Working Paper 15
2024

Orchestrating Equality of Opportunities

Sex segregation and gender bias in decision-making

Maxime Parodi, OFCE, Sciences Po
Hélène Périvier, OFCE, Sciences Po
Audrey Etienne, Univ. Rouen Normandie
Reguina Hatzipetrou-Andronikou, Nantes université, CENS
Hyacinthe Ravet, Sorbonne Univ., IReMus

Published : 2024-09-20
Modified : 2025-09-25

CONTACT

OFCE
10 place de Catalogne
75014 Paris, FRANCE
Tel : +33 1 44 18 54 24
https://www.ofce.sciences-po.fr

Orchestrating Equality of Opportunities

Sex segregation and gender bias in decision-making

Inspired by the seminal paper of Goldin and Rouse (2000), we use a similar quasi-experimental framework and data from actual auditions for French orchestras. While their study documented the impact of blind auditions on the likelihood of women being hired, we extend their approach by investigating the underlying mechanisms that explain why blind auditions matter. In particular, we focus on the role of gender stereotypes in jury decisions. The high degree of sex segregation among instruments and the heterogeneity of orchestras auditions provide a unique context for assessing how stereotypes shape decision-making. We compare the decisions made in blind auditions to those made in nonblind auditions, for female and male musicians and for different levels of sex segregation by instrument. Using a triple differences design, we estimate a mixed-effects logistic model to account for the clustering structure of the data. When a screen is used to hide the candidates, a woman (a man) is more likely to be selected for male (female) instruments than a man (a woman). Conversely, when the jury sees the candidates perform, they tend to select musicians whose gender is in the majority among the instrumentalists. By comparing the probability of success for women or men in blind and nonblind auditions, we show that the sex segregation of instruments impairs the impartiality of judges and prevents them from selecting the best musician. (JEL codes: J7, J16)

Maxime Parodi, maxime.parodi@sciencespo.fr

Hélène Périvier, helene.perivier@sciencespo.fr

Audrey Etienne, audrey.etienne3@univ-rouen.fr

Reguina Hatzipetrou-Andronikou, reguina.hatzipetrouandronikou@univ-nantes.fr

Hyacinthe Ravet, hyacinthe.ravet@sorbonne-universite.fr

Remerciements

This research was funded by the French Agency for Research (ANR), project ANR-17-CE41-0010 PRODIGE and by the Alliance de Recherche sur les Discriminations-Domaine d'intérêt majeur « Genre, inégalités, discriminations », Région Ile-de-France (2013-2014). We thank the French Association of Orchestras (AFO) and the orchestras without whom this research would not have been possible. We also thank our colleagues from the OFCE seminar, as well as David Card, Anne Boring and Grégory Verdugo for their insightful comments.

Table of contents

1 Introduction

Sex segregation in the labour market influences the gender representation of jobs by recruiters. This leads to stereotypes, which can be defined as « mental representations of real differences between groups » (Hilton and Von Hippel, 1996). The decision to hire someone can be influenced by this biased view. In return, gender bias reinforces sex segregation. To assess this unfair treatment, it is crucial to identify the role of gender stereotypes in decision-making. However, it is extremely difficult to prove empirically how the judgment to select or reject a candidate is influenced by such bias. In this regard, correspondence and experimental methods are widely used in the literature to measure the role played by stereotypes (Bertrand and Duflo, 2017 ; Bertrand and Mullainathan, 2004 ; Booth and Leigh, 2010 ; Bordalo *et al.*, 2016, 2019 ; Carlsson, 2011 ; List, 2004 ; Neumark, Bank and Van Nort, 1996). In the education literature, quasi-experiments based on administrative data from both anonymous and non-anonymous tests and exams have also been used to shed light on how gender biases affect decision-making (Breda and Hillion, 2016 ; Breda and Ly, 2015 ; Lavy, 2008). From another perspective, Goldin and Rouse (2000) uses the recruitment process in orchestras as a quasi-experiment and show that blind auditions significantly increased the probability that women would advance or be hired in US orchestras between the 1950s and the 1990s.

Inspired by this seminal paper, we examine the hiring of musicians in permanent French orchestras to investigate the mechanisms underlying this result. The music industry offers specific advantages from this perspective. First, the judges can evaluate the quality of the candidates without seeing the musicians. In these processes, the candidates are hidden from the jury during their individual performances. This obscuring usually takes the form of a screen or a curtain or a room divider being placed in front of the musician. We use the term « blind auditions » to refer to these procedures. In this way, characteristics (such as gender) are then separated from objective criteria (such as the sound of a musical performance), which then become the only ones considered in the decision-making process. Second, the employment structure within orchestras is stable, with both the composition of the workforce by instrument type and the total number of musicians remaining largely unchanged. Third, the sex segregation of instruments is pronounced, deeply rooted, and well known by various actors in the field of classical music, including the professionals serving on selection committees. The heterogeneity of instruments regarding the proportion of female musicians in orchestras can thus be used to determine how sex segregation affects the representations of what constitutes a good musician in the eyes of the jury and subsequently their decisions.

We have built an original individual dataset from actual recruitment competitions organized by 13 French orchestras between 2001 and 2021. We use blind and non-blind auditions as a quasi-experiment to test the role of gender stereotypes in decision-making. Following a triple difference design (DDD), we estimate a mixed-effects logistic model to account for the clustering structure of the data (orchestras, competitions, rounds within the same competition). We compare the decisions made in blind auditions, where a screen hides the identity of the candidate from the panel of judges, to those made in non-blind auditions, for female and male musicians as well as for instruments with different proportions of female musicians. We then assess how gender bias affects the judgement of decision makers.

Relative to Goldin and Rouse's paper, our contribution is to leverage the heterogeneity in the degree of instrument feminization to identify the role of gender stereotypes in jury bias. More broadly, we contribute to the literature by precisely measuring how gender bias, reinforced by sex segregation in the workplace, shapes decision-making. Compared with the existing literature, our approach provides a framework with several advantages. Firstly, in contrast to correspondence and experimental methods, at each round the jury makes a costly, binding and real decision about who continues in and who is eliminated from the process. Secondly, blind auditions in the music sector provide a framework in which the counter-factual is not a similar candidate of the opposite gender but rather a candidate for whom this information is not available, that is, a neutral reference. Thirdly, most studies about decision bias and discrimination do not address the possible past selection of the two pools of candidates. We avoid this *survivorship bias*, documented by Wald (1943) and Eldridge (2024), by comparing the probability of women and men being selected when the audition is blind. Fourthly, unlike some studies in the education literature (Breda and Hillion, 2016 ; Breda and Ly, 2015 ; Lavy, 2008), our empirical framework provides a context for assessing how the decisions change for comparable and similar performances.

Our findings are twofold. By comparing the probability of success for female and male musicians in blind auditions, we show that musicians who do not comply with gender norms by playing an instrument in which their gender is under-represented outperform those who do. We interpret this result as a « survival effect » throughout the training process that helps shape the quality of the candidate pool by gender prior to the hiring process. By comparing the probability of success in blind and non-blind auditions, we show that the gender segregation of instruments impairs the impartiality of judges and prevents them from selecting the best candidate. When a screen is used to hide the candidates, a woman (a man) is more likely to be selected for male (female) instruments than a man (a woman). When fewer than 40% of musicians playing a given instrument are women, the jury exhibits a gender bias in favor of men. Notably, a large number of instruments within orchestras fall below this threshold. Blind auditions promote impartiality by neutralizing these biases. Reducing sex segregation mitigates the influence of gender stereotypes on decision-making. We estimate that a 10% increase in the proportion of women within an instrument raises the odds of a woman, relative to a man, passing a non-blind audition by a factor of 1.47.

2 Descriptive statistic

2.1 Orchestral Structure and Recruitment Procedures

In France, there are approximately 36 permanent orchestras, employing nearly 2,500 musicians. These ensembles are funded by public authorities (the state and/or local governments) and are spread across the country, often located in major regional cities or integrated into opera houses. The employment structure of orchestras remains stable, both in the number of positions and in the types of instruments represented. Therefore, changes in the share of female musicians within an orchestra cannot be explained by variations in its composition or size.

To select the best musicians, orchestras organize recruitment competitions for each instrument and position. The orchestras announce that a competition will be held. These competitions attract musicians from across the country and abroad. Each competition is organized into several rounds, most often three (preliminary, semi-final and final), although a competition can last up to five rounds. In each round, musicians play their instrument individually in front of the selection committee. Most of the music that candidates are expected to perform during the live auditions is provided in advance. The composition of the selection committee is the same for each round. The preliminary round is mainly used to eliminate unqualified candidates. Candidates selected for the semi-final can be considered potentially suitable for hiring. The final round generally results in a hire; in some cases, however, no candidate is selected if none reach the level expected for hiring.

The principle of blind auditions in the recruitment of musicians has gradually been introduced in French orchestras to limit the effects of co-optation within the circle of former students and acquaintances of the conductor and/or assistant conductor (Ravet, 2011). It has become a tool to guarantee the impartiality of the judges and to reduce the influence of gender stereotypes. During individual performances, candidates are hidden from the jury, typically by a screen, curtain, or room divider placed between the musician and the jury. Under these conditions, the judges can assess the quality of the performance and compare candidates without seeing them. Characteristics such as gender are then separated from objective criteria (such as the sound of a musical performance), which are then the only ones considered in the decision-making process.

Even when a musician is not visible, the jury might still pick up cues about gender—for instance, from the sound of walking to the stage, if a female musician is wearing heels[1]. Most orchestras take measures to ensure the effectiveness of blind auditions—for example, by placing carpets to minimize this risk.

[1] In the early 1980s, when blind auditions were introduced in French orchestras, some female musicians avoided this issue by wearing large shoes so as not to be recognized (Ravet, 2011).

