Do Grades Have Absolute Meaning? An Experiment on Majority Judgment

Théo Delemazure, Roberto Brunetti, Antoinette Baujard, Sylvain Bouveret

Do Grades Have Absolute Meaning? An Experiment on Majority Judgment

Théo Delemazure[*], Roberto Brunetti[†], Antoinette Baujard[‡], Sylvain Bouveret[§]

November 7, 2024[¶]

Abstract

Whether in education, performance reviews, or elections, grades serve as a tool for assessment, yet the universality of their meanings remain an open question. When voting under majority judgment, voters assign verbal grades such as "excellent, very good, good, fairly good, acceptable, insufficient, to reject" to each candidate. The meaning of these grades should be clear and consistent to every voter. Balinski and Laraki (2011) call it "universal language" and claim that the grade labels convey absolute meaning. This paper explores the concept of "absolute meaning". We conducted an online experiment (N=1955) where participants voted for French presidential candidates under majority judgment with different grade scales. We find that the grade distributions obtained by candidates are strongly impacted by the grade scales used by voters. Therefore, the data rejects the assertion that grades convey absolute meaning.

Keywords: Voting behavior, Majority Judgment, Ballot information, Online Experiment, Framing Effect

JEL Codes: D72, C93

[*]Université Paris Dauphine, 75 016 Paris, France. LAMSADE (UMR CNRS 7243)

[†]Université Lumière Lyon 2, CNRS, Université Jean Monnet Saint-Etienne, emlyon business school, GATE, 69007, Lyon, France, and Univ Rennes, CNRS, CREM-UMR6211, F-35000 Rennes, France.

[‡]Corresponding author: Université Jean Monnet, GATE Lyon Saint-Etienne UMR 5824, F-42 023 Saint-Etienne, France. Orcid: 0000-0002-4156-7527

[§]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

[¶]We are grateful to Jérôme Lang, Jean-François Laslier, and Sonia Paty for their feedback on the first draft of this paper, as well as to participants at the Social Choice and Welfare Society conference (Paris), and the CREST/Waseda Workshop for insightful comments and discussions, in particular Ulle Endriss, Guillaume Hollard, Yukio Koriyama, Laurent Linnemer, Oliwia Sczupska, Honorata Sosnowska, Radu Vranceanu, and Bill Zwicker.

1 Introduction

Physical phenomena like mass, distance, and time are measured using specific units such as grams, meters, or seconds. These measurements are reliable and universally consistent because they are based on internationally recognized metrological standards (*e.g.*, the standard meter bar). By contrast, mental states, such as emotions or preferences, are not publicly observable nor straightforwardly measurable and comparable across individuals (Robbins, 1932; Barrett, 2006). Indeed, there is no such thing as a natural measure of citizens' views or the common good. Therefore, specific devices must be designed to assess individual preferences and aggregate them into collective preferences. In electoral democracies, voting serves as the device through which citizens express their preferences, ultimately resulting in a ranking of candidates. However, correctly interpreting and comparing voters' preferences rely on two crucial assumptions: First, that a true individual preference exists, which can be interpreted without bias (epistemic property); second, that the voting rule in place does not distort the measurement of individual preferences (invariance property).

Majority judgment (MJ) has been proposed to eliminate distortions caused by other voting rules and more accurately represent voters' assessments of candidates (Balinski and Laraki, 2007). Under MJ, voters assess every candidate using verbal grades.[1] The winner is the candidate with the highest median grade, with a specific tie-breaking rule used to rank candidates who have the same median grade. One intention of proposing this best median rule is to reduce strategic voting (Balinski and Laraki, 2007, 2011, 2020). Then, the messages used as inputs for voting should be communicated in a common language shared among voters. This ensures that all votes can be interpreted consistently, allowing collective meaning to be derived from individual preferences. Balinski and Laraki (2011, page 161) sought some labels to evaluate candidates that "faithfully represent the merit of candidates, the excellence of performances, and the quality of competitors"; they also assert that "scales or measures constitute common languages of words that have absolute meanings, clearly understood by those who use them." A first field experiment on MJ in Orsay, France, aimed to test whether voters use verbal grades in a homogeneous way across subsamples. However, this is not necessarily a proper test of common language, as people may share a common language but use it differently, and vice versa (Fleurbaey, 2014).

This paper investigates whether grades in MJ have absolute meanings by examining whether the choice of grade by a voter for a candidate is impacted by framing effects related to the grade scale.[2] We conducted an online experiment (N=1955) where partic-

[1] MJ has already been used in practice. Among others, it has been used to choose the left-wing candidate at the 2016 and the 2022 French *primaire populaire* and to select the local representatives of the centrist party *La République en Marche* in 2019. The use of verbal grades also characterizes the broad consultation on the proposals of the French Citizen Climate Convention in 2020, and the voting rule used for the Paris participatory budget in 2021, 2022, and 2023.

[2] Another implication of the assumption of absolute meaning of grades would be that a voter's assigned

ipants evaluated 2022 French presidential candidates using MJ. Participants were randomly assigned to one of two treatment groups, each with a different grade scale. In one treatment (MJ7), seven grades were available ("excellent," "very good," "good," "fairly good," "acceptable," "insufficient," "to reject"). In the other treatment (MJ5), only five grades were available, excluding "excellent" and "to reject." Notably, the hypothetical nature of the vote reduces, if not eliminates, incentives for strategic behavior, which could otherwise influence median-based voting rules (Laslier, 2019) and confound the analysis of voting behavior.

We observe that grading behavior strongly depends on the grade scale. More precisely, when the two extreme grades are unavailable in MJ5, voters do not simply shift their evaluations to the closest grades (the new lowest or highest) but make greater use of intermediate grades. The result is remarkably robust for voters' evaluations of every presidential candidate and holds also in the case of highly disliked candidates. As an example, in the MJ5 treatment, 7.27% of participants assigned the grade "Acceptable" to Marine Le Pen (far right), compared to only 3.32% in the MJ7 treatment. The higher incidence of the "Acceptable" grade is surprising given that it has arguably a greatly different meaning than the always-present "Insufficient". Additionally, the shift to a 5-grade scale can also shift candidates' final evaluations. For instance, in the case of Emmanuel Macron (center), the absence of the "To reject" grade in the MJ5 treatment shifts his median evaluation from "Insufficient" (in MJ7) to "Acceptable", even though the "Insufficient" grade is still available to voters. These results suggest that the meaning of grades is not absolute. Instead, grades convey ordinal, relative assessments of candidates.

Related literature Our paper relates to three strands of literature, beginning with the social choice literature on alternative voting rules. The debate on the merits of various voting rules is broad and primarily grounded in theory (Sen, 1995; Brams and Fishburn, 2002). Within this literature, MJ has been proposed and analyzed theoretically in a series of articles and a book (Balinski and Laraki, 2007, 2011, 2014, 2020). Other academic contributions have challenged MJ's properties and assumptions (Felsenthal and Machover, 2008; Brams, 2011; Laslier, 2019; Fabre, 2021). These alternative voting rules need empirical testing and, in the case of MJ, there are only a few exceptions of experiments (Balinski and Laraki, 2010; Baujard et al., 2024). We contribute to this literature by conducting a large online experiment demonstrating that votes cast under MJ depend on the scale of grades used, thereby rejecting the assertion that grades have an absolute meaning.

Second, our main result—that the grade scale affects people's votes—connects our paper to a more recent strand of literature that investigates how preferences observed through voting depend on the voting rule used. The first issue in the literature is that, given that voting is meant to aggregate preferences and generate a collective choice,

grade to a candidate is independent of the set of other candidates in the election. Investigating this implication could be the topic of another paper.

strategic voting is unavoidable (Gibbard, 1973; Satterthwaite, 1975). Even best median rules, such as MJ, can be manipulated by voters (Laslier, 2019). Using data from a laboratory experiment, Baujard et al. (2024) find that participants vote in an equally strategic way under both MJ and evaluative voting. The latter voting rule is an ideal benchmark as it has been theoretically identified as manipulable (Núñez and Laslier, 2014). Thus, individual choices are likely to differ from genuine preferences in a strategic context also under MJ.

Another issue arises when voting rules allow voters to assess every candidate on a fixed, cardinal scale. Indeed, even without strategic voting, there is a "calibration" issue arising from the translation of individuals' rankings of candidates and the grades they assign. The framing effect arising from the response scale is a well-known issue in questionnaires in psychology (Schwarz et al., 1991, 2012), marketing (Weijters et al., 2010; Moe and Schweidel, 2012; Tsekouras, 2017), health (McDowell, 2006), and welfare (Fleurbaey and Blanchet, 2013).[3] Under evaluative voting, where voters assign numerical grades to each candidate and votes are aggregated by the sum, voters' behaviors and electoral outcomes have been shown to be impacted by the grading scale considered (Baujard et al., 2018, 2021; Darmann et al., 2019). However, no study focused on potential framing effects under MJ, where verbal grades might hold absolute meanings compared to numerical grades. In our experiment, the hypothetical nature of the vote does not incentivize strategic behaviors; thus, we can draw conclusions on the relation between the grades used and the framing effect abstracting from the strategic voting confounder.

