ofce

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 : 2024-11-21

SciencesPo

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 also use data from actual auditions for French orchestras to identify gender bias in jury decisions. The high degree of sex segregation among instruments and the heterogeneity of orchestras auditions provide a context for assessing the role of gender stereotypes in 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 labor 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 educational literature, quasi-experiments based on administrative data from both blind and nonblind tests and exams are also a way to shed light on how gender biases affect decision-making (Breda and Hillion, 2016 ; Breda and Ly, 2015 ; Lavy, 2008).

Inspired by the seminal paper of Goldin and Rouse (2000), we use the hiring of musicians in permanent orchestras to explore this question. 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 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 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. 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 unprecedented individual dataset from actual recruitment competitions organized by 13 French orchestras between 2001 and 2021. We use blind and nonblind 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 nonblind 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.

Our paper contributes to the literature by precisely measuring the role of gender bias fed by sex segregation in the workplace, in 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 counterfactual is not a similar candidate of the opposite sex 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 underrepresented 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 sex segregation of instruments impairs the impartiality of judges and prevents them from selecting the best candidate. Blind auditions promote impartiality by neutralizing these biases. We show that reducing the sex segregation mitigates the influence of gender stereotypes in 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.

2 Descriptive statistics

2.1 Data

In France, there are approximately 30 permanent orchestras that have a total of 2,500 permanent musician positions. 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, which increased to 33% in 2010 and 36% in 2016. However, there are still differences from one instrument to another as the percentage of women ranges from 1% to 85% depending on the instrument. To select the best musicians, orchestras organize recruitment competitions for each instrument[1] and position. Each competition is organized in several rounds, most often three rounds, but a competition can last up to 5 rounds.

The Ardige database is a the collection of information on competition organized by permanent French orchestras between 2001 and 2021, mainly during the 2010s. Of the 30 AFO orchestras, only 13 of them have kept usable archives of the competitions they organised to recruit their musicians. The database contains information related to 322 competitions. We have collected detailed information on all rounds of each competition for each 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 first 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 second rounds, 884 candidates in the third rounds and 283 candidates who are ultimately hired in total. Some competitions do not end with a hire, if the jury is ultimately unconvinced by the candidates' performances.

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.

[1] In some cases, the same competition is used to recruit several musicians who play either the same instruments, or similar instruments (such as flute and flute piccolo).

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.

2.2 Sex segregation between instruments

The current representation of male and female musicians holding tenured positions in orchestras is characterized by a high degree of sex segregation. This sex 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 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 first 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).

2.3 Blind or not blind

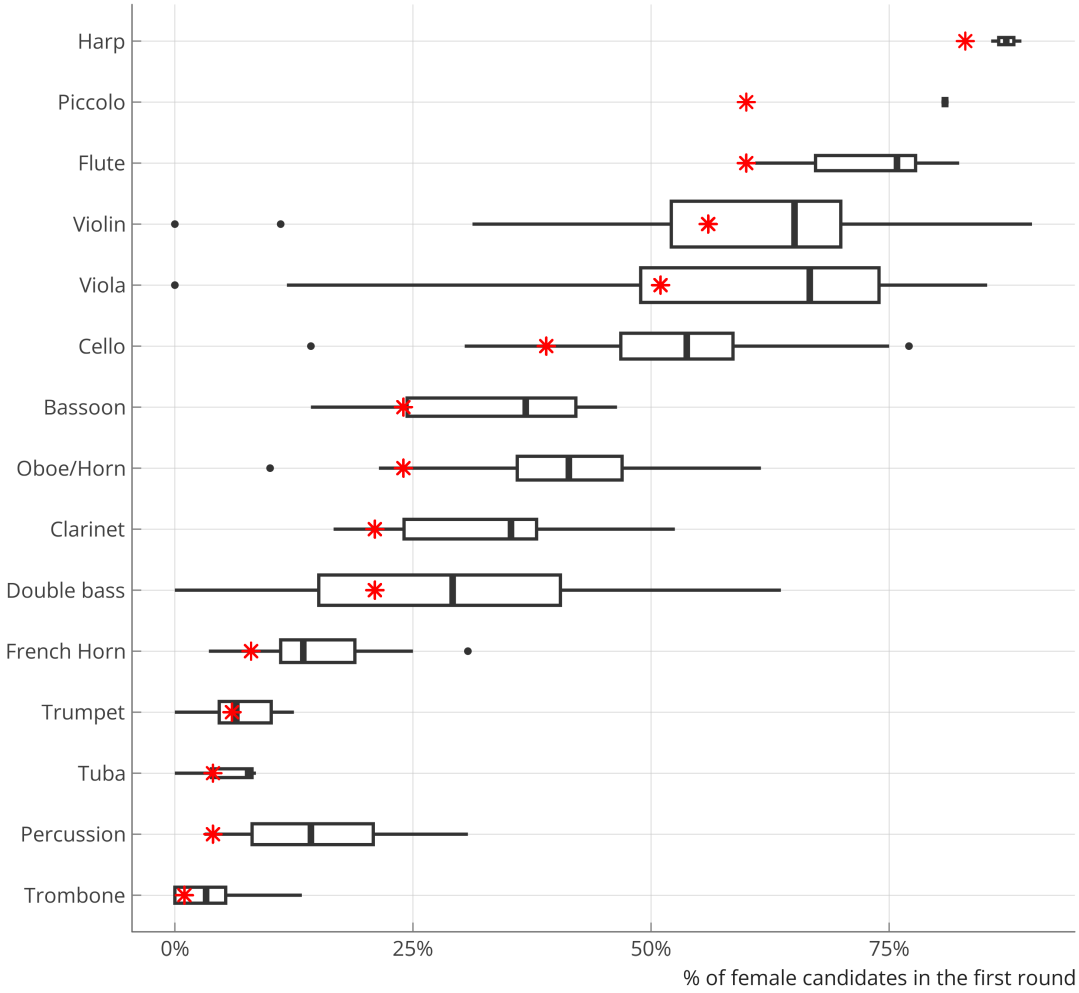
The use of a « screen » in the recruitment of musicians has gradually been introduced in French orchestras to limit the effects of co-optation within the circle of former students and acquaintances of the conductor and/or assistant conductor (Ravet, 2011). It has become a tool to guarantee the impartiality of the judges and to reduce the influence of gender stereotypes. During individual performances, the candidates are hidden from the jury, typically by a screen or curtain placed in front of the musician. Under these conditions, the judges can assess the quality of the 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.

