Social Choice and Welfare (2021) 56:803–834
https://doi.org/10.1007/s00355-020-01300-z

ORIGINAL PAPER

Some regrettable grading scale effects under different versions of evaluative voting

Antoinette Baujard[1] · Herrade Igersheim[2] · Isabelle Lebon[3]

Received: 18 December 2019 / Accepted: 2 November 2020 / Published online: 27 November 2020

Abstract
Many voters seem to appreciate the greater freedom of expression afforded by alternative voting rules; in evaluative voting, for example, longer grading scales and/or negative grades seem desirable in so far as, all other things being equal, they allow greater expressivity. The paper studies to what extent the behavior of voters, and the outcomes of elections, are sensitive to the grading scale employed in evaluative (or "range") voting. To this end, we use voting data from an experiment conducted in parallel with the 2017 French presidential election, which aimed to scrutinize the negative grade effect and the length effect in grading scales. First, this paper confirms that the introduction of a negative grade disfavors "polarizing" candidates, those whose political discourse provokes divisive debate, but more generally we establish that it disfavors major candidates and favors minor candidates. Second, under non-negative scales, polarizing candidates may be relatively disfavored by longer scales, especially compared with candidates who attract only infrequent media coverage and who are little known among voters. Third, longer scales assign different weights to the votes of otherwise equal voters, depending on their propensity to vote strategically. Overall, we observe that the benefits of the expressivity provided by longer scales or negative grades need to be balanced against the controversial advantage these give to minor candidates, and their tendency to undermine the principle that each vote should count equally in the outcome of the election.

1 Introduction

Voting rules are defined by the combination of a balloting system and a system for aggregating votes. In multi-nominal voting rules, voters are able to independently evaluate every candidate, but various different balloting systems—i.e., ways to express individual preferences—can be used in that evaluation; these include "Yes or no", "approval" or "non-approval", numbers on a given grading

✉ Antoinette Baujard
antoinette.baujard@univ-st-etienne.fr

Extended author information available on the last page of the article

scale, and expressions such as “excellent, very good, good, passable, mediocre, inadequate”, or “excellent, good, fair, poor”. There is an open question concerning the relative desirability of different balloting systems, and above all what criteria should be employed in selecting from among the various available devices. It has been argued that, ceteris paribus, voters derive greater satisfaction from voting rules which allow greater expressivity (e.g., Baujard and Igersheim 2007); but then the claim that more expressive grading scales are more desirable relies crucially on that “ceteris paribus” clause, and it needs to be shown that all other things can indeed be held equal even while the grading scale changes. Another argument is that offering more balloting options is likely to disfavor “polarizing candidates,” i.e., those whose speech creates strong divisions in the electorate, and who are associated with populism and corruption (see, for example, the arguments in favor of the Janacek method from the Institute H21); but this bias remains to be formulated precisely and confirmed.

The present paper tackles this issue from an empirical point of view, drawing on a specifically designed field experiment on the use of evaluative voting. Evaluative voting is a single-winner voting method that asks each voter to evaluate each candidate on a pre-determined numerical scale, and elects the one who receives the largest total. It is also known as utilitarian voting, range voting, score voting, or point voting. Such an additive method seems very simple and natural to use in many settings; as an aggregation rule, it satisfies the normative properties of utilitarian aggregation; in particular, it is immune to the logical difficulties or “paradoxes” associated with the aggregation of qualitative preferences that do not admit interpersonal comparisons of utilities (Arrow 1951).

A better understanding of evaluative voting is necessary for many reasons. First, as a summation mechanism, this rule is very widely used—for example, to compute average grades in schools or average scores in sports. Second, a number of official political elections employ such a multi-nominal system, i.e., allowing voters to assess every candidate. Such an evaluation is not possible under the more traditional families of single-name voting rules in which just one candidate is selected (the first-past-the-post or two-round system). Many European systems employ multinominal rules (Farrell 2001), such as the municipal elections in France and certain Swiss elections (Lachat et al. 2017). Third, evaluative voting rules enable voters to nuance their support depending on the scale they are dealing with. This second characteristic of evaluative voting rules is evident for instance in Latvia (Laslier et al. 2015), or in Germany and Luxembourg where “cumulative voting” systems are employed. Finally, it has been shown that voters appreciate the possibility for expression afforded by evaluative rules, which extends further than mono-nominal voting rules, or even approval voting (Baujard and Igersheim 2007). Voters prefer relatively longer scales, and they prefer scales with negative and positive grades as compared to scales with positive grades only. One might suppose that, from among a range of different scales, evaluative voting rules differ only as a result of a direct or indirect labeling effect: the results or relative scores of candidates vary due to the specific ways in which the labels of grades are interpreted by voters (Baujard et al. 2018). Up to now, the state of the art has thus focused on the labeling effect, suggesting that length would only count through its indirect consequences on the labels of grades, and protocols thus far have not been able to go beyond this vague assertion.

In this paper we focus on balloting systems that use the evaluative voting rule, and ask the following questions: Is a longer and more fine-grained scale preferable to a shorter one? Should the scale offer negative grades or not? We thus shed light on the effects induced by negative grades and scale length under evaluative voting. First, beyond the confirmation that negative grades disfavor polarizing candidates, the findings are quite unexpected: a striking feature is that the introduction of negative grades essentially favors minor candidates, and only indirectly disfavors polarizing candidates. Secondly, although at first sight it seems that length broadly speaking does not matter, the data support that a length effect might exist for specific types of candidates, and would indeed be significant under other circumstances. We find that longer scales favor minor, unknown candidates. Thirdly, the greater potential expressivity induces different behaviors by different voters, such that certain voters are now likely to depart more substantially from strategic behavior, and as a consequence their vote will count for less in the outcome of the election. Although our paper will not develop this issue, these heterogeneous behaviors may be induced by voters' specific psychological traits: while some will continue to reason and act strategically, some others will take these alternative voting methods as an opportunity to express more fully their real political preferences. In fact these different ways of approaching the act of voting have already been addressed in the literature devoted to the voting paradox, though this work has not focused on evaluative voting rules in particular (see in particular Schram and Sonnemans 1996a, b; Blais and Young 1999; Grosser and Schram 2006; Feddersen et al. 2009; Dittman et al. 2014). These three features, brought out distinctively in this paper, might be considered reasons to reject longer scales with negative grades, contrary to the layman's appreciation of more expressive grading scales.

Section 2 describes the experimental protocol and our data. Section 3 introduces the theoretical notions on which the data analysis rests. Section 4 presents the observed consequences of grading scales of different lengths with or without negative grades, in order to disentangle the length and labeling effects. Section 5 presents the observed consequences of varying the length of the scale when all offered grades are non-negative numbers. Section 6 concludes.

2 Experimental design for comparing grading scales

During the first round of the 2017 French presidential election, voters from four polling stations were invited to take part in an experiment which took place immediately after they had voted in the official ballot. One of the objectives of this experiment was to study the features of evaluative voting, and more precisely the possible effects that the chosen grading scale can have on the behavior of the voters and thus on the result of the election. Two dimensions of the voting scale were studied in particular: the length of the scale, and the presence or absence of negative grades.

The voting rules tested are the following.[1]

- *Approval Voting (AV)*. Under approval voting, for each candidate the voter indicates whether she is in favor of the election of the candidate. The candidate whose election is approved by the largest number of voters is elected. Note that AV is formally equivalent to range voting with the binary scale $\{0, 1\}$. A non-response is counted as a 0.
- *Evaluative voting (EV)* with different grading scales:
  - *EV3*. Evaluative voting with the positive grading scale $\{0, 1, 2\}$. A non-response is counted as a 0.
  - *EV3neg*. Evaluative voting with the positive and negative grading scale $\{-1, 0, 1\}$. A non-response is counted as a -1.
  - *EV4*. Evaluative voting with the positive grading scale $\{0, 1, 2, 3\}$. A non-response is counted as a 0.
  - *EV4neg*. Evaluative voting with the positive and negative grading scale {-1,0,1,2}. A non-response is counted as a -1.
  - *EV6*. Evaluative voting with the positive grading scale $\{0, 1, 2, 3, 4, 5\}$. A non-response is counted as a 0.

The organization of the experimental votes essentially replicates the protocol described in Laslier et al. (2002); it is similar to that of the official vote and, in particular, it guarantees the anonymity of the answers. Although replication of previous results is of genuine interest, we should however highlight two major differences with the protocols used previously, which allow us to explore novel questions.

First, in the previous protocol, and in particular that used by Baujard et al. (2018) for the 2012 French presidential election, non-response grades are always equal to zero, even in the presence of a negative grade, in order to capture the layman's intuition about a neutral grade. As a consequence, any change of score in switching from scales with non-negative grades to scales with negative grades could be due to non-responses by the voters. In particular, the average grade of unknown candidates likely to attract many non-responses must automatically increase. It was therefore hard to disentangle the actual negative grade effect on scores from the simple consequence of non-response. On the contrary, in the 2017 protocol, non-responses were recorded as the lowest grades, such that this issue cannot arise. Any increase of scores of the non-covered (or unknown) candidates hence could not be due to the non-response rule, but to the voters' desire to give them a grade higher than the lowest grade. The negative grade effect can thus be studied more closely. Let us add to this that the non-response rule was clearly expressed, written on the ballots and voters were reminded of it by volunteers. The data confirmed the voters did indeed understand the rule: a substantial number of them never used the lowest grade and only focused on higher grades—since, indeed, to do otherwise would

[1] Note that not rating a candidate is systematically considered as giving the lowest grade, 0 or -1 depending on the chosen scale, and the participants were clearly notified about this point.

Table 1 Experimental design for range voting scales

	Voters	Participation (%)	Sample	Scales: AV and...
Hérouville	1180	56	661	EV4, EV6
Strasbourg	1874	54	1016	EV3, EV3neg, EV4, EV4neg

involve uselessly ticking a box, whether the lowest or median grade, whereas the same purpose could be served by doing nothing.[2]

Second, the standard protocol used in 2007 and in 2012 implied that samples gathered in different polling stations were not directly comparable: as voters using different voting rules came from different cities, it was difficult to disentangle the fact that voters belonged to distinct sociological backgrounds, from how different rules were actually used by every voter. In order to handle this sample bias and increase the internal validity of the experiment, the 2017 protocol was adapted to propose different rules to a given sample of voters. While maintaining the conditions of anonymity and free participation, it proposes a randomized choice of different ballots in given voting stations. A paper ballot allowed every voter to vote experimentally twice: they were offered approval voting and another modality of evaluative voting; and in addition every voter would be asked to set down on the ballot paper their first-round vote done with the official rule. The randomization in each site allows us to study the effects of the variations of the grading scales. We are then able to study each of the issues independently in the given experimental polling stations: the scale length effect for positive grades on the one hand, and the negative grade effect under different lengths on the other.

The external validity of this new protocol might be debated, in as far as each result is generated by observations from one location only. To address this issue, we have chosen to apply comparable treatments in different locations that are sufficiently remote from each other. In as much as we eventually obtain consistently similar results in the different locations, we believe that external validity should not be considered as a serious issue.

The experimental design was as follows. We worked in four polling stations: two in Hérouville-Saint-Clair (Normandy—now called Hérouville) and two in Strasbourg (Alsace). In some stations, we randomly allocated to each participant one experimental ballot with positive grading scales drawn from among ballots of different lengths; in other stations, we randomly allocated one experimental ballot drawn from among ballots with or without negative grades and under different lengths. Technically, all participants tested approval voting and one form of evaluative voting, randomly chosen according to the design described in Table 1.