2.2 The data

We use two types of datasets. To identify the share of female musicians in French orchestras, we use data collected by the *Association Française des Orchestres* (AFO). Most French orchestras are members of the AFO. We also use an original dataset, the Ardige database, which is a collection of information on competitions organized by permanent French orchestras that are members of the AFO. Of the around 30 AFO orchestras, only 13 have maintained usable archives of the competitions they held to recruit their musicians. These competitions took place between 2001 and 2021, mainly during the 2010s. The database covers 322 competitions, we have collected detailed data on all rounds of each competition for every candidate. In each round, musicians play their instrument individually in front of the selection committee : the Ardige database contains 12,521 « individual performances ». The same musician can apply for several competitions. We observe 9,170 total candidates in the preliminary rounds and 5,298 musicians. Thus, the candidates are applied an average of 1.7 times. We observe a total 2,377 candidates in the semi-final rounds, 884 candidates in the final rounds and 283 candidates who are ultimately hired in total. As mentioned previously, some competitions do not result in a hire if the jury is ultimately unconvinced by the candidates' performances. In addition to quantitative data, we draw on ethnographic and qualitative evidence from interviews with candidates, orchestra administrators, and jury members (Hatzipetrou-Andronikou *et al.*, 2025).

The Ardige database includes 23 different instruments contained within a symphony orchestra, organized into 15 categories. These are, in decreasing order of the number of competitions observed in the database (Table 2.1) : violin, viola, double bass, cello, oboe & horns, French horn, trumpet, percussion, clarinets, bassoons, trombone, flute, tuba, piccolo and harp. Most of the competitions are for violinists, with 82 auditions observed (i.e. 25% of the competitions observed), followed by violists, with 47 competitions (15% of the competitions observed), then double bassists with 36 competitions (i.e. 11% of the competitions observed). 60% of the competitions are for strings, 19% are for winds, 16% are for brass instruments, and 5% are for percussion. Women represent 47.5% of the candidates and 43% of the hires.

2.3 Sex segregation of instruments

Since the 1970s, the number of female musicians hired in French symphony orchestras has increased (Ravet, 2011). On average, according to data from the French Association of Orchestras (AFO), women represented 32% of permanent orchestra musicians in 2000, increasing to 33% in 2010 and 36% in 2016. However, significant differences remain between instruments, with the proportion of women ranging from 1% to 85% depending on the instrument. The current representation of male and female musicians holding tenured positions in orchestras is characterized by a high degree of sex segregation. This segregation is deeply rooted in the history of this sector. Different factors have been proposed as possible causes of instrument gendering such as *appearance of an instrument, its manner of playing, approximation of pitch range of instruments to player's vocal range, educational and training opportunities, attitudes of teachers and music directors, and lack of female role models in secondary and higher education* ... (Sergeant and Himonides, 2019). Women are particularly represented among

Table 2.1: Descriptive statistics on competitions

	All rounds				After 1st round	
	Nber of competitions	Nber of ind. perf.	% of female perf.	% of ind. perf. with screen	Nber of ind. perf.	% of ind. perf. with screen
Violin	82	3543	64.5%	72.5%	952	44.0%
Viola	47	1467	65.2%	57.5%	401	31.7%
Double bass	36	809	35.0%	67.1%	260	28.5%
Cello	26	1234	53.6%	81.6%	314	45.1%
Oboe/Horn	22	841	37.8%	69.1%	236	41.1%
French Horn	20	670	17.0%	83.8%	163	52.5%
Trumpet	18	573	7.0%	64.4%	140	30.7%
Percussion	17	653	13.7%	54.5%	210	22.4%
Clarinet	15	751	30.5%	71.9%	181	33.7%
Bassoon	13	394	33.8%	55.3%	123	43.9%
Trombone	10	362	4.4%	59.4%	97	53.6%
Flute	9	861	72.3%	91.4%	172	57.0%
Tuba	3	135	5.9%	85.2%	28	28.6%
Piccolo	2	151	79.5%	90.7%	38	63.2%
Harp	2	77	89.6%	27.3%	21	0.0%
Total	322	12521	47.5%	70.8%	3336	39.9%

Source: Ardige
NB:Each competition consists of several rounds in which musicians perform individually, either behind a screen (blind) or without a screen (non-blind). In Ardige, we observe 82 competitions for violin, with a total of 3,543 individual performances across all rounds, of which 64.5% are given by female musicians, and 72.5% of these individual performances are blind auditions.

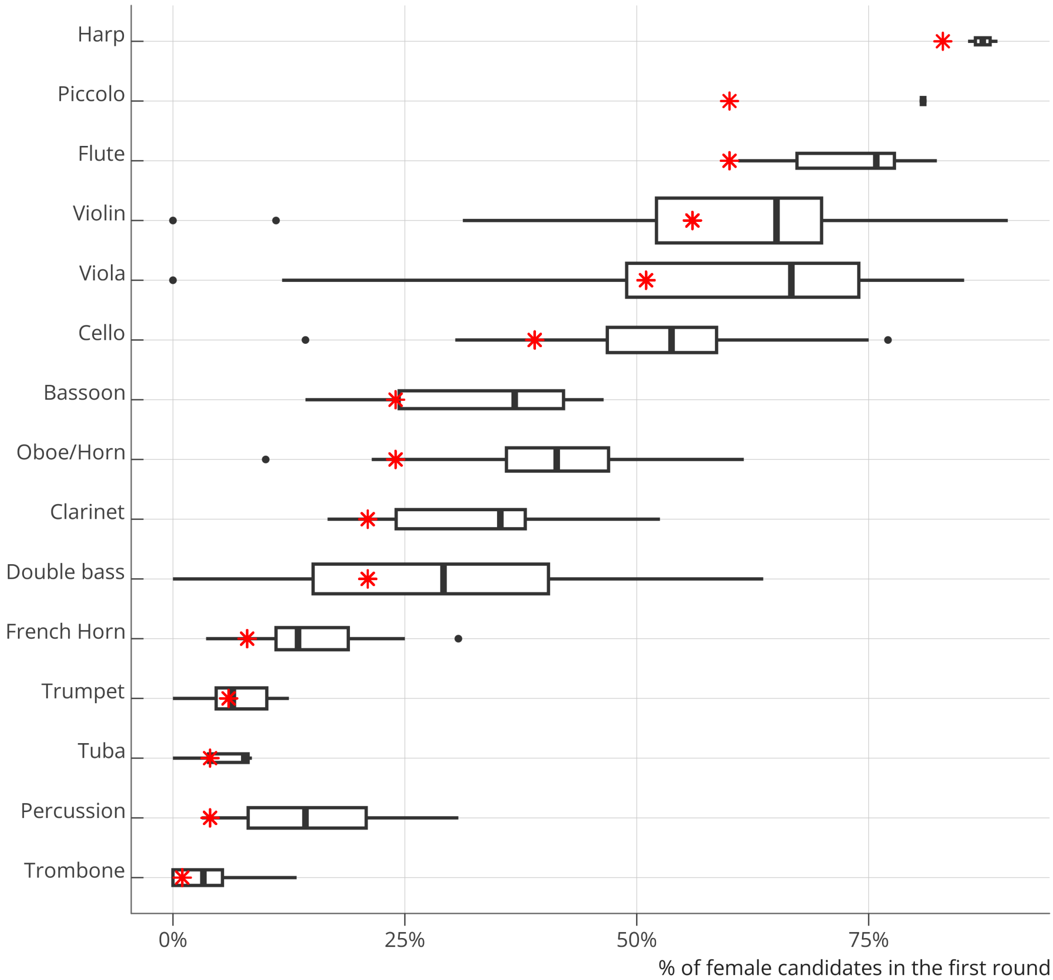
flutists, violists and violinists, whereas men outnumber women among trombonists, tuba players and horn players.

The proportion of female candidates in the auditions organized by the 13 orchestras within the Ardige database varies dramatically from one instrument to another (Table 2.1). As expected, women are particularly well represented in the string sections but almost absent in the percussion and the brass sections. To determine whether this segregation is representative of the overall situation in French orchestras, we compared our data to the representation of women per instrument in the 30 French orchestras published by the AFO for the 2016-2017 season (red stars, Figure 2.1). We observe that the ranking of instruments according to the percentage of female musicians is the same in the AFO data as in the Ardige database; i.e., trombone, percussion, tuba and trumpet are the least feminized instruments whereas harp, piccolo, flute and violin are the most feminized instruments. In the Ardige database, the proportion of women among the candidates is greater than it is observed in French orchestras for most instruments. This illustrates the ongoing feminization of the orchestras. Some auditions are oversubscribed in the preliminary round, such as for flute or piccolo whereas others are characterized by a small pool of candidates, such as for bassoon or double bass. Women are overrepresented in these overcrowded auditions (Figure 2.2).

More than 70% of the competitions rounds in the Ardige database take place behind a screen (Table 2.1). However, most orchestras use a screen only during the preliminary round. From the second round onward, only 40% of the rounds examined are conducted as blind auditions. Over the period considered, some orchestras have modified their practices[2]. Some orchestras, two in particular, use blind auditions throughout the recruitment process. Within the same competition, the decision to use a screen may vary from round to round; however, all candidates appearing in the same round of a given competition are subject to the same rule, whether blind audition or not. Considering the different rounds of all the observed auditions, the Ardige database provides enough variations in procedures to evaluate the impact of blind auditions on the gender of the selected musicians.