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 onl for the first round. From the second round on, only 40% of the rounds examined are blind auditions. Some orchestras have changed their prac-

tices[2]. In the Ardige database, two orchestras 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. 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. We then assume that applicants do not self-select into competitions according to whether or not they use the screen. In addition the use of blind auditions is determined at the orchestra level for all type of instruments.

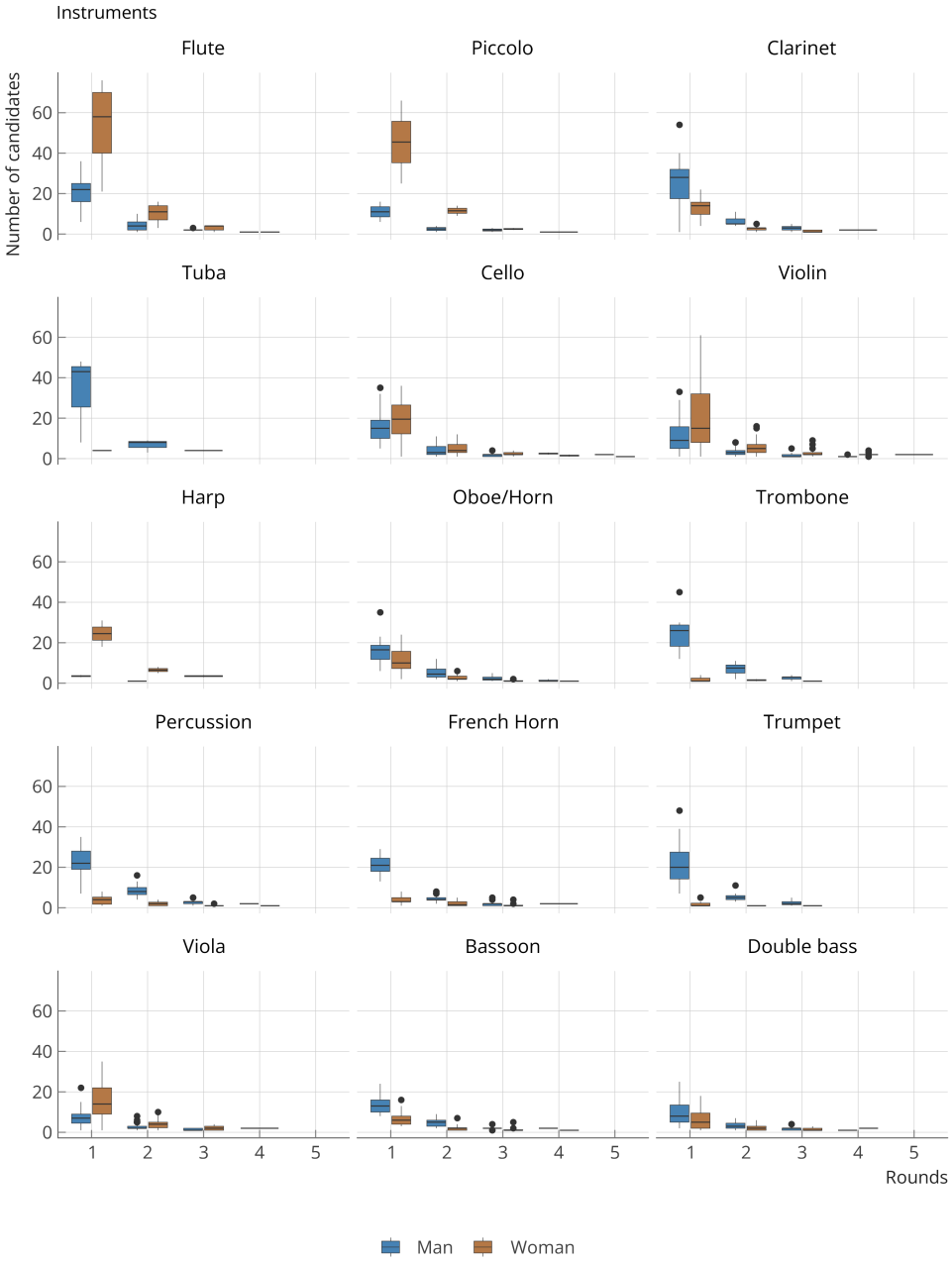
[2]While two orchestras have recently added blind auditions to their process, a few have dropped a blind auditions from their competitions.