[2] We can assert with certainty that the voters had assimilated the rule that the default grade was the lowest one: when they do not give a grade, they positively intended to attribute the lowest one and this is a conscious choice. As evidence for this, if we compute the impact of the change of default rule from lowest to median grade, the results of all the candidates arousing rejection attitudes would increase. The score of Marine Le Pen, who arouses strong opinions and no indifference, would increase dramatically.

It should further be stressed that no direct political interpretation of the results is possible, since these four polling stations are not representative of the national vote and, moreover, the samples resulting from the voluntary participation of the voters in the experiment do not match the distribution of the official votes in these polling stations. These biases can be corrected, however, so we can study how different outcomes are induced by different rules. Appendix 7.2 sets out the results corrected in accordance with the method proposed by Baujard et al. (2014) for these voting rules. However, we will not interpret these different outcomes in the paper. As the aim of the paper is not to draw political conclusions, it does not need to rely on corrected outcomes. Hence we focus on the comparison of voting behavior according to different grading scales, for which the raw data are reliable and undebatable.

3 Theoretical background

3.1 The representation issue

The theory underlying evaluative voting has been developed only quite recently. It focuses mainly on studies of axiomatic characterizations of evaluative voting—see, for instance, Smaoui and Lepelley (2013), Pivato (2013), Macé (2018)—or similar rules—such as Aleskerov et al. (2007), Gaertner and Xu (2012), Alcantud and Laruelle (2014), or Gonzalez et al. (2019). But rather than characterizing the general properties of evaluative voting, this paper aims at studying how preferences are expressed in individual ballots. This representation problem is not yet well developed in the particular case of evaluative voting (with the notable exception Ceron and Gonzalez 2019), such that one must still refer to the general theory of measurement.

Discussing how strategic or sincere voting occurs in evaluative voting allows us to introduce the issue of the translation of preferences into their expression in balloting. The theoretical prediction of strategic behavior under evaluative voting is straightforward: a strategic voter would only use extreme grades. This corresponds to the intuition of “maximizing the effect of my vote,” and it is no surprise that this is also the outcome of the game-theoretical analysis (Núñez and Laslier 2014). However, the empirical analysis of evaluative voting in the context of real elections contradicts such theoretical predictions (Igersheim et al. 2016). It has been shown that voters do use intermediate grades in real contexts whenever the evaluative grading scales allow them to, i.e., for any kind of evaluative voting (except approval voting, for obvious reasons). This use of the intermediate grades clearly corresponds to a sincere vote, and further investigation of this behavior is needed in order to elicit alternative predictions.

The core of the concept of sincere voting is that there exists, in the voter’s mind, some underlying genuine evaluation of the candidates, and that under each rule the voter chooses to express a vote that matches this reference as far as possible. Under single-name voting, the consequence is obvious: voters vote for their preferred candidate. But under evaluative voting the consequences are not so clear, because it is difficult to know which grades best reflect voters’ opinions: this raises a problem of *representation* in the sense of the theory of measurement (Narens 1985). For evaluative voting, this issue has been termed the *calibration* problem by Baujard et al. (2018).

The calibration problem can be described using the technical apparatus of relative utilitarianism (Dhillon and Mertens 1997). Voter i has a utility function that associates a real number $u_i(c)$ to each candidate c. Given the (finite) set of candidates, the maximum and minimum utilities are denoted $u_i^{\max}$ and $u_i^{\min}$. The utilities are linearly transformed so that the worst candidate is evaluated at 0 and the best one at 1. This yields the "relative utilitarian" evaluation:

$$v_i(c) = 1 - \frac{u_i^{\max} - u_i(c)}{u_i^{\max} - u_i^{\min}} \in [0, 1].$$

Now, given a finite set of possible grades, the sincere voter may reason as follows: she will use the extreme grades for her best and worst candidates and will accommodate the rating of the others taking into account the intensity of her preferences and the coarseness of the scale.

But the notion of sincerity embodied in the previous proposal may seem too relative. For instance, a voter who thinks that all candidates are bad ones might not wish to give the best grade to any of them. A simple variant, then, is to imagine that, on top of the existing candidates, the voters postulate two additional hypothetical candidates—one absolutely good and one absolutely bad—and calibrate linearly as before.

These ideas are sufficient to provide non-trivial implications that are testable with our experimental design. For instance, linear calibration implies that, when comparing the scales 0, 1, 2, 3 and 0, 1, 2, 3, 4, 5 on two samples of the same population, the observed fraction of 0s and 1s in the first will be statistically identical to the fraction of 0s, 1s, and 2s in the second. This is one of the basic methodologies we implement for analyzing our data.

3.2 Classification of candidates, state of the art

Following Baujard et al. (2014) and Darmann et al. (2017), in drawing distinctions between the election outcomes for different rules or different versions of one rule, we suppose that assertions such as "this type of candidate is favored or unfavored" make sense since it allows us to refer to general categories of candidates independently of their political color.

The first distinction concerns the importance of candidates in this election. We can describe those who have reasonable chances to win the election as the 'main' or 'major' candidates. They are called 'viable' by Cox (1997), or 'serious' by Myerson (2002). These concepts should obviously be adapted to the voting rule in force. In the case of evaluative votes, the definition of viable candidates must take into account the fact that the result of the vote is not likely to be affected by the existence of close or clone candidates. The set of viable candidates can therefore only be widened compared to the one defined in official single-name voting systems. The non-viable candidates, i.e., those who have no chance of being elected, will be called minor candidates. The viable/non-viable candidates distinction could have influenced the behavior of voters. In the end, being sincere towards a non-viable candidate is inconsequential. As such a candidate is not likely to be in first position, the grade assigned to him/her by a voter does not influence the outcome of the election. Conversely, strategic considerations are likely to be significant and relevant for viable candidates.

To complete the typology, Baujard et al. (2014) distinguish the "exclusive and inclusive" candidates among the major candidates in a paper which analyzed the 2012 French presidential election, as follows: "candidates who arouse strong feelings, whether positive or negative, among voters, are called 'exclusive' candidates; while candidates who are liked by a large number of voters, but not necessarily strongly liked, nor in a way that excludes support for others, are called 'inclusive' candidates". In this paper, other candidates, called 'the little candidates', were described as 'unknown' because of their scant media coverage.

In the same vein, based on an exit poll conducted after the 2015 parliamentary election in the Austrian federal state of Styria, Darmann et al. (2017) analyze political preferences data and derive a fourfold classification of parties. They introduce the category of 'unpopular' parties as those who "have a strong support from only a small group and are seen negatively by a large proportion of society". In addition, they distinguish 'polarizing', 'popular', and 'medium' parties. Polarizing parties "get both strong support from a certain, significantly large, part of society as well as strong negative support from another, significantly large, proportion of society". Popular parties "have a strong support from a specific segment of society, and are seen positively by a large proportion of society". The medium parties "are acceptable to a large proportion of society and induce strong (positive or negative) views in only small groups".

Notice that these classifications require measurement not only of the intensity of support for the different candidates or the different parties, but also of the extent of their electoral basis. This implies being able to link the experimental observations to the national representation of the candidates. Considering such categories would then require that we first correct the raw data for participation bias, since voters come to the experiment on a free and voluntary basis and are therefore not politically representative of the set of official voters. Yet although we would be able to perform such a correction (Baujard et al. 2014), we don't want to, since such correction can also induce other biases by artificially overweighting specific behaviors. The 2017 experiment was thus designed especially to study the effect of grading scales independently of the political dimension of the election, as described above in the experimental design. Therefore the classification of candidates must be presented differently.

3.3 Classification of candidates, novel definitions

Without relying on corrected data, we may assess the intensities of support and rejection expressed by voters by comparing the outcomes from approval voting and other evaluative voting rules. Such a comparison allows us to establish the following principle: strong rejection of a candidate supposes that a non-approval translates into the lowest grade, and strong support supposes that an approval becomes the highest grade. The systematic differences in the use of intermediate grades in the case of approval on the one hand, and in the case of non-approval on the other, may eventually allow us to refine the classification of candidates. Although there could be peculiarities for each grading scale, we still observe some regularities.

However, as noted above, voter behavior may also depend on whether the candidates are viable or not, the consequences of voters' choices not being the same in the one case as opposed to the other. In the 2017 French election, four candidates might be called viable candidates under the standard single-name rule: F. Fillon, M. Le Pen, E. Macron, and J.-L. Mélenchon. Notably, based on our knowledge of the context of the election, we should consider that B. Hamon, whose score suffered from the proximity with J.-L. Mélenchon and E. Macron, is among the viable candidates in a multi-nominal rule. Five candidates belong thus to the category of major candidates: F. Fillon (FF), B. Hamon (BH), M. Le Pen (MLP), E. Macron (EM), and J.-L. Mélenchon (JLM). The 2017 French election also included six other candidates which we can consider as minor: N. Arthaud (NA), F. Asselineau (FA), J. Cheminade (JC), N. Dupont-Aignan (NDA), J. Lassalle (JL), and P. Poutou (PP).

The major candidates are divided into two groups, depending on their good or bad evaluation by voters who do not approve them, for which we adapt the notions of popular candidates and polarizing candidates. When they are non-approved, the popular candidates, B. Hamon, E. Macron, and J.-L. Mélenchon, attract a significant proportion of intermediate grades, because they are considered positively or as acceptable by most voters, including those who do not support them. On the contrary, the polarizing candidates, F. Fillon and M. Le Pen, are strongly rejected by those who do not approve them, these voters almost invariably giving them the lowest grade, whatever it is. In other words, they attract few intermediate grades. Contrary to the definitions of Baujard et al. (2014) and Darmann et al. (2017), these two types of candidates are not distinguished by the intensity of their electors' support, hence the following novel definitions for two sub-categories of major candidates: first, the popular candidates are the ones who attract positive feelings, including from voters who do not approve them; second, the polarizing candidates are the ones who are sharply rejected by those who do not approve them.

Among minor candidates, we also need to add a further distinction. Because they are unknown by voters, some minor candidates cannot be seen either positively or negatively, for lack of knowledge from the voters: one reason for this is that they are covered by the national media only in periods of electoral campaigns, and are barely part of the public debate in other periods. Let us call these candidates the "non-covered candidates". F. Asselineau, J. Cheminade, and J. Lassalle may be called "non-covered".

In contrast, the political lines of unpopular candidates are sufficiently well known by voters for them to form an opinion. We shall be able to check whether such candidates benefit from strong support by certain voters and whether they are seen negatively by the majority of others. Among the minor candidates, N. Arthaud, N. Dupont-Aignan, and P. Poutou may be described as "unpopular".

A description of the candidates with respect to the political frame of the election is provided in Appendix 7.1.

4 Negative grades matter for certain types of candidates

This section focuses on the negative grade effect on the basis of the Strasbourg data. The Strasbourg protocol was specifically designed to compare scales of identical lengths with or without negative grades. Three- and four-step scales are tested.

As defined in the protocol described above, in Strasbourg the voters were randomly assigned one ballot and could vote twice. All the participants tested approval voting plus another kind of evaluative voting using one of the four grading scales: {-1,0,1} (EV3neg), {0,1,2} (EV3), {-1,0,1,2} (EV4neg), and {0,1,2,3} (EV4). We obtained respectively 247, 251, 236, and 282 experimental ballots. From these data, we can study the specific effects of negative grades when comparing EV3 and EV3neg on the one hand, and EV4 and EV4neg on the other.