Third, we contribute to the debate on the distinct views of democracy (Elster, 1986, 1998; Girard, 2019; Landemore, 2020). These views include electoral democracy — studied by the social choice tradition that focuses on aggregating individual preferences — deliberative democracy — valuing equal respect for autonomous agents with evolving views when confronted with the reasons of others — and epistemic democracy — where the involvement of many individuals is instrumental in uncovering a common existing truth. Voting rules, whether applied to voters or juries, share the same formal patterns and are analyzed using the same methods in social choice theory (Laslier, 2004). Accordingly, Balinski and Laraki (2007) approach electoral democracy and epistemic democracy indistinctly, attempting to identify the best methods applicable in both contexts. However, they correspond to two distinct issues. Questions of electoral democracy address procedures for aggregating individual preferences, each considered independently legitimate with no obligation to conform to any norms; this aligns with Arrow's theorem. Conversely, questions of epistemic democracy seek to collectively discover an existing truth on which each individual holds incomplete beliefs or knowledge; this is typical of

[3]For instance, the marketing literature has extensively studied the importance of response scale labels in understanding how consumers evaluate products, especially in online ratings (Weijters et al., 2010; Moe and Schweidel, 2012; Tsekouras, 2017). Given the clear differences in response styles across scales, Weijters et al. (2010) concludes that "interpreting levels of agreement with Likert items in an absolute sense (*e.g.*, 'the majority of respondents agree') is necessarily a tentative exercise at best".

the jury theorem (Laslier, 2010). Girard (2014) also claim these different contexts correspond to drastically different philosophical properties. Among others, that there exists a political truth is a debated assumption (Reiss, 2019, 2020). For those who question the idea of political truth, the goal of finding the uniquely best decision at the collective level does not make any sense; the aim of electoral democracy is rather reconsidered as an attempt to find a compromise and take seriously individual preferences. We argue that the desired properties of rules should therefore differ, depending on whether they are used for electoral or epistemic purposes. To our knowledge, few papers have tackled the issue that a procedure could be more or less adapted to electoral or epistemic democracy (see Procaccia and Shah, 2015 and Allouche et al., 2022 for approval voting). Our results indicate that framing effects can significantly influence the language used to articulate assessments, supporting the claim that participants under MJ express ordinal political preferences for candidates rather than absolute assessments of their merit. Hence, the quest of objectively evaluating the merit of each candidate is misleading and MJ might be more adapted to epistemic issues than to electoral democracy.[4]

The remainder of the paper is organized as follows. In Section 2, we identify the assumption that will be tested. In Section 3, we present the experimental design and our participant sample. In Section 4, we present our results. Finally, in Section 5, we provide our conclusions and possible avenues for future research.

2 The tested hypothesis

Let us consider the assertion that candidates' assessments under MJ are made in a universal language. A corollary of this assertion is that the grades used in the assessments are *associated with absolute meanings*. Asserting that grades have absolute meanings involves two different assumptions, depending on whether they are considered from an *inter-personal* or *intra-personal* perspective.

From an *inter-personal* perspective, assuming that there is a common language in grading implies that grades have the same meaning for all individuals. Thus, if two individuals share the same assessment of a candidate, they should assign the same grade to that candidate. In contrast, if there is no common language, imagine that voter A consistently overstates their evaluations of every candidate, while voter B consistently understates theirs. As a result, candidates supported by A might be favored over those supported by B, even when their genuine assessments are identical. In our case, testing the inter-personal perspective of absolute meaning requires comparing grades from individuals with exactly the same preferences. However, such preferences are inherently difficult to quantify. Therefore, we do not examine the *inter-personal* perspective in this paper.

From an *intra-personal* perspective, assuming that the grades have absolute meanings

[4]For instance, Balinski and Laraki (2007) discuss the case of wine judges, in which the evaluation of wine quality relies on well-defined and clearly understood grades.

implies that individuals consistently provide the same assessments in different contexts. If an individual's genuine assessment remains constant over time, modifying the grade scale should not affect the expression of their preferences. For example, if an individual assigns a grade to a candidate, they would select the same grade even if the lowest grade were removed, unless the assigned grade itself was eliminated. Hence, reducing the scale would consistently reallocate the removed assessments to the closest ones. In contrast, if every other assessment is altered by reducing the scale, this indicates that the grades do not convey absolute meaning but only convey ordinal relative meaning. The reallocation of grades may follow a smooth relative pattern (as in Dhillon and Mertens, 1999) or result in discontinuities. Hence, we test the following consequence of the "absolute meaning" assumption:

Tested hypothesis: Consider a given grade scale where both the lowest and highest grades are removed. After this reduction, a voter shifts only the former lowest and highest grades to the new extremes, leaving the rest of the grade distribution unchanged.

One way to test this hypothesis would be to examine how the same individuals vote using different versions of MJ. However, a within-subject design could lead to carry-over effects, where participants choose the same grades to appear consistent, or demand effects, where they act based on what they believe researchers expect (Charness et al., 2012). To avoid these biases, we opted for a between-subject design. That is, we test the hypothesis of absolute meaning by studying differences in the distribution of grades assigned by individuals in two comparable populations, each voting under MJ with different grading scales (more detail in Section 3).

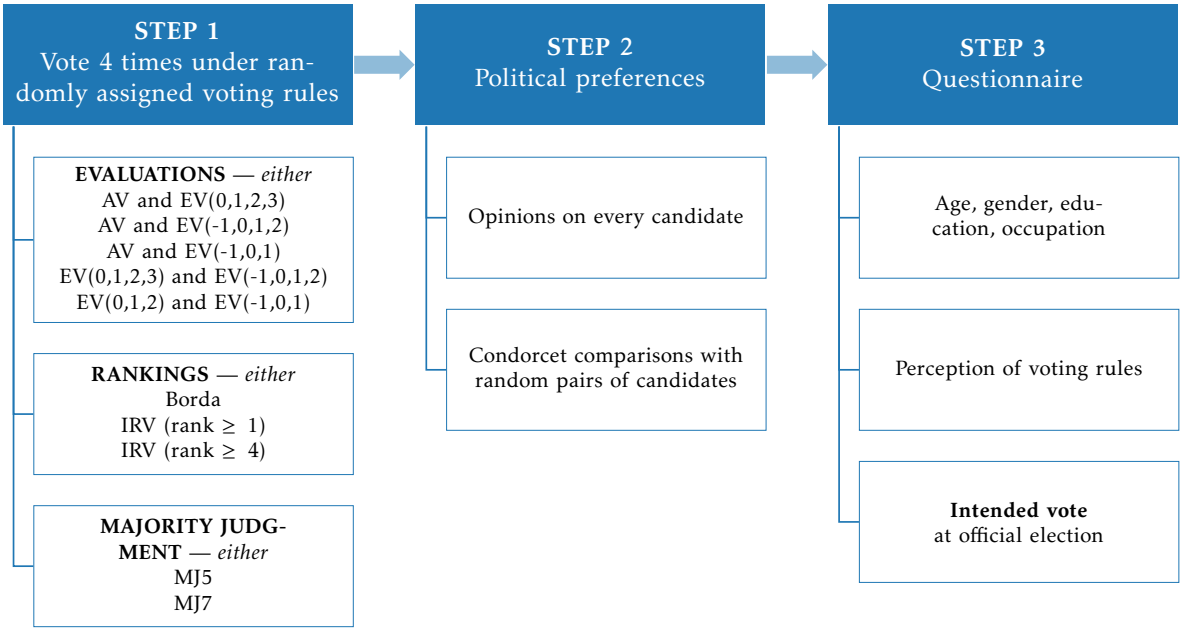
3 Experimental design and collected data

We now present the experimental protocol (for a complete description, see Delemazure and Bouveret 2024). The experiment, carried out using an online web application, is divided into three parts, as shown in Figure 1. After the introduction screen, in the first step of the experiment, each participant is asked to vote for the 12 official candidates in the 2022 French presidential election using three different voting rules sequentially. Each voting rule is one-round and randomly selected from the ten rules listed in the note of Figure 1. The voting rules in this part differ for each participant and are randomly assigned to ensure that each sequence includes two scoring rules, one ranking rule, and one version of MJ, presented in this order.

All participants voted under MJ after they learn how votes are aggregated, even though we do not elaborate on the tie-breaking rule.[5] Some participants were randomly assigned

[5]We provided the link to the Wikipedia page that explains how the tie-breaking rule works (in French).

Figure 1: Sequence of the experiment



Note: AV stands for approval voting; EV(0,1,2,3), EV(-1,0,1,2), EV(-1,0,1), and EV(0,1,2) stand for evaluative voting, with the numbers in parenthesis indicating the grades available; Borda stands for Borda ranking of 4 candidates; IRV stands for instant runoff voting. The rules used during the first step are not the same for all participants: they are randomly chosen so that the set always contains two scoring rules, one ranking rule, and one version of MJ, presented in this order.

to the MJ7 treatment, where they voted on a 7-label scale: "To reject," "Insufficient," "Acceptable," "Fairly good," "Good," "Very Good," and "Excellent." In contrast, the rest of the participants were assigned to the MJ5 treatment and voted on a 5-label scale, such that the scale contains only the 5 middle labels, excluding the lowest and highest grades, "To reject" and "Excellent." These grades partially mirror those used in French schools for the high school terminal exam (*Baccalauréat*), making them familiar to the participants of this experiment.