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 orchestras 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 find out whether the sex segregation of instruments observed in French orchestras influences the jury's decision beyond the quality of the musical performance during an audition. We use blind and nonblind auditions as a quasi-experiment to test for the role of gender stereotypes on the perception of the jury regarding the quality of individual performances. In a blind audition, we assume that the jury is not able to identify the gender of the musicians. The judge's decisions in a blind audition are then unbiased and reflect the objective quality of the performance. With a blind audition, the musicians selected can be considered objectively better than those who are rejected. Under these conditions, if we observe a gender gap in the probability of being selected in blind auditions, one gender outperforms the other. This can be explained by selections made throughout the training of musicians. Blind auditions allow us to identify this upstream selection and prevent us from the survivorship bias, which consists of focusing on only individuals who have already passed a selection process, while overlooking those who have 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).

To capture the role of gender stereotypes, we use the data published by the AFO on the percentage of female musicians per instrument in 2016-2017 (Association Française des Orchestres, 2018). 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. The heterogeneity of instruments in terms of the proportion of female musicians in orchestras can then be used to determine how the sex segregation affects the representation of the jury decisions.

Following Gruber (1994), we use a triple difference design (DDD). This method is widely used to identify the causal effect of a treatment (Olden and Møen, 2022). We compare the relative outcome (i.e., the probability of passing the round) between women and men within the treatment group (i.e., musicians performing behind a screen) to the relative outcome of a set of control performances (i.e., those who performed without a screen) and measure the change in the treatments' relative outcomes for different types of instruments (i.e., with respect to the proportion of female musicians in AFO orchestras). We control for the number of candidates in the round of a competition.

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) becomes the basic statistical unit. The structure

of the dataset is characterized by 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 several times during the same competition due to he or she passing certain rounds. The same musician can also apply for different competitions[1]. Moreover, two performances by different musicians are not statistically independent if they belong to the same round of the same competition, since the jury makes its decision by comparing these two performances; thus, an excellent performance can lead to the elimination of the other candidate. Finally, the orchestras have their own requirements, practices and reputations: each of the 13 orchestras is indexed as m. Each competition (indexed as l) is structured with different rounds (indexed as k), during which individuals (indexed as j) are auditioned. In each round, the jury eliminates the candidates who are judged not to be at the required level -relatively or/and absolutely-, and keeps those who are qualified to be auditioned in the following round.