4.1 Negative grades favor minor candidates: comparisons of scores

Theoretically, for a given scale length, voters should not be sensitive to the specification of grades. This means that, for each candidate, the distribution of the grades would simply be translated by one unit when comparing EV3 with EV3neg, and EV4 with EV4neg.

Identical behavior of voters with EV3 and EV3neg on the one hand, with EV4 and EV4neg on the other, should result in similar scores for each of the candidates once corrected by the 1-point shift on the grades. The first hypothesis to be tested regarding negative grades is therefore the following:

Hypothesis 1 (*AN1*) For each candidate, the score is translated by one unit when comparing EV3 with EV3neg, and EV4 with EV4neg, which neutralizes a possible effect of the length of the scale.

The scores obtained by the candidates in EV3neg and EV4neg presented in Table 2 are increased by one. These are normalized scores.

For the quasi-uninamity of minor candidates, which are those located at the bottom of the table, the Student's T-tests show that the AN1 hypothesis must be rejected with a level of significance of 5% or even 1%. For the 3-step scales EV3 and EV3neg as well as for the 4-step scales EV4 and EV4neg, the introduction of the negative grade significantly increases the normalized scores of all the minor candidates except for N. Dupont-Aignan under EV4 and EV4neg.

For the viable candidates, the results are different. In both cases, EV3 versus EV3neg and EV4 versus EV4neg, the scores of the two polarizing candidates (M. Le Pen and F. Fillon) are not significantly modified by the introduction of the negative grade (with the exception of F. Fillon under EV3 and EVneg, but with a level of significance of 10%): the label of the grading scale does not seem to

Table 2 Comparison of scores in scales with and without negative grades (Strasbourg data)

Candidates	EV3	EV3neg normalized	T-Test EV3-EV3neg	EV4	EV4neg normalized	T-Test EV4-EV4neg
EM	1.01	1.11	− 1.324 (0.186)	1.37	1.40	− 0.169 (0.866)
MLP	0.15	0.18	− 0.695 (0.487)	0.30	0.23	0.996 (0.320)
FF	0.31	0.42	− 1.829 (0.068)*	0.55	0.55	0.137 (0.891)
JLM	1.22	1.21	0.116 (0.908)	1.57	1.79	− 2.094 (0.037)**
BH	1.09	1.28	− 2.685 (0.008)***	1.48	1.74	− 2.523 (0.012)**
NDA	0.27	0.52	$-4.461(1.03\times10^{-}5)$ ***	0.44	0.54	− 1.429 (0.154)
JL	0.30	0.44	− 2.625 (0.009)***	0.45	0.62	− 2.518 (0.012)**
PP	0.57	0.74	− 2.705 (0.007)***	0.86	1.07	− 2.458 (0.014)**
FA	0.20	0.36	− 3.482 (0.001)***	0.26	0.48	$-3.804(1.612\times10^{-}5)$ ***
NA	0.47	0.63	$-5.459(8.055\times10^{-}8)$ ***	0.62	0.91	$-3.958(8.61\times10^{-}5)$ ***
JC	0.14	0.37	− 2.648 (0.008)***	0.25	0.49	$-4.513(8.267\times10^{-}6)$ ***

The table displays the normalized scores for EV3, EV3neg, EV4, and EV4neg as well as the t statistics when one compares EV3 and EV3neg on the one hand, and EV4 and EV4neg on the other. The p-values are in parentheses. The symbol "*" (resp. "**" and "***") means that the hypothesis H_0 is rejected at the level of significance of 10% (resp. 5% and 1%)

have a real influence on the expression of opinion by voters regarding them. On the other side, the case of popular candidates (E. Macron, J.-L. Mélenchon, and B. Hamon) is not uniform. The scores of all these candidates increase when the scales begin with −1 rather than 0. However, these increases are not always sufficient to be significant, with only those of B. Hamon and J.-L. Mélenchon (for EV4 to EV4neg) sufficient to be considered as such.

Overall, the data suggest that the presence of negative grades favors minor candidates, but does not make any significant difference to the scores of polarizing candidates. Meanwhile, the increase in the scores of the popular candidates is weaker than that of the minor candidates and sometimes too weak to be taken into account. Thus an improvement of the relative position of the minor candidates compared to all the major ones necessarily follows.

4.2 Negative grades especially favor non-covered candidates: use of extreme grades

The analysis of voters' behavior regarding approved and non-approved candidates should improve our understanding of these phenomena. Our data shed light on the proportion of lowest or highest grades with negative or positive scales. First, one can consider the proportion of lowest grades assigned to each candidate when he/she is not approved. A priori, one might expect that this proportion would remain unchanged regardless of the candidate. So let us test the following hypothesis:

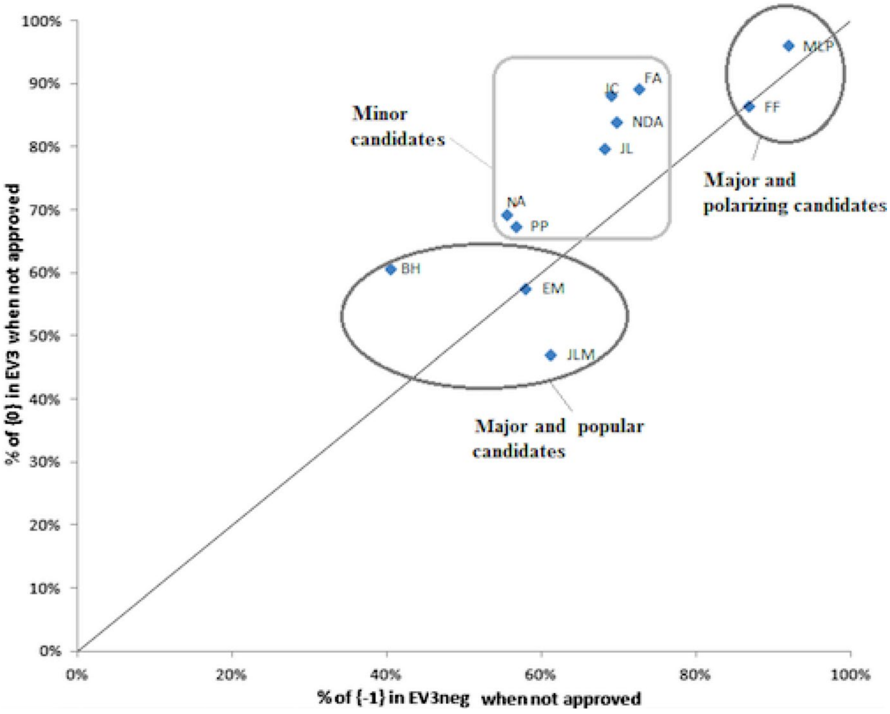


Fig. 1 Use of lowest grade for non-approved candidates in 3-step scales

Hypothesis 2 (*AN2*) Up to a 1-point translation, the different scales with and without negative grades generate the same proportion of lowest grades for each candidate when he/she is not approved.

Symmetrically, we consider the proportion of the highest grades for each candidate when he/she is approved. We assume again that the label change should not modify voters' behavior. We thus test the following hypothesis:

Hypothesis 3 (*AN3*) Up to a 1-point translation, the different scales with and without negative grades generate the same proportion of highest grades for each candidate when he/she is approved.

The figures 1 and 2 clearly show that whether hypothesis AN2 is validated or rejected depends on the kind of candidates. In the two figures, all the observations regarding minor candidates are situated clearly above the first bisector. This means that the introduction of the negative score −1 clearly modifies the behavior of voters vis-à-vis minor candidates, and this is especially the case for non-covered candidates (see J. Cheminade or F. Asselineau). Even though they do not approve of them, voters give all of them a lower number of −1 under EV3neg

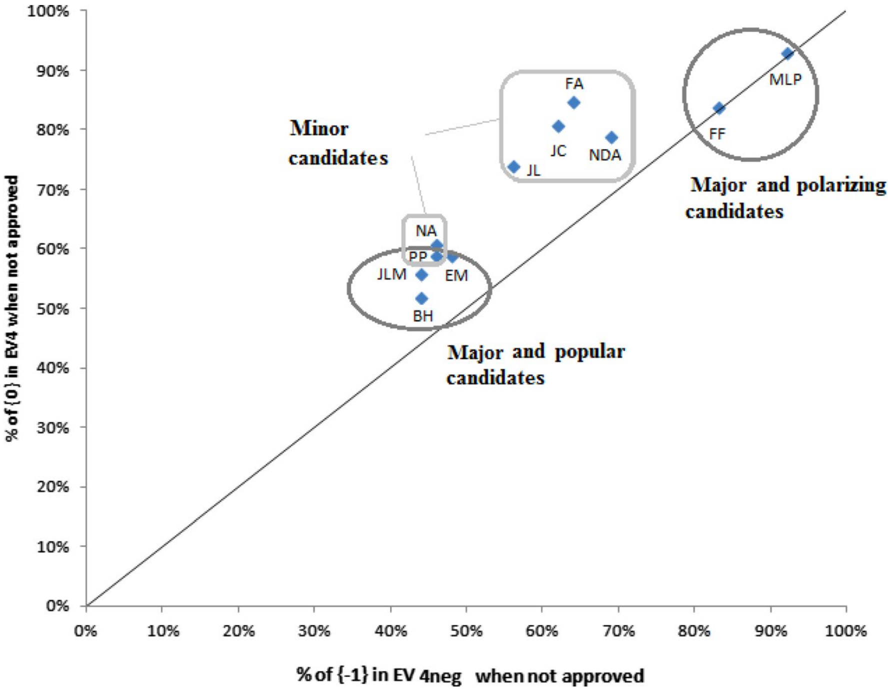


Fig. 2 Use of lowest grade for non-approved candidates in 4-step scales

or EV4neg than 0 under EV3 or EV4. For these minor candidates, it seems that the symbolic dimension of the negative score is crucial. Although voters do not want to support them, they do not want to punish them either; they thus opt for 0 instead of −1 in EV3neg or EV4neg, without taking the risk that they win since they are non-viable candidates. Consequently, hypothesis AN2 must be rejected for them.

The case of polarizing candidates is very different. For M. Le Pen and F. Fillon, the vast majority of voters who do not support them give them the lowest grade, 0 or −1. For polarizing candidates, the hypothesis AN2 is validated. These behaviors clearly explain why the relative scores of the polarizing candidates and the minor candidates are modified by the introduction of the negative grade to the advantage of the latter.