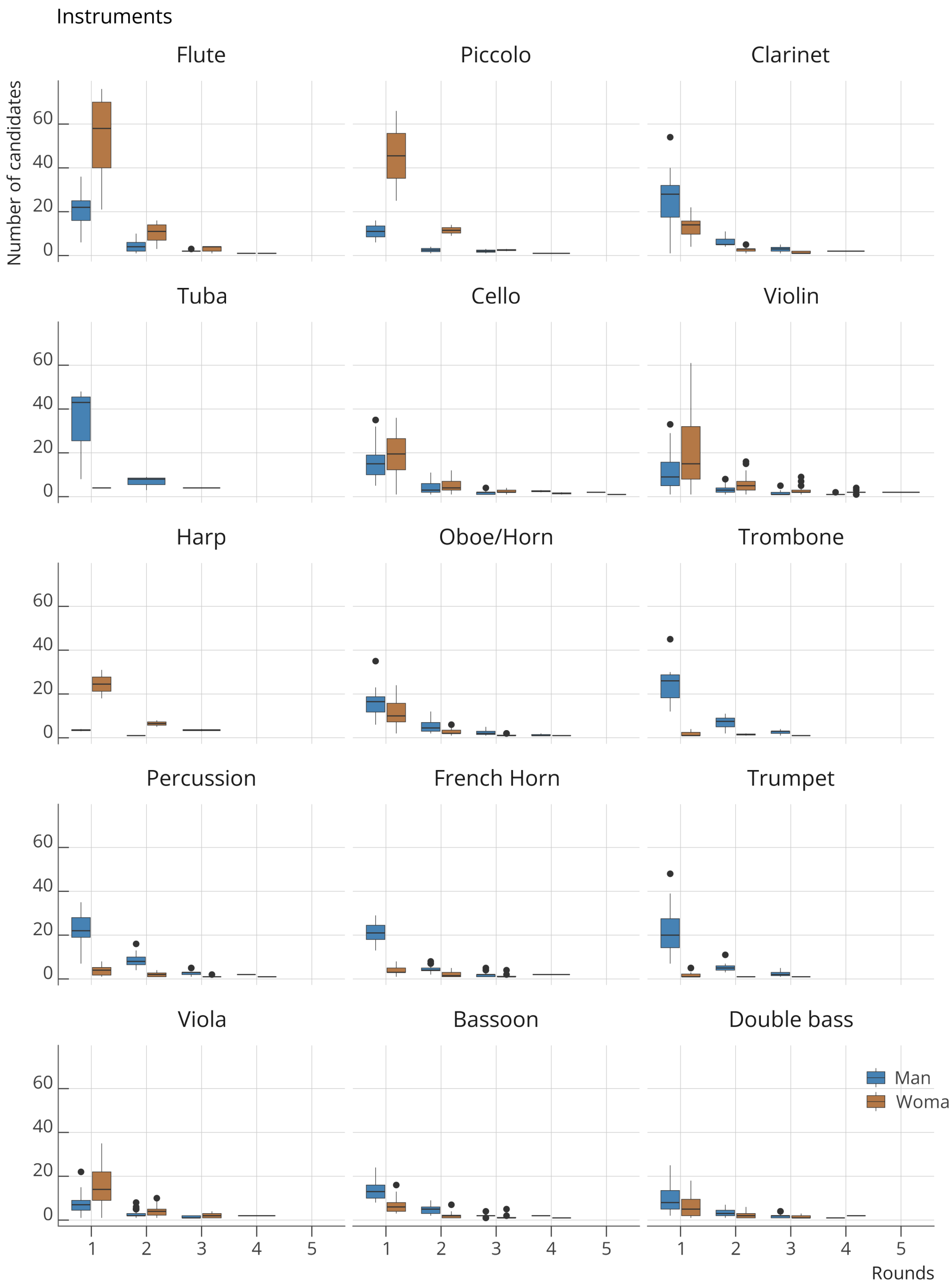
[2]While two orchestras have recently added blind auditions to their process, a few have dropped a blind auditions from their competitions.

Fig. 2.1: Distribution of competitions according to % of women among the candidates and % of female musicians observed in the French orchestras



Source: Ardige and Association française des orchestres (AFO).
NB: Red stars correspond to the average % of female musicians observed in orchestra belonging to the French Association of Orchestras in 2016-2017.

Fig. 2.2: Numbers of candidates in each round by gender and instruments



source: Ardige

3 Study design

3.1 The empirical strategy

Our aim is to determine whether gender stereotypes influence the jury's decision to pass a candidate to the next round. We use blind and non-blind auditions as a quasi-experimental design to test the role of gender stereotypes in shaping the jury's perception of individual performance quality. Each audition belongs to a set S of competing auditions, corresponding to a given round in an orchestra contest.

The use of blind auditions during some or all rounds of the selection process is mainly determined at the orchestra level, and applies to all instrument types. The selection process is the same for all the candidates of a given competition. We assume, firstly, that applicants do not self-select into competitions according to whether or not the screen is used in such or such rounds (assumption $A1$). Indeed, permanent positions in orchestras are rare, so musicians enter every possible competition to maximise their chances of being hired. Candidates cannot afford not to apply to one competition rather than another because of the presence or absence of blind auditions. In addition, once musicians enter a competition, they cannot avoid participating in a blind audition if one is scheduled as part of the process and most competition are anyway a mix of blind and non blind auditions[1].

Secondly, in a blind audition, we assume that the jury is unable to identify the musician's gender (assumption $A2$, further details can be found in the description of the recruitment procedure above.) Therefore, the judges' decisions are presumed to reflect the objective quality of the performance. Under this assumption, the musicians selected in blind auditions can be considered objectively better than those who are rejected.

Finally, it is possible that the nature of the audition, blind or non-blind, influences a musician's performance. We assume that, on average, the relative performance of female and male musicians within a set of auditions S is not affected by the presence or absence of a screen (Assumption $A3$). This hypothesis serves as the standard parallel trends assumption required for the validity of the difference-in-differences design.

Moreover, if we observe a gender gap in the probability of being selected in blind auditions, it would suggest that one gender outperforms the other. This disparity may reflect differences arising from selection processes during musicians' training. Blind auditions make it possible to identify such upstream selection effects and help prevent survivorship bias, which occurs when analyses focus only on individuals who have already passed a selection process while overlooking those who have

[1] The qualitative data support this finding, as candidates report that they do not choose competitions based on whether the audition procedure is blind or non-blind.

not. This is an important bias to consider, as it is in other fields such as finance (Elton, Gruber and Blake, 1996) and medical research (Ioannidis, 2005).

By comparing the probability of selection for female versus male musicians in blind and non-blind auditions, we can assess whether jury decisions are influenced by gender stereotypes. We identify these stereotypes by exploiting variation in female representation across instruments in French orchestras. The sex segregation of instruments is both deeply rooted and well known by various actors in the field of classical music, including by the professionals who sit on the selection committees. Therefore, this segregation may reinforce stereotypes that instruments with low female (male) representation are inherently less suited to women (men), thereby biasing the jury's perception of talent and fit.

3.2 The sample

We organize the Ardige dataset in such a way that each individual performance i (i.e., the performance of a musician in a given round in a given competition or person-round) becomes the basic statistical unit. Each competition is structured into multiple rounds during which individual musicians are auditioned. In each round, the jury eliminates candidates who are judged not to meet the required standard—either in absolute or relative terms—and retains those deemed qualified to advance to the next round. As a result, the structure of the dataset exhibits different levels of clustering: a cluster of individual performances by the same musician, a cluster of rounds of auditions embedded in competitions for a given instrument, which are themselves grouped by orchestra.

Therefore, the hypothesis of statistical independence between observation units may not be fully respected. Indeed, the same musician can be auditioned multiple times during the same competition, as they advance through different rounds. The same musician can also apply for different competitions[2]. Moreover, two performances by different musicians are not statistically independent if they occur in the same round of the same competition, since the jury evaluates them comparatively; an outstanding performance by one candidate may lead to the elimination of another. But, the main dependency is on the number of competitors in a round of a contest. The more competitors there are, the less chance there is of getting through. Finally, the orchestras have their own requirements, practices and reputations. That said, since each audition takes place in a specific context, not all these dependencies are self-evident: a candidate may be a good candidate for a particular competition, round, or stage of their career, but may not be at the required level for the next round.

The preliminary round of the competition differs from subsequent rounds for several reasons. First, some musicians who already work with the orchestra as non-permanent members may be allowed to bypass this stage. Second, the pool of candidates in the preliminary round is more heterogeneous and larger than in later stages of the competition, as some applicants apply despite not meeting the minimum required standard and are quickly eliminated by the jury. Consequently, there is little room for bias in the decision-making process during this round, since it primarily involves removing candidates who fail to meet the required criteria. Third, musicians who fail in the preliminary round

[2] In the Ardige database, 2/3 of the individuals apply for only one contest.

of a given competition—which occurs frequently due to the high selection standards at this stage—are likely to fail in the preliminary rounds of other competitions as well. Consequently, individual performances i cannot be considered independent, as evidenced by a high intraclass correlation coefficient (ICC) of 33% for performances at the preliminary round.

For these reasons, our statistical analysis focuses on individual performances observed after the preliminary round. In these subsequent rounds, the jury evaluates candidates whose skill levels are closely matched. Their relative performance becomes more important, and judgment bias may therefore play a more decisive role in the selection process. At that stage, individual performances can be considered quasi-independent, as the ICC is low (4.7%). Indeed, a musician's performance may vary from one audition to another due to factors such as the level of its competitors, the stakes and specific demands of the audition, their stress level, and so forth. Additionally, some orchestras have a set repertoire for certain rounds, and a musician's familiarity with this repertoire can vary. Thus, each round presents a distinct challenge with its own requirements.