In the second step, participants provide their opinions on each candidate using a continuous scale with 100 values and compare random pairs of candidates. In the third step, participants complete a short questionnaire about their age, socio-professional category, and perceptions of the voting rules they encountered. We also ask which candidate they voted for (or intend to vote for) in the official election.

Our sample Our online experiment was conducted during the French Presidential election in 2022. Anyone could participate freely in the experiment by visiting the website[6] and answering the questions. We primarily advertised the experiment through academic and social networks, particularly Twitter. The website opened on April 8th, 2022. In total, a total of 2,308 people participated before May 7th, including 687 who participated before the first round of the official election (on April 10th), and 2,229 before the second

[6] https://vote.imag.fr/

round (on April, the 24th).[7] Among these participants, we excluded participants younger than 18 years, those who voted abroad, those who were not on the electoral rolls for the presidential elections, those who skipped the vote under MJ and the socio-demographic and political questions. The final dataset includes 1,995 participants. The political context ensured that participants were familiar with all the candidates and the stakes of the collective decision. This aimed to prevent the typical "rational ignorance" and "under-engagement" that can lead to unusual and non-interpretable outcomes (Fishkin, 2011).

Table 1 shows the summary statistics of our sample divided into the two treatments and the same statistics for the French general population. Two important points must be highlighted. First, our sample is not representative of the French population; it is skewed toward male, young, well-educated, and left-wing participants. Additionally, only a few participants abstained from voting, while the actual abstention rate in the elections was significantly higher. This selection bias is not surprising, considering that the sample was not obtained through a survey institute. To address our sample's lack of representativeness, we apply post-stratification weights in a robustness check and show that results are robust. Second, there are no significant differences between MJ7 and MJ5 in terms of socio-demographic and political preferences, indicating successful randomization. Nevertheless, in all regressions, we control for participants' socio-demographics and political preferences to account for any minor differences between the two treatments.[8]

4 Results

4.1 Descriptive analysis

We first display the complete dataset on MJ5 and MJ7 to visualize how behaviors comply with the "assumption of absolute meaning" in case of a change in scale. Figure 2 presents the observed distribution of the grades for the scales MJ7 and MJ5 when we pool the votes for every candidate. The assumption of absolute meaning is represented as Expected MJ5. In Expected MJ5, the grades "To reject" and "Excellent" disappear and are respectively converted to "Insufficient" and "Very good". Hence, compared to MJ7, the expected MJ5's "Insufficient" exactly covers both "To reject" and "Insufficient", while the expected MJ5's "Very good" exactly covers both "Very good" and "Excellent". The proportion of other grades is not impacted and the limits separating the different grades are exactly the same.

If the assumption of absolute meaning holds, the distribution Expected MJ5 and the distribution observed under MJ5 should be identical. However, this is not the case. The proportion of "To reject" and "Insufficient" for MJ7 is higher than that of "Insufficient" for MJ5, meaning that by removing the extremes, some people who would have given the

[7] Note that although the website remains available for practice, the data used in this article is limited to participants who participated before May 7th .

[8] To analyze the hypothetical results of the elections, adjusted data are provided on the website `https://vote.imag.fr/results/online-2022`.

Table 1: Summary statistics

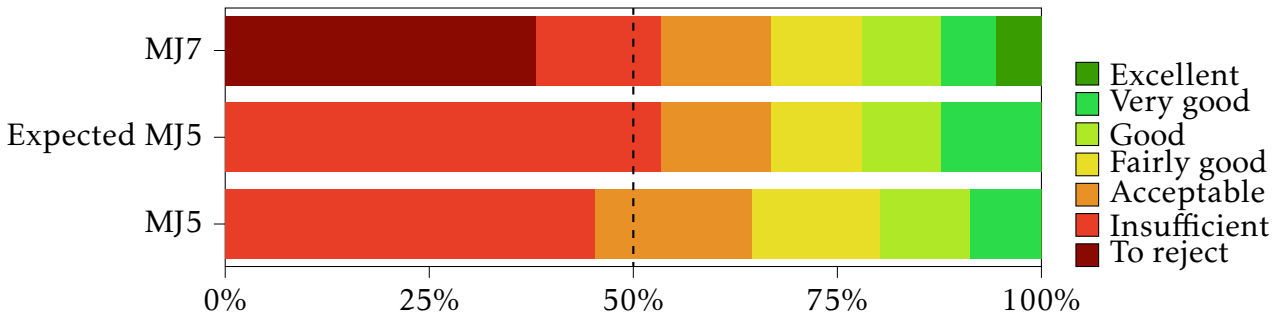
	Our sample		General population
	JM7 (N=965)	JM5 (N=990)	
Gender			
Male	0.68	0.68	0.48
Female	0.32	0.32	0.52
Age			
18-29	0.44	0.42	0.17
30-39	0.27	0.30	0.15
40-49	0.18	0.15	0.16
50+	0.11	0.13	0.51
Education			
High school diploma	0.05	0.05	0.68
Higher education	0.95	0.95	0.32
Vote			
Emmanuel Macron (Center)	0.12	0.11	0.28
Marine Le Pen (Far right)	0.02	0.02	0.23
Jean-Luc Mélenchon (Far left)	0.61	0.64	0.22
Eric Zemmour (Far right)	0.02	0.02	0.07
Valérie Pécresse (Right)	0.00	0.01	0.05
Yannick Jadot (Left)	0.11	0.07	0.05
Jean Lassalle (Center)	0.01	0.02	0.03
Fabien Roussel (Far left)	0.02	0.03	0.02
Nicolas Dupont-Aignan (Right)	0.00	0.01	0.02
Anne Hidalgo (Left)	0.02	0.02	0.02
Philippe Poutou (Far left)	0.01	0.01	0.01
Nathalie Artaud (Far left)	0.00	0.00	0.01
Blank vote	0.00	0.00	0.01
Abstained	0.00	0.00	0.26

Note: The data on gender, age, and education of the French general population was retrieved from the census data (INSEE). The data on the official results of the first round of the 2022 presidential election was retrieved from the French Ministry of Interior.

grade "Insufficient" instead gave "Acceptable". Furthermore, the proportions of "Acceptable" and "Fairly good" are greater for MJ5 than for expected MJ5. We derive the same conclusion by observing the proportion of "Very Good" and "Excellent".

In Appendix A, we present the grade distribution for each of the main candidates, confirming that the distribution varies across the two treatments. Notably, for Emmanuel Macron (center) and Jean Lassalle (center), the median grade shifts from "Insufficient" in MJ7 to "Acceptable" in MJ5, even though the "Insufficient" grade remains available. Similarly, Fabien Roussel's (far left) median grade moves from "Acceptable" to "Fairly good," while "Acceptable" remains an option. For Jean-Luc Mélenchon (far left), the median grade changes from "Very good" to "Good," even though "Very good" is still available. These shifts suggest that the grade scale can impact the final assessments candidates receive.

Figure 2: Distribution of grades for MJ5, MJ7, and expected MJ5



Notes: The bar labelled MJ7 shows the observed distribution of grades under MJ7. Among the participants in MJ7, 38.10% assigned a grade of "To reject," while 15.33% gave a grade of "Insufficient". The bar labeled Expected MJ5 describes the expected distribution of grades under MJ5, considering observed behaviors under MJ7, should the assumption of absolute meaning of grades hold. 53.43% (38.10%+15.33%) of participants should give a grade of "Insufficient". The bar labeled MJ5 describes the observed distribution of grades under MJ5. 45.36% of participants have given the grade "Insufficient". Expected MJ5 and MJ5 should be the same under the assumption of the absolute meaning of grades.

Additionally, Appendix B provides further visualizations and interpretations of the results, replicating the main findings for candidates previously approved or not under approval voting and based on grades received in evaluative voting. Across cases, we confirm that the meaning associated with grades varies between treatments.

4.2 Regression analysis

We now further test the assumption of absolute meaning and make statistical inference by comparing individual voting behavior under MJ7 and MJ5. To do this, we estimate the following linear probability model

$$y_{ij} = \kappa + \beta \text{MJ5}_{ij} + \gamma \mathbf{X_i} + \epsilon_{ij}, \tag{1}$$

where y_{ij} is a dummy variable representing the grade assigned by voter i to candidate j under either treatment MJ5 or MJ7. y_{ij} is equal, in turn, *1)* to 1 when the voter uses a low grade, namely "To reject" or "Insufficient", and to 0 otherwise; *2)* to 1 when the voter uses a high grade, namely "Very good" or "Excellent"; *3)* to 1 when the voter uses an intermediate grade, that is, either "Acceptable," "Fairly good," or "Good". $MJ5_{ij}$ is a dummy variable, which is equal to 1 if the participant i who assesses candidate j is assigned to the MJ5 treatment, and 0 if assigned to MJ7. $\mathbf{X_i}$ represents the control variables, *e.g.*, participants i's gender, age, education level, and political preferences, and ϵ_{ij} represents the error term.