The case of popular candidates is a bit more complex. As we noted previously, they receive relatively few lowest grades from voters who do not support them, even when the lowest grade is 0. When the scale is long enough, we observe a slight decrease of this proportion—already low—from EV4 to EV4neg (see Fig. 2). Note, however, that this decline is more limited than we observed for the minor candidates. With the three-grade scales (EV3 and EV3neg), the effect of the introduction of the negative score for popular candidates is less homogeneous than for minor or polarizing candidates, since the proportion of the lowest grade decreases strongly for B. Hamon, remains stable for E. Macron, and increases for J.-L. Mélenchon. We

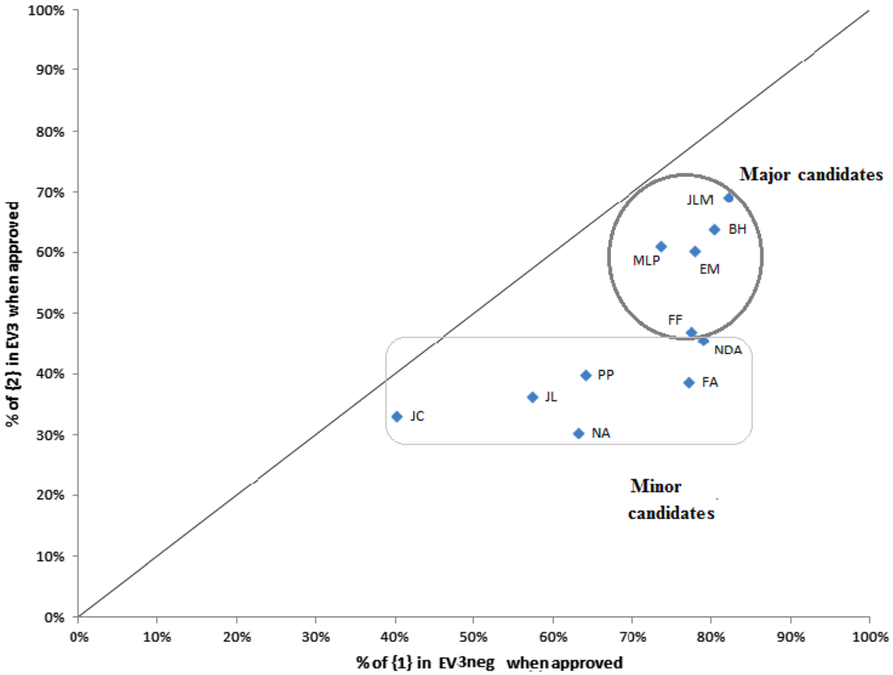


Fig. 3 Use of highest grade for approved candidates in 3-step scales

cannot exclude that this heterogeneity derives simply from our particular dataset, or that it reveals a peculiarity of the perception of each of the three candidates. The first hypothesis, that of a particularity of the sample, is reinforced by the fact that the main characteristic of Fig. 1 for popular candidates is constituted by the position of JLM, but the position of this candidate mainly results from the proportion of 0 with EV3 (around 45%). This proportion is lower than that observed for EV4 (greater than 55%), while the extension of the scale should rather have the opposite effect. Nonetheless, for lack of a clear trend, it is difficult to conclude whether or not the AN2 assumption holds for popular candidates. But for the candidates as a whole, this hypothesis is clearly rejected.

Figures 3 and 4 illustrate the use of the highest grade for candidates when they are approved depending on the presence or absence of a negative grade.

Considering only the 3-step scales EV3 and EV3neg, we conclude that overall the introduction of the negative grade significantly increases the use of the highest grade in the case of approval, irrespective of the candidate concerned. However, the comparison of EV4 and EV4neg in Fig. 4 leads to a reconsideration of this first observation. When one considers the transition from EV4 to EV4neg, the proportion of the highest grade increases for some candidates, while it decreases for others, among both major and minor candidates.

We still need to explain the difference in the use of a negative grade between the two scale lengths. It is simply wrong to believe that the presence of the negative

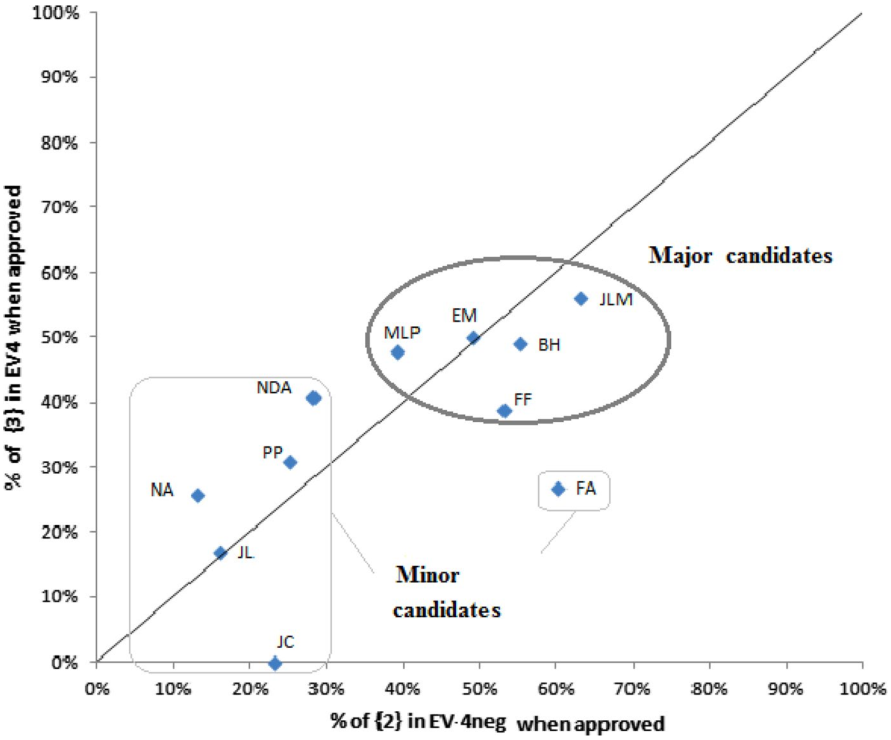


Fig. 4 Use of highest grade for approved candidates in 4-step scales

grade systematically increases the use of the highest grade. In fact, the existence of a single strictly positive grade in {-1,0,1} does encourage voters to use it much more often than 2 in {0,1,2}. But this is no longer the case if one compares the scale { -1,0,1, 2} with {0,1,2,3}. This observation refers to the impact of scale length on the results obtained. What we observe for EV3 and EV3neg leads us to reject hypothesis AN3, which assumed the stability of the proportion of highest grades for scales with the same length when a negative grade is introduced. However, a comparison between EV4 and EV4neg suggests that this effect could disappear with longer scales.

The comparisons between the use of the highest grade also depend on the different types of candidates. Comparing EV4 and EV4neg, we observe that unpopular candidates and non-covered candidates fall on either side of the first bisector; but minor candidates are too rarely approved to be able to derive any conclusive interpretation from this feature. Comparing the four tested scales, the figures show distinctly that, among approved candidates, major candidates more often benefit from highest grades than minor candidates. They also confirm that polarizing candidates do not attract more highest grades than popular candidates. The distinction between polarizing and popular candidates among the major candidates does not concern the intensity of their support but rather the intensity of their rejection.

To summarize, the introduction of a negative grade in a voting scale changes the behavior of voters but in a way that is not homogeneous from one type of candidate to another. The minor candidates benefit the most from the presence of a negative grade, and the non-covered candidates even more so. Since voters do not know very much about the latter, a fortiori they do not strongly reject them, and a significant proportion of voters who do not support them prefer to give them a zero rather than a negative grade. On the other hand, this kind of "no rejection" behavior does not occur for polarizing candidates who are disadvantaged by the introduction of a negative grade. As for the popular candidates, the behavior towards them is so conciliatory that many voters who do not support them (1) avoid giving them 0 on a positive scale, (2) and avoid all the more giving them -1 on a scale with a negative grade. All these observations lead to a possible distortion of the scores and, potentially, of the ranking of the candidates, and the origin of this distortion is mainly to be found in the behavior of voters towards the candidates they do not approve.

5 Evaluative voting is not stable under different scale lengths

This section focuses on the scale length effect on the basis of the Hérouville data. The Hérouville protocol was specifically designed to compare the effects of the range of available grades when there are no negative grades. As defined in the protocol described above, the voters were randomly assigned one paper ballot on which all the participants tested approval voting plus another kind of evaluative voting: either the $\{0, 1, 2, 3\}$ scale (termed EV4, for 53% of the participants) or the $\{0, 1, 2, 3, 4, 5\}$ one (termed EV6). 661 ballots from the Hérouville experiment are usable for this analysis.[3]

5.1 Scale length alone seemingly does not matter: distribution of grades and comparison of scores

The use of even-length scales, AV, EV4, or EV6, enables the partition of the scales into two intervals of identical size. This allows us to analyze voters' behavior when they face positive scales of different lengths. Voters may have sincere or strategic concerns, but in either case a reasonable guess would be that they associate lower grades (in the lower half of the scale) to candidates they do not approve and higher grades to candidates they do approve (in the upper half of the scale). In such case, the distribution of grades into two intervals of same length should remain stable whatever the scale. This leads us to the following assumption.

[3] Note that a similar analysis based on the Strasbourg data under AV vs. EV4 provides similar conclusions as those presented here.

Table 3 Reduction to two classes: grades distribution (Hérouville data)

Candiates	AV (N = 661)		EV4 (N = 350)			EV6 (N = 311)		
	$\{0\}$ (%)	$\{1\}$ (%)	$\{0, 1\}$ (%)	$\{2, 3\}$ (%)	McNemar test	$\{0, 1, 2\}$ (%)	$\{3, 4, 5\}$ (%)	McNemar test
EM	51.59	48.41	52.29	47.71	0.022 (0.883)	49.20	50.80	0.735 (0.391)
MLP	87.90	12.10	86.86	13.14	0.00 (1)	88.10	11.90	0.364 (0.547)
FF	80.79	19.21	80.29	19.71	0.00 (1)	81.35	18.65	0.00 (1)
JLM	46.44	53.56	44.00	56.00	2.703 (0.100)	47.59	52.41	0.543 (0.461)
BH	47.81	52.19	49.71	50.29	0.00 (1)	50.16	49.84	3.12 (0.077)*
NDA	86.99	13.01	87.43	12.57	0.00 (1)	86.82	13.18	0.00 (1)
JL	94.55	5.45	93.14	6.86	1.714 (0.190)	91.96	8.04	1.25 (0.264)
PP	74.74	25.26	80.00	20.00	7.521 (0.006)***	76.21	23.79	0.108 (0.742)
FA	95.31	4.69	96.86	3.14	0.00 (1)	94.21	5.79	0.00 (1)
NA	85.33	14.67	86.86	13.14	2.56 (0.110)	88.10	11.90	0.64 (0.424)
JC	95.01	4.99	96.57	3.43	0.00 (1)	94.53	5.47	1.067 (0.302)
Average	77	23	78	22		77	23	

The table displays the distribution of grades into two classes for AV, EV4, and EV6 as well as the McNemar statistics when one compares the (paired) AV and EV4 on the one hand, and AV and EV6 on the other. The p-values are in parentheses. The symbol "*" (resp. "**" and "***") means that the hypothesis H_0 is rejected at the level of significance of 10% (resp. 5% and 1%)

Hypothesis 4 (*AL1*) The distribution of the grades associated with each candidate remains stable for the scales AV, EV4, and EV6 when reduced linearly to two classes.

The obtained partitioned results are shown in Table 3. As one can see from Table 3, the distributions of the grades in two intervals are the same up to two exceptions when comparing AV, EV4, and EV6.[4] This observation leads us to accept hypothesis AL1.

We now focus on the candidates' scores. If voters' behavior is not biased, lengthening the grading scale should induce a proportional enhancing of scores for every candidate. Let us then consider the following hypothesis:

Hypothesis 5 (*AL2*) The normalized scores of each candidate are unchanged under various scales.