The first round of the competition differs from the subsequent rounds for several reasons. First, some musicians who already work with the orchestra but are not permanent members may be allowed to skip this first round. Second, some candidates do not meet the minimum required standard at this stage and the jury quickly and easily eliminates them. The pool of candidates in this first round is then more heterogeneous than that in the subsequent rounds. Third, the *intraclass correlation coefficient* (ICC) of individual performances i at the first round is high (33%). Musicians who fail in the first round of a given competition (which happens frequently due to the high level of selection at this stage) are likely to fail in the first round of other competitions. As a result, the individual performances i cannot be considered as independent. For these reasons, we focus the statistical analysis on individual performances observed after the first round. In these subsequent rounds, the jury selects between candidates whose levels are close together. Their *relative* performance becomes more important[2] and judgment bias could thus be more decisive. At such times, individual performances can be considered as quasi-independent. The ICC is low (4.7%). Indeed, the same musician performs differently from one audition to another owing to his/her level of rehearsal; the stakes, the specifics and demands of the audition; his/her level of stress, and so on. In addition, some orchestras have a set repertoire for some of the rounds, and a musician may be more or less familiar with that repertoire. Each round is then a specific 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 without a screen. We then compare rounds where the pool of candidates was selected through blind auditions in the previous rounds. For this selected sample (excluding the first rounds and including only rounds for which the previous round was blind), we check the effect of the nesting of the data. 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 com-

[1] In the Ardige database, 2/3 of the individuals apply for only one contest.
[2] However, the absolute performance is still important, as orchestras may not recruit if the jury decides that none of the candidates are at the required level.

petition, 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. In the end, the number of observations in the relevant sample for the analysis is 1,339 individual performances, of which 44.7% are women.

3.3 The model

We estimate a logistic model and a mixed-effects logistic model, which also known as a multilevel logistic model, to consider the nesting due to the orchestra. We note that Y_i is the binary variable which worth 1 when the performance i is good enough for the individual j to pass the following round and 0 if he or she fails. We consider the following independent variables:

- gender of the candidate j: $gender_j$;
- number of candidates competing during the round k: $Ncand_k$;
- presence or absence of a screen during the round k: $blind_k$;
- share of female musicians playing instrument l within French orchestras (SW_l);
- orchestra m carrying out the selection and recruitment: $Orch_m$

For each individual performance i, the probability for individual j to pass round k playing instrument l in orchestra m, is estimated as follows:

$$\begin{aligned} logit\Big(Prob(Y_i = 1)\Big) = \alpha_0 + \alpha_{0m} \\ + \alpha_1 \times Ncand_k \\ + \beta_1 \times gender_j + \beta_2 \times SW_l + \beta_3 \times blind_k \\ + \gamma_1 \times gender_j \times SW_l \\ + \gamma_2 \times gender_j \times blind_k \\ + \gamma_3 \times SW_l \times blind_k \\ + \delta \times gender_j \times SW_l \times blind_k \end{aligned} \tag{3.1}$$

$$with\ \alpha_{0m} \sim \mathcal{N}(O, \sigma^2)\ random\ effect\ for\ an\ orchestra$$

Equation 3.1 differs from a simple logistic model in that the error term on the orchestras is used. The equation for the simple logistic model is the same without the random effect α_{0m}. The DDD estimator is given by the estimation of the parameter δ.

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. Once the number of candidates is considered, the probability of passing a round of auditions does not depend on what happens to the other individual performances, wherever they are in the clusters.

