender gap in leadership aspirations in male-dominated contexts compared to female-dominated ones, and that this gap was primarily driven by women being less willing to become leaders in male-majority teams. One of the important factors behind this was exactly that, that women, on average, were discouraged by having lower relative beliefs about their performance, and low expectations about receiving electoral support from male-dominated teams. The mechanism on self-limiting behavior has further been supported by Coffman (2014), who conducted an experiment where she found that women are less confident in gender incongruent areas, and that when faced with this incongruity, they tend to contribute with less input to team, driven by this low self-assessment.

2.3 Prejudice in Teaching Evaluations

Previous research on gender (in)equality in the football context has, to a large degree, been conducted through qualitative case studies, where interviews over a small sample of leaders have been performed and then analyzed with management theories. To our knowledge, there exists little to no previous research on gender bias in the evaluation of football coaches, examined through an experimental treatment vs control study.

Even though our paper focuses on gender bias, and more specifically prejudice, in the evaluation of football coaches. As motivated in the Introduction, we consider our work to be closely related to previous research on gender bias in teaching evaluations.

In a recent paper, Mengel, Sauermann, and Zölitz (2019) studies whether there is a gender bias in university teaching evaluations by using a quasi-experimental sample of 19,952 student evaluations. The authors exploit that students are randomly allocated to female and male instructors within each course. In this way, they hope to identify the causal effect of gender on teaching evaluations. They found that female instructors receive constantly lower evaluations compared to male instructors, holding students' grades and study hours constant. Female instructors receive worse evaluations even in components that they cannot control, such as course material. More specifically, they find male students' rates to be the driver of the low evaluations for female university teachers. The gender bias against women is also considerably larger for math-related teaching content. Moreover, they show that the gender bias holds independently of the amount of female and male instructors within each course, suggesting that the gender bias favoring male instructors is general.

In a similar setting, Boring (2017) uses data from a university in France to study gender biases in student teaching evaluations. By using a fixed effects and generalized ordered logit regression analysis, the author finds that male students favor male professors. In addition, the paper finds that students value different teaching dimensions in male and female teachers, and that these dimensions match with gender stereotypes. For instance, despite identical student learning outcomes, men are perceived by both female and male students to be better leaders and more knowledgeable. However, given the fact that Boring (2017) does not use any randomization, the results should be considered with care.

Using a similar experimental setting to ours, MacNell, Driscoll, and Hunt (2015) conduct an online course experiment in which they manipulate the information about the gender of the instructor revealed to the students. In line with previous findings, they find that students evaluate male instructors significantly higher compared to female instructors, no matter the instructor's gender. By hiding

information about the gender identity of the instructor, the authors are able to hold teaching quality and style constant. However, one obvious drawback of the results is the small sample size of 43 students assigned to 4 different instructors, which raises concerns regarding statistical power.

Despite some methodological concerns about parts of the literature mentioned, and the fact that most of the previous literature on evaluations of leaders has been conducted in the academic context, we find it to be highly relevant literature for our case setting (football). In both the previous literature and in our case study, the subjects are young, high achieving individuals living in some of the most gender equal countries of the world (World Economic Forum 2022). The contexts share similarities in that they traditionally have been mostly associated with men (Bagilhole 2002), and as in football, the fraction of females enrolling in graduate programs have steadily increased over the years, while the numbers of female professors are lagging behind (few women chooses to pursue a career within academia). Thus, we find there to be highly relevant synergies between female leadership in the academic- and football context.

2.4 Our Contribution to the Literature

Our study contributes to the literature on gender bias in leadership. As leadership is highly context based, we have narrowed our focus to the football setting, answering the call of Burton and Leberman (2017) for more research on gender (in)equality in managerial leadership positions in football. We do so by designing a framed field experiment (Harrison and List 2004) that examines a channel of prejudice that, to our knowledge, has not been examined in an experimental setting in football before.

The way we contribute to the literature on gender bias in leadership is two-fold. First, we build onto previous research on gender bias towards leaders in the football domain, expanding knowledge on whether prejudice from players towards coaches may be a factor of the inequality. The justification for our research is that no previous studies have conducted an experiment to test this channel. Further justification is that the waste majority of the literature is conducted in North America, and that performing our research in one of the most gender equal countries of the

world may drive differences in gender bias.

Secondly, we add on to the literature on gender bias in evaluations of leaders. The sheer scarcity of experimental studies in how gender bias are expressed in evaluations of leaders, are also justification for our study. Expanding the research from the academic context to that of football, a broader image of gender bias in evaluations can appear.

3 Method

The methodology part of our thesis is organized as follows: i) we begin by describing the experimental design, the pool of participants, and the procedure of the experiment. We then move on to describe and motivate ii) our data, iii) our statistical method, models, hypotheses, subanalyses, and robustness checks. Finally, we wrap up the section by looking at iv) further statistical considerations and v) our hypotheses.

3.1 Setup and Conditions

The statistical methods were decided in a pre-analysis plan before collecting all data and submitted to osf.io. This was done to minimize the risk of p-hacking and researcher's degree of freedom (Simmons, Nelson, and Simonsohn 2016). It was submitted on 28 September, when 56 % ($N_1 = 138$ and $N_2 = 144$) of the responses were recorded. The pre-analysis plan is followed unless clearly stated otherwise.

3.1.1 Experimental design

The participants watch a video of a technical football skill - a volley shot - explained with a voice-over from a coach. Within each team, the participants are randomly assigned to one out of two videos, where one video has a female voice-over and the other has a male voice-over. After watching the video, each participant answers a survey. The chosen design is a between subject experiment. A within subject design would have forced us to show both videos to the participants, with the disadvantaged that they would have noticed the differences between the videos, and the model had picked up other behaviors than solely a (potential) gender bias. Links to the videos are found in Appendix A.

3.1.2 Pilot study

Before starting to collect data, a pilot study was conducted to ensure that the decided procedure of the experiment works in a practical setting. This includes, but is not limited to, handing out QR-codes, answering potential questions that might arise, and ensuring that the video had the right level of difficulty for our chosen sample group. The pilot study indicated that the difficulty level of the video was satisfactory and that only minor changes to the survey were required.

The pilot study was conducted on 17 and 18 August 2022 with 37 players ($N_1 = 20, N_2 = 17$) from two teams, a team of boys aged 18 and under, and a division 1 women's team. Neither teams are considered "elite" by the Swedish Football Association. Data from the pilot study are not included in our analyses since they are not part of the selected population.

3.1.3 Population and sample size

The experiment was conducted on players from all teams in the Svenska Spel f19 A and B fall series, and all teams except one from p19 Allsvenskan. These are the highest divisions for players 19 years and under in Sweden. The boys division consists of one series, while the girls division is split into two series. There are 32 teams in total, of which 31 teams participated in the experiment. The missing team dropped out due to practical issues and time constraints. Each team consists of 13 to 21 players, and the total number of observations is 505, divided into two test groups ($N_1 = 250$ and $N_2 = 255$). The sessions were conducted between 27 August and 1 November 2022.

The number of players who participated from each team depended on the number present on each respective experiment day. This varied depending on whether the experiment was conducted in conjunction with a game or in conjunction with a practice. If the experiment was conducted before/after a game, the participant pool depended on the selection of players for that specific game, which in turn depended on the skills, illness, and injuries of the players. If the experiment was conducted in conjunction with a practice, the group of participants depended to a large extent on sickness. We encouraged injured players to participate as well. Five players did not speak Swedish and were therefore excluded from the experiment.

Whether a team participated or not came down to logistical concerns and their willingness to participate in the experiment. To increase the willingness of the teams, we collaborated with the two organizations in charge of the series - Elite Football Women (EFD) and Swedish Professional Football Leagues (SEF). As mentioned in the previous paragraph, all but one team eventually chose to participate, and the willingness was generally high. The one issue we encountered was logistics. As the chosen series are national, that means the teams are spread out

all over Sweden. With a limited time schedule, we were unable to travel to each team. Thus, six teams conducted the experiment digitally, while the rest were conducted at their home arena (thirteen teams), or in conjunction with an away game in Stockholm or Uppsala (twelve teams). Only one team was unable to make time for the experiment, even though they expressed a willingness to participate. We assess that this will not affect the validation of our experiment.

The reasons we chose this particular sample of elite youth football players were threefold. Firstly, the players in the sample are highly skilled and have had enough coaching to be able to assess the quality of the instructions in the video competently. This should make the video evaluations less noisy. Secondly, by surveying this group, we expected to get a fairly balanced sample with respect to gender. If all the teams had participated with all their players, we expected the sample to consist of 53 % boys and 47 % girls. Thirdly, at this level, we expect almost all players to have grown up in Sweden - an advantage as the experiment focuses on gender bias in the Swedish culture. If we had surveyed the top senior teams, there would have been a proportion of foreign players and, therefore, a risk of picking up norms and values from other cultures than the Swedish one. Thus, our chosen sample results in a fairly homogeneous participant pool when it comes to cultural background, experience, age, and gender. This is an advantage when attempting to isolate a (potential) gender bias.

Given this sample size ($N = 505$), 80 % power and a significance level at 5 %, we have the power to detect an effect size of Cohen's d [1] $= 0.2498$, which is considered a small effect size. This ex-ante effect size calculation is not part of our pre-analysis plan.

3.1.4 Procedure

The experiment sessions started with an introduction of the thesis and the authors. The same introduction was given in all sessions; see Appendix A. The participants were not informed about the purpose of the experiment. Thereafter, the participants were randomly given one QR-code each, which they scanned using

1. Cohen's d is an unitless, standardized measure of effect size for measuring the difference between two group means. The effect size is considered small if Cohen's d = 0.2, medium = 0.5 and large = 0.8 (Carson 2012).

their own phone and which directed them to the survey. Half of the distributed codes in each session had a link to a video with a female voice-over and the other half had a link to a video with a male voice-over. The codes were shuffled before distribution, to achieve stratified randomization at the team level. This is useful and important in the cases of small trials (Kernan et al. 1999). After watching the video, the participants independently answered the survey. The full survey and a translated version can be found in Appendix A. The whole session lasted about 15 minutes. The whole team had to watch the video and answer the survey during the same time period, so the participants could not discuss the video or the questions between sessions.