To make them directly comparable, the candidates' scores under the two longer scales are normalized by dividing their initial scores by the maximum grade, namely 3 for EV4 and 5 for EV6. Table 4 shows the normalized scores.

[4] The tests carried out relate to the average frequencies of the grades divided into two segments. Individual comparisons are not possible because no participant used EV4 and EV6 simultaneously.

Table 4 Comparison of scores under positive grading scales (Hérouville data)

Candidates	AV	EV4 normalized	EV6 normalized	T-Test AV-EV4	T-Test AV-EV6	T-Test EV4-EV6
EM	0.48	0.48	0.45	− 0.345 (0.730)	− 1.823 (0.069)	− 0.179 (0.858)
MLP	0.12	0.13	0.11	− 0.307 (0.759)	0.240 (0.811)	0.755 (0.451)
FF	0.19	0.19	0.19	− 0.616 (0.538)	0.171 (0.864)	0.309 (0.757)
JLM	0.54	0.53	0.49	0.00 (1)	− 2.862 (0.004)***	1.239 (0.216)
BH	0.52	0.48	0.48	− 1.633 (0.103)	− 3.060 (0.002)***	− 0.303 (0.762)
NDA	0.13	0.14	0.15	1.317 (0189)	1.372 (0.171)	− 0.179 (0.858)
JL	0.05	0.1	0.12	4.530 (0.00)***	4.496 (0.00)***	− 1.140 (0.255)
PP	0.25	0.24	0.25	− 1.269 (0.205)	0.076 (0.939)	− 0.546 (0.585)
FA	0.05	0.06	0.09	2.586 (0.010)***	2.856 (0.005)***	− 2.325 (0.020)**
NA	0.15	0.18	0.17	1.677 (0.094)*	2.433 (0.016)**	0.409 (0.682)
JC	0.05	0.06	0.08	2.892 (0.004)***	1.205 (0.229)	− 1.993 (0.047)**

The table displays the scores for AV, (normalized) EV4, and (normalized) EV6 as well as the t statistics when one compares AV and EV4, AV and EV6, and EV4 and EV6. The p-values are in parentheses. The symbol "*" (resp. "**" and "***") means that the hypothesis H_0 is rejected at the level of significance of 10% (resp. 5% and 1%)

With the two exceptions, Table 4 makes it clear that the normalized scores of all major candidates are assumed to be equivalent regardless of the length of the scale considered: Hypothesis AL2 is verified for all major candidates in almost every case.

Regarding the minor candidates the results are more nuanced, for several differences in the scores prove to be significant. Lengthening the scale leads to an increase in the normalized scores of the minor non-covered candidates (F. Asselineau, J. Cheminade, and J. Lassalle). On the other hand, the minor unpopular candidates (P. Poutou, N. Dupont-Aignan, and, to a lesser extent, N. Arthaud) do not benefit significantly from the lengthening of the scale. With the scale lengthening, a part of voters may be tempted to allocate points to most unknown candidates, as in a protest vote, without running the risk of actually electing them. As a result, non-covered candidates can benefit from a longer scale: the AL2 hypothesis is thus only partly satisfied for minor candidates.

At first glance, the data lead us to accept hypothesis AL2, consistent with previous results in the literature (Baujard et al. 2018; Laslier 2019). But when we look in detail, we see that we are not able to assert that this hypothesis holds for all candidates. Beside this first exception concerning non-covered candidates, a further look provides a more nuanced view of the length effect, as will be shown in the next section.

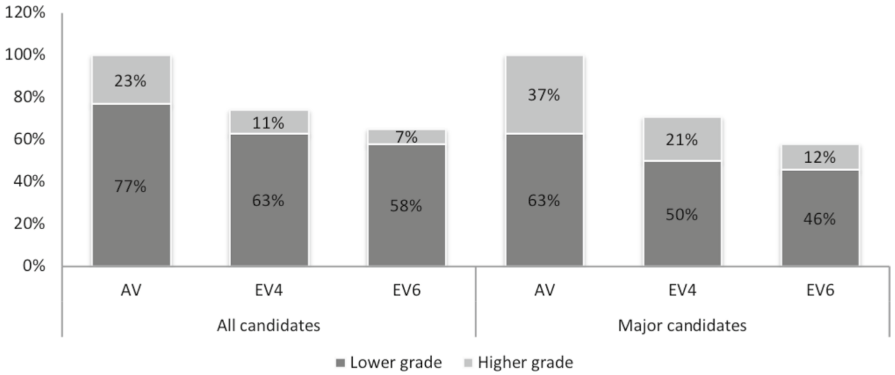


Fig. 5 Use of extreme grades for each scale length

5.2 Scale length matters for minor and popular candidates: use of extreme and intermediate grades

The significant use of the intermediate grades, which we have previously observed, prohibits us from thinking that voters behave in a purely strategic way, not even towards the viable major candidates. An analysis by voting ballot and then by candidate confirms this impression and reveals the consequences of a lengthening in the grading scale.

The proportion of voters who only use extreme grades is indeed very low. Considering the set of all candidates, only 11.7% of the ballots under EV4 only use extreme grades, and this proportion falls to 4.5% under EV6. If one focuses on major candidates only, the proportion of "strategic" ballots increases slightly to 18% and 8.7%, under EV4 and EV6 respectively.

This decline in the share of strategic ballots suggests that we should make a more systematic study of the use of intermediate grades for each of the candidates. The attribution of these scores ought not to be modified by the extension of the scale, which brings us to formulate the following hypothesis:

Hypothesis 6 (*AL3*) The length of the scale does not change the use of the extreme grades, for any candidate.

It seems that the AL3 hypothesis can also be rejected, since the proportion of intermediate grades increases with the extension of the scale (see Figure 5); but for more precision it is necessary to look at the grades received by each candidate in order to understand whether the behavior of the voters is modified whenever they are asked to grade one or the other candidate.

Figure 6 makes it possible to compare the use of intermediate grades under the EV4 and EV6 scales for each candidate. It appears that intermediate grades are used for all candidates, and that in almost all cases the share of intermediate grades tends

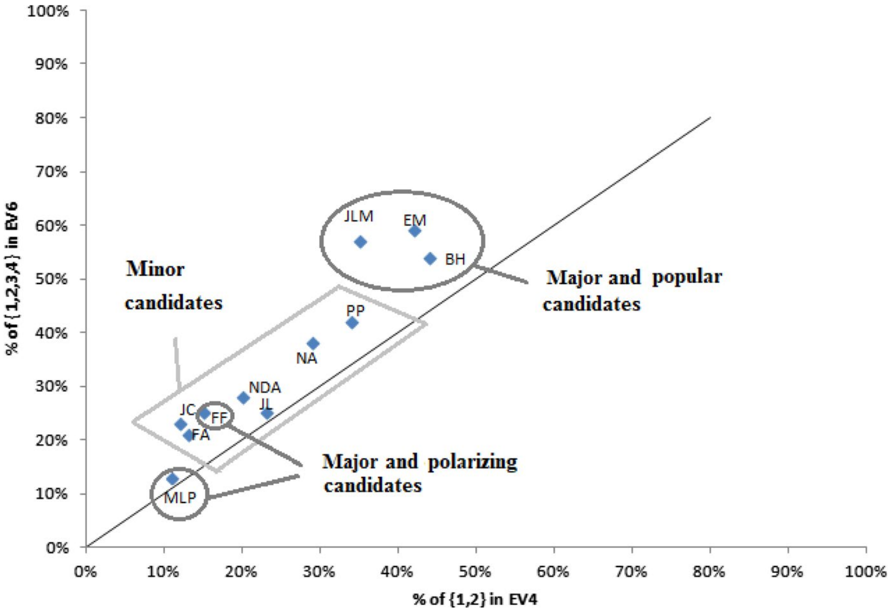


Fig. 6 Use of intermediate grades for each candidate

to increase with the extension of the scale, i.e., if we compare EV6 with EV4. Hence we can definitively reject hypothesis AL3.

Figure 6 further shows that there are differences in the use of intermediate grades from one candidate to another, thus meaning that the voters adapt their behavior depending on the kind of candidates they have to grade and the length of the scale. Two viable candidates received only a few intermediate grades under both EV4 and EV6: M. Le Pen and, to a lesser extent, F. Fillon, as shown in Figure 6. The three remaining viable candidates in this election—the popular ones E. Macron, J.-L. Mélenchon, and B. Hamon—both garner the most intermediate grades under EV4 and EV6. They also receive more intermediate grades under EV6 than EV4: the spots representing them in Figure 6 are more distant from the first bisector than the spots standing for other candidates.

As such, this shows that, as a rule, the behavior of most voters is not strategic, but is rather guided by the intensity of their feelings of sympathy or antipathy for the candidates. To gain greater understanding of the behavior of voters depending on the kind of candidate, Table 5 displays the proportions of approvals and non-approvals which turn into intermediate grades for each candidate.

Some trends are clearly visible. When candidates have been approved (AV = 1), the proportion of voters who give them an intermediate grade rather than the maximum one depends on the length of the scale and on the kind of candidate. Indeed, from EV4 to EV6, the proportion of approvals which turn into intermediate grades

Table 5 Transformation of approvals (AV = 1) / non-approvals (AV=0) into intermediate grades (Hérouville data)

Candidates	EM	MLP	FF	JLM	BH	NDA	JL	PP	FA	NA	JC
Grades 1 and 2 in EV4 for AV=0	40%	7%	9%	33%	36%	16%	17%	23%	11%	22%	9%
Grades 1 and 2 in EV4 for AV=1	44%	46%	40%	36%	52%	67%	71%	67%	50%	67%	82%
Grades 1, 2, 3, and 4 in EV6 for AV=0	45%	7%	16%	44%	45%	20%	24%	30%	20%	31%	16%
Grades 1, 2, 3, and 4 in EV6 for AV=1	73%	59%	58%	66%	63%	68%	79%	77%	74%	83%	82%
Chi-squared test for AV=0 in EV4-EV6	0.747 (0.387)	0.020 (0.888)	5.098 (0.024)**	3.474 (0.062)*	2.2 (0.138)	0.812 (0.368)	4.103 (0.043)**	3.333 (0.068)*	7.749 (0.005)***	5.210 (0.022)**	4.337 (0.037)**
Chi-squared test for AV=1 in EV4-EV6	26.383 (2.8 × 10^{-7})***	0.882 (0.348)	3.345 (0.067)*	30.616 (3.145 × 10^{-8})***	3.434 (0.064)**	0.00 (1)	(0.706)	1.552 (0.213)	(0.255)	2.426 (0.119)	(1)

The table displays the proportions of approvals and non-approvals which turn into intermediate grades for each candidate depending on the scale (EV4 or EV6). To compare EV4 and EV6, Pearson's Chi-squared tests or Fisher's exact tests are conducted. The p-values are in parentheses. The symbol "*" (resp. "**" and "***") means that the hypothesis H_0 is rejected at the level of significance of 10% (resp. 5% and 1%)

Table 6 Proportion of voters using the full extent of the grading scale and frequency difference tests

AV	EV4	EV6	Chi-squared Test AV-EV4	Chi-squared Test AV-EV6	Chi-squared Test EV4-EV6
100%	83.52%	64.95%	56.017 (7.184×10^{-4}) ***	108.01 (2.2×10^{-16}) ***	29.432 (5.791×10^{-8}) ***

The table displays the proportion of voters who used its full extent for AV, EV4, and EV6. To compare them, McNemar's or Pearson's Chi-squared tests are conducted. The p-values are in parentheses. The symbol "*" (resp. "**" and "***") means that the hypothesis H_0 is rejected at the level of significance of 10% (resp. 5% and 1%)

tends to increase for the three popular candidates (E. Macron, J.-L. Mélenchon, and B. Hamon) (see Table 5), which shows that their scores under EV4 and EV6 tend to decrease as compared to under AV. The proportion of non-approvals (AV=0) which turn into intermediate grades also significantly increases with the length of the scale for every minor candidate, except for N. Dupont-Aignan. Consequently, their scores under EV4 and EV6 increase as compared to under AV.