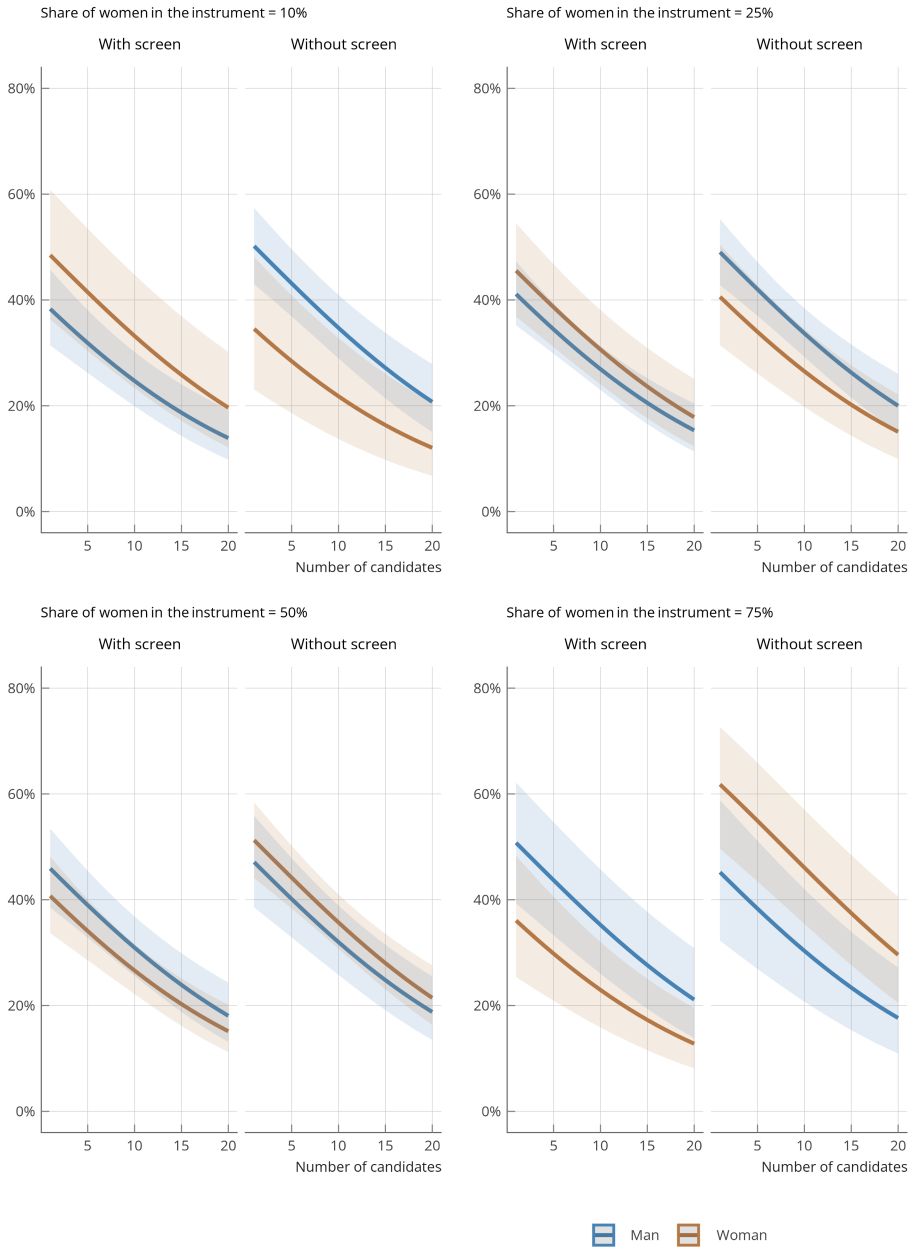
As detailed in Table 4.1, the more candidates there are, the lower the chance that each candidate has passed is. The DDD parameter (δ) is significantly different from zero, meaning that the probability of passing depends on a combination of the candidate's gender, the use of a screen, and the degree of feminization of the instrument. Rather than summing the parameters for each combination, we illustrate the compound 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 greater chance of passing. 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 there is no screen, the odds change dramatically : female musicians are more likely to pass if they play a female instrument ; and male musicians are more likely to pass if they play a male instrument. Finally, if the instrument is gender balanced, then the odds are the same for women and men. The appendix details different tests to check the

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

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)
Blind audition (ref. yes)	0.61** (0.28)	0.59** (0.28)
Share of Women (AFO)	0.75 (0.53)	0.78 (0.53)
Gender * blind audition	−1.41** (0.61)	−1.42** (0.61)
Gender * Share of Women	−1.53 (0.93)	−1.57* (0.94)
Blind audition * Share of Women	−1.08 (0.79)	−1.09 (0.79)
Gender * blind audition * 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.	

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



Source: Ardige and AFO

5 Discussion and conclusion

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 underrepresented 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 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 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 favor 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 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 nonblind 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 behavior or attendance. Similarly, a written test may not capture the same skills as an oral exam. Moreover, the stakes in blind and nonblind 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 labor 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 : 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.

6 Appendix

6.1 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).