For the teams that conducted the experiment online, one of the authors was somewhat "present" in the room over an online meeting. The introduction was the same as in the physical session, but a leader from the team in the digital sessions got the responsibility of shuffling and distributing the QR-codes. They did not know which QR-code were directed to which video. All the teams doing the experiment online got the same instructions; see Appendix A.

No financial incentives were provided because once the team accepted to be a part of the experiment, the participants had no reason not to participate, and thus needed no incentives. Furthermore, since the team had assigned time to the experiment, and every participant had to remain in their seats until the whole team had finished the survey, the risk of drop-outs due to low patience should have been minimized. This was later on supported by our data, which showed only three drop-outs out of 505 participants in total. Also, if we had individual incentives, we would have needed to collect personal information, which would have led to a GDPR issue.

3.1.5 Design of video

The decision to use a video design to find a (potential) gender bias was inspired by the paper of MacNell, Driscoll, and Hunt (2015) where they use online classes to keep quality of teaching constant. However, they hide the gender of the teacher, where our experiment, more like Mengel, Sauermann, and Zölitz (2019), lets the participants know the gender of the teacher or coach while rating. The problem

Mengel, Sauermann, and Zölitz (2019) faces is to hold teaching quality constant, which they solve by controlling for, for example, study hours and having a large sample. Since there is so much that can vary between two coaches, we chose to use the video format for this experiment.

A practical advantage of having everyone watch the same video is that we had very low demands on the teams when conducting the experiment. There were no issues with participants sitting close to each other because if they had a look at a team-mates screen, they would see the same thing as on their own screen, which minimized the risk of them figuring out what we were testing and possibly biasing their answers.

The length of the video and the length of the treatment was a major factor in this experiment. The videos are four minutes long, and there are three minutes of talking/"coaching". The practical advantages of making a short video are many. By making the video short, we hoped to ensure that the players remained focused throughout the experiment. Furthermore, a short video made it easier for the teams to free up time to participate in the experiment. The previous mentioned papers that examine gender bias in evaluations have much longer treatment, for example MacNell, Driscoll, and Hunt (2015) who used a five week course. However, studies using videos as short as 3 minutes have managed to find an effect (Schnall, Benton, and Harvey 2008). Another argument for it being enough treatment with a four minute video is based on the subject pool itself and their usage of technology and video clips. The average length of a YouTube video in 2018 was 11.7 minutes, (Ceci 2021) however, 12 % of the content is today made up of videos less than a minute (Conviva 2022). The length of top performing videos on Facebook in 2019 was 1.4 minutes (Dixon 2022) and the optimal time for a TikTok video at the end of 2021 was 21-34 seconds (Stokel-Walker 2022). Therefore, it is argued that the participants are used to the format of short video clips.

The video was produced only for this purpose. A girl and a boy are seen in the video with the same amount of screen time, similar technical skills, and similar looks; see figure 2, in Appendix A. The purpose was to minimize any gender impact or identification with the players seen in the video. For example, if there would only have been a boy in the video, the female participants might identify

less with the player, and hence experience that they learned less. The video presents a technical instruction rather than a tactical one, as a technical scenario is appropriate to all players, whilst a tactical scenario would by nature have to be more team specific since tactics vary across teams. The motivation for choosing a volley-shot as the technical skill was that it is relevant for every player on the field and because the skill rarely gets attention in the regular team practices. As it is a quite advanced skill, we hoped that the players would learn something from watching the video and therefore pay more attention to it, compared to if it were an easier skill that everybody already knew. Since 70 % of the participants answered that they learned something from the video, we consider this goal to be achieved.

The video material was produced by two professional and educated football coaches. The dialect, sociolect, and intonation of the male and female narrators were kept as similar as possible. The instructions in the video were the same and the script of the video can be found in Appendix A. Setting up the experiment this way should allow us to isolate the effect that the gender of the narrator has on the players' assessments of the instruction videos.

3.1.6 Design of survey

The survey started with an evaluation of the video, which was placed first in the survey to ensure that the participants had a fresh memory of the video. After evaluating the video, the players answered an attention question: *"what technical skill was shown in the video?"* This to ensure that they paid attention when watching the video. After that, they got to evaluate the video with their own words. The next section included demographic questions such as age, whether they learned anything from the video, and a question to see if they paid attention to the gender of the coach, that is, whether they noticed the treatment they were assigned to. There was an option *"I do not know"* for participants who did not remember or did not notice the gender of the coach.

We also collected data to test some of the mechanisms that have been found to have an effect on gender bias in the previous literature. First, to investigate whether a potential bias is driven by male or female participants, as in Mengel, Sauermann,

and Zölitz (2019), participants were asked to fill in which team they belonged to. We asked this question instead of collecting the gender of the participant, so our models are based on whether the participant is a member of a female or a male team. Second, in accordance with Beaman et al. (2009) we wanted to see how previous exposure to female football coaches affects a potential gender bias. Thus, we asked the players if they were currently having a male or female head or assistant coach and if they had ever had a female head or assistant coach. If they answered *"yes"* on having had one, they had to specify at which football level (5v5, 7v7, 9v9, 11v11).

By asking participants to evaluate the video first and answer questions about the instructor and current/previous coaches afterward, we ensured that participants did not have gender issues top of mind when assessing the videos. It was not possible to jump between questions or skip questions.

3.1.7 Classification of experiment

Harrison and List (2004) presents a taxonomy of field experiments. Based on six different factors, they define field contexts of experiments and classify these into four different groups. Based on this taxonomy, we argue that our experiment is a framed field experiment.

We classify it as a framed field experiment because of the subject pool (non-standard) and the nature of the information set that the subjects bring to the task (football knowledge). But, since the participants know that they are part of an experiment, we cannot rate this experiment as a natural field experiment. Thus, our experiment fulfills the main criteria for being classified as a framed field experiment.

3.2 Data

Here we present the coding of the dependent and independent variables.

3.2.1 Dependent variables

The dependent variable is the mean of the responses of the respondents to the five questions below, on a scale of 1-6 stars. Based on the Cox III (1980) framework,

the most efficient scale is between 5 and 9, and if one wants to avoid the neutral answer, the scale needs to be an even number.

- On a scale of 1-6 where 6 is the best, how instructive did you find the video?
- On a scale of 1-6 where 6 is the best, how did you find the coach speaking in the video?
- On a scale of 1-6 where 6 is the best, how did you find the instructions in the video?
- On a scale of 1-6 where 6 is the best, how professional did you find the coach in the video?
- On a scale of 1-6 where 6 is the best, do you think the video showed how a good player would perform a volley shot?

3.2.2 Independent variables

There are three independent variables used in the main regression and the sub-analyses that need further explanation.

Firstly, $femC$ is the gender of the coach, the voice-over in the video, where female ($femC = 1$) or male ($femC = 0$). This is thus our treatment.

Secondly, whether the participant is part of a male or female team, $femP$, where female ($femP = 1$) or male ($femP = 0$).

Lastly, a dummy variable to determine whether the participant has had relevant exposure to a female coach. We define the relevant exposure as currently having a female head or assistant coach and/or if the participant has had a female head coach on 11v11. We chose to include current assistant coaches because we were afraid that the sample would be too small if we only included head coaches. The variable exposure, Ex, takes ($Ex = 1$) if the participant has a relevant exposure or ($Ex = 0$) if there is no relevant exposure.

3.2.3 Outliers and exclusion

An outlier is defined as "an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins et al. 2002), which none of our answers are. The most extreme results we can get are if someone presses "1" on all or "6" on all the questions, which we do not consider outliers by the mentioned definition. There were zero players rating the video with "1" on all questions and 41 participants who gave the video a full score.

Participants who miss the control question *"what technical skill was shown?"* are removed. The answer we keep is *"volley shot"* and the answers we exclude are *"throw-in"*, *"header"* and *"slide tackle"*. This is because if they got that question wrong, they clearly did not watch the video and therefore would not have been able to evaluate the video. As previously mentioned, the five participants who did not speak Swedish were excluded from the experiment.

3.3 Statistical Methods

In this section, we present our statistical methods. First, the main regression together with our hypothesis and statistical tests. After that, we present and motivate our subanalyses. A series of robustness checks are presented and finally some further statistical considerations on our statistical methods. We run an ordinary least squares regression on our data.

3.3.1 Main regression

We start by comparing the means of the treatment groups to test for a raw gender bias. This is done by comparing the means of the two treatment groups with a non-parametric Mann-Whitney U test. Then we run the regression 3.1 for further investigation.

$$y_i = \alpha_i + \delta_1 femC + \epsilon_1 \tag{3.1}$$

A two sided t-test is used to test the null hypothesis below.

$$H_0 : \delta_1 = 0 \text{ against } H_1 : \delta_1 \neq 0 \tag{3.2}$$

If we can reject the null hypothesis of 3.2 at the statistically significant 5 % level

we interpret this as evidence of a gender bias. The size and direction of the bias are estimated by δ_1.

3.3.2 Subanalyses

Participants member of male of female team

The first subanalysis, 3.3, test our first mechanism inspired by Mengel, Sauermann, and Zölitz (2019). By including an interaction term, we investigate whether the rating of the video is affected by whether the participant belongs to a male or female team. This generates four hypotheses to test.

$$y_i = \alpha_i + \delta_1 femC + \delta_2 femP + \delta_3 femC \cdot femP + \epsilon_i \tag{3.3}$$

Hypothesis 3.4. No gender differences with respect to whether participants belong to a male or female team or in coach's gender, tested with an F-test.

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0 \text{ against } H_1 : \delta_1 \neq \delta_2 \neq \delta_3 \neq 0 \tag{3.4}$$

The null hypothesis implies that there are no gender differences when evaluating the video, neither with respect to the coach nor with respect to the participants belonging to a male or female team. If we can reject H_0 that would suggest that there is a gender difference.

Hypothesis 3.5. Participants in a female team do not evaluate the male and female coach differently, tested with a two sided t-test.

$$H_0 : \delta_1 + \delta_3 = 0 \text{ against } H_1 : \delta_1 + \delta_3 \neq 0 \tag{3.5}$$

Hypothesis 3.6. Participants in a male team do not evaluate the male and female coach differently, tested with a two sided t-test.

$$H_0 : \delta_1 = 0 \text{ against } \delta_1 \neq 0 \tag{3.6}$$

H_0 from 3.5 implies that participants in a female team make no difference in how they evaluate the male or female coach. H_0 from equation 3.6 implies that participants in a male team do not evaluate the male and female coach differently.