Our goal is to test for any difference in the probability that participants use high, low, or intermediate grades between treatments MJ5 and MJ7. This difference is captured by

the coefficient β associated with the MJ5 treatment variable. If the assumption of absolute meaning holds, we would expect no difference in the probability of using these grades across the two treatments (*i.e.*, $\beta = 0$). Indeed, voters who would use the extreme grades in MJ7 ("To reject" and "Excellent") should simply shift to the corresponding extreme grades in MJ5 ("Insufficient" and "Very good"), leaving the distribution of the intermediate grades unchanged. In contrast, if the assumption of absolute meaning does not hold, the coefficient β should significantly differ from zero. The higher β in absolute level, the higher the discrepancy between grading behaviors when the grade scale varies. A negative (positive) β implies that the probability of using a *1)* low, *2)* high or *3)* intermediate grade is lower (higher) under MJ5 than under MJ7.

We estimate model (1) by OLS and cluster standard errors at the individual level. Results are shown in Table 2. First, we observe that the probability of using the lowest grade (column 1) and highest grade (column 2) in MJ5 is lower than that of using the two lowest and two highest grades in MJ7. Both effects are statistically ($p < 0.01$) and economically significant, with the treatment leading to an 8 p.p. decrease (-15%) in the use of the low grades and a 3.6 p.p. decrease (-29%) in the use of the high grades. The decrease in the use of extreme grades in the MJ5 treatment shifts grading behavior towards the center, resulting in a 12 p.p. higher probability (+35%) of using intermediate grades than in the MJ7 treatment. Even in this case, the effect is statistically significant ($p < 0.01$). These results reveal that the grades' meaning depends on the relative position on the scale used.

Table 2: Regression analysis - Probability of using low, high, or intermediate grades

	Low grades	High grades	Intermediate grades
	(1)	(2)	(3)
MJ5	-0.080***	-0.036***	0.12***
	(0.006)	(0.004)	(0.007)
Constant	0.54***	0.14***	0.32***
	(0.02)	(0.009)	(0.02)
Dep. var. mean in MJ7	0.53	0.12	0.34
Observations	23460	23460	23460
Clusters	1955	1955	1955
Controls	✓	✓	✓
R-squared	0.010	0.011	0.015

Notes: OLS regression. The dependent variable is a dummy for whether participants assigned to a candidate i) "To reject" or "Insufficient" (column 1), ii) "Very good" or "Excellent" (column 2), and iii) "Acceptable", "Fairly good", and "Good" (column 3). MJ5 is a dummy equal to 1 if participants voted with the 5-label scale (without "To reject" and "Excellent") and 0 if they voted with a 7-label scale. Controls include participants' gender, age, education level, and political preferences. Standard errors clustered at the participant level in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.

Vote for each candidate. We also estimate the model (1) considering participants' votes for each of the main candidates representing the different sides of the French political spectrum.[9] We plot the estimated β coefficients for each regression in Figure 3, along with the estimated coefficient from the main regressions where all votes are pooled. Our result on using the low (panel (a)) and high grades (panel (c)) is robust when we consider the vote for each candidate, apart from a null effect on the use of high grades for Pécresse (Right) and Le Pen (Extreme right). The latter null effect is explained by the fact that in our sample, virtually no voter gave the highest grades available to either candidate (2.4% to Le Pen and 0.92% to Pécresse).

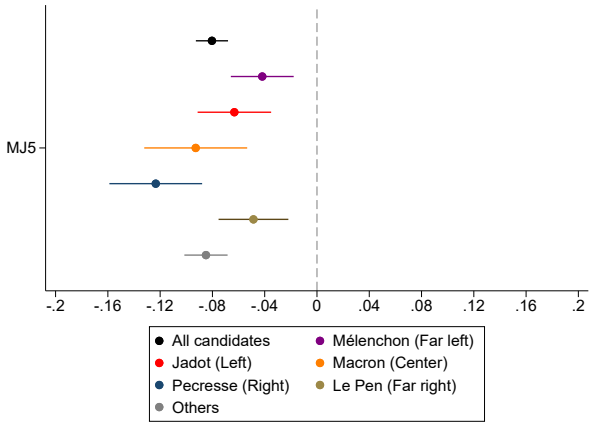
In Panel (c), we further find that our results on the use of intermediate grades remain valid and similar in magnitude when we consider the vote for each candidate (in all cases, $p < 0.01$). However, when considering the vote for Marine Le Pen (extreme right), the coefficient on *MJ5* has a lower yet highly significant effect ($p < 0.01$). In the MJ7 treatment, 83.83% of participants assign "To reject," 5.39% "Insufficient," and 3.32% "Acceptable" to her. This low score is easily interpreted in light of the left-wing composition of our sample. Yet, in the MJ5 treatment, 84.85% of participants assign "Insufficient" to her and 7.27% "Acceptable," suggesting that also for highly disliked candidates, the absence of a low-grade shifts grades to the middle. Hence, more participants declare her as "Acceptable," which has an extremely different meaning than "Insufficient."

Robustness checks We finally conduct a series of robustness checks displayed in appendix C: first, we apply post-stratification weights using the iterative proportional fitting (or raking) method (Kolenikov, 2014) to account for the representation bias in our data. In one regression, the weights account for the gender, age, and education differences between our sample and the French general population. In a different regression, the weights are built to account for differences in participants' vote and the official election outcome. Second, we run the main estimations using Probit regression to check whether our results are robust enough to use different estimation methods. Third, we use randomization inference (Young, 2019) to recompute the p-values of our main regressions, aiming to test whether the observed effect of the MJ5 treatment stems from the randomization process or if different random distributions of participants to the treatment would result in similar outcomes. Finally, in Table C.0.3, we replicate the main results while controlling for the rules participants tried before voting under MJ to control for possible carry-over effects. In all these robustness exercises, the effect of the MJ5 treatment remains strongly significant, confirming the validity of our results.

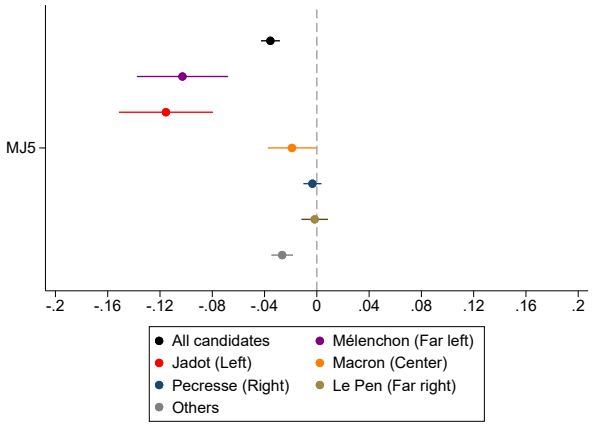
[9] We replicate this analysis considering also other minor candidates in Figure B.0.5.

Figure 3: Treatment difference between MJ7 (Baseline) and MJ5 - Main candidates

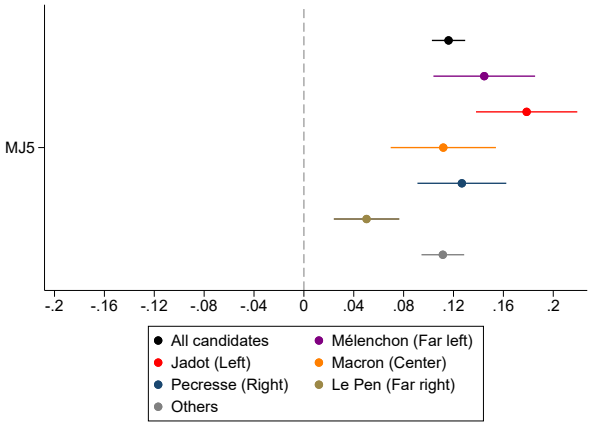
(a) Low grades ("To reject" or "Insufficient")



(b) High grades ("Very good" or "Excellent")



(c) Intermediate grades ("Acceptable", "Fairly good", or "Good")



Notes: Dots with horizontal lines indicate point estimates of the coefficient on the MJ5 treatment with 95% confidence intervals from linear least squares regression. Controls include gender, age, education, and political preferences. $N = 23,460$ in the regression with all candidates, $N = 1,955$ in the regression with the single candidates, $N = 13,685$ in the regression with the other candidates. Standard errors clustered at the individual level in the regression with all candidates and the other candidates. When considering the vote for individual candidates, we use robust standard errors.

5 Conclusion

In this paper, we scrutinized the interpretation of the "absolute meaning of grades" under MJ. More specifically, we tested whether the use of grades in fictitious elections with MJ is impacted by a framing effect due to the grade scales. With this aim, we have run an online experiment where participants were randomly assigned to a treatment in which they voted with different versions of MJ, whether MJ5, with a 5-scale of grades, or MJ7, with a 7-scale of grades.

Our results suggest that the meaning of grades is significantly relative to scales, even in cases in which participants had strong preferences over candidates. Two striking cases

are the vote for Emmanuel Macron (center), for which the median grade shifts from "Insufficient" in the MJ7 treatment to "Acceptable" in the MJ5 treatment, even though "Insufficient" is still available to voters; and the vote for Marine Le Pen (far right), who obtained more "Acceptable" grades in the MJ5 treatment only because the scale shrinked. These results reject the assumption of absolute meaning of grades.