The pool of candidates in a given round results from the selection of the previous round. Consequently, if the previous round was not blind, then the pool of candidates for the following round may already have been affected by a possible bias. To avoid a double counting, which would lead to an overestimation of the potential bias, we exclude from our sample rounds where candidates were previously selected through a non-blind audition. We then compare rounds where the pool of candidates was selected through blind auditions in the previous rounds. For this selected sample (excluding the preliminary rounds and including only rounds for which the previous round was blind), the ICC of the individual performances remains low at 6.6%. If we assume that the dependency of the individual performances extends only over several rounds of the same competition, and not over several competitions, then the ICC decreases to 5.6%. The ICC for orchestras is only 1.2%. The ICC for the instruments is less than 1%. We also calculate the ICCs of the three nesting variables together, which are 3.9% for individuals (all auditions), 0.7% for instruments and 1.1% for orchestras.

Although each performance of the same musician can be considered as quasi independent, we randomly select one performance for each individual among all the performances he or she has given. The relevant sample for the analysis includes 1,339 individual auditions, 44.7% of which were performed by women.

3.3 The model

Goldin and Rouse (2000) adopt a DiD strategy and estimate the change in the probability that a woman will be selected for the next round, or hired, when a screen is used compared to auditioning without one. Our data and quasi-experimental framework are similar, but we additionally incorporate the share of female musicians playing each instrument to identify gender bias. We then compare the probability of a woman versus a man being selected with a screen relative to without, across instruments characterized by their degree of feminization. We then adopt a triple-difference approach, following Gruber (1994). This method is widely used to identify the causal effect of a treatment (Olden and Møen, 2022).

We denote by Y_i the outcome of musician i's audition performance: either success (the musician is selected for the next round) or failure. The instrument is predetermined and fixed for each competition. We consider the use of a screen as the control condition, as during a blind audition, the jury's decision is unbiased. Conversely, the absence of a screen is considered as the treatment, the jury observes the candidate's gender while they are performing and may be influenced by gender bias. The performance during the audition i of a musician whose gender is G (female or male), playing an instrument $instr$, could be treated or not :

$$T_i = \begin{cases} 0 & \text{if the audition is blind (screen used)} \\ 1 & \text{if the audition is not blind (no screen)} \end{cases}$$

Assumption $A1$ ensures that candidates cannot anticipate the treatment, while assumption $A3$ requires that the relative performance of those who will be treated and those who will not follows the same average trend. Under these conditions, and following the formal approach proposed by Rubin (1974), we obtain:

$$\mathbb{E}(Y_i^{g,instr,0} | T_i = 1) = \mathbb{E}(Y_i^{g,instr,0} | T_i = 0)$$

The *Average Treatment effect on the Treated* (ATT) for each gender is :

$$ATT^{g,instr} = \mathbb{E}(Y_i^{g,instr,1}) - \mathbb{E}(Y_i^{g,instr,0} | T_i = 1)$$

To account for gender differences, we compute the difference-in-differences-in-differences (DDD):

$$ATT^{instr} = ATT^{female,instr} - ATT^{male,instr} \tag{3.1}$$

It can be written as follows :

$$\begin{aligned} ATT^{instr} &= \overbrace{\mathbb{E}(Y_i^{female,instr,1} - Y_i^{male,instr,1})}^{Gender\ and\ survival\ biases} - \overbrace{\mathbb{E}(Y_i^{female,instr,0} - Y_i^{male,instr,0})}^{Survival\ bias} \\ &= \underbrace{\mathbb{E}(Y_i^{female,instr,1} - Y_i^{female,instr,0})}_{Female\ stereotype} - \underbrace{\mathbb{E}(Y_i^{male,instr,1} - Y_i^{male,instr,0})}_{Male\ stereotype} \end{aligned}$$

The ATT measures the causal effect of being visible while performing, for the musicians who actually performed without the screen.

We denote by P_i the probability that the musician's performance audition i is successful, that is, that the musician is selected for the next round. We estimate the term P_i using either a logistic regression or a mixed-effects logistic model with each observations is an individual performance (a person-round)[3]. The estimation of the mixed-effects logistic model, also known as a multilevel logistic model, allows us to account for the nesting of observations within orchestras.

We consider the following independent variables:

- Gender of the candidate performing the audition i: $gender_i \in \{0, 1\}$ with 1 if female musician
- Number of candidates competing during the same set auditions s: N_s;
- T_i is the treatment (non-blind) or not (blind) of the audition $i : T_i \in \{0, 1\}$
- Share of female musicians within all French orchestras playing the same instrument as candidate performing the audition i: $genderInstr_i$.

This last variable is specific to each instrument and allows us to control for instrument effects by incorporating information on its degree of feminization. The data come from statistics published by the AFO on the percentage of female musicians per instrument in 2016–2017(Association Française des Orchestres, 2018).

The model is then specified as follows:

$$\begin{aligned} logit(P_i) = \alpha_0 &+ \alpha_1 \times N_s \\ &+ \beta_1 \times gender_i + \beta_2 \times genderInstr_i + \beta_3 \times T_i \\ &+ \gamma_1 \times gender_i \times genderInstr_i + \gamma_2 \times gender_i \times T_i + \gamma_3 \times genderInstr_i \times T_i \\ &+ \delta \times gender_i \times genderInstr_i \times T_i \\ &+ (optional)\ \theta_{orchestra} \\ with\ \theta_{orchestra} \sim \mathcal{N}(O, \sigma^2)\ & random\ effect\ for\ an\ orchestra \end{aligned} \tag{3.2}$$

Equation 3.2 represents a simple logistic model, which becomes a multilevel logistic model when the random effects $\theta_{orchestra}$ is included.

Following Equation 3.1, the estimated *Average Treatment effect on the Treated* , $\widehat{ATT}$, is then given by :

$$\begin{aligned} \widehat{ATT} = &\hat{\gamma_2}^{female,1} - \hat{\gamma_2}^{male,1} - \hat{\gamma_2}^{female,0} + \hat{\gamma_2}^{male,0} \\ &+ \hat{\delta}^{female,1}_{genderInstr.} - \hat{\delta}^{male,1}_{genderInstr.} - \hat{\delta}^{female,0}_{genderInstr.} + \hat{\delta}^{male,0}_{genderInstr.} \end{aligned}$$

Taking into account the different dummies, the estimated ATT can be expressed as a function of the share of women playing the instrument, as follows:

$$\widehat{ATT} = \hat{\gamma}_2 + \hat{\delta}^{female,1}_{genderInstr.} \times genderInstr \tag{3.3}$$

[3] With a binary outcome variable (here: pass or fail), two estimation methods are recommended: probit and logit. Our results are robust regardless of the method used, but we favour logit primarily for its simplicity and the ease of interpreting results, especially in applied predictive contexts.

$\widehat{ATT}$ measures the net effect of candidate visibility (non-blind) on the probability of success for women relative to men, while accounting for instrument-level segregation. If $\widehat{ATT} > 0$, the bias is in favour of women (or against men), and vice versa.

4 Results

The results of the mixed-effects logistic regression are similar to those of the standard logistic regression. This confirms that despite the clustering design of the data, observations i can be considered as independent. In other words, once the number of candidates is taken into account, the probability of passing a round of auditions does not depend on the outcomes of other individual performances, regardless of their position within the clusters.

Table 4.1: Regression results

	Probability to be qualified for next round	
	logistic	*generalized linear mixed-effects*
Number of candidates	-0.07^{***} (0.01)	-0.07^{***} (0.01)
Gender (ref. Men)	0.57 (0.41)	0.57 (0.41)
Treatment (ref. no)	0.61^{**} (0.28)	0.59^{**} (0.28)
Share of Women (AFO)	0.75 (0.53)	0.78 (0.53)
Gender * Treatment	-1.41^{**} (0.61)	-1.42^{**} (0.61)
Gender * Share of Women	−1.53 (0.93)	-1.57^{*} (0.94)
Treatment * Share of Women	−1.08 (0.79)	−1.09 (0.79)
Gender * Treatment * Share of Women	3.55^{***} (1.38)	3.60^{***} (1.38)
Constant	-0.49^{**} (0.22)	-0.48^{**} (0.23)
Observations	1,339	1,339
Log Likelihood	−826.49	−826.10
Akaike Inf. Crit.	1,670.97	1,672.19
Bayesian Inf. Crit.		1,724.19
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

As detailed in Table 4.1, the more candidates there are, the lower the chance that each candidate has passed is. The parameter δ is significantly different from zero, showing that the probability of passing depends jointly on the candidate's gender, the use of a screen and the instrument's degree of feminization. We illustrate the compound effect in a graph showing the predicted probability of passing the round for different degrees of feminization of the instruments. Figure 4.1 shows the predicted probability of passing the audition for women and men, according to the use of a screen, the number of candidates and the degree of feminization of the instrument observed in the orchestras (ranging from 10% of female musicians up to 75%[1]). When the audition is blind, the model predicts that female musicians playing a male instrument have a higher likelihood of being selected. Conversely, male musicians who play a female instrument are more likely to pass the round when the audition is blind. For gender-neutral instruments, the odds of passing the round are similar for men and women. When the audition is not blind, the odds change dramatically: female musicians are more likely to succeed if they play a female-dominated instrument, while male musicians are more likely to succeed if they play a male-dominated instrument. Finally, if the instrument is gender balanced, then the odds are the same for women and men. The Appendix provides details of various tests conducted for verification.