Hypothesis 3.7. Differences in video evaluations between the male and female coach do not depend on participants belonging to a male or female team, tested with a two sided t-test.

$$H_0 : \delta_3 = 0 \text{ against } \delta_3 \neq 0 \tag{3.7}$$

H_0 from equation 3.7 states that neither the players belonging to a male or female team evaluates the male or female coach differently.

Relevant exposure

The second subanalysis, 3.8, test our second mechanism inspired by Beaman et al. (2009). We include an interaction term to investigate how the player's relevant exposure (Ex) to a female coach affects the rating of the video. This generates four hypotheses to test.

$$y_i = \alpha_i + \delta_1 femC + \delta_2 Ex + \delta_3 femC \cdot Ex + \epsilon_i \tag{3.8}$$

Hypothesis 3.9. Having or not having relevant exposure to a female coach does not have an effect on the evaluation of the video with respect to the coach's gender. Tested with an F-test.

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0 \text{ against } \delta_1 \neq \delta_2 \neq \delta_3 \neq 0 \tag{3.9}$$

The null hypothesis implies that there are no differences between relevant exposure to a female coach and no exposure when evaluating the video, with respect to the coach's gender. If we can reject H_0 that would suggest that there is a difference.

Hypothesis 3.10. Having relevant exposure does not have an effect on the evaluation of the video, with respect to the gender of the coach in the video. Tested with a two sided t-test.

$$H_0 : \delta_1 + \delta_3 = 0 \text{ against } \delta_1 + \delta_3 \neq 0 \tag{3.10}$$

Hypothesis 3.11. Not having relevant exposure does not have an effect on the evaluation of the video, with respect to the gender of the coach in the video. Tested with a two sided t-test.

$$H_0 : \delta_1 = 0 \text{ against } \delta_1 \neq 0 \tag{3.11}$$

H_0 from 3.10 implies that players with relevant exposure make no difference in how they evaluate the male or female coach. H_0 from equation 3.11 implies that players without exposure do not evaluate the male and female coach differently.

Hypothesis 3.12. Differences in the evaluation of the male and female coach in the video do not depend on relevant exposure to a female coach. Tested with a two sided t-test.

$$H_0 : \delta_3 = 0 \text{ against } \delta_3 \neq 0 \tag{3.12}$$

H_0 from equation 3.12 states that neither player with or without relevant exposure evaluates the videos differently.

3.3.3 Robustness checks

On each question separately

To make sure our results are robust and that potential results are not driven by only one of the questions, the main regression and the subanalyses are run on each of the five evaluation questions separately. For example, a potential gender bias could be driven by the professionalism question only. This could result in that a statistical significant result found when running the regressions on the mean is misinterpreted and more significance is attributed to gender bias in perception of the instructions. Another potential problem that could be avoided through running this robustness check is if the three regressions, when run on the mean, show no statistical significance but there is a gender bias on one of the questions individually.

High and low dispersion

The second robustness check is on high and low dispersion within the questionnaire. This should rule out the possibility that the results are driven by 'careless' participants who 'always tick the same box' when filling out the survey. To define individuals as 'low dispersion' or 'high dispersion', respondents, we calculate the standard deviation of a player's answers across all five questions. Low dispersion (high dispersion) is defined as evaluations with a below-median (above-median) standard deviation. Thus, we create two samples and run the regressions on those samples, respectively. We compare the regressions using a t-test to see if the

coefficients differ between the two groups.

$$H_0 : \delta_{1;high} = \delta_{1,low} \text{ against } H_1 : \delta_{1,high} \neq \delta_{1,low} \tag{3.13}$$

Treatment question

The third robustness check is performed with respect to whether the players answered the treatment question correctly. *"Was there a male or female coach speaking in the video"* with the three answer options *"female"*, *"male"* or *"I do not remember"*. If the player correctly answers this question, they are assigned ($NotCorrectTr = 0$), if they answer wrong or *"I do not remember"* they take the value ($NotCorrectTr = 1$). We start by running the regression below with an interaction variable, and if the results are statistically significant, we do exploratory analyses to see if there are patterns to be detected, for example, if players tend to miss the treatment more when there is a female coach.

$$y_i = \alpha_i + \delta_1 femC + \delta_2 NotCorrectTr + \delta_3 femC \cdot NotCorrectTr + \epsilon_i \tag{3.14}$$

Differences in video evaluations between male and female instructors do not depend on participants' answer on the treatment question, tested with a two sided t-test.

$$H_0 : \delta_3 = 0 \text{ against } H_1 : \delta_3 \neq 0 \tag{3.15}$$

We recognize that there are more tests that could have been conducted, such as running the regressions for type of session (digital, physical, post-game, post-training, etc.), but choosing not to do it due to two main reasons. First, the lack of subjects in the different groups makes it difficult to obtain statistical significance, and second, many tests increase the risks of false positives. Furthermore, since we randomize within each team, effects across sessions are likely to be small.

3.3.4 Further statistical considerations

We have strata fixed effects on all our regressions, stratified on the team level, which in our setting is also the session level. This is done through a categorical variable included in our linear regressions. The goal is to remove session-specific variance; for example, if a team lost a game they might rate both videos lower than a team after a win.

Our definition of statistical significance is $p < 0.05$. To detect what the minimum detectable effect is, given the sample size that in the end turned out to be available to us, we run a power analysis with the power 80 %.

Since heteroskedasticity-robust standard errors tend to have a downward bias i.e. more likely to get false positives, we first test for heteroskedasticity. The Breusch-Pagan test and the White test help us find whether there is a relationship between residuals and explanatory variables. If there is, we use robust standard errors in all regressions to minimize the risk of bias due to heteroskedasticity.

Due to the fact that we do not assign treatment to clusters of units, but rather to individuals (player level), there is no need to cluster our standard errors (Abadie et al. 2022).

All of our empirical methods have been predetermined in our pre-analysis plan.

3.4 Hypotheses

Given the previous research and the statistical methods described, we predict the results of our analyses.

In line with the research from Mengel, Sauermann, and Zölitz (2019), Boring (2017), and MacNell, Driscoll, and Hunt (2015) we argue that it is reasonable to expect that the participants rate the video with a female voice-over lower than the video with the male voice-over in our main regression 3.1.

As stated in our first subanalysis 3.3, we investigate whether belonging to a female or male team is a mechanism of a potential gender bias. Still, we see no reason why our results should differ from Mengel, Sauermann, and Zölitz (2019) and thus hypothesize that there will be a difference in how participants belonging to a male or female team rate the videos, with respect to the coach's gender. We hypothesize that the groups will rate the video with a female voice-over lower than the video with a male voice-over, but participants from a male team will rate the video with a female voice-over the lowest.

Our second subanalysis 3.8 investigates whether relevant exposure to a female coach is a mechanism of a potential gender bias. Despite the fact that the re-

search from Beaman et al. (2009) is set in India, we hypothesize that our experiment should find a similar effect, although we expect it to be smaller in the Swedish setting. Specifically, we expect there to be a difference in rating of the video between the group that has relevant exposure and the group that has not, with respect to the coach's gender. We hypothesize that the groups will rate the video with a female voice-over lower than the video with a male voice-over. We hypothesize that participants without relevant exposure rate the video with a female voice-over the lowest.

4 Results

The result section of our thesis presents our results, and further analysis of the results is made in the discussion in Section 5. Our results are presented in the following manner: i) we begin by presenting the descriptive data and then move on to ii) our main regression. Thereafter the section with iii) the subanalyses, followed by iv) the robustness checks. Finally, v) a short summation of the results.

4.1 Descriptive Data

Table 1 shows the descriptive statistics of the primary data collected through the experiment. In total there are 505 observations, split up into two test groups ($N_1 = 250$ and $N_2 = 255$). There are $N = 241$ participants from a male team and $N = 264$ participants from a female steam. 3 of these participants answered the attention question wrong, and their results are not a part of our regressions. Of the 505 participants, 41 % ($N = 207$) have the relevant exposure. 31 participants answered the treatment question with *"I do not know"* or answered wrong. The age varies between 14 and 20 years, with an average age of 17.5 years. The average of the participants mean of ratings is 4.9 on a scale of 1-6, with a minimum of 2.2. Since the standard deviation is almost one star on the rating-scale (0.7) we can already see that the variation is quite high compared to the ratings.

Table 1: Descriptive statistics

Statistic	N	N=1	Mean	St. Dev.	Min	Max
Mean of ratings	505		4.9	0.7	2.2	6.0
Attention question	505	3	0.01	0.1	0	1
Did you learn	505	359	0.7	0.5	0	1
Exposure	505	207	0.4	0.5	0	1
Member of female team	505	264	0.5	0.5	0	1
Female coach in video	505	255	0.5	0.5	0	1
Age	505		17.5	1.0	14	20
NotCorrectTR	505	31	0.1	0.2	0	1

In Table 2 the relevant descriptive statistics are grouped by treatment, and a balancing test has been conducted to investigate if the distribution between the treatment groups is similar. The F-statistics are small and not significant on the 5 % level, there is no statistical significance between the two groups.

Table 2: Balancing test

Voice-over :	Male			Female			
Statistic	N	Mean	St. Dev.	N	Mean	St. Dev.	Test
Mean of ratings	250	4.873	0.74	255	4.896	0.732	F=0.131
Attention question	250	0.004	0.063	255	0.008	0.088	F=0.315
Did you learn	250	0.716	0.452	255	0.706	0.457	F=0.063
Exposure	250	0.396	0.49	255	0.424	0.495	F=0.394
Member of female team	250	0.52	0.501	255	0.525	0.5	F=0.015
Age	250	17.436	1.074	255	17.471	1.003	F=0.14
NotCorrectTr	250	0.06	0.238	255	0.063	0.243	F=0.016

Note: *p<0.1; **p<0.05; ***p<0.01

4.2 Main Regression

Before discussing the results of our main regression and subanalyses, we need to present our results of the Breusch-Pagan test to test for heteroskedasticity, Table 3. The test statistics are $25.17; 25.27; 26, 94$ and the corresponding p-value is 0.76 for the three models. Since the p-value is larger than 0.05, we fail to reject the null hypothesis and we do not have sufficient evidence to say that heteroskedasticity is present in the regression model. Also, when relaxing the assumption of normally distributed standard errors and conduction a White-test, we fail to reject the null hypothesis. Therefore, no robust standard errors are used.