Note that the argument of grade meaning relativity does not disrupt MJ election outcomes, which only depend on a given unique scale. It would, should we establish that different people who vote together may associate inconsistent meanings to the same grade. The interpersonal comparability issue as to how individuals with exactly the same preferences use the grade labels has been discussed in the case of evaluative voting and utilitarianism in general, but it is yet to be addressed in the case of MJ. A potential venue for future research, for instance, would be to analyze this question in the lab where individuals are assigned exogenous preferences.

This analysis has been conducted in a political context, and its results should not be generalized *a priori* to non-political contexts. In particular, we cannot infer from this study whether grades used with MJ in an epistemic context would be absolute or scale-dependent.

Finally, we should acknowledge a potential limitation of our study as the self-selected and unrepresentative sample may raise concerns regarding the external validity of our results. However, we believe this limitation does not undermine our findings for two reasons. First, a robustness analysis using post-stratification weights indicated that the results remained consistent when we accounted for differences between our sample and the general population. Second, there is no compelling reason to believe that framing effects would be absent in a more representative sample. Specifically, in our case, the framing effect observed in a sample predominantly composed of engaged participants likely represents a lower bound for the effect of interest, as these participants are expected to have firmer opinions on candidates.

Declarations

5.1 Funding

This research was funded by the French National Research Agency under the Citizens project "ANR-22-CE26-0019".

5.2 Competing interests

The authors declare that they have no relevant or material financial interests that relate to the research described in our paper entitled "Do Grades Have Absolute Meaning? An Experiment on Majority Judgment".

References

Allouche, T., Lang, J., and Yger, F. (2022). Truth-Tracking via Approval Voting: Size Matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):4768–4775.

Balinski, M. and Laraki, R. (2007). A Theory of Measuring, Electing, and Ranking. *Proceedings of the National Academy of Sciences*, 104(21):8720–8725.

Balinski, M. and Laraki, R. (2010). Election by majority judgement: Experimental evidence. In *In situ and Laboratory Experiments on Electoral Law Reform: French Presidential Elections*, chapter 2, pages 13–54. Springer, Heidelberg.

Balinski, M. and Laraki, R. (2011). *Majority Judgment: Measuring, Ranking, and Electing*. The MIT Press.

Balinski, M. and Laraki, R. (2014). Judge: Don't Vote! *Operations Research*, 62(3):483–511. Publisher: INFORMS.

Balinski, M. and Laraki, R. (2020). Majority Judgment vs. Majority Rule. *Social Choice and Welfare*, 54(2):429–461.

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, 1(1):28–58. Publisher: SAGE Publications Inc.

Baujard, A., Brunetti, R., Lebon, I., and Marsilio, S. (2024). How People Understand Voting Rules. *Working paper*.

Baujard, A., Gavrel, F., Igersheim, H., Laslier, J.-F., and Lebon, I. (2018). How voters use grade scales in evaluative voting. *European Journal of Political Economy*, 55:14–28.

Baujard, A., Igersheim, H., and Lebon, I. (2021). Some regrettable grading scale effects under different versions of evaluative voting. *Social Choice and Welfare*, 56(4):803–834.

Brams, S. J. (2011). Grading Candidates.

Brams, S. J. and Fishburn, P. C. (2002). Chapter 4 Voting procedures. In *Handbook of Social Choice and Welfare*, volume 1 of *Handbook of Social Choice and Welfare*, pages 173–236. Elsevier.

Charness, G., Gneezy, U., and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1–8.

Darmann, A., Grundner, J., and Klamler, C. (2019). Evaluative voting or classical voting rules: Does it make a difference? Empirical evidence for consensus among voting rules. *European Journal of Political Economy*, 59:345–353.

Delemazure, T. and Bouveret, S. (2024). Voter Autrement 2022 - The Online Experiment ("Un Autre Vote"). Dataset and companion article on Zenodo.

Dhillon, A. and Mertens, J.-F. (1999). Relative utilitarianism. *Econometrica*, 67(3):471–498.

Elster, J. (1986). The market and the forum: Three varieties of political theory. In *Foundations of social choice theory*, pages 103–132. Cambridge University Press, Cambridge.

Elster, J. (1998). Deliberative democracy. *Cambridge University of Pennsylvania*.

Fabre, A. (2021). Tie-breaking the Highest Median: Alternatives to the Majority Judgment. *Social Choice and Welfare*, 56(1):101–124.

Felsenthal, D. S. and Machover, M. (2008). The Majority Judgement Voting Procedure: A Critical Evaluation. *Homo oeconomicus*, 25 (3/4):319–334.

Fishkin, J. S. (2011). *When the people speak. Deliberative democracy and public consultation*. OUP.

Fleurbaey, M. (2014). Review of Majority judgment. Measuring, ranking, and electing. *Social Choice and Welfare*, 42(3):751–755. Publisher: Springer.

Fleurbaey, M. and Blanchet, D. (2013). *Beyond GDP: Measuring Welfare and Assessing Sustainability*. Oxford University Press.

Gibbard, A. (1973). Manipulation of voting schemes : a general result. *Econometrica*, 41:587–601.

Girard, C. (2014). La règle de majorité en démocratie : équité ou vérité ? *Raisons politiques*, 53(1):107–137.

Girard, C. (2019). *Délibérer entre égaux. Enquête sur l'idéal démocratique*. L'esprit des lois. Vrin.

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*.

Landemore, H. (2020). *Open democracy: reinventing popular rule for the twenty-first century*. Princeton University Press, Princeton.

Laslier, J.-F. (2004). *Le vote et la règle majoritaire: analyse mathématique de la politique*. CNRS.

Laslier, J.-F. (2010). In Silico Voting Experiments. In Laslier, J.-F. and Sanver, M. R., editors, *Handbook on Approval Voting*, pages 311–335. Springer, Berlin, Heidelberg.

Laslier, J.-F. (2019). The strange "Majority Judgment". *Revue economique*, 70(4):569–588.

McDowell, I. (2006). *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press USA - OSO, New York, 3rd ed edition.

Moe, W. W. and Schweidel, D. A. (2012). Online Product Opinions: Incidence, Evaluation, and Evolution. *Marketing Science*, 31(3):372–386. Publisher: INFORMS.

Núñez, M. and Laslier, J.-F. (2014). Preference intensity representation: Strategic overstating in large elections. *Social Choice and Welfare*, 42(42):313–340.

Procaccia, A. D. and Shah, N. (2015). Is Approval Voting Optimal Given Approval Votes? *Advances in neural information processing systems*, 28.

Reiss, J. (2019). Expertise, Agreement, and the Nature of Social Scientific Facts or: Against Epistocracy. *Social Epistemology*, 33(2):183–192.

Reiss, J. (2020). Why Do Experts Disagree? *Critical Review*, 32(1-3):218–241.

Robbins, L. (1932). *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions : Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L. (1991). Rating Scales Numeric Values may change the Meaning of Scale Labels. *Public Opinion Quarterly*, 55(4):570–582.

Schwarz, N., Knäuper, B., Oyserman, D., and Stich, C. (2012). The Psychology of Asking Questions. In *International Handbook of Survey Methodology*. Routledge.

Sen, A. (1995). How to Judge Voting Schemes. *The Journal of Economic Perspectives*, 9(1):91–98.

Tsekouras, D. (2017). The Effect of Rating Scale Design on Extreme Response Tendency in Consumer Product Ratings. *International Journal of Electronic Commerce*, 21(2):270–296. Publisher: Routledge _eprint: https://doi.org/10.1080/10864415.2016.1234290.

Weijters, B., Cabooter, E., and Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247.