Following Equation 3.3 and the Table 4.1, we represent the ATT expressed as a function of the share of female musicians in an instrument in Figure 4.2 :

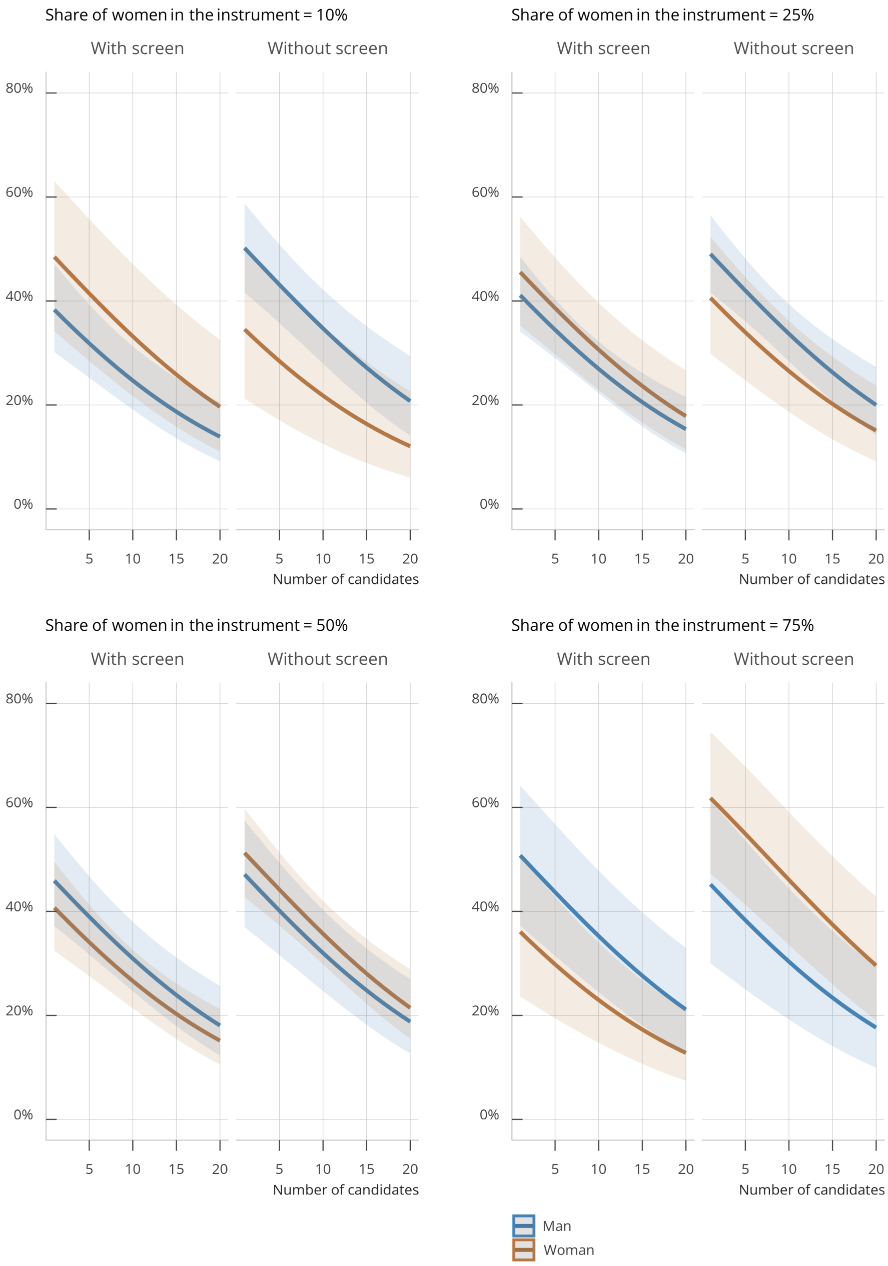
$$\widehat{ATT} = -1.41 + 3.55 \times share\,of\,women\,in\,instrument$$

Female musicians face a disadvantage when playing male-dominated instruments while visible, but this pattern reverses as the instrument becomes more feminized. The turning point is around 40.4% female musicians, not significantly different from parity. Notably, several instruments are highly male-dominated, whereas only the harp is distinctly female-dominated.

Our findings show that jury decisions in the selection process are shaped by gender bias. Blind auditions mitigate this effect by fostering impartiality and limiting the influence of such biases. Similarly, reducing sex segregation weakens the impact of gender stereotypes in the jury's decision-making: we estimate that if the proportion of women within an instrument increases by 10%, the odd ratio for a woman versus a man to pass an unblind round is multiplied by 1.47 (with a 90% confidence interval [1.17;1.84]). In the long run, this approach is expected to reduce not only the degree of sex segregation and the inequality of opportunity, but also, ultimately, the level of discrimination.

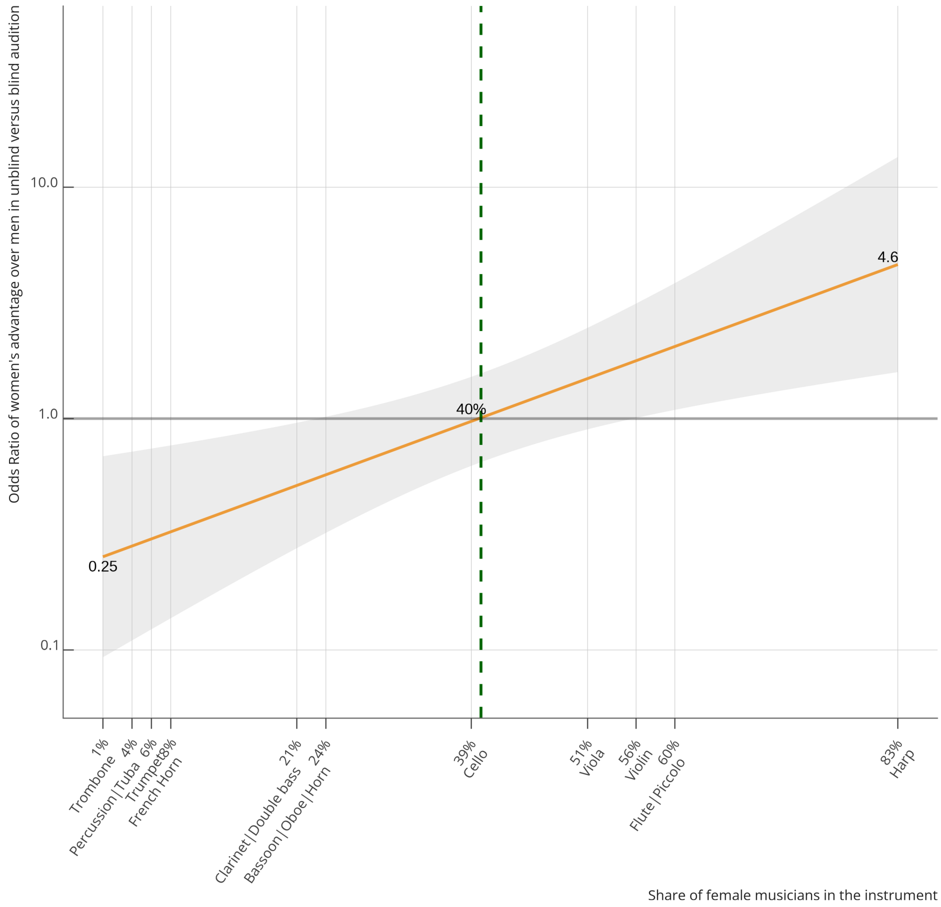
[1] There is no instrument for which women represent more than 90% of players.

Fig. 4.1: Predicted probability to be qualified for the next round (excluding the 1st round)



Source: Ardige and AFO

Fig. 4.2: Average Treatment effect on the Treated (ATT)



Source: Ardige and AFO

5 Discussion

First, we show that the probability of being selected in a blind audition differs for men and women according to type of instrument played. When a screen is used, musicians playing an instrument for which their gender is under-represented have a better chance of passing the audition round than do those whose gender is overrepresented among the players of the instrument. Assuming that blind auditions reveal the quality of play of musicians, this implies that individuals who do not conform to gender norms in their choice of instrument outperform the other gender. This is due to a selection or survivorship bias that occurs prior to the competition. In the case of musical careers, years of training constitute a very long selection process (Sergeant and Himonides, 2019). The choice of an instrument is limited by the association of gender with instrument with « girls' instruments » such as flute or harp versus « boys' instruments » such as drums, trumpets, and trombones (Abeles, 2009 ; Abeles and Porter, 1978). Musicians whose gender is rare among the instrumentalists had to face the difficulties associated with their « atypical » choice of instrument, as it is the case in other fields (Heilman and Wallen, 2010). As a result, those who are not the best may give up more often than those who are not the best but whose gender is highly represented among the students. In addition, among the gender-nonconforming musicians, those who are good enough to persist in their choice may work harder to prove their legitimacy to play a certain instrument despite their gender. This phenomenon can be called the « survivor effect ». It explains why the level of these individuals is higher than that of the other gender. As a result, they have a better chance of passing an audition than the other gender does, all else being equal.

Second, we find that when an instrument is roughly gender-balanced, the probability of being selected is the same for women and men regardless of the procedure (blind or not). When the audition is not blind, the jury is more likely to select a candidate whose gender is in the majority among the instrumentalists. Relative to their decisions in blind auditions, the jury tends to overestimate the quality of gender conformists and underestimate the quality of nonconformists. Interestingly, the greater the degree of sex segregation is, the stronger the jury bias is: at high levels of segregation, the jury bias in favour of the gender majority more than compensates for the *survivor effect*. Not accounting for the *survivor effect* would lead to an underestimation of the extent of the gender bias in the decision-making. The observation of sex segregation reinforces gender stereotypes, which influence the jury's representation of what makes a good musician (Heilman, Caleo and Manzi, 2024).

Our results support assumption $A3$, which states that, on average, the relative performance of female and male musicians is not affected by the presence or absence of a screen. One might argue that female musicians are more affected by stress than their male counterparts when they are visible. However, because gender bias depends on the instrument, this argument is not convincing. It could still be suggested that such stress varies depending on whether a musician belongs to the gender minority within a given instrument. Our findings regarding survivorship bias indicate that these musicians have already been highly selected upstream; for this reason, it is highly unlikely that they are more affected by stress within an instrument, even when they are visible.