4.2.1 Main regression

Our main regression investigates whether there is a raw gender bias between the two treatment groups. Let us first look at the distributions with the boxplot in

Table 3: Breusch-Pagan test

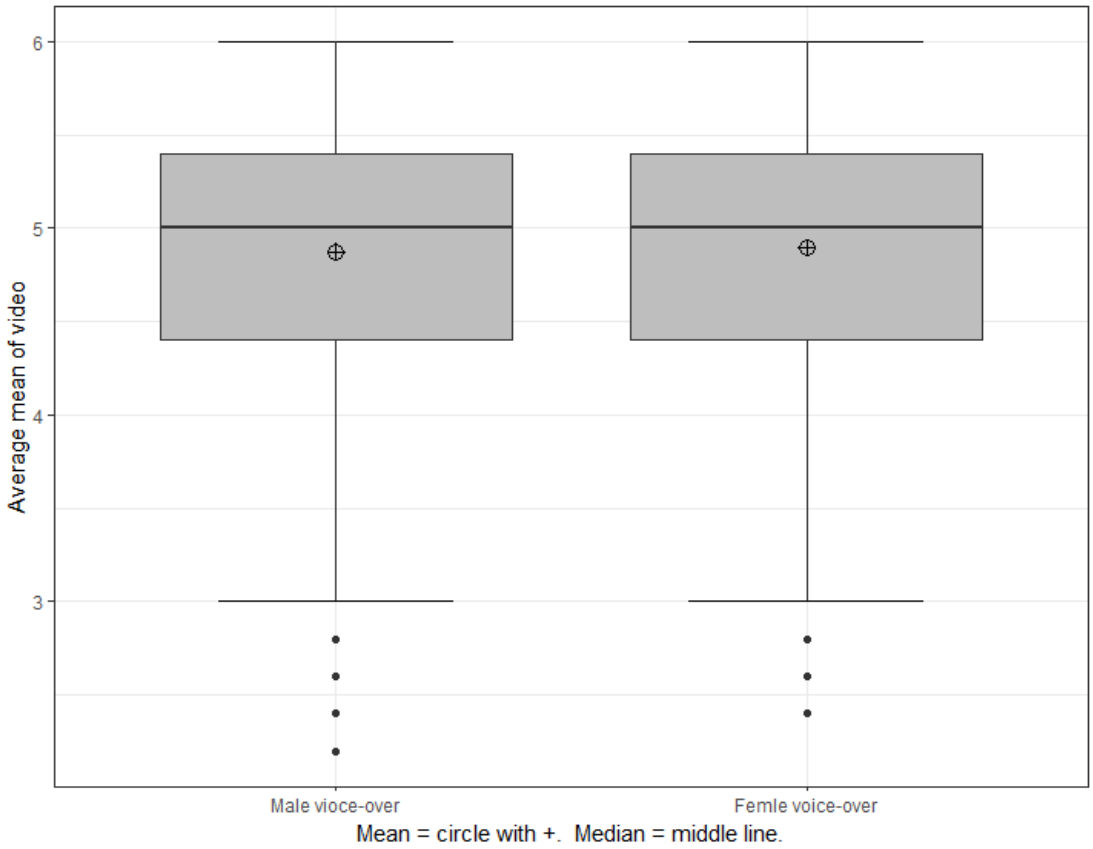
	statistic	p.value	parameter	method
Reg (1)	25.17	0.76	31.00	studentized Breusch-Pagan test
Reg (2)	25.17	0.76	31.00	studentized Breusch-Pagan test
Reg (3)	26.94	0.76	33.00	studentized Breusch-Pagan test

Figure 1. We can see that the distributions overlap and the means are very close to each other. To investigate whether the difference between these two means is statistically significant, we look at the results from our non-parametric Mann-Whitney U test and our parametric t-test from the main regression, column (1), in Table 4. The Mann-Whitney U test results in a two-sided test with $p-value = 0.76$. This indicates that we cannot reject the null hypothesis that the distributions are equal and conclude that we cannot determine a significant difference between treatment groups.

The parametric t-test in Table 4 for model (1) shows a similar result. The large standard errors, bigger than the point estimate, indicate a lack of statistical significant effect, and the t-test fails to reject the null hypothesis. The minimum detectable effect size with this sample size, if $p = 0.05$ and a power of 80 %, is 0.168 stars on our scale of 1-6 stars.

To conclude, the variance in the mean rating of the video does not appear to differ significantly between the treatment groups.

Figure 1: Box plot of treatment groups



4.3 Subanalyses

4.3.1 Subanalysis one

Table 4, column (2) shows the results of our first subanalysis. The first subanalysis includes an interaction term to investigate whether the participant is part of a male or female team affects the rating of the video. Four hypotheses have been presented with respect to the first subanalysis, (see Section 3.3.1). We see that hypotheses 3.5, 3.6, and 3.7 cannot be rejected at the 5 % level. Thus, in this sample, there is no evidence of a gender bias from members of a male or female team. Regarding hypothesis 3.4, whether or not the model is predictive as a whole, we see that the F-statistic is significant on a 1 % level. However, this result seems to be driven by the statistical significance of our constant, rather than by any of the coefficients. It suggests that our explanatory variables do not explain the

variation on a statistically significant level.

Since R^2 does not change from the main regression to the first subanalysis, adding the variable of members of a female team does not explain more of the variance than does model (1). The minimum detectable effect size given 80 % power and a significance level at $p = 0.05$ for δ_1 in model (2) is 0.252 stars on our scale of 1-6 stars, for δ_3 in model (2) is 0.364 stars, and for $\delta_1 + \delta_3$ is 0.252 stars.

4.3.2 Subanalysis two

Table 4, column (3) shows the results of our regression on model (3). The second subanalysis includes an interaction term to investigate how the player's relevant exposure to a female coach affects the rating of the video. Four hypotheses have been presented with respect to the second subanalysis, see Section 3.3.1. We see that hypothesis 3.10, 3.11 and 3.12 cannot be rejected at the 5 % level. Thus, there is no evidence of a gender bias with respect to whether the participants in this sample have had relevant exposure or not. Regarding hypothesis 3.9, whether or not the model is predictive as a whole, we see that the F-statistic is significant on a 1 % level. However, like in model (2), this result seems to be driven by the statistical significance of our constant rather than any of the coefficients.

Since R^2 does not change from the main regression to the second subanalysis, adding the variable of exposure does not explain more of the variance than the model (1) does. The minimum detectable effect size given 80 % power and a significance level at $p = 0.05$ for δ_1 in model (3) is 0.224 stars on our scale of 1-6 stars, for δ_3 in model (3) is 0.392 stars, and for $\delta_1 + \delta_3$ is 0.308 stars.

4.4 Robustness Checks

As described in Section 3.3.3 three robustness checks were conducted.

4.4.1 On each question separately

In Appendix B the results of our first robustness checks are presented; see Table 5, Table 6 and Table 7. The main regression and the subanalyses were run on each of the questions independently. The main regression, Table 5, tells us the same story as when run on the mean of ratings, no gender effect can be found.

Table 4: Main regression an sub-analyses

	Dependent variable:		
	Mean rating of video		
	Model (1)	Model (2)	Model (3)
Female coach	0.04	0.05	0.09
	(0.06)	(0.09)	(0.08)
Female team member		0.08	
		(0.25)	
Exposure			0.10
			(0.14)
Female coach*Female team member		−0.02	
		(0.13)	
Female coach*Exposure			−0.13
			(0.13)
Constant	5.11***	5.03***	5.06***
	(0.19)	(0.17)	(0.22)
Session fixed effects	Yes	Yes	Yes
Observations	502	502	502
R^2	0.15	0.15	0.15
Adjusted R^2	0.09	0.09	0.09
Residual Std. Error	0.70 (df = 470)	0.70 (df = 469)	0.70 (df = 468)
F Statistic	2.60*** (df = 31; 470)	2.51*** (df = 32; 469)	2.47*** (df = 33; 468)
$\delta_1 + \delta_3$		0.03	−0.05
		(0.09)	(0.11)

Note[1]*:* *p<0.1; **p<0.05; ***p<0.01

Note[2]*:* *Standard errors in brackets*

Neither the two subanalyses, Table 6 and Table 7, run on each question separately, can reject the null hypotheses on a 5 % significance level. However, a peculiar finding that arouse is the statistical effect on the survey question "*On a scale of 1-6 where 6 is the best, how did you find the instructions in the video?*". Yet, this estimate pointed in the opposite direction of previous research, has a statistical significance below our limit, and should be interpreted as a null result.

4.4.2 High and low dispersion

Secondly, to see if there is a difference between the participants with a high or low dispersion in their responses. In Table 8, Appendix C, the results are shown on the main regressions run on the divided sample. Calculating the p-value of the comparison of the two coefficients, we get $p = 0.883$ which would mean that we cannot reject the null of the two being equal. This suggests that the difference in dispersion does not have an effect on the results.

4.4.3 Treatment question

Finally, the robustness check with respect to whether or not the players answered the treatment question correctly. 31 players answered that they did not know the gender of the coach in their video or answered the question incorrectly. In Table 9, Appendix D, the results of a regression with the treatment coefficient as an intercept are shown. Due to the small sample group, it is expected not to get a significant result, which is exactly what the result is showing.

The results of the robustness checks strengthen our conclusion from the previous result, we cannot find a gender effect in this sample between the two videos.

4.5 Summary of Results

One main regression has been conducted, estimating whether there is a difference between the mean ratings of the two videos. Based on the results of the non-parametric and the parametric test we cannot, on any significance level, reject the null hypothesis of the two means being the same. Due to the high power of the model, if there was an economically significant effect to be detected, it is likely that we would have found it.

In subanalysis one, an interaction term has been included to see if the mean rating of videos varies depending on whether the participant is a member of a male or a female team. The results show that we cannot reject the null hypothesis that this does not vary with participants belonging to a male or female team.