Given these observations, how can we explain the relative stability of the scores of most candidates apart from the non-covered ones? We can observe effects in opposite directions: the translation of approvals into intermediate grades lowers the scores, while the translation of non-approvals into intermediate grades increases the scores. These effects eventually compensate each other, such that differences are often considered as non-significant. However, we cannot ensure that this precarious equilibrium always occurs. Table 5 pinpoints the trends more specifically. With scale lengthening, there might be an evolution of scores: higher scores for minor candidates, and lower scores for all major candidates, the popular ones in particular. Lengthening the grading scales might thus benefit the former to the detriment of the latter, and hence an inversion of rankings cannot in principle be excluded.

5.3 Scale length matters for different voters: Comparative use of the distribution of grades

In an evaluative voting system, the influence of an elector on the outcome depends on her use of the grading scale offered to her. Thus, an elector who gives the same grade to all the candidates, whatever it is, would have no effect on the outcome, since it would increase the overall score of each candidate by the same amount. Beyond this extreme case, the maximum impact an elector can have on the difference of scores between two candidates, and therefore on the outcome, would be to give the maximum grade to her favorite, while giving the minimum grade to the other.

Strategic voters should want to influence the outcome of the election as much as they can. From this perspective, each voter should therefore use the full extent of the grading scale. We test the following hypothesis:

Hypothesis 7 (*AG1*) For all scale lengths, all voters use the entire scale of grades.

Considering the possible impact of the lengthening of the scale on voters' behavior, we test a supplementary assumption:

Hypothesis 8 (*AG2*) With the same minimum grade, the lengthening of the scale does not modify the proportion of voters who use the full extent of the grading scale.

In practice, with few exceptions, all voters use the minimum grade for a large part of the candidates. The difference between the highest and the lowest grades given by each elector therefore depends exclusively on the highest grade this elector has given. Table 6 displays the proportion of voters who used its full extent for each of the three scales tested in Hérouville, given that for AV the proportion is necessarily 100%.

With EV4 and EV6, the proportion of voters who do not use the entire scale extent is significantly lower than 100%. Assumption AG1 is thus rejected. When a grading scale has strictly more than two steps, a significant proportion of voters do not use the full extent of the grading scale. They do not want to give the maximum grade to any of the candidates.

The effect of scale extension is unambiguous. The extension of the scale from EV4 to EV6 leads to a significant decrease of the proportion of voters who use the entire extent of the grading scale (see Table 6).[5] Assumption AG2 is thus rejected. In the context of evaluative voting rules, the lengthening of the grading scales introduces a growing difference in the influence exerted by different voters: similar voters have a dissimilar influence on the outcome, depending on their different usage of the proposed scales.

This result should be interpreted as follows: a major part of voters favor expressive voting over strategic voting. Although greater expression reduces their influence on the vote outcome, many voters exclude the possibility of giving a high grade to a candidate, even though she or he is their favorite, if they do not consider that this candidate "deserves" a high grade. In as much as a longer scale automatically allows more variety and adaption of grades, voters behave as if a grade captures the distance from or degree of conformity with their ideas. This captures that they behave in a way intended to express their personal views on the candidates, rather than to influence the result.[6]

This interpretation seems to be shared by Sylvain Spinelli, the president of the "Vote de valeur" Association (Value voting), who favors the evaluative voting rule

[5] Note that the same result appears when we consider the different lengths of scales tested in Strasbourg to the same minimum grade, so avoiding other biases: there is significantly less use of the whole extent of the scale from EV3 to EV4 and from EV3neg to EV4neg.

[6] Note that our observation holds for voters' behavior when testing a new voting rule in an experimental setting. If an evaluative voting rule were to be implemented, these aspects are likely to be covered by the media and political parties, such that we could expect more strategic behavior.

with the specific grading scale $(-2, -1, 0, +1, +2)$. In an interview with Laslier (2019), Spinelli insists on the voters' inclination toward expressive rather than strategic voting. In as much as longer scales also offer an opportunity that would not exist otherwise, longer scales do allow expression, and this may be considered as valuable per se: "When taking value voting seriously, it is necessary to concur that voters do not grade but express themselves. Voters use the whole scope of values to influence the outcome; they could also choose to express themselves less sharply by using just −1, 0, 1 or 0, 1, 2 as if they were 'abstaining just a bit', leaving the other the opportunity to express and determine the outcome." The desire to express "partial abstention" is more easily satisfied with longer scales.

Spinelli also acknowledges that this may impose a limit on the equal representativeness of all voters. "However, irrespective of whether value voting is perfectly egalitarian at the formal level, voters could misinterpret its principles and confuse it with the standard grading system used in schools. While grading students, teachers have no reason to focus only on extreme grades." The unequal influence of voters results from their confusion between two different exercises: grading, which is a strictly expressive device (allowed by longer scales), and voting, which implies participation in a collective choice.

We have thus demonstrated that, in so far as scales are longer, a higher proportion of voters use the more intermediate grades and depart from strategic behavior. As a result, another regrettable property of longer scales is that some voters weigh less in the results.

6 Conclusion

This paper concerns evaluative voting, that is, voting systems wherein voters can grade every candidate under a given grading scale: this contrasts with majority or plurality voting where they can pick just one among all, or any other rules where they are asked to rank the various candidates. We study to what extent the behavior of the voters, and thus the influence of different voters on the outcome of the election, is sensitive to the range of the grading scale. Our scrutiny relies on experimental data collected in parallel with the 2017 French presidential election: on the day of the first round, alternative evaluative voting rules were proposed in situ to the voters, after casting their official vote, using different scales. Based on a novel protocol with randomized allocation of ballot papers in every polling station, the results illuminate the two effects studied in this paper: the length of the grading scale and the introduction of a negative grade.

First, in accordance with results previously established in the literature, the introduction of a negative grade clearly distorts the scores, relatively disfavoring the polarizing candidates. But this paper additionally establishes that the introduction of negative grades relatively favors minor candidates, rather than favoring major popular candidates.

Second, for non-negative scales of different lengths, we have shown that we cannot confirm the assumption of stability of voting outcomes, for there is a significant change concerning minor candidates. We show in particular that those who benefit more from longer scales are the non-covered candidates, i.e., those who are little known by the voters due to their limited media coverage. We also provided evidence that, although it is potentially compensated (as indeed happens in our data set), there might in principle be a significant length effect in favor of minor candidates: the share of intermediate grades relatively decreases for polarizing candidates whenever the length of the grading scale increases. Regarding the biases between voters which the chosen grading scale might create, it has been stressed that the longer the grading scale (with positive grades only), the less is the full extent of grades used by voters. We suggest this derives more from expressive than from strategic behavior. For longer scale lengths, this feature is likely to lead to different influences being exerted by different kinds of voters on the outcome of the election.

A study of this kind aims to offer a comparative analysis of the properties of rules, notably including their capacity for expression, or their tendency to favor a certain type of candidate. The introduction of negative grades and the use of longer scales are often regarded as desirable in so far as, all other things being equal, they allow more expressivity—where the possibility of expression may be considered to be valuable in its own right. This view required further examination; in particular, the underlying assumption that more expressive rules do not change the voting outcome needed scrutiny. All in all, our results highlight that the stability of outcomes under evaluative voting rules using different scale lengths is by no means uncontroversially established, and that the different grading scales are not all equivalent. Longer scales, or scales with negative grades, essentially favor minor, unpopular, or even non-covered candidates in the senses we have defined above. Further, otherwise equal voters then come to have unequal weights in the election, depending on their propensity to use extreme grades. We suggest that these two properties are unlikely to be desirable for a democracy.

Acknowledgements This paper has been presented at the INDEPTH conference held at the University of Saint-Etienne in March 2019; the “Individual Preferences and Social Choice” workshop held at the University of Graz in April 2019; the Conference on Economic Design, Budapest, June 2019; the 3rd ILLC Workshop on Collective Decision Making, University of Amsterdam, June 2019; and the “Colloque sur l’histoire et la philosophie des théories économiques, autour du travail de Philippe Mongin”, University of Paris 2, December 2019. We are grateful to Jean-François Laslier, Mostapha Diss, Ulle Endriss, Umberto Grandi, Markus Brill, Hans Peters and two anonymous referees for insightful remarks that have helped to improve the paper. This research has benefited from financial support by the DynaMITE project (ANR13-BSH1-0010), the IDEXLYON at the University of Lyon (project INDEPTH) within the Programme Investissements d’Avenir (ANR-16-IDEX-0005), the CoCoRICo-CoDec Project (ANR-14-CE24-0007), and the University of Strasbourg Foundation. We would also like to thank the whole team of “Voter autrement 2017”, and the public officials who helped us to implement the experiment: see http://www.gate.cnrs.fr/vote/merci.

Appendix

Description of the candidates

There were 11 candidates at the 1st round of the French Presidential election in 2017. We here present them in the official order, randomly chosen by the Constitutional Council, and specify their initials, as they are presented in the figures and tables, and results in % in the first round of the Presidential Election.