6.2 Goodness-of-fit

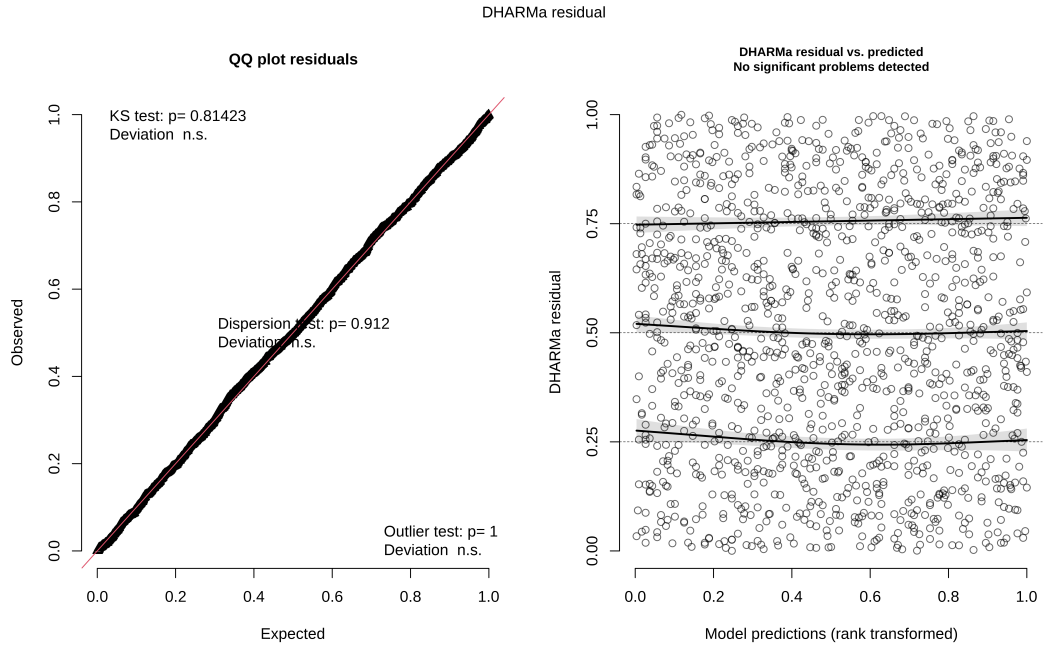
The McFadden's pseudo-R2 for the logistic model is 0.05. It is the same as the one for the multilevel model. Since this value is not self-explanatory, we prefer the Tjur's pseudo-R2. For the logistic model, it is equal to 0.03, which is the difference between the predicted probabilities for those who 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.

6.3 Diagnostics of the fit

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

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

Fig. 6.1: Residuals for the standard model

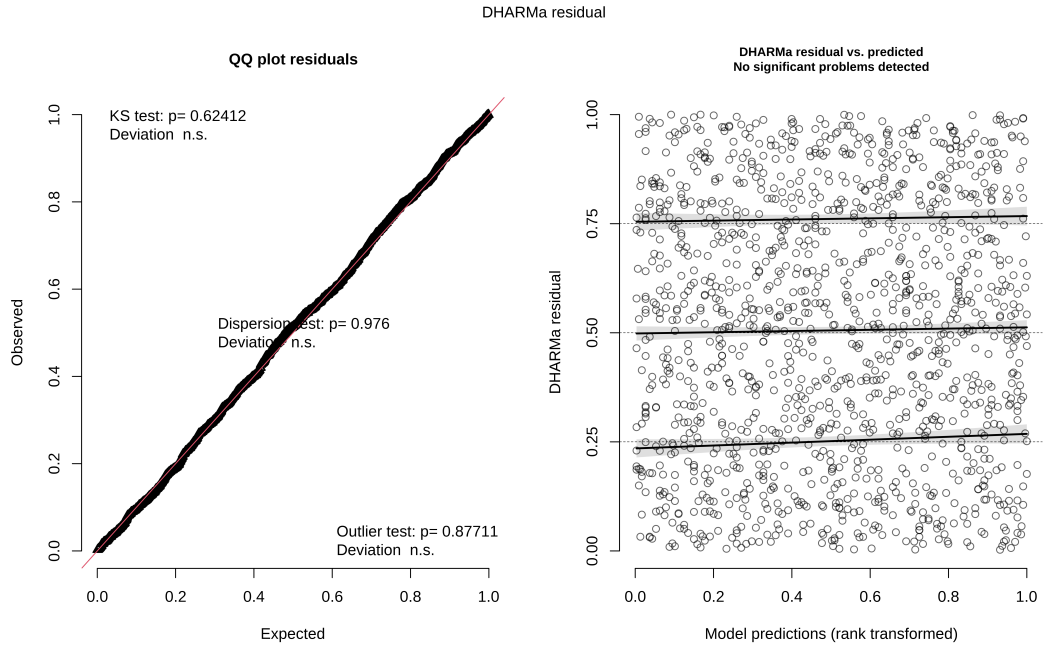



```
data:  Mod_base_clean
outliers at both margin(s) = 10, observations = 1339, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.007968127
95 percent confidence interval:
 0.003586961 0.013691381
sample estimates:
frequency of outliers (expected: 0.00796812749003984 )
                                            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
```

Fig. 6.2: Outlier test for the multilevel model




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

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. 6.3: Outlier test for the standard model