In subanalysis two, we have instead let the mean ratings of the video vary depending on whether the participants currently or recently had a female coach, relevant exposure. The results support the notion that there is no difference in the mean rating between the videos, at any significance level.

To ensure that our results hold for each question by it self, all three models were run on each question separately. We also cannot find support for rejecting the null hypothesis here, supporting the results of the main regressions.

The following two robustness checks are estimated to ensure that our findings hold in more settings. First, we test if there is a difference depending on the dispersion of the participants' responses, and second, we test if the participants who answered the treatment question wrong have a different result from our other findings. Our robustness checks support our findings of no effect.

5 Discussion

As our research aimed to investigate why there are so few female football coaches at the elite level in Sweden, we cannot find evidence that this is driven by gender bias in players evaluation of coaches. We cannot reject the null hypothesis of no gender bias based on our experiment with a video of a technical football skill in our selected sample.

In our two subanalyses we control for the participants belonging to a male or female team and whether or not they have relevant exposure with a female coach. From these we conclude that we cannot say that gender bias varies with the participant belonging to a male or female team, and there is no support of it varying with the participant having relevant exposure with a female football coach. Thus, the results support our main regression.

In addition, three robustness checks were conducted with the goal of strengthening our results. First, the main regression and the subanalyses were run separately on each question, and on neither of them there was evidence of a gender bias. Second, the robustness check that controlled for high and low dispersion in the answer choices could not reject the null hypothesis, and finally, the robustness check on whether or not the participant answered the treatment question correctly could not reject the null hypothesis. We conclude that the robustness checks support our lack of evidence of a gender bias. The reason for this lack off effect could be due to two major reasons, which will be discussed in the next section.

5.1 Analysis of Results

As stated above, we cannot reject the null hypothesis that there does not exist a gender bias in the evaluation of football coaches in our sample. Thus, we cannot with certainty state whether i) there does not exist a gender bias at all or ii) there exists a gender bias, but we were not able to pick up the effect with our experiment design.

What should be emphasized when considering the two scenarios is that we are powered to detect a small effect size, as our Cohen's d = 0.2498, given a 5 % significance level and 80 % power. Also, the minimum detectable effect size for our

main regression is 0.168 stars, on our scale of 1-6 stars, given a 5 % significance level and 80 % power, which is on the limit of what could be considered economically significant. Thus, we can conclude that we are underpowered to detect a very small effect size but powered to detect an economically significant effect size. With the chosen experimental design, we should have been able to isolate the specific channel of a raw gender bias in evaluations by holding everything but the gender constant. Thus, we believe it to be likely that our results mirror the reality - that there does not exist a gender bias in our sample at our level of treatment.

There is a risk, naturally, that our treatment is too weak and that varying the gender of the voice in the video is not sufficient in itself to detect a bias, even in the scenario that a true gender bias actually does exist in evaluations of football coaches. The level of treatment was something that was thoroughly discussed prior to the experiment, and we believe that the chosen treatment level has several advantages. The most important one is that if we had found a significant effect on our dependent variable, we could have said with high power that it was driven by a gender bias. Were we to choose a larger treatment - for example, having a female and male coach instructing the technical skill in person (not only through voice), it would have been more difficult to say with certainty that this effect was driven by a gender bias alone, as multiple other characteristics of the coaches would have showcased than solely their gender, for example, their expressed confidence, their posture, body language, etc., even when aiming to hold these as similar as possible. To control for other factors like these, we would have needed to cluster our standard errors on team level (opposed to on player level) and collect a greater amount of observations to get the same degree of power. As we tested all but one team of our chosen sample (the top youth division for females/males), this would have required us to test other groups of football players, for example non-elite players. For all these reasons, we are confident in the experimental design chosen.

Although we did not find evidence for a *raw* gender bias, there could still exist a gender bias that we could not detect. Since our treatment design is limited in scope, it may be the case that we have eliminated *gendered factors*, which contributes to a bias towards women. For example, a potential bias could be driven by something related to women's appearance, how female coaches lead the team during a match, or other unobservables. As stated in the literature section,

it could also be driven by, for example, female coaches that showcase feminine traits that are incongruent with the prototypical traits of a football coach (Eagly and Karau 2002). As we kept the manuscripts gender neutral, holding everything except the actual gender of the coach constant, we eliminated the possibility of gendered factors playing a role in the evaluation of the coaches. Thus, if gendered factors and not the raw gender itself would be a driver of a gender bias towards coaches, we would not be able to detect it with our experiment.

It might very well be the case that scenario 1 holds true - that there exists no gender bias in our sample, and that our results mirror the reality. Even though we cannot state this with certainty, we cannot reject that this is not the case.

Mengel, Sauermann, and Zölitz (2019) - our main source of inspiration when designing our research - found a significant gender bias in the academic context in the Netherlands. As the Netherlands and Sweden are ranked very similarly on the World Gender Equality Ranking (World Economic Forum 2022), and there are obvious synergies between the academic field and the one of elite sports (hierarchical, high-performing, and pedagogical environments), our results could be interpreted as a bit surprising with this in mind. Even MacNell, Driscoll, and Hunt (2015), who used a similar treatment as we did - with online teaching - found a raw gender bias in his study. However, we do not believe that this means that students are more gender biased in the Netherlands. Mengel, Sauermann, and Zölitz (2019) ran a large natural experiment where the treatment was huge (spanning over the whole duration of a study semester with almost 20,000 observations). Thus, when a true gender bias did exist, his study was more prone to pick up the effect.

Interesting enough, our (non)findings do somewhat contradict the previous research that has been made on female leadership in the football domain. According to Burton (2015), the masculine environment that football still is can lead to women being evaluated as less capable leaders in football administrations, regardless of their traits or characteristics (raw gender bias). Cunningham found the same indication in his 2015 paper (Cunningham and Chelladurai 2015). However, since no observational study on the existence of a gender bias has been conducted in the football domain specifically before, and the previous literature in this domain has been conducted in the US, it may not be too far fetched to believe we

actually found the true (null) effect.

5.2 Internal Validity

With our framed field experiment, we are able to examine whether there is a raw gender bias in an isolated setting, with almost everything except gender being constant. We have the power to detect a small effect size, a thorough pre-analysis plan made before conducting the study, and a pilot study. Therefore, we argue that the internal validity of this thesis is high.

However, the threats to internal validity that we believe might have affected our results or weakened confidence in our lack of causal relationship are related to the procedure of conducting the experiment. One consideration of this would be if a considerable share of the participants would have had a low attention span and not fully absorbing the treatment due to lack of focus on the video. While conducting the experiment, we did observe a few participants with seemingly low attention span while watching the video. However, only 0.6 % answered the attention question incorrectly, 6 % answered the treatment question incorrectly, and 70 % answered that they learned something from the video, suggesting that a possible effect of lack of attention would be small.

Further threats to internal validity that we do not consider major, but nevertheless are worth mentioning, are that the two of us, conducting all the experiments, were female. In addition, teams that conduct the experiment after a game could potentially be affected by the gender of the referee. If the referee was female and the game went poorly, it could work as a negative prior.

Most issues that could have arisen and be a threat to internal validity are solved through the randomization in our experiment, as well as through using fixed effects in our regressions. Examples of such problems that could have affected our result are if evaluations have been affected by the mood of the players, for example, after losing a game, they might be more negative and rating both videos lower, compared to a team that just won a game. Furthermore, participants in the experiment who knew the female football player in our video could have been more positive to our video, as well as the few players who knew one of the experimenters. Additionally, the evaluations could differ in the digital sessions due to not meeting

the examiners in real life. However, even if these circumstances would affect some of the individual ratings, the treatment was orthogonal, which means that we randomized the videos at the team level.

When it comes to the video, one factor that could have influenced the evaluation of the coaches, even though we tried to keep everything else constant, is that some players might have perceived one of the coaches to be older than the other, potentially assigning him/her a greater confidence for that reason. However, we believe that this effect, if it exists at all, would be very small and would not affect our result. When designing the experiment, we took this into consideration and chose two coaches with the same age to minimize the risk of age bias.

5.3 External Validity

There are two ways of looking at the external validity of this research. Firstly, through how well our results are applicable in other settings of football, such as in non-elite environments, on a senior level or in other countries. Secondly, if our results can be applied in other settings than football.

Starting with the context of football, it is interesting to discuss if and how our findings could be applied to different subject groups. Starting with the non-elite environment in Sweden - a much bigger part of Swedish football than the elite environment - one defining difference is that, even though it is still heavily skewed, in the non-elite environment there are more female coaches (Svensk Fotboll 2022). According to (Beaman et al. 2009) a possible gender bias should be reduced by that increase, and thus it is not too much of a stretch to insinuate that the lack of a raw gender bias in an elite environment is probably applicable to the non-elite environment as well. When it comes to a senior and abroad setting, one must keep in mind the special context of Sweden as a country, as we are one of the top countries in terms of gender equality in the world. Therefore, even if there is a lack of raw gender bias in Sweden, it does not mean that there is none in other countries in the world. This reasoning can be applied for the senior teams in Sweden as well, players from all over the world are playing in the teams which could possibly affect a gender bias. We would thus argue that our results are not applicable outside of Sweden or in a senior environment.

The second dimension of external validity is whether our finding(s) can be translated to domains other than football. As mentioned in the introduction, sports are an integral part of society. Therefore, (in)equality in the football setting should arguably reflect other parts of society as well. In this research, we were unable to reject the null hypothesis that it does *not* exist a gender bias from players in the evaluation of football coaches. Does this then mean that there does not exist a gender bias in any form of leadership evaluation in the Swedish society? We argue that the external validity of our thesis is weak in this parameter, making it difficult for us to draw any conclusion about any other domain than football. As leadership is highly contextual (Ayman 2004), and our sample group was specifically chosen as well as the level of treatment, the transversality of the result should be treated with caution when discussing other domains. However, we argue that there might be synergies between the (non-)gender bias of our sample pool with a potential (non-)gender bias in other settings where this group operates, for example, in other youth teams in sports similar to football or in a high school environment. With regards to the latter, it is worth having in mind that, in the football setting, we are testing for a gender bias in a male dominated context. As Mengel, Sauermann, and Zölitz (2019) found evidence for in her research, in courses that reflected more strongly stereotypical masculine characteristics, such as mathematics, gender bias was greater than in less male stereotypical courses, such as social sciences. With that in mind, the football setting examined in this research may be more likely reflected in a male dominated category of education.