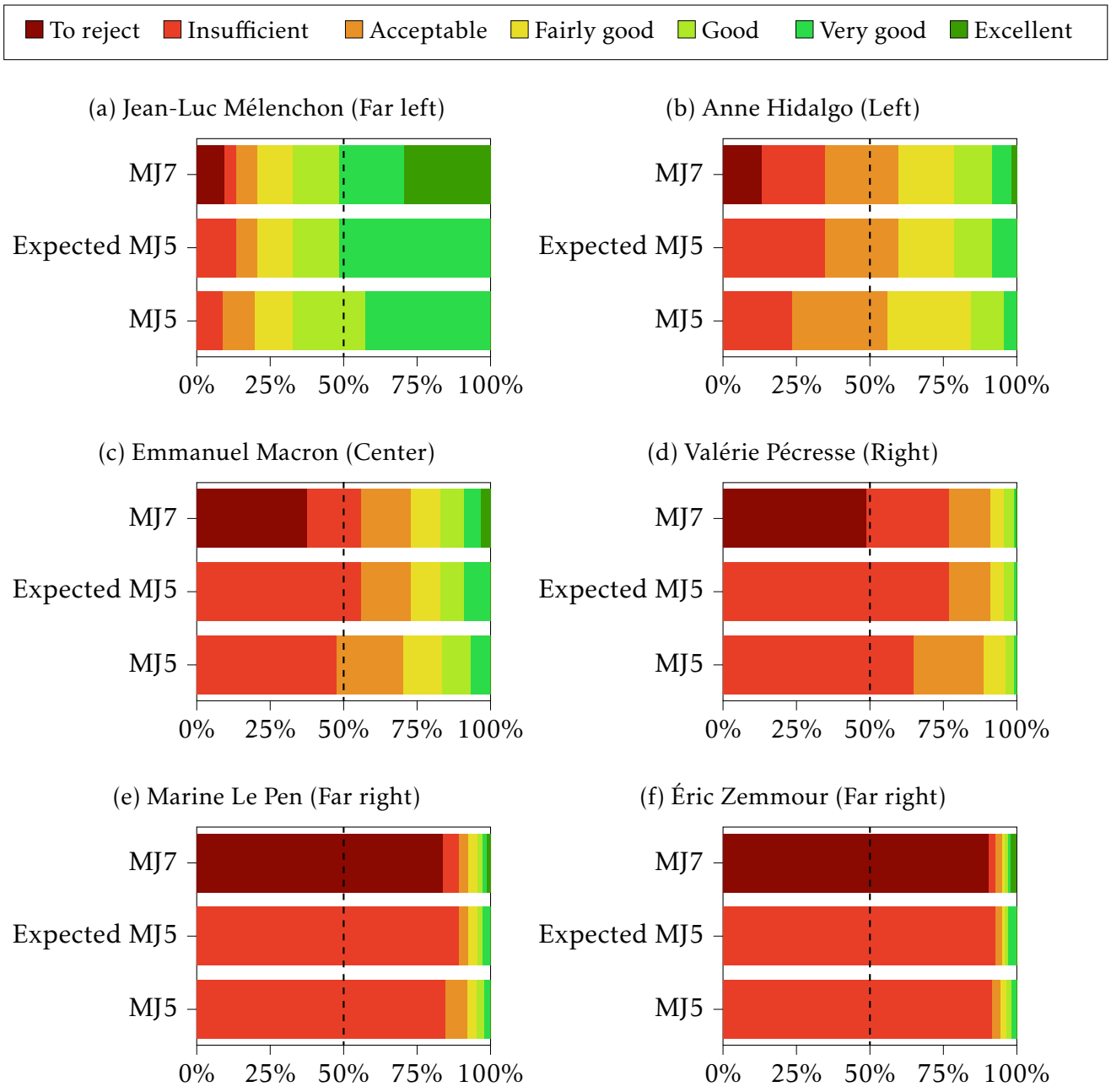
Young, A. (2019). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *The Quarterly Journal of Economics*, 134(2):557–598.

Appendix

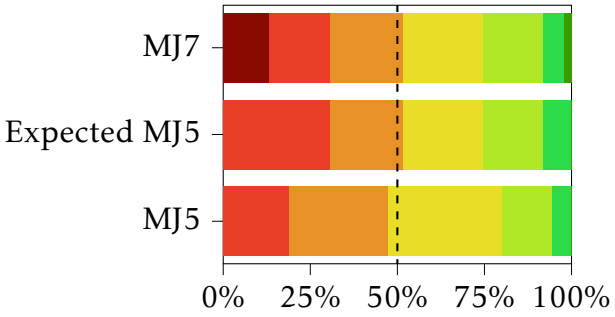
Table of Contents

A Results for specific candidates

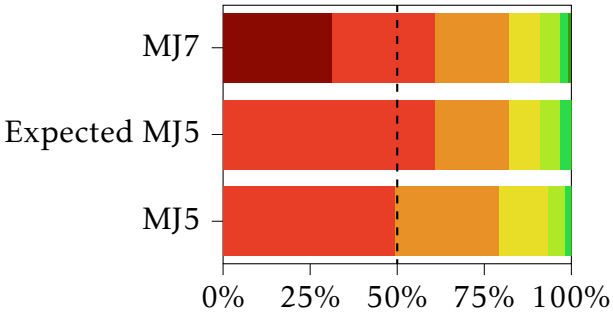
Figure A.0.1: Distribution of grades for each candidate



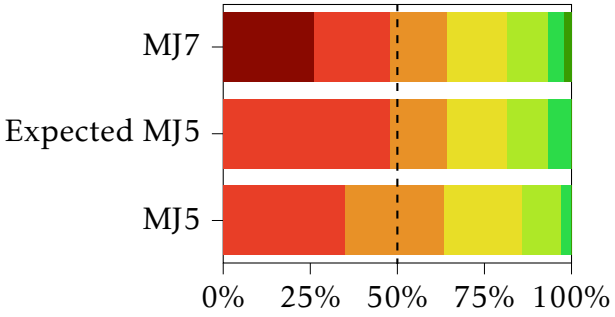
(g) Fabien Roussel (Far left)



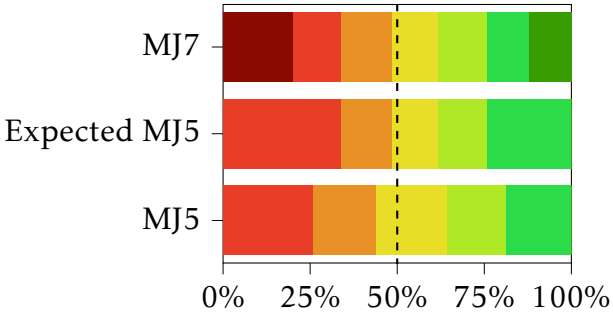
(h) Jean Lassalle (Center)



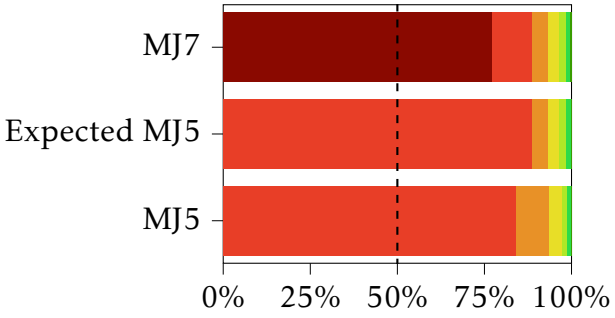
(i) Nathalie Arthaud (Far left)



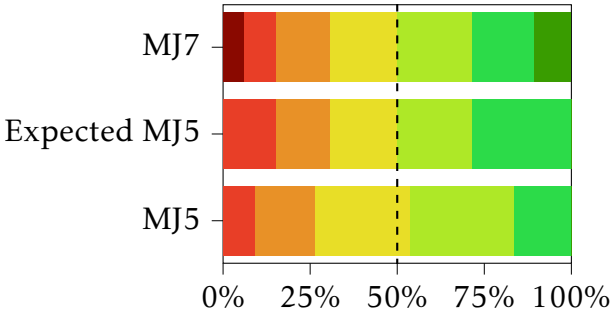
(j) Philippe Poutou (Far left)



(k) Nicolas Dupont-Aignan (Right)



(l) Yannick Jadot (Left)



B Additional results

Figures B.0.1a (resp. Figure B.0.1b) shows, like Figure 2, the distribution of grades on candidates, but this figure is restricted to the participants that tested approval voting, and to the candidates who were approved (resp. disapproved) by those participants.

Figure B.0.1: Distribution of the grades of approved and unapproved candidates for MJ5 and MJ7