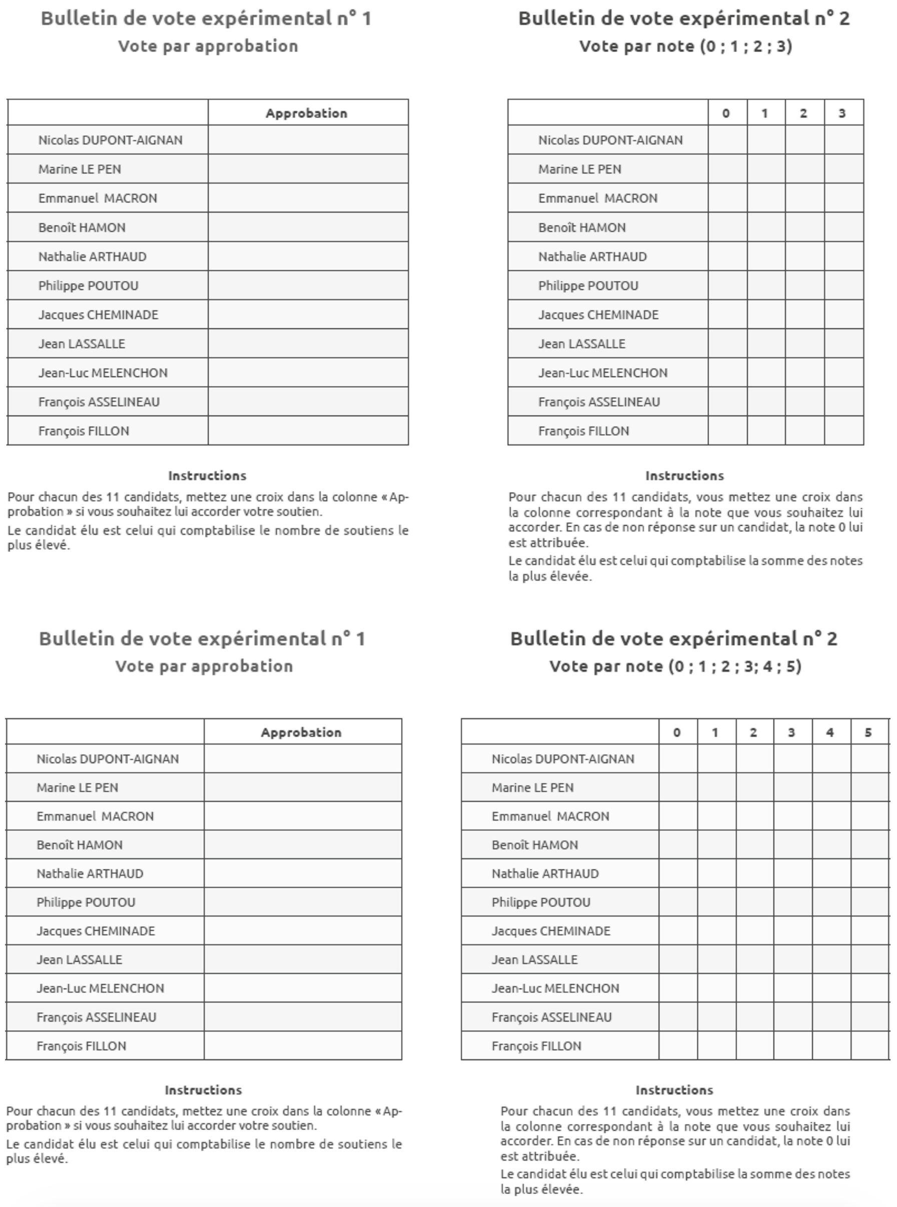
NDA	Nicolas Dupont-Aignan	Debout la France! (DLF)	4.70
MLP	Marine Le Pen	Front national (FN)	21.30
EM	Emmanuel Macron	En marche! (EM)	24.01
BH	Benoît Hamon	Parti socialiste (PS)	6.36
NA	Nathalie Arthaud	Lutte ouvrière (LO)	0.64
PP	Philippe Poutou	Nouveau parti anticapitaliste (NPA)	1.09
JC	Jacques Cheminade	Solidarité et progrès (S& P)	0.18
JL	Jean Lassalle	Résistons	1.21
JLM	Jean-Luc Mélenchon	La France insoumise (FI)	19.58
FA	François Asselineau	Union populaire républicaine (UPR)	0.92
FF	François Fillon	Les Républicains (LR)	20.01

Ballot papers used in the protocol

We here display the different ballot papers used in the protocol (see Figs. 7, 8, 9).

Corrected scores

It must be stressed that both Strasbourg and Hérouville voting stations do not accurately reflect the composition of the French electorate at the national level and, because participation was free and open, participants in the experiment are not representative of their voting station. In order to be able to compare the different experimental results between voting stations, we made a primary adjustment to the rough data in order to correct both representation and selection biases. In a questionnaire added to experimental ballots, participants were asked about their official votes. Each ballot has been weighted by the ratio between the score of the corresponding candidate in the official election and the share of participants who declared to have voted in his/her favor (Tables 7, 8).



Bulletin de vote expérimental n° 1

Vote par approbation

	Approbation
Nicolas DUPONT-AIGNAN	
Marine LE PEN	
Emmanuel MACRON	
Benoît HAMON	
Nathalie ARTHAUD	
Philippe POUTOU	
Jacques CHEMINADE	
Jean LASSALLE	
Jean-Luc MELENCHON	
François ASSELINEAU	
François FILLON	

Instructions

Pour chacun des 11 candidats, mettez une croix dans la colonne « Approbation » si vous souhaitez lui accorder votre soutien.

Le candidat élu est celui qui comptabilise le nombre de soutiens le plus élevé.

Bulletin de vote expérimental n° 2

Vote par note (0 ; 1 ; 2 ; 3)

	0	1	2	3
Nicolas DUPONT-AIGNAN				
Marine LE PEN				
Emmanuel MACRON				
Benoît HAMON				
Nathalie ARTHAUD				
Philippe POUTOU				
Jacques CHEMINADE				
Jean LASSALLE				
Jean-Luc MELENCHON				
François ASSELINEAU				
François FILLON				

Instructions

Pour chacun des 11 candidats, vous mettez une croix dans la colonne correspondant à la note que vous souhaitez lui accorder. En cas de non réponse sur un candidat, la note 0 lui est attribuée.

Le candidat élu est celui qui comptabilise la somme des notes la plus élevée.

Bulletin de vote expérimental n° 1

Vote par approbation

	Approbation
Nicolas DUPONT-AIGNAN	
Marine LE PEN	
Emmanuel MACRON	
Benoît HAMON	
Nathalie ARTHAUD	
Philippe POUTOU	
Jacques CHEMINADE	
Jean LASSALLE	
Jean-Luc MELENCHON	
François ASSELINEAU	
François FILLON	

Instructions

Pour chacun des 11 candidats, mettez une croix dans la colonne « Approbation » si vous souhaitez lui accorder votre soutien.

Le candidat élu est celui qui comptabilise le nombre de soutiens le plus élevé.

Bulletin de vote expérimental n° 2

Vote par note (0 ; 1 ; 2 ; 3; 4 ; 5)

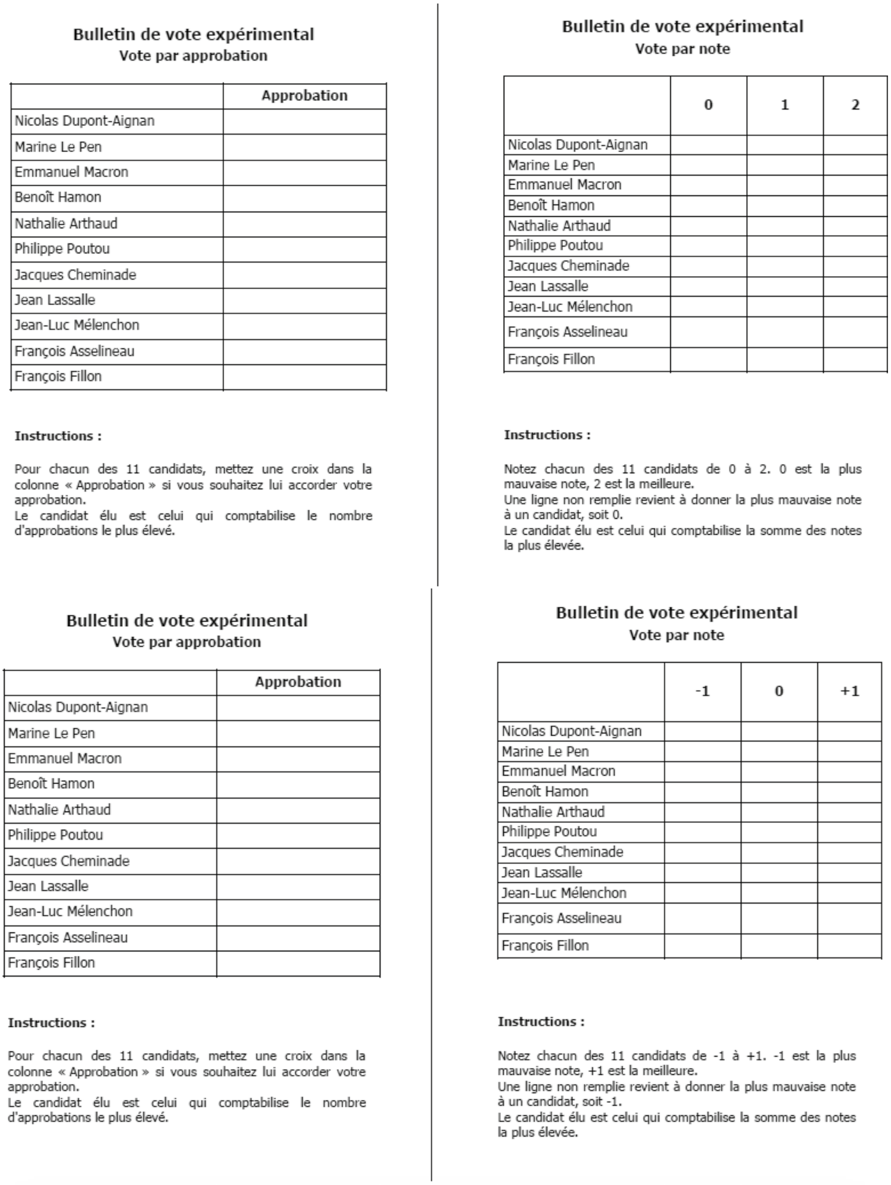
	0	1	2	3	4	5
Nicolas DUPONT-AIGNAN						
Marine LE PEN						
Emmanuel MACRON						
Benoît HAMON						
Nathalie ARTHAUD						
Philippe POUTOU						
Jacques CHEMINADE						
Jean LASSALLE						
Jean-Luc MELENCHON						
François ASSELINEAU						
François FILLON						

Instructions

Pour chacun des 11 candidats, vous mettez une croix dans la colonne correspondant à la note que vous souhaitez lui accorder. En cas de non réponse sur un candidat, la note 0 lui est attribuée.

Le candidat élu est celui qui comptabilise la somme des notes la plus élevée.

Fig. 7 Ballot papers used in Hérouville, where voters are asked to vote under approval voting and either EV4 or EV6



Bulletin de vote expérimental
Vote par approbation

	Approbation
Nicolas Dupont-Aignan	
Marine Le Pen	
Emmanuel Macron	
Benoît Hamon	
Nathalie Arthaud	
Philippe Poutou	
Jacques Cheminade	
Jean Lassalle	
Jean-Luc Mélenchon	
François Asselineau	
François Fillon	

Instructions :

Pour chacun des 11 candidats, mettez une croix dans la colonne « Approbation » si vous souhaitez lui accorder votre approbation.
Le candidat élu est celui qui comptabilise le nombre d'approbations le plus élevé.

Bulletin de vote expérimental
Vote par note

	0	1	2
Nicolas Dupont-Aignan			
Marine Le Pen			
Emmanuel Macron			
Benoît Hamon			
Nathalie Arthaud			
Philippe Poutou			
Jacques Cheminade			
Jean Lassalle			
Jean-Luc Mélenchon			
François Asselineau			
François Fillon			

Instructions :

Notez chacun des 11 candidats de 0 à 2. 0 est la plus mauvaise note, 2 est la meilleure.
Une ligne non remplie revient à donner la plus mauvaise note à un candidat, soit 0.
Le candidat élu est celui qui comptabilise la somme des notes la plus élevée.

Bulletin de vote expérimental
Vote par approbation

	Approbation
Nicolas Dupont-Aignan	
Marine Le Pen	
Emmanuel Macron	
Benoît Hamon	
Nathalie Arthaud	
Philippe Poutou	
Jacques Cheminade	
Jean Lassalle	
Jean-Luc Mélenchon	
François Asselineau	
François Fillon	

Instructions :

Pour chacun des 11 candidats, mettez une croix dans la colonne « Approbation » si vous souhaitez lui accorder votre approbation.
Le candidat élu est celui qui comptabilise le nombre d'approbations le plus élevé.