5.4 Future Research

As stated before, this experiment has its limitations because it tries to isolate the raw gender effect. Any future research should focus on increasing the treatment, either through length or through a more intense treatment, or maybe both. An experiment with increased length could, for example, be designed in the same way as ours, though with more videos of different technical skills, every week for two months. Or, to intensify the treatment, have the female or male coach on site to instruct the skill.

To further investigate why there are so few female football coaches in Sweden today, future research should branch out from players and look for prejudice from

other parts of the industry as well. For example, are the female coaches hired based on the same criteria as male coaches and do their colleagues evaluate them the same as male equivalences? An interesting dimension would be to investigate what the chief executive managers at elite football clubs in Sweden believe the result of our study (of which they are acquainted) would be. Could it be that the managers hiring the coaches believe there is a bias from players and favors male coaches based of faulty assumptions? Unfortunately, this was beyond the scope of our thesis.

Another way of deepening the knowledge of gender bias within sports is to look outside of the football context. For instance, similar experimental methods in other male dominated sports, such as ice-hockey, could support or debunk our results. Also, since we know (Born, Ranehill, and Sandberg 2018) that women prefer to lead in female dominated contexts, while men happily lead in both, it would be interesting to further investigate this question in a female dominated sport. Is the same skewed distribution present and is it driven by similar mechanisms?

5.5 Going Forward

Based on the results of this research, we cannot find any support for the existence a gender bias from players when evaluating football coaches. Therefore, we cannot support that gender bias is a driving factor behind why there are so few female football coaches in the elite environment in Sweden today. Thus, for the Swedish Football Association (SvFF) to reach its goal of 40 % female football coaches by 2025, efforts are proposed not to be directed toward this specific channel - as there is no clear evidence that there exists a bias in the evaluations of female coaches.

However, there is an inequality in the ratio of female/male coaches and the reasons for this are likely tenfold. As mentioned in the literature review, there exist multiple theories and research on different channels that explain why there are so few female leaders in general, and these channels may likely translate from the societal level to the sports level. It may be due to women having lower willingness to lead (Born, Ranehill, and Sandberg 2018) that women impose self-limiting behavior due to having less confidence in male-dominated contexts (ibid.). It may also be due to women incurring a loss in utility from deviating from their

gender role if taking on a leadership role that is described in masculine terms (Coffman 2014). It may also be due to the existence of a gender bias that we could not detect in this research, or the beliefs of whether a gender bias exists or not. Maybe our findings mirror the reality - that no gender bias in evaluation exists from players - but that the sport managers hiring the coaches believe this is the case and hire male coaches, thinking they will get higher evaluated by the players of the teams. This would be an interesting avenue for future research, as mentioned in the previous section.

6 Conclusion

What once was seen as a man's world is now opening up to women and the skewed distribution between men and women in leadership positions is beginning to change (Grant Thornton 2022; Economic co-operation and Development 2018). However, not in every industry: football is falling behind (Acosta and Carpenter 2012) and this thesis aims to investigate why. In Sweden's top divisions for men and women (OBOS Damallsvenskan, Allsvenskan, Elitettan, and Superettan), only 6 % of the head coaches were women in the 2022 season, and if the Swedish Sports Confederation wants to achieve its goal of having the share of male / female coaches at least 40 % by 2025 within each sport at the youth level, as well as at the national team level (Riksidrottsförbundet 2022), something needs to change.

Due to the similarities between academia and the elite football environment (hierarchical, high-performing, and pedagogical environments), we investigated whether the established gender bias from students to teachers (Mengel, Sauermann, and Zölitz 2019) can also be found between players and coaches.

Through a framed field experiment (Harrison and List 2004) we have tested 505 players between 14 and 20 years of age at the elite youth level to see if they would rate a female football coach lower than a male football coach, solely because of her gender. The participants watched a video of a technical skill and then evaluated it. Within each team, half of the participants were instructed by a female coach and the other half by a male coach. In our main regression, there was no evidence of a difference between the two treatment groups, a result supported by subanalyses and robustness checks. Despite what could be argued as a low treatment and a small sample size, we believe that we are powered to detect an economically significant effect and that our internal validity is high.

Does this mean that there is no gender bias to be found within football? Although we would love to claim this, we believe that is not the case. There are still many things that a woman need to overcome to break into male dominated environments, and even more to lead in them. However, we would argue that youth players at an elite level at least will not dismiss a female coach before even meeting her, solely based on her gender.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge. 2022. "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics* 138 (1): 1–35.

Acker, J. 1990. "Hierarchies, jobs, bodies: A theory of gendered organizations." *Gender & society* 4 (2): 139–158.

Acosta, R. V., and L. J. Carpenter. 2012. "Women in Intercollegiate Sport: A Longitudinal, National Study. Thirty-Five Year Update, 1977-2012." *Acosta-Carpenter.*

Akerlof, G. A., and R. E. Kranton. 2000. "Economics and identity." *The quarterly journal of economics* 115 (3): 715–753.

Anderson, E. D. 2009. "The maintenance of masculinity among the stakeholders of sport." *Sport management review* 12 (1): 3–14.

Ayman, R. 2004. "Situational and contingency approaches to leadership."

Bagilhole, B. 2002. "Challenging equal opportunities: Changing and adapting male hegemony in academia." *British journal of sociology of education* 23 (1): 19–33.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova. 2009. "Powerful women: does exposure reduce bias?" *The Quarterly journal of economics* 124 (4): 1497–1540.

Boring, A. 2017. "Gender biases in student evaluations of teaching." *Journal of public economics* 145:27–41.

Born, A., E. Ranehill, and A. Sandberg. 2018. "A man
s world?–The impact of a male dominated environment on female leadership."

Bowles, H. R., L. Babcock, and L. Lai. 2007. "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask." *Organizational Behavior and human decision Processes* 103 (1): 84–103.

Brewer, M. B. 2007. "The importance of being we: human nature and intergroup relations." *American psychologist* 62 (8): 728.

Burton, L. J. 2015. "Underrepresentation of women in sport leadership: A review of research." *Sport management review* 18 (2): 155–165.

Burton, L. J., and S. Leberman. 2017. "An evaluation of current scholarship in sport leadership: Multilevel perspective." *Women in sport leadership,* 16–32.

Carson, C. 2012. "The effective use of effect size indices in institutional research." In *31st Annual Conference Proceedings,* vol. 41.

Ceci, L. 2021. *YouTube average video length by Category 2018,* April. https://www.statista.com/statistics/1026923/youtube-video-category-average-length/.

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

Conviva. 2022. *Conviva's State of Streaming Q4 2021.*

Cox III, E. P. 1980. "The optimal number of response alternatives for a scale: A review." *Journal of marketing research* 17 (4): 407–422.

Cunningham, G. B., and P. Chelladurai. 2015. *Diversity & inclusion in sport organizations: A multilevel perspective.* Routledge.

Cunningham, G. B., and M. Sagas. 2002. "The differential effects of human capital for male and female Division I basketball coaches." *Research Quarterly for Exercise and Sport* 73 (4): 489–495.

———. 2007. "Examining potential differences between men and women in the impact of treatment discrimination." *Journal of Applied Social Psychology* 37 (12): 3010–3024.

Dixon, S. 2022. *Length of top performing videos on Facebook worldwide from 2017 to 2019, by reaction type,* April. https://www.statista.com/statistics/1001943/duration-top-performing-videos-facebook-reaction-types-worldwide//.

Duehr, E. E., and J. E. Bono. 2006. "Men, women, and managers: are stereotypes finally changing?" *Personnel psychology* 59 (4): 815–846.

Eagly, A. H., and S. J. Karau. 2002. "Role congruity theory of prejudice toward female leaders." *Psychological review* 109 (3): 573.

Economic co-operation, O. for, and Development. 2018. *Is the Last Mile the Longest?: Economic Gains from Gender Equality in Nordic Countries.* OECD.

FIFA, .-. 2019. *A breakthrough year for women's football.* https://www.fifa.com/tournaments/womens/womensworldcup/france2019/news/2019-a-breakthrough-year-for-women-s-football.

Fink, J. S. 2008. "Gender and sex diversity in sport organizations: Concluding comments." *Sex Roles* 58 (1): 146–147.

Grant Thornton, .-. 2022. *Women in business 2021.* https://www.grantthornton.global/en/insights/women-in-business-2021/.

Grappendorf, H., A. Pent, L. Burton, and A. Henderson. 2008. "Gender Role Stereotyping: A Qualitative Analysis of Senior Woman Administrators' Perceptions Regarding Financial Decision Making." *Journal of Issues in Intercollegiate Athletics.*

Harrison, G. W., and J. A. List. 2004. "Field experiments." *Journal of Economic literature* 42 (4): 1009–1055.

Haudenhuyse, R. P., M. Theeboom, and E. A. Skille. 2014. "Towards understanding the potential of sports-based practices for socially vulnerable youth." *Sport in Society* 17 (2): 139–156.

Hawkins, S., H. He, G. Williams, and R. Baxter. 2002. "Outlier detection using replicator neural networks." In *International Conference on Data Warehousing and Knowledge Discovery,* 170–180. Springer.

Hovden, J. 2010. "Female top leaders–prisoners of gender? The gendering of leadership discourses in Norwegian sports organizations." *International Journal of Sport Policy and Politics* 2 (2): 189–203.

Kernan, W. N., C. M. Viscoli, R. W. Makuch, L. M. Brass, and R. I. Horwitz. 1999. "Stratified randomization for clinical trials." *Journal of clinical epidemiology* 52 (1): 19–26.

Kiesler, S. B. 1975. "Actuarial prejudice toward women and its implications." *Journal of Applied Social Psychology* 5 (3): 201–216.

Knoppers, A., B. B. Meyer, M. Ewing, and L. Forrest. 1991. "Opportunity and work behavior in college coaching." *Journal of Sport and Social Issues* 15 (1): 1–20.

Kozlowski, S. W., and K. J. Klein. 2000. "A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes."

LaVoi, N. M., and A. Baeth. 2018. "Women and sports coaching." In *The Palgrave handbook of feminism and sport, leisure and physical education,* 149–162. Springer.