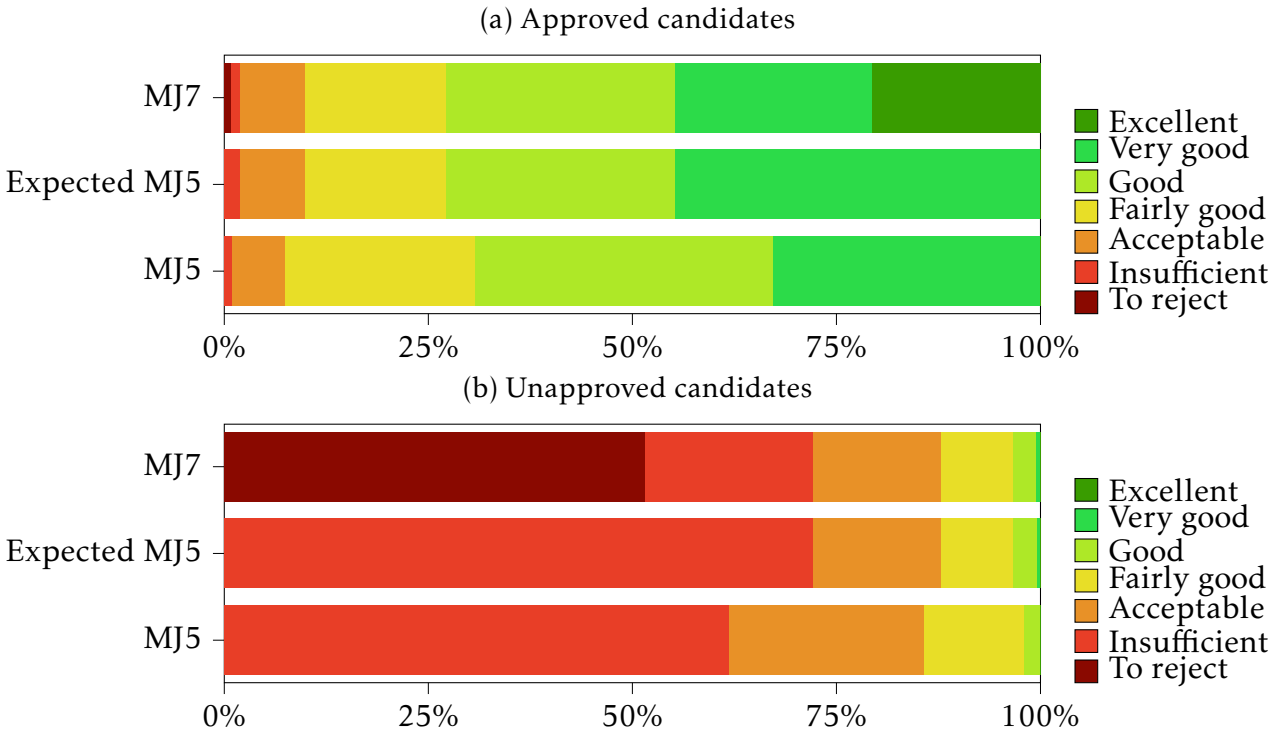


Figure B.0.2 presents approval rates of candidates depending on which grade participants assigned them. Under the assumption of absolute meaning, these approval rates should not vary between MJ5 and MJ7 for the middle grades, namely "Acceptable", "Fairly good" and "Good". Indeed, if a participant approves a candidate but gives this candidate "Acceptable" under MJ5, she or he should also give this candidate the same grade under MJ7. However, that is not what we observe in Figure B.0.2. In particular, the relative variation for "Acceptable" and "Good" are not negligible. For instance, the approval rate of candidates with the grade "Acceptable" goes from 9% with MJ5 to 15% with MJ7.

Figure B.0.3 proposes a similar representation of experimental data, considering the possible scores of evaluative voting -1, 0, and 1: This scale contains one objectively negative score, one neutral, and one positive. We also note that the proportion of participants who assign an extreme grade of MJ5 ("Insufficient" or "Very Good") increases for score 0 when we go to MJ7. Note that this proportion also increases for scores -1 and 1.

Participants were also asked to provide their "true" assessment of each candidate, independent of any collective voting, by assigning a number between 0 and 100. This

Figure B.0.2: Approval rate of candidates having a grade assigned

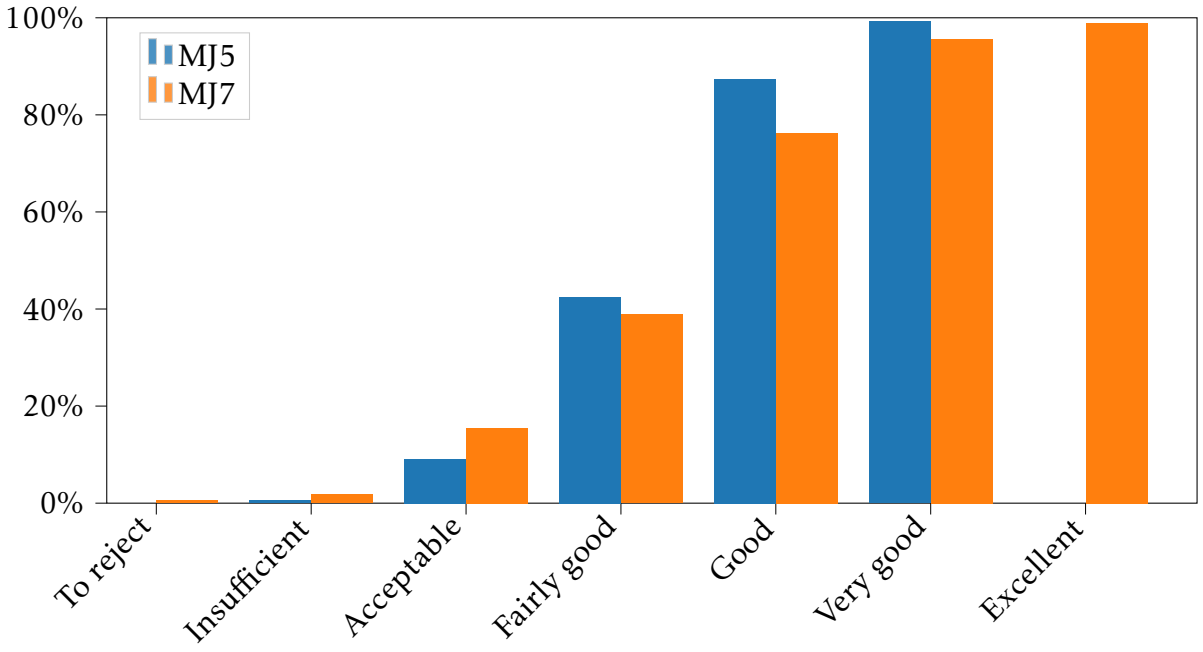
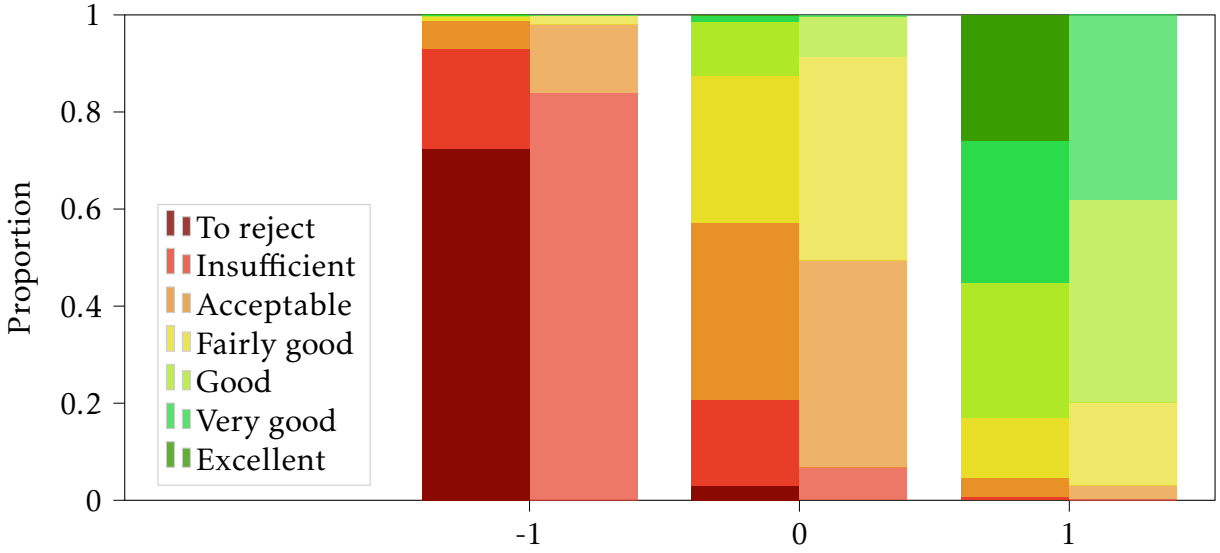


Figure B.0.3: Distribution of the grades for scores -1, 0 and 1 for MJ5 and MJ7



continuous score represents the cardinal evaluation of candidates as perceived by the participants. We can analyze the distribution of participants' use of different grades for candidates based on their opinions. Once more, the assumption of absolute meaning predicts that the proportion assigned to the middle grades "Acceptable," "Fairly good," and "Good" should remain consistent for both MJ5 and MJ7. Figure B.0.4 shows that it is not the case for "Acceptable" and "Good", for which the deviation is important.