6 Conclusion

Our findings are in line with those of Goldin and Rouse (2000), showing that blind auditions increase impartiality in recruitment procedures in orchestras. We show that gender bias is associated with the sex segregation of instruments influence jury's decision. Observing this segregation influences the jury's perception of what constitutes a good musician for a given instrument, a judgment that is deeply rooted in prevailing representations. Notably, there are few feminized instruments, whereas many instruments in orchestral positions are male-dominated.

Our results are consistent with other studies showing that individuals are less likely to be selected in gender congruent fields (Bordalo *et al.*, 2019 ; Clarke, 2020 ; Coffman, Exley and Niederle, 2021). Other studies in the education literature find opposite results. Lavy (2008) compares the grades given by teachers to male and female students throughout the year with their scores at the final anonymous national test and finds that male students are discriminated against in both female and male subjects. Breda and Hillion (2016) and Breda and Ly (2015) use competitions in different disciplines in higher education. The quasi-experiment is based on comparing the ranking of female and male candidates in the first stage of the competition, an anonymous written test (used as a blind test), and in the final stage, an oral test (used as a non-blind test). They find that in male-dominated fields (such as mathematics or physics), the ranking of female candidates increases between the two stages. In both cases, the two parts of the quasi-experiment are not perfectly comparable. Unlike the final exam, the grades given by the teacher may reflect not only the students' skills and knowledge in different subjects but also aspects such as classroom behaviour or attendance. Similarly, a written test may not capture the same skills as an oral exam. Moreover, the stakes in blind and non-blind tests are different, which may affect the performance of women and men differently (Azmat, Calsamiglia and Iriberri, 2016). Additionally, in a pool of highly selected candidates, the jury may wish to favour the under-represented gender when the selection is not blind at the final oral stage. Lastly, these opposite results could be due to the fact that the stereotypes take different forms in the educational and in the classical performance sectors.

In line with the literature on stereotypes and decision bias, we identify of the role of gender bias in the labour market. Gender stereotypes operate at two different levels. First, they help shape the career choices of young people. Second, sex segregation in jobs affects hiring decisions and prevents recruiters from selecting the best workers. In highly demanding jobs, such as professional musicians, selecting the best performers is a complex task, especially when such selection is made from a pool of candidates with high and narrow ability levels. This explains why gender stereotypes are used as shortcuts for complex decisions. Blind auditions promote impartiality by neutralizing these biases. Reducing the sex segregation mitigates the influence of gender stereotypes in decision-making In the long run, this approach is expected to reduce not only the degree of sex segregation and the inequality of opportunity, but also, ultimately, the level of discrimination.

7 Appendix

7.1 Difference in differences on aggregated data

A triple-difference approach can always be reduced to a standard DiD by appropriately transforming the dependent variable (Roth *et al.*, 2023). This involves capturing the first difference in a new outcome variable. For this purpose, we have aggregated the Ardige data by sets of competitors within each round and contest S. As for the DDD, we have only kept the sets whose musicians have not be seen in a previous round and dismissed the preliminary round. Unlike DDD, the DiD framework requires that the same individual be considered multiple times.

We note Z_s^t the relative advantage of female candidates versus male candidates in the set of s when they play in condition t, with $t = 0$ when the performance i is behind a screen (no treatment) and $t = 1$ t when the audition is not blind (treatment).

$$\begin{aligned} Z_s^{instr,t} &= \mathbb{E}_i(Y_s^{female,instr,t}) - \mathbb{E}_i(Y_s^{male,instr,t}) \\ &= \text{log}\left(\text{OddsRatio}_s^{female|male,instr,t}\right) \end{aligned}$$

The *Average Treatment effect on the Treated* (ATT) in this DiD framework is :

$$ATT_{DiD}^{instr} = Z_s^{instr,1} - Z_s^{instr,0}$$

The DiD method consists in estimating the aggregated outcome Z_s^t on a dataset where each observation corresponds to a set s of competitors. However, the method faces one main challenge, that is the presence of sampling zeros. Not all sets contain a sufficient number of individuals to fill the four cells (male or female crossed with fail or pass) required to calculate the log odds ratios Z_s^0 or Z_s^1. A common practice (Agresti (2002), p.392) is to fill these cells with a small value (0.1 or less). To minimize the impact of this choice, which could produce arbitrary extreme values, we then perform a robust regression.

The model is specified as follows:

$$Z_s^t = \alpha_0 + \beta_1\, genderInstr + \beta_2\, T + \gamma\,(genderInstr \times T)$$

We estimate an MM-type regression (Yohai, 1987). As shown in Table 7.1, the results are very similar to those obtained from the DDD approach. There are no significant differences between the two lines representing the estimated ATT by the share of women in the instrument.

Table 7.1: Robust Regression result

	Robust regression M-M
Constant	0.72* (0.31)
Share of Women (AFO)	-2.15** (0.74)
Treatment (ref. no)	-1.30** (0.50)
Share of Women (AFO):Treatment (ref. no)	3.71** (1.19)
N weighted	2389
R^2	0.15

$$\widehat{ATT}_{DiD} = -1.30 + 3.71 \times share\,of\,women\,in\,instrument$$

Female musicians face a disadvantage when playing male-dominated instruments while visible, but this pattern reverses as the instrument becomes more feminized. The turning point with the DiD is around 35% female musicians, and as for the DDD, not significantly different from parity.

7.2 Comparison of the models

We check that there is no significant difference between the standard logistic model and the multilevel model. Their *Akaike Information Criterion* (AIC) are very close : 1671 for the standard model and 1672.2 for the multilevel model. The difference is too small to discriminate between the two models.

If we apply a likelihood ratio test between the two models (considering that the standard model is only a special case of the multilevel model), we get : 0.78 for 1 degree of freedom, which gives a probability of 0.38. This means that the inclusion of the random effect was not necessary (but it was mandatory to check this point).

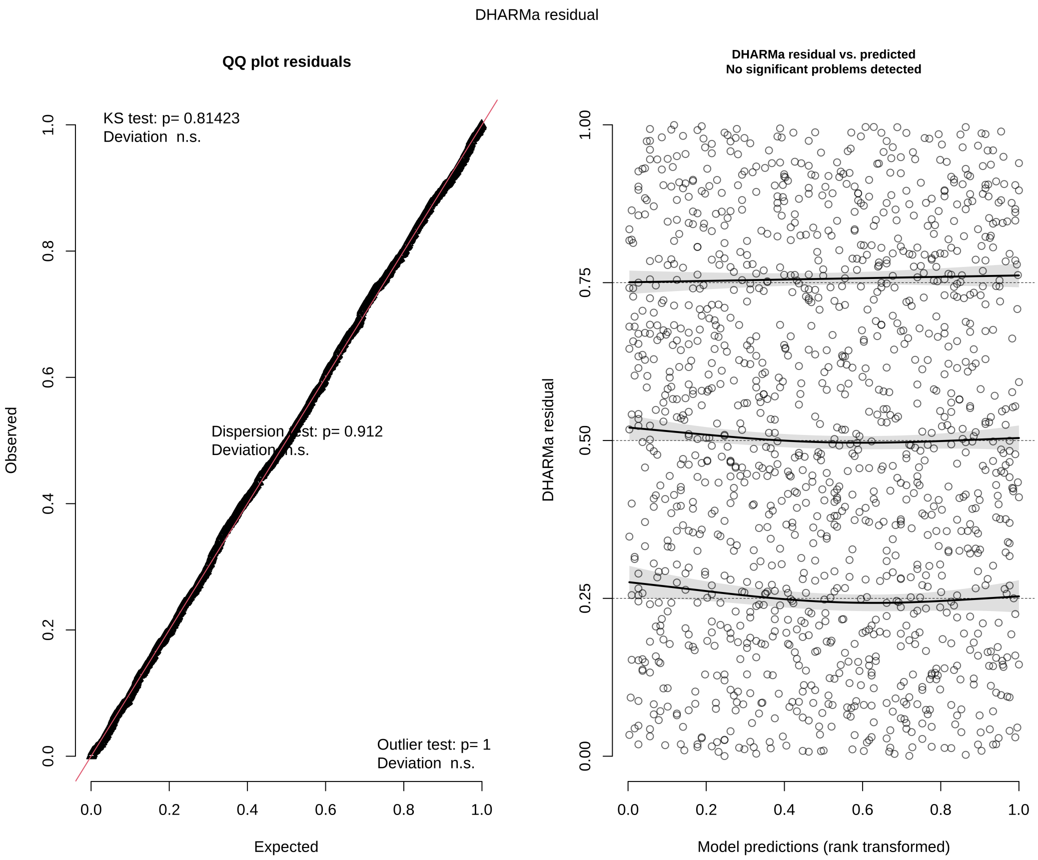
7.3 Goodness-of-fit

The McFadden's pseudo-R2 for the logistic model is 0.03. It is the same as the one for the multilevel model. Since this value is not self-explanatory, we prefer the Tjur's pseudo-R2. For the logistic model, it is equal to 0.03, which is the difference between the predicted probabilities for those who passed and those who failed. This difference is not large, probably because the model only takes into account gender bias and not the main criterion, which is the objective quality of candidate's play. The area under the curve (ROC) is another criterion of goodness-of-fit. For the standard logistic model, it is 0.6 and for the multilevel model, 0.61. These models have some predictive power, better than random, even if they are far from perfect predictions. The Hosmer-lemeshow goodness-of-fit is another criterion. For the logistic model it is : 2.94 with 8 degrees of freedom and a p-value of 0.94. For the multilevel model : 5.19 with 8 degree of freedom and a p.value of 0.74. As we can infer from these values, the two models fit well.

7.4 Diagnostics of the fit

We use a simulation based-approach to make a diagnostics on the residuals of the models (see Hartig (2022)).