MacNell, L., A. Driscoll, and A. N. Hunt. 2015. "What's in a name: Exposing gender bias in student ratings of teaching." *Innovative Higher Education* 40 (4): 291–303.

Mengel, F., J. Sauermann, and U. Zölitz. 2019. "Gender bias in teaching evaluations." *Journal of the European economic association* 17 (2): 535–566.

MMS. 2021. *OS i Tokyo 2021.* https://mms.se/wp-content/uploads/_dokument/rapporter/tv-tittande/evenemang/2021/Rapport,+%5C%20OS%5C%20Tokyo%5C%202021.pdf.

Powell, G. N., and D. A. Butterfield. 2003. "Gender, gender identity, and aspirations to top management." *Women in management review.*

Riksidrottsförbundet, (2022. *Idrotten i siffror 2021.* https://www.rf.se/globalassets/riksidrottsforbundet/nya-dokument/nya-dokumentbanken/idrottsrorelsen-i-siffror/2021-idrotten-i-siffror---rf.pdf?w=900&h=700.

Sartore, M. L., and G. B. Cunningham. 2007. "Explaining the under-representation of women in leadership positions of sport organizations: A symbolic interactionist perspective." *Quest* 59 (2): 244–265.

Schnall, S., J. Benton, and S. Harvey. 2008. "With a clean conscience: Cleanliness reduces the severity of moral judgments." *Psychological science* 19 (12): 1219–1222.

Shaw, S., and W. Frisby. 2006. "Can gender equity be more equitable?: Promoting an alternative frame for sport management research, education, and practice." *Journal of sport management* 20 (4): 483–509.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2016. "False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant."

Sportstatistik. 2022. *Medlemmar.* https://idrottsstatistik.se/foreningsidrott/medlemmar/.

Stokel-Walker, C. 2022. *Tiktok wants longer videos-whether you like it or not,* February. https://www.wired.co.uk/article/tiktok-wants-longer-videos-like-not?utm_source=twitter&utm_medium=social&utm_campaign=onsite-share&utm_brand=wired-uk&utm_social-type=earned.

Svensk Fotboll, .-. 2022. *Endast sex procent av fotbollstränare på elitnivå är kvinnor, nu lanseras Tränarlyftet.* https://aktiva.svenskfotboll.se/nyheter/2022/05/endast-sex-procent-av-fotbollstranare-pa-elitniva-ar-kvinnor--nu-lanseras-tranarlyftet/.

Van Velsor, E., S. Taylor, and J. B. Leslie. 1993. "An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness." *Human Resource Management* 32 (2-3): 249–263.

Whisenant, W. A. 2008. "Sustaining male dominance in interscholastic athletics: A case of homologous reproduction... or not?" *Sex Roles* 58 (11): 768–775.

World Economic Forum, .-. 2022. "Global Gender Gap Report 2022."

A Appendix: The Experiment

Links to Videos

Link to video with a male voice-over Click This

Link to video with a female voice-over Click This

Figure 2: Screenshot from Video

Script for the Video

This video is about volley shots. We will focus on when the ball comes from the side and in the air. A technical moment that every player needs to master but rarely gets any attention in the ordinary team practice.

When the ball comes toward you, the most important thing is to hit the ball, not the force. Since the ball is already coming towards you with force, you should only focus on hitting the ball and getting it on goal. It is easy to want to use all your force, but then you only risk missing the ball instead.

There are two things to focus on to succeed. First, follow the ball with your eyes through its whole path; if you look away for only a second, you risk missing the ball or hitting the ball incorrectly. And be on your toes. Even if you know where

the ball is going to land, the wind or another player can affect the path and you need to be able to adjust your position quickly.

As stated earlier, the speed of the ball is not created by the force you hit the ball with, but instead where the ball hits the foot and the foot hits the ball. On the foot there is a hard section, it feels like a lump. This hard section makes the ball go faster, if you hit a little to high or low on your foot, the ball will go slower. Point your toes so the foot tightens, is it crucial for speed and precision.

The place where the ball ends up is determined by where on the ball you strike. If you strike high, the ball will go down, if you strike in the middle the ball will go straight and if you strike below the ball it will go up. One of the most common mistakes is to hit the ball low, making it go above the goal. Thus, you should strike it in the middle or slightly higher up to press the ball forward or slightly downward.

To get a smooth and controlled pendulum movement as possible, the body needs to create a straight line from where you hit the ball to your head. To succeed with your pendulum movement, it is an advantage to strike the ball as late in its path as possible, thus as close to the ground as possible. That lowers the demands of balance, mobility, and strength in your core and supporting leg. If you instead try to hit the ball early in its path, when the ball is high up from the ground, the risk becomes great that your body bends at the core. Then, you might hit the ball incorrectly and bot speed and precision will deteriorate.

The goal the the volley is to change direction of the ball so that it goes where you want it to go. The ball will get the same direction as your body. Your body's power-direction thus becomes important. The first step is to direct your supporting leg where you want the ball to go. The body will follow and the ball will follow the body. The second step is where to hit the ball, as we have discussed earlier. And the final step is to follow through your motion after striking the ball. You do that by letting go of your supporting leg and continue moving forward. It will give you a more natural pendulum movement and it makes it easier for you to be part of the next moment in the game.

Script Before a Session

Hello,

we are currently writing our master thesis in economics at Stockholm School of Economics, we are doing this together with Elite Football Women (EFD) and Swedish Elite Football (SEF). This survey is conducted on every team in Svenska Spel f19 och p19 Allsvenskan, which in why you are doing this today and we are very grateful for that. We are investigating how to best teach a technical skill through video. This is what is going to happen now: you will get one QR-code each, which you scan with your phone. That will direct you to a site with a YouTube-link. It is important that you click on the link before you click "next", because if you press next the link will disappear and you will not be able to return. Use your own headphones, your own phone and your own QR-code.

So, click the link, watch and listen to the video. After that, go back to the survey and complete it. If you happen to press "next" to quick anyway, scan the code through Snapchat instead.

Please be quiet throughout the whole experiment so everyone can focus.