Figure B.0.4: Average note given to candidates having a given grade

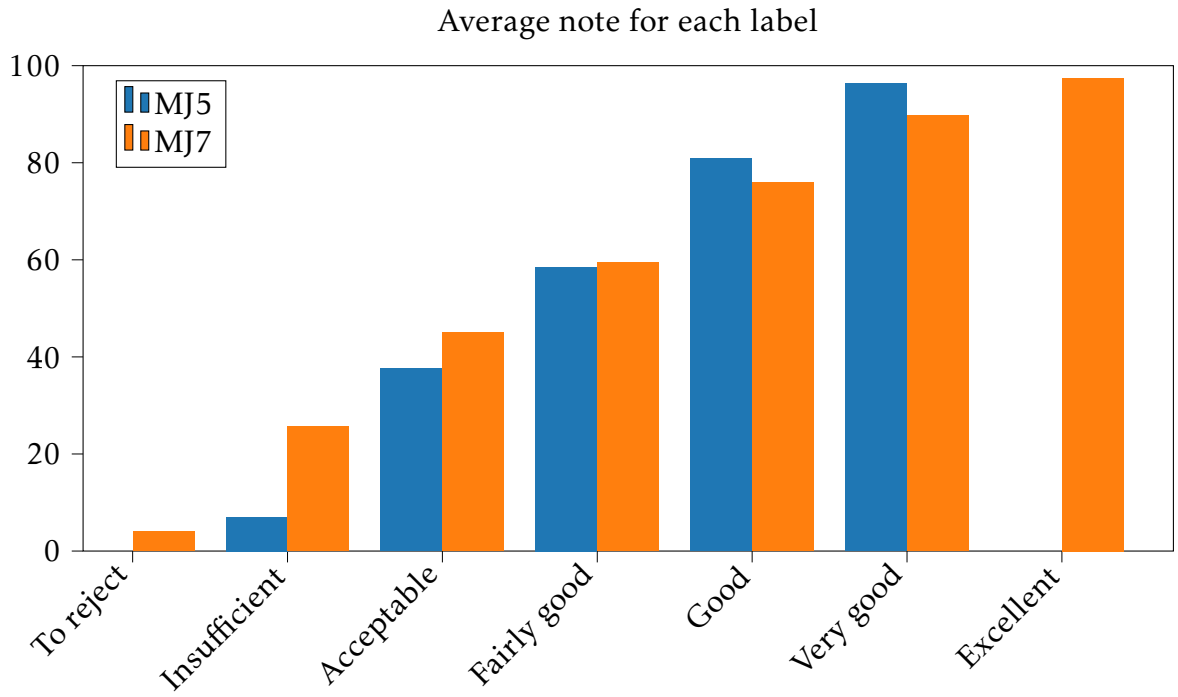
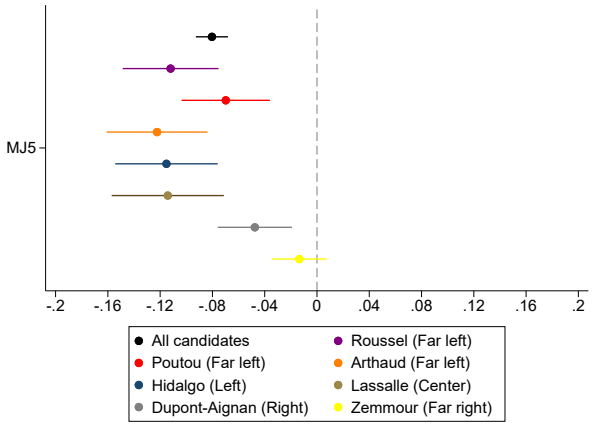
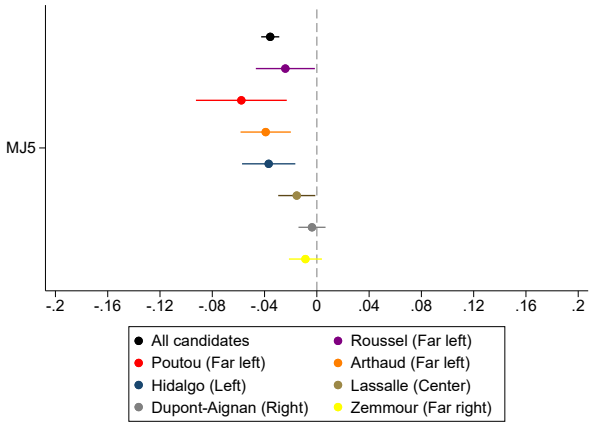


Figure B.0.5: Treatment difference between MJ7 (Baseline) and MJ5 - Other candidates

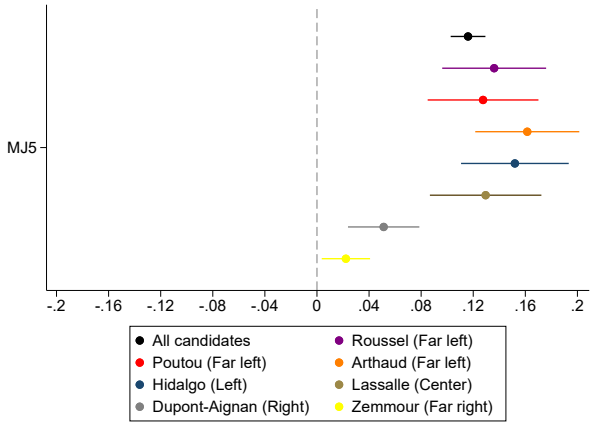
(a) Low grades ("To reject" or "Insufficient")



(b) High grades ("Very good" or "Excellent")



(c) Intermediate grades ("Acceptable", "Fairly good", or "Good")



Notes: Dots with horizontal lines indicate point estimates of the coefficient on the MJ5 treatment with 95% confidence intervals from linear least squares regression. Controls include gender, age, education, and political preferences. $N = 23,460$ in the regression with all candidates, $N = 1,955$ in the regression with the single candidates. Standard errors clustered at the individual level in the regression with all candidates. When considering the vote for individual candidates, we use robust standard errors.

C Robustness checks

Table C.0.1: Regression analysis - Robustness checks

	Weights:Demographics			Weights:Political pref.			Probit		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Low grades	High grades	Intermediate grades	Low grades	High grades	Intermediate grades	Low grades	High grades	Intermediate grades
MJ5	-0.088***	-0.034***	0.12***	-0.067***	-0.021***	0.088***	-0.20***	-0.20***	0.30***
	(0.03)	(0.01)	(0.03)	(0.03)	(0.006)	(0.03)	(0.02)	(0.02)	(0.02)
Constant	0.55***	0.15***	0.29***	0.49***	0.11***	0.40***	0.096**	-1.08***	-0.46***
	(0.03)	(0.01)	(0.03)	(0.04)	(0.01)	(0.04)	(0.04)	(0.05)	(0.05)
Dep. var. mean in MJ7	0.53	0.12	0.34	0.53	0.53	0.53	0.53	0.12	0.34
Observations	23460	23460	23460	22728	22728	22728	23460	23460	23460
Clusters	1955	1955	1955	1894	1894	1894	1955	1955	1955
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
R-squared	0.027	0.022	0.036	0.013	0.029	0.015			

Notes: The dependent variable is a dummy for whether participants gave to a candidate "To reject" or "Insufficient" (columns 1,4,7), "Very good" or "Excellent" (columns 2,5,8), and "Acceptable", "Fairly good", and "Good" (columns 3,6,9). In columns (1-3), we build post-stratification weights accounting for the gender, age, and education differences between our sample and the French general population. In columns (4-6), the weights are built to account for differences in participants' vote and the official election outcome. Specifically, we build four relevant categories, left, center, right, and abstained to have a correspondence between our participants and the election results. Additionally, we exclude 61 participants unsure about whom to vote. We control for participants' gender, age, education level, and political preferences. Standard errors clustered at the participant level in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table C.0.2: Regression analysis - Randomization inference

	(1)	(2)	(3)
	Low grades	High grades	Intermediate grades
MJ5	-0.080	-0.035	0.116
	(0.000)	(0.000)	(0.000)
	[0.000]	[0.000]	[0.000]
Constant	0.531	0.142	0.327
	(0.000)	(0.000)	(0.000)
	[0.000]	[0.000]	[0.000]
Dep. var. mean in MJ7	0.34	0.34	0.34
Observations	22932	22932	22932
Controls	✓	✓	✓
R-squared	0.01	0.01	0.02

Notes: The dependent variable is a dummy for whether participants gave to a candidate "To reject" or "Insufficient" (columns 1), "Very good" or "Excellent" (columns 2), and "Acceptable", "Fairly good", and "Good" (columns 3). Original p-values are in parentheses, and randomization inference p-values are in squared brackets. The latter p-values are computed by resampling using 1,000 iterations. We control for participants' gender, age, education level, and political preferences. Standard errors clustered at the participant level in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table C.0.3: Regression analysis - Controlling for rules used before MJ

	Low grades	High grades	Intermediate grades
	(1)	(2)	(3)
MJ5	-0.080***	-0.036***	0.12***
	(0.006)	(0.004)	(0.007)
Evaluative rules (base: AV and EV(0,1,2,3))			
AV and EV(-1,0,1,2)	0.00062	-0.00056	-0.000059
	(0.010)	(0.006)	(0.01)
EV(-1,0,1)	0.016	-0.0065	-0.0092
	(0.01)	(0.006)	(0.01)
EV(-1,0,1) and EV(0,1,2)	-0.0060	0.00029	0.0057
	(0.010)	(0.006)	(0.01)
EV(-1,0,1,2) and EV(0,1,2,3)	0.00037	-0.00039	0.000017
	(0.01)	(0.006)	(0.01)
Ranking rules (base:Borda)			
IRV(rank $\geq$ 1)	-0.0017	-0.0034	0.0051
	(0.008)	(0.004)	(0.008)
IRV(rank $\geq$ 4)	-0.0040	0.0053	-0.0013
	(0.008)	(0.004)	(0.008)
Constant	0.54***	0.14***	0.32***
	(0.02)	(0.010)	(0.02)
Dep. var. mean in MJ7	0.53	0.12	0.34
Observations	23460	23460	23460
Clusters	1955	1955	1955
Controls	✓	✓	✓
R-squared	0.011	0.011	0.015

Notes: The dependent variable is a dummy for whether participants gave to a candidate "To reject" or "Insufficient" (columns 1), "Very good" or "Excellent" (columns 2), and Acceptable, Fairly good, and Good (columns 3). The category of *Evaluative rules* and *Ranking rules* include the four pairs of rules and of three ranking rules that participants could try before votin under MJ. Controls include participants' gender, age, education level, and political preferences. Standard errors clustered at the participant level in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.