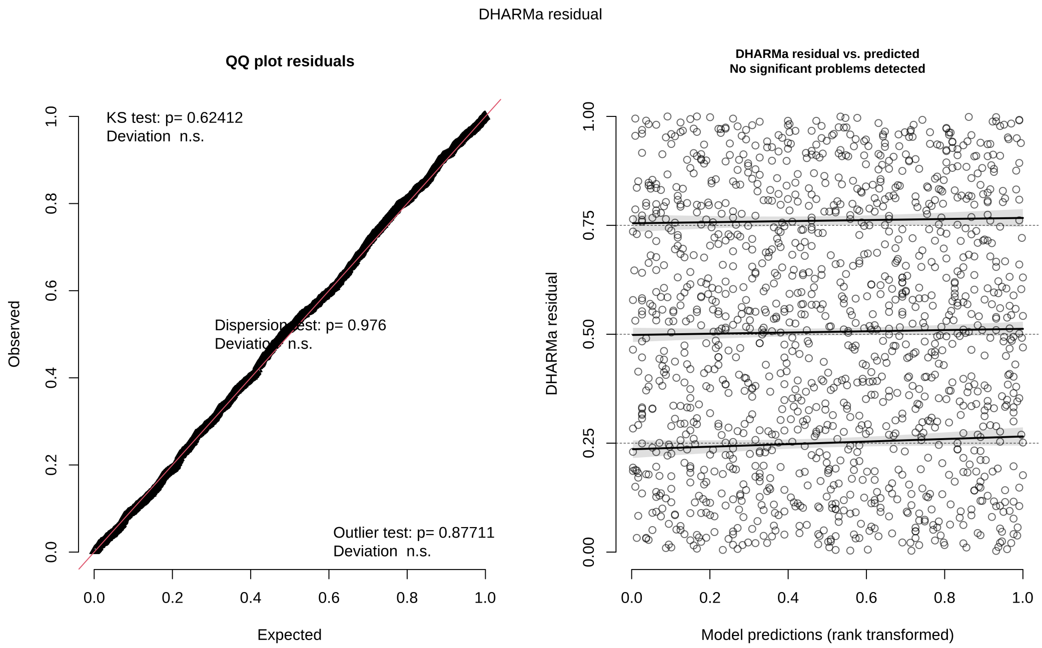
Fig. 7.1: Residuals for the standard model




```
    DHARMa outlier test based on exact binomial test with approximate
    expectations

data:  Mod_base_clean
outliers at both margin(s) = 10, observations = 1339, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.007968127
95 percent confidence interval:
 0.003586961 0.013691381
```

Fig. 7.2: Outlier test for the multilevel model



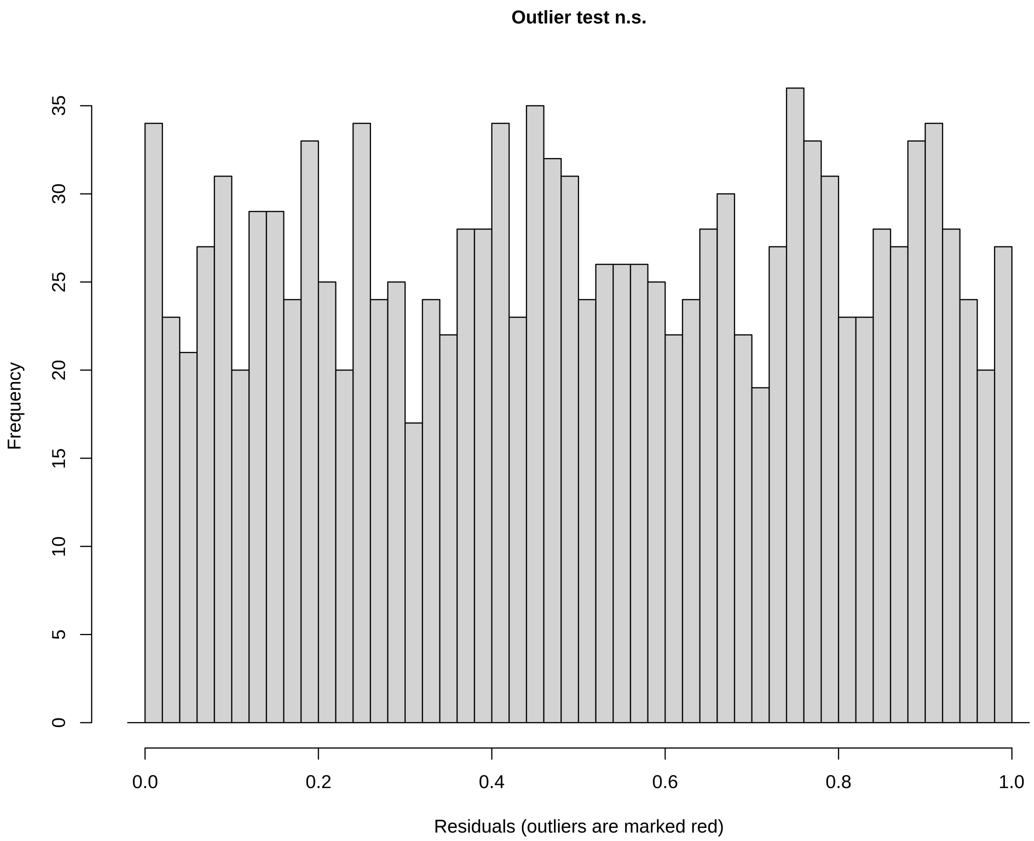

```
sample estimates:
frequency of outliers (expected: 0.00796812749003984 )
                                           0.00746826


    DHARMa outlier test based on exact binomial test with approximate
    expectations

data:  Mod_rand_clean
outliers at both margin(s) = 11, observations = 1339, p-value = 0.8771
alternative hypothesis: true probability of success is not equal to 0.007968127
95 percent confidence interval:
 0.004107869 0.014651416
sample estimates:
frequency of outliers (expected: 0.00796812749003984 )
                                          0.008215086


    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated
```

Fig. 7.3: Outlier test for the standard model




```
data:  simulationOutput
dispersion = 1.0007, p-value = 0.912
alternative hypothesis: two.sided



    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 0.99879, p-value = 0.976
alternative hypothesis: two.sided
```