Survey as Seen by the Participants

Figure 3: Survey as seen by participants

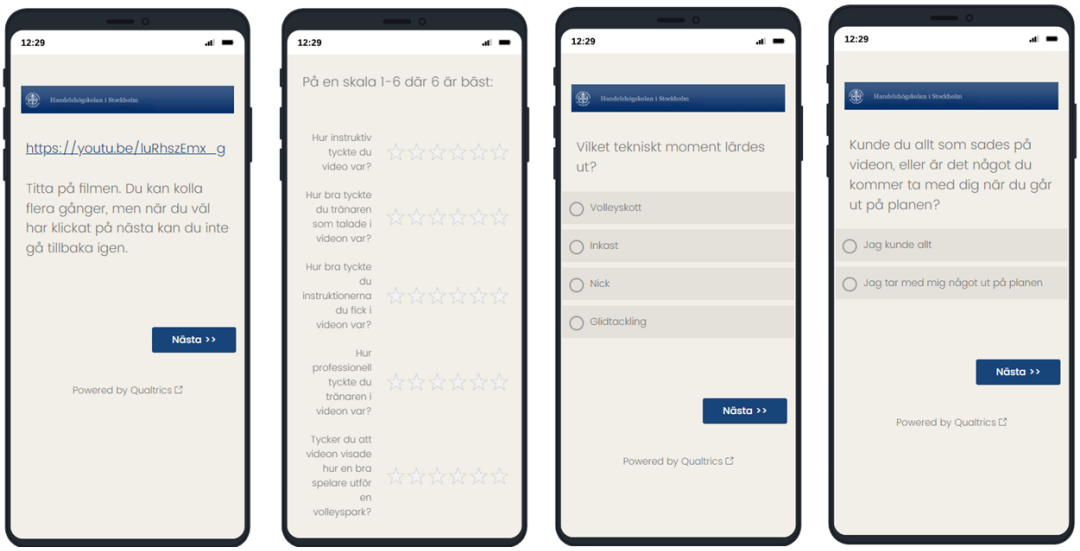


Figure 4: Survey as seen by participants

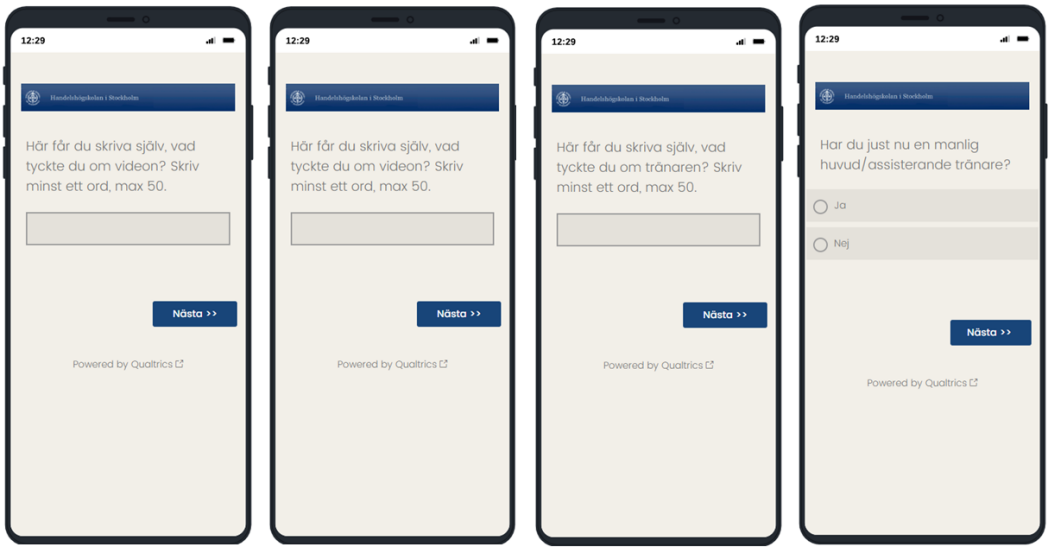


Figure 5: Survey as seen by participants

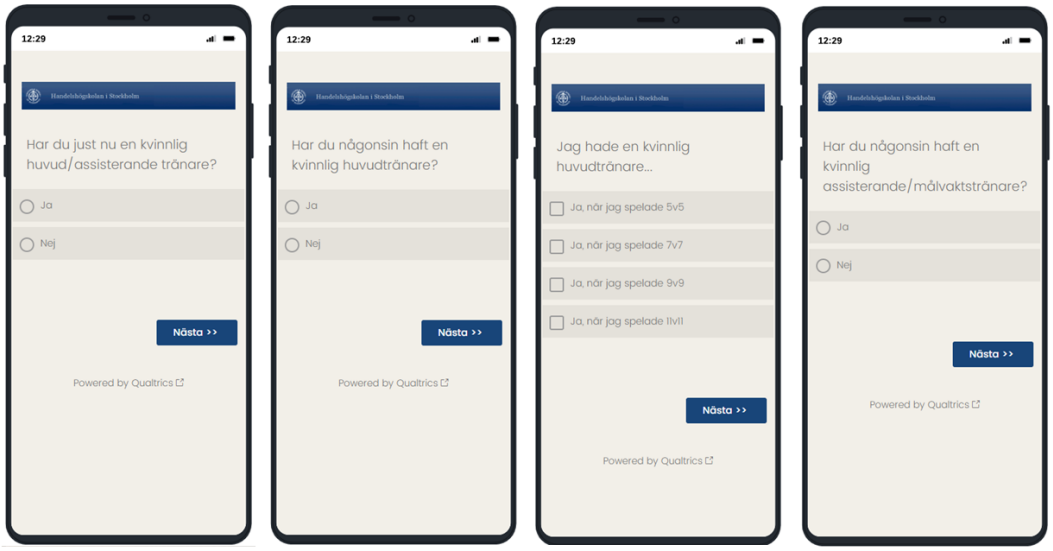
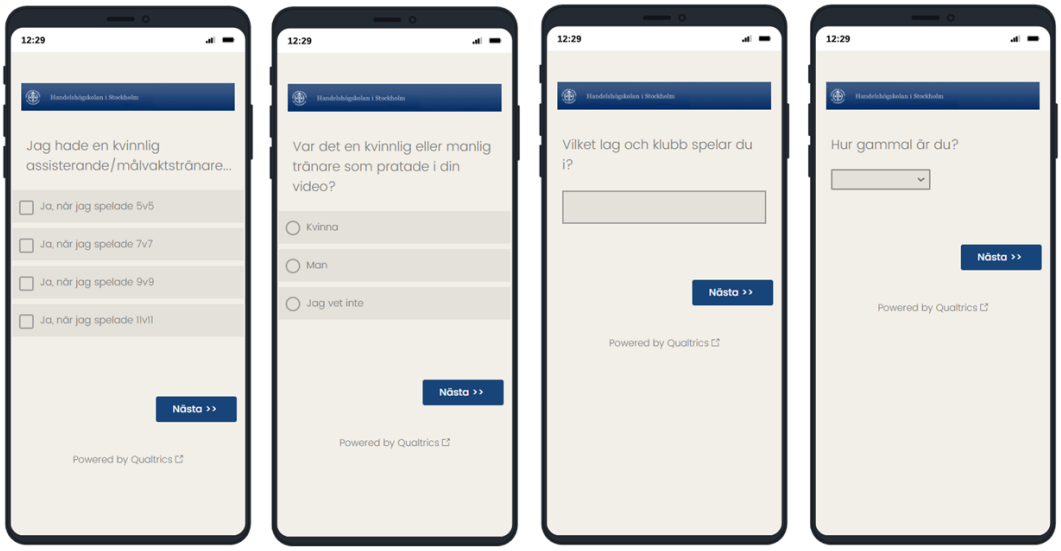


Figure 6: Survey as seen by participants



Survey - Translated

1. Watch the video. You can watch as many times as you want, but when you have pressed "next" you cannot go back again.

2. On a scale of 1-6 where 6 is the best:

    - how instructive did you find the video?
    - how did you find the coach speaking in the video?
    - how did you find the instructions in the video?
    - how professional did you find the coach in the video?
    - do you think the video showed how a good player would perform a volley shot?

3. What technical skill was taught? Volleyshot/throw-in/header/slide tackle.

4. Did you know everything in the video, or is there something you will take with you out on the field? I knew everything/I take something with me.

5. Here you get to write by yourself, what did you think of the video? Write at least one word, maximum 50.

6. Here you get to write by yourself, what did you think of the coach? Write at least one word, maximum 50.

7. Do you currently have a male head/assistant coach? Yes/no.

8. Do you currently have a female head/assistant coach? Yes/no.

9. Have you ever had a female head coach? Yes/no.

10. I had a female head coach...

    - Yes, when I played 5v5.
    - Yes, when I played 7v7.
    - Yes, when I played 9v9.
    - Yes, when I played 11v11.

11. I had a female assistant/goalie coach...

- Yes, when I played 5v5.
- Yes, when I played 7v7.
- Yes, when I played 9v9.
- Yes, when I played 11v11.

12. Were there a female or male coach talking in your video? Male/femle/I do not remember.
13. What club and team do you play in?
14. How old are you?

Instructions to Coaches for Teams where the Experiment is Conducted Digitally

Hello and again, thank you for participating in our study.

It is very important that the player do not know what we are testing, sp if you need to tell them something tell them that we are testing "how to best teach a technical skill through video".

Instructions before the survey: - print the two pages I have sent you and cut out the QR-codes. - remind the players to bring headphones and their phone. Maybe bring an extra pair if someone forgets.

Instructions during the survey: - hand out the QR-codes to the players. It is important that they get one each, they cannot share it. - everyone must use their own headphones, they cannot share. - if someone has forgotten they cannot participate, since they cannot do it with on speaker or share with someone. - it is important that they are quiet throughout the whole experiment.

We will describe the experiment during our introduction. Thank you again, "see" you on xx!

B Appendix: Robustness Check - Separated by Question

Table 5: Robustness check of main regression

	Dependent variable:				
	Instructive	The Coach	The instructions	Professional	Execution
	(1)	(2)	(3)	(4)	(5)
Female coach	0.07	−0.02	0.06	0.04	0.04
	(0.07)	(0.08)	(0.07)	(0.10)	(0.11)
Constant	5.29***	5.41***	5.43***	4.98***	4.44***
	(0.21)	(0.23)	(0.21)	(0.29)	(0.31)
Session FE	Yes	Yes	Yes	Yes	Yes
Observations	502	502	502	502	502
R^2	0.10	0.09	0.12	0.10	0.14
Adjusted R^2	0.04	0.03	0.06	0.04	0.09
Residual Std. Error (df = 470)	0.78	0.88	0.81	1.10	1.19
F Statistic (df = 31; 470)	1.69**	1.52**	2.06***	1.69**	2.56***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: Robustness check of sub-analysis 1

	Dependent variable:				
	Instructive	The Coach	The instructions	Professional	Execution
	(1)	(2)	(3)	(4)	(5)
Female coach	0.10	−0.04	0.18*	0.03	−0.02
	(0.10)	(0.11)	(0.11)	(0.14)	(0.15)
Female team member	0.31	0.33	0.64**	0.63	0.10
	(0.28)	(0.31)	(0.29)	(0.39)	(0.42)
Female coach*Female team member	−0.06	0.04	−0.23	0.02	0.11
	(0.14)	(0.16)	(0.15)	(0.20)	(0.21)
Constant	5.00***	5.07***	4.86***	4.34***	4.31***
	(0.18)	(0.20)	(0.19)	(0.26)	(0.28)
Session FE	Yes	Yes	Yes	Yes	Yes
Observations	502	502	502	502	502
R^2	0.10	0.09	0.12	0.10	0.15
Adjusted R^2	0.04	0.03	0.06	0.04	0.09
Residual Std. Error (df = 469)	0.78	0.88	0.81	1.10	1.19
F Statistic (df = 32; 469)	1.64**	1.47**	2.08***	1.63**	2.49***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: Robustness check of sub-analysis 2

	Dependent variable:				
	Instructive	The Coach	The instructions	Professional	Execution
	(1)	(2)	(3)	(4)	(5)
Female coach	0.12	0.05	0.17*	0.05	0.06
	(0.09)	(0.10)	(0.09)	(0.13)	(0.14)
Exposure	0.18	0.17	0.29*	−0.12	0.001
	(0.16)	(0.18)	(0.16)	(0.22)	(0.24)
Female coach*Exposure	−0.13	−0.18	−0.27*	−0.03	−0.03
	(0.14)	(0.16)	(0.15)	(0.20)	(0.22)
Constant	5.17***	5.32***	5.26***	5.10***	4.45***
	(0.24)	(0.27)	(0.25)	(0.34)	(0.37)
Session FE	Yes	Yes	Yes	Yes	Yes
Observations	502	502	502	502	502
R^2	0.10	0.09	0.13	0.10	0.14
Adjusted R^2	0.04	0.03	0.07	0.04	0.08
Residual Std. Error (df = 468)	0.78	0.88	0.81	1.10	1.19
F Statistic (df = 33; 468)	1.63**	1.47**	2.08***	1.59**	2.40***

Note: *p<0.1; **p<0.05; ***p<0.01

C Appendix: Robustness Check - High and Low Dispersion

Table 8: Robustness check: High or Low Dispersion

	Dependent variable: Mean rating of video	
	Mean rating of video	
	(High dispersion)	(Low dispersion)
Female coach	0.08	0.06
	(0.08)	(0.09)
Constant	5.25***	4.95***
	(0.23)	(0.24)
Session fixed effects	Yes	Yes
Observations	254	248
R^2	0.20	0.27
Adjusted R^2	0.09	0.16
Residual Std. Error	0.60 (df = 222)	0.66 (df = 216)
F Statistic	1.80*** (df = 31; 222)	2.52*** (df = 31; 216)
Note:		*p<0.1; **p<0.05; ***p<0.01

D Appendix: Robustness Check - Treatment Question

Table 9: Robustness check: Treatment Question on Mean Rating

	Dependent variable:
	Mean rating of video
Female coach	0.06
	(0.06)
Not correct on treatment question	0.33*
	(0.19)
Female coach*Not correct on treatment question	−0.39
	(0.28)
Constant	5.10***
	(0.19)
Session fixed effects	Yes
Observations	502
R^2	0.15
Adjusted R^2	0.09
Residual Std. Error	0.70 (df = 468)
F Statistic	2.53*** (df = 33; 468)
Note:	*p<0.1; **p<0.05; ***p<0.01