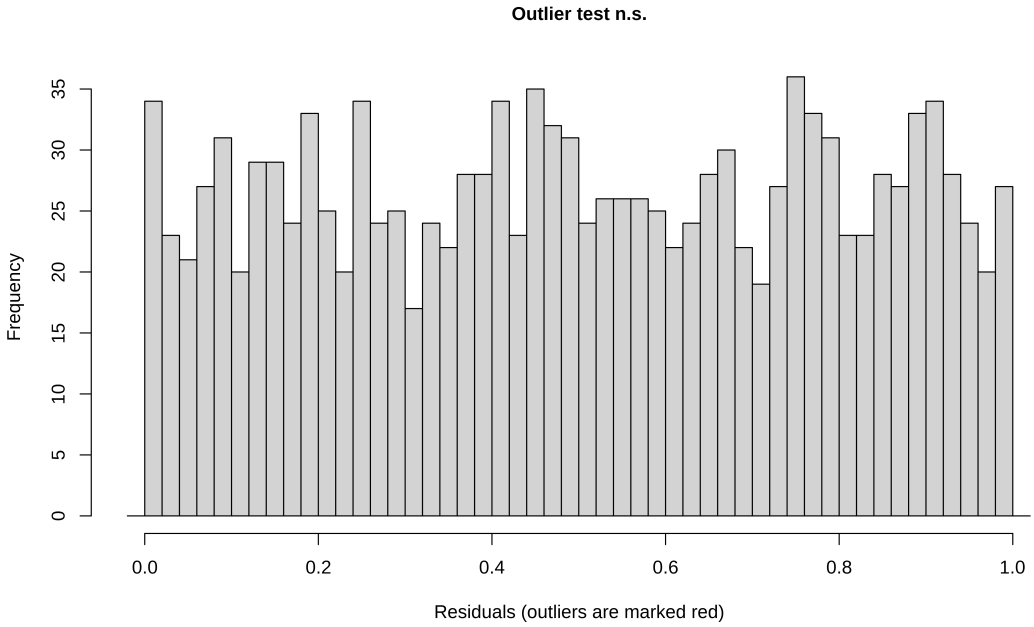


Fig. 6.4: Outlier test for the multilevels model

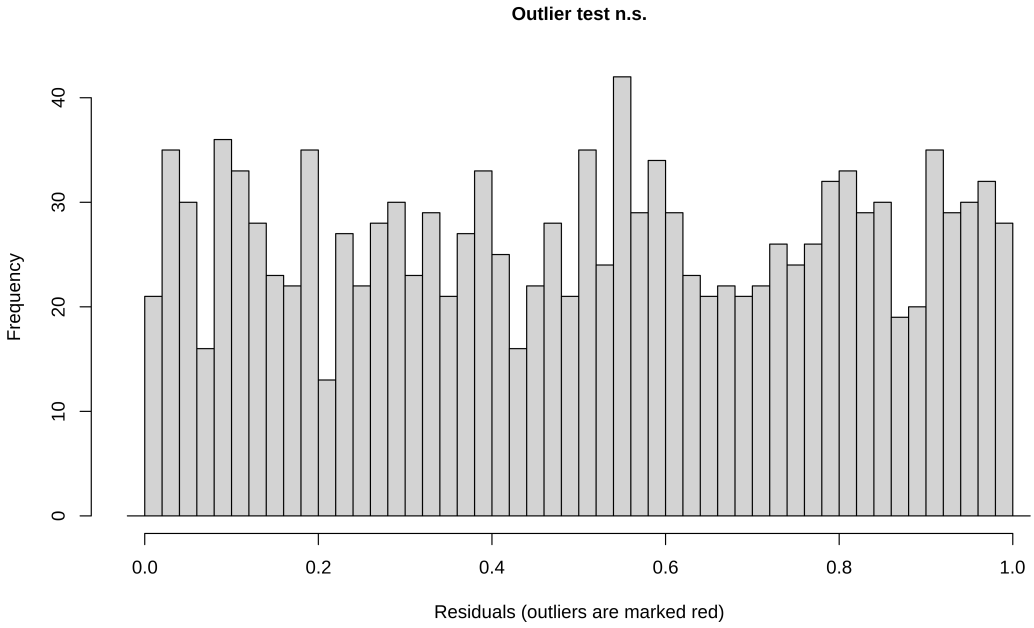


Fig. 6.5: The standard model

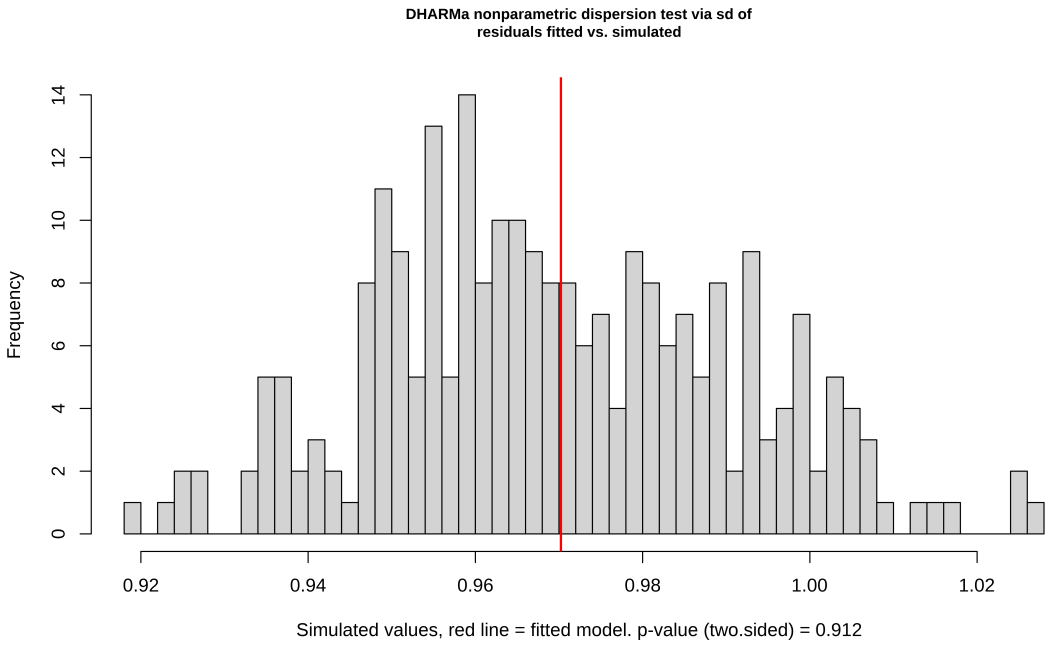
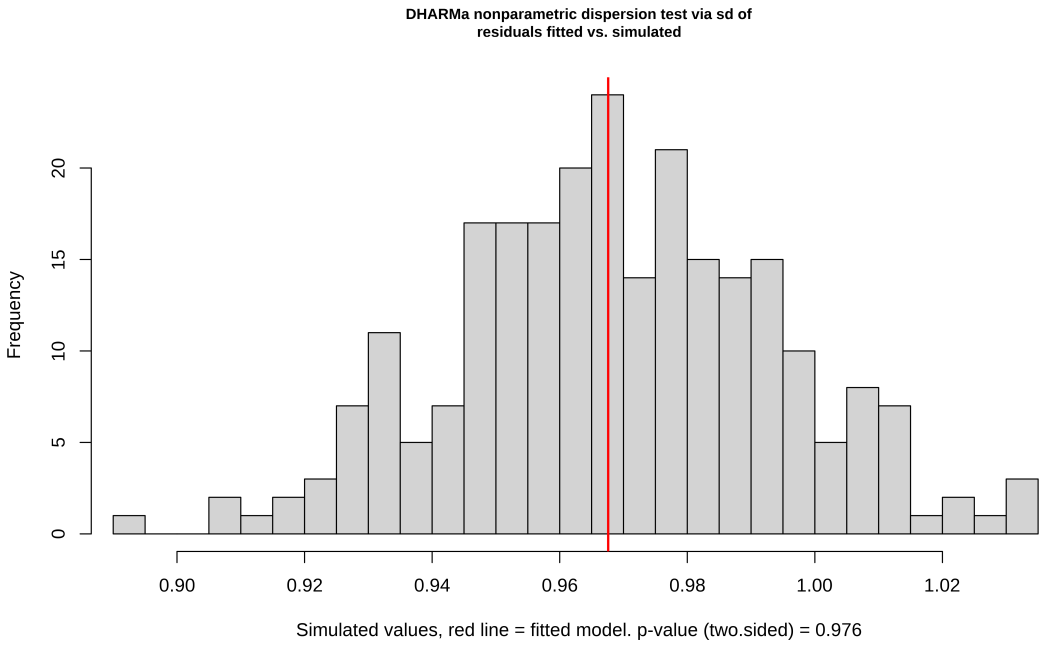


Fig. 6.6: The multilevels model



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.

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*.

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).

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*.