Fig. 7.4: Outlier test for the multilevels model

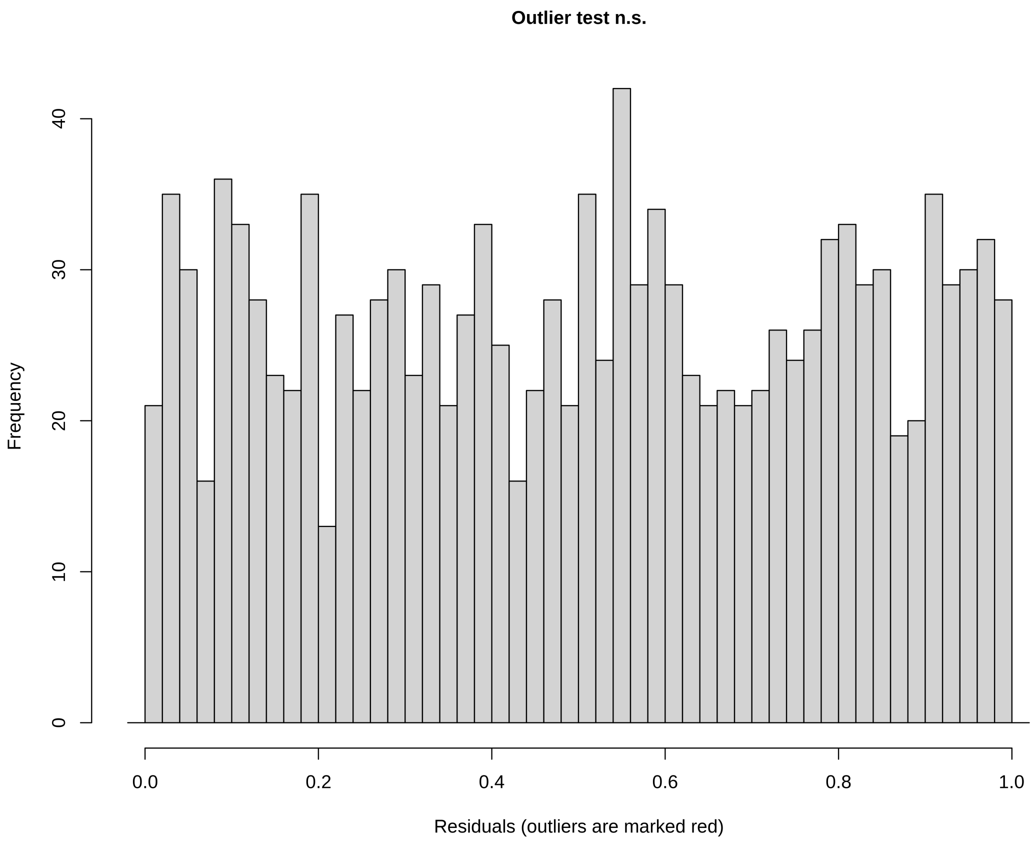


Fig. 7.5: The standard model

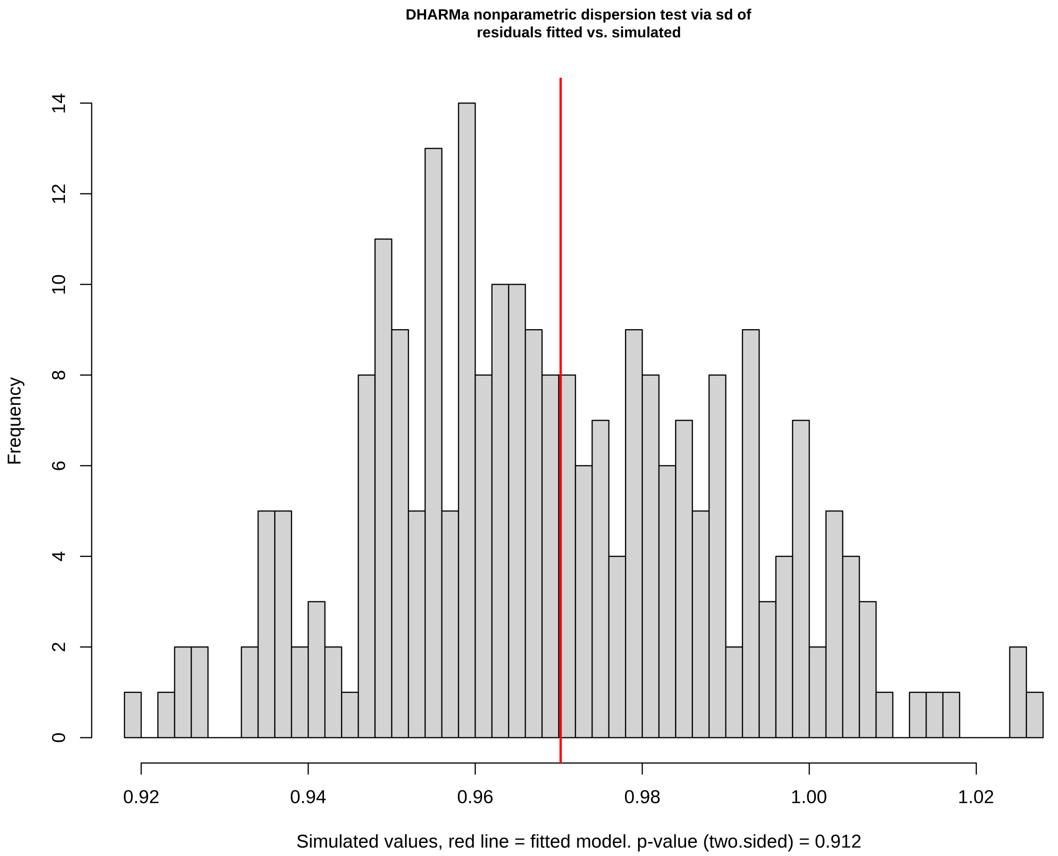
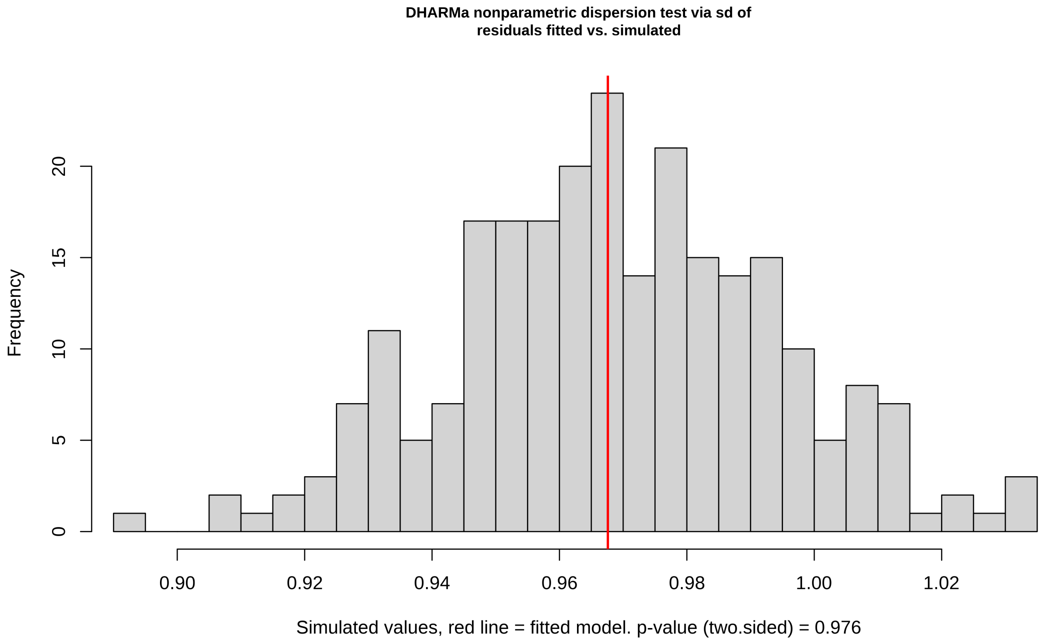


Fig. 7.6: The multilevels model



References

Abeles H. (2009). "Are Musical Instrument Gender Associations Changing?" *Journal of Research in Music Education*, *57*, n° 2, pp. 127–139.

Abeles H.F., Porter S.Y. (1978). "The Sex-Stereotyping of Musical Instruments," *Journal of Research in Music Education*, *26*, n° 2, pp. 65–75.

Agresti A. (2002). *Categorical data analysis*, John Wiley & Sons (Wiley series in probability and statistics.).

Association Française des Orchestres (2018). "L'égalité femmes/hommes dans les orchestres membres de l'AFO."

Azmat G., Calsamiglia C., Iriberri N. (2016). "Gender Differences in Response to Big Stakes," *Journal of the European Economic Association*, *14*, n° 6, pp. 1372–1400.

Bertrand M., Mullainathan S. (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, *94*, n° 4, pp. 991–1013.

Bertrand M., Duflo E. (2017). "Field Experiments on Discrimination," *Handbook of Economic Field Experiments*, pp. 309–393.

Booth A., Leigh A. (2010). "Do employers discriminate by gender? A field experiment in female-dominated occupations," *Economics Letters*, *107*, n° 2, pp. 236–238.

Bordalo P., Coffman K., Gennaioli N., Shleifer A. (2016). "Stereotypes," *The Quarterly Journal of Economics*, *131*, n° 4, pp. 1753–1794.

Bordalo P., Coffman K., Gennaioli N., Shleifer A. (2019). "Beliefs about Gender," *American Economic Review*, *109*, n° 3, pp. 739–773.

Breda T., Hillion M. (2016). "Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France," *Science*, *353*, n° 6298, pp. 474–478.

Breda T., Ly S.T. (2015). "Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France," *American Economic Journal: Applied Economics*, *7*, n° 4, pp. 53–75.

Carlsson M. (2011). "Does Hiring Discrimination Cause Gender Segregation in the Swedish Labor

Market?" *Feminist Economics*, *17*, n° 3, pp. 71–102.

Clarke H.M. (2020). "Gender Stereotypes and Gender-Typed Work," dans Zimmermann K.F. (ed.), Springer International Publishing, Cham, pp. 1–23.

Coffman K.B., Exley C.L., Niederle M. (2021). "The Role of Beliefs in Driving Gender Discrimination," *Management Science*, *67*, n° 6, pp. 3551–3569.

Eldridge S. (2024). "Survivorship bias | Definition, Meaning, Examples | Britannica,"

Elton E.J., Gruber M.J., Blake C.R. (1996). "Survivor bias and mutual fund performance," *The Review of Financial Studies*, *9*, n° 4, pp. 1097–1120.

Goldin C., Rouse C. (2000). "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, *90*, n° 4, pp. 715–741.

Gruber J. (1994). "The incidence of mandated maternity benefits," *The American Economic Review*, *84*, n° 3, pp. 622–641.

Hartig F. (2022). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models*.

Hatzipetrou-Andronikou R., Ravet H., Etienne A., Parodi M., Périvier H. (2025). "À l'aveugle ?" *Travail, genre et sociétés*, *53*, n° 1, pp. 97–115.

Heilman M.E., Caleo S., Manzi F. (2024). "Women at Work: Pathways from Gender Stereotypes to Gender Bias and Discrimination," *Annual Review of Organizational Psychology and Organizational Behavior*, *11*, n° Volume 11, 2024, pp. 165–192.

Heilman M.E., Wallen A.S. (2010). "Wimpy and undeserving of respect: Penalties for men's gender-inconsistent success," *Journal of Experimental Social Psychology*, *46*, n° 4, pp. 664–667.

Hilton J.L., Von Hippel W. (1996). "Stereotypes," *Annual Review of Psychology*, *47*, n° 1, pp. 237–271.

Ioannidis J.P.A. (2005). "Why most published research findings are false," *PLoS Medicine*, *2*, n° 8, p. e124.

Lavy V. (2008). "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment," *Journal of Public Economics*, *92*, n° 10-11, pp. 2083–2105.

List J.A. (2004). "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *The Quarterly Journal of Economics*, *119*, n° 1, pp. 49–89.

Neumark D., Bank R.J., Van Nort K.D. (1996). "Sex Discrimination in Restaurant Hiring: An Audit Study," *The Quarterly Journal of Economics*, *111*, n° 3, pp. 915–941.

Olden A., Møen J. (2022). "The triple difference estimator," *The Econometrics Journal*, *25*, n° 3, pp. 531–553.

Ravet H. (2011). *Musiciennes: enquête sur les femmes et la musique*, Éd. Autrement, Paris (Collection Mutations).

Roth J., Sant'Anna P.H.C., Bilinski A., Poe J. (2023). "What's trending in difference-in-differences? A synthesis of the recent econometrics literature," *Journal of Econometrics*, *235*, n° 2, pp. 2218–2244.

Rubin D.B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, *66*, n° 5, pp. 688–701.

Sergeant D.C., Himonides E. (2019). "Orchestrated sex: The representation of male and female musicians in world-class symphony orchestras," *Frontiers in Psychology*, *10*, p. 1760.

Wald A. (1943). "A method of estimating plane vulnerability based on damage of survivors." *Statistical Research Group, Columbia university, CRC}*, *432*.

Yohai V.J. (1987). "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, *15*, n° 2.