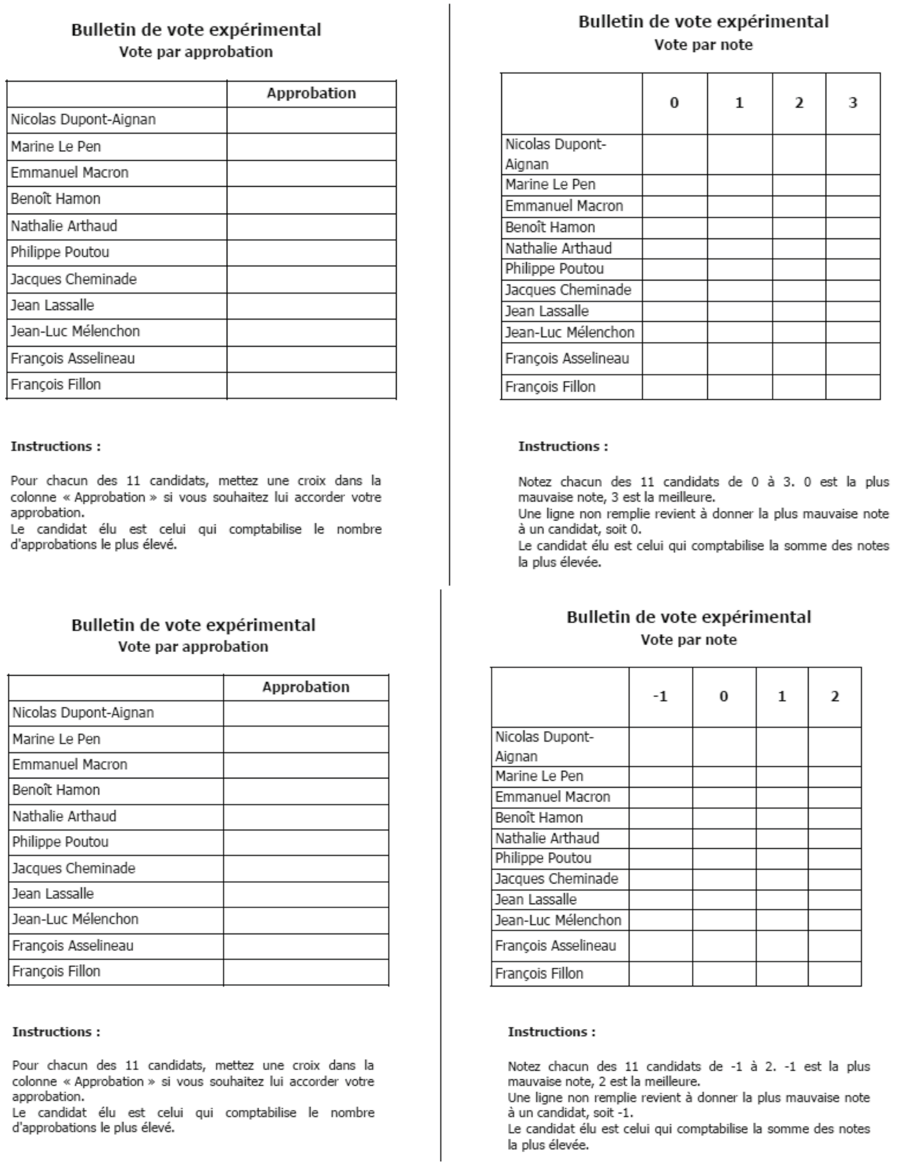
Bulletin de vote expérimental
Vote par note

	-1	0	+1
Nicolas Dupont-Aignan			
Marine Le Pen			
Emmanuel Macron			
Benoît Hamon			
Nathalie Arthaud			
Philippe Poutou			
Jacques Cheminade			
Jean Lassalle			
Jean-Luc Mélenchon			
François Asselineau			
François Fillon			

Instructions :

Notez chacun des 11 candidats de -1 à +1. -1 est la plus mauvaise note, +1 est la meilleure.
Une ligne non remplie revient à donner la plus mauvaise note à un candidat, soit -1.
Le candidat élu est celui qui comptabilise la somme des notes la plus élevée.

Fig. 8 Ballot papers used in Strasbourg, where voters are asked to vote under approval voting and either EV3 or EV3neg



Bulletin de vote expérimental

Vote par approbation

	Approbation
Nicolas Dupont-Aignan	
Marine Le Pen	
Emmanuel Macron	
Benoît Hamon	
Nathalie Arthaud	
Philippe Poutou	
Jacques Cheminade	
Jean Lassalle	
Jean-Luc Mélenchon	
François Asselineau	
François Fillon	

Instructions :

Pour chacun des 11 candidats, mettez une croix dans la colonne « Approbation » si vous souhaitez lui accorder votre approbation.
Le candidat élu est celui qui comptabilise le nombre d'approbations le plus élevé.

Bulletin de vote expérimental

Vote par note

	0	1	2	3
Nicolas Dupont-Aignan				
Marine Le Pen				
Emmanuel Macron				
Benoît Hamon				
Nathalie Arthaud				
Philippe Poutou				
Jacques Cheminade				
Jean Lassalle				
Jean-Luc Mélenchon				
François Asselineau				
François Fillon				

Instructions :

Notez chacun des 11 candidats de 0 à 3. 0 est la plus mauvaise note, 3 est la meilleure.
Une ligne non remplie revient à donner la plus mauvaise note à un candidat, soit 0.
Le candidat élu est celui qui comptabilise la somme des notes la plus élevée.

Bulletin de vote expérimental

Vote par approbation

	Approbation
Nicolas Dupont-Aignan	
Marine Le Pen	
Emmanuel Macron	
Benoît Hamon	
Nathalie Arthaud	
Philippe Poutou	
Jacques Cheminade	
Jean Lassalle	
Jean-Luc Mélenchon	
François Asselineau	
François Fillon	

Instructions :

Pour chacun des 11 candidats, mettez une croix dans la colonne « Approbation » si vous souhaitez lui accorder votre approbation.
Le candidat élu est celui qui comptabilise le nombre d'approbations le plus élevé.

Bulletin de vote expérimental

Vote par note

	-1	0	1	2
Nicolas Dupont-Aignan				
Marine Le Pen				
Emmanuel Macron				
Benoît Hamon				
Nathalie Arthaud				
Philippe Poutou				
Jacques Cheminade				
Jean Lassalle				
Jean-Luc Mélenchon				
François Asselineau				
François Fillon				

Instructions :

Notez chacun des 11 candidats de -1 à 2. -1 est la plus mauvaise note, 2 est la meilleure.
Une ligne non remplie revient à donner la plus mauvaise note à un candidat, soit -1.
Le candidat élu est celui qui comptabilise la somme des notes la plus élevée.

Fig. 9 Ballot papers used in Strasbourg, where voters are asked to vote under approval voting and either EV4 or EV4neg

Table 7 Strasbourg data—average corrected scores, for the different voting rules

Candidates	AV	EV3	EV3neg	EV4	EV4neg
EM	44.20	0.93	0.15	1.27	0.38
MLP	24.56	0.38	− 0.28	0.75	− 0.32
FF	25.77	0.46	− 0.30	0.70	− 0.26
JLM	43.25	0.96	0.16	0.36	0.44
BH	35.55	0.70	0.15	1.02	0.17
NDA	19.83	0.45	− 0.07	0.74	− 0.32
JL	7.63	0.31	− 0.23	0.36	− 0.55
PP	17.47	0.42	− 0.38	0.47	− 0.31
FA	7.53	0.21	0.16	1.20	0.44
NA	8.39	0.33	− 0.38	0.46	− 0.41
JC	3.64	0.12	− 0.25	0.73	− 0.31

The table displays the average corrected scores for AV, EV3, EV3neg, EV4, and EV4neg on the basis of Strasbourg data

Table 8 Hérouville data—average corrected scores, for the different voting rules

Candidates	AV	EV4	EV6
EM	42.75	1.26	2.10
MLP	26.12	0.81	1.32
FF	27.26	0.85	1.24
JLM	39.02	1.22	1.94
BH	34.42	1.06	1.69
NDA	17.76	0.62	0.98
JL	5.57	0.30	0.60
PP	17.08	0.50	0.96
FA	5.12	0.20	0.44
NA	9.88	0.42	0.63
JC	3.83	0.17	0.41

The table displays the average corrected scores for AV, EV4 and EV6 on the basis of Hérouville data

References

Alcantud JCR, Laruelle A (2014) Dis-&approval voting: a characterization. Soc Choice Welfare 43(1):1–10. Previously in: Ikerlanak Discussion Paper IL6312, Departamento de Fundamentos del Análisis Económico I, Basque Country University UPV-EHU, Bilbao, Spain

Aleskerov F, Yakuba V, Yuzbashev D (2007) A threshold aggregation of three-graded rankings. Math Soc Sci 53:106–110

Arrow KJ (1951) Social choice and individual values. Wiley, New York

Baujard A, Igersheim H, Lebon I, Gavrel F, Laslier J-F (2014) Who's favored by evaluative voting? An experiment conducted during the 2012 French presidential election. Elect Stud 34:131–145

Baujard A, Gavrel F, Igersheim H, Laslier J-F, Lebon I (2018) How voters use grade scales in evaluative voting. Eur J Political Econ 55:14–28

Baujard A, Igersheim H (2007) Expérimentation du vote par note et du vote par approbation pendant les élections présidentielles françaises le 22 avril 2007. Analyses. Rapports et documents du Centre d'Analyse Stratégique.

Blais A, Young R (1999) Why do people vote? an experiment in rationality. Public Choice 99:39–55

Ceron F, Gonzalez S (2019) A characterization of approval voting without the approval balloting assumption. WP GATE Lyon Saint-Etienne 1938

Cox GW (1997) Making votes count: strategic coordination in the world's electoral system. Cambridge University Press, Cambridge

Darmann A, Grundner J, Klamler C (2017) Election outcomes under different ways to announce preferences: an analysis of the 2015 parliament election in the Austrian federal state of Styria. Public Choice 173(1):201–216

Dhillon A, Mertens J-F (1997) Relative utilitarianism. Econometrica 67(3):471–498

Dittman I, Kubler D, Maug E, Mechtenberg L (2014) Why votes have value: instrumental voting with over confidence and overestimation of others' errors. Games Econ Behav 84:17–38

Farrell D (2001) Comparing electoral systems. Palgrave, Besingtoke

Feddersen T, Gailmard S, Sandroni A (2009) Moral bias in large elections: theory and experimental evidence. Am Political Sci Rev 103:175–192

Gaertner W, Xu Y (2012) A general scoring rule. Math Soc Sci 63(3):193–196

Gonzalez S, Laruelle A, Solal P (2019) Dilemma with approval and disapproval votes. Soc Choice Welf 53(3):497–517

Grosser J, Schram A (2006) Neighborhood information exchange and voter participation: an experimental study. Am Political Sci Rev 100:235–248

Igersheim H, Baujard A, Lebon I, Laslier J-F, Gravel F (2016) Individual behavior under evaluative voting. A comparison between laboratory and in situ experiment. In: Blais A, Laslier J-F, van der Straeten K (eds) Voting experiments. Springer, Berlin

Lachat R, Laslier J-F, Van der Straeten K (2017) Strategic voting in multi-winner elections with approval balloting: an application to the 2011 regional government election in Zurich, chapter 6. In: Blais A (ed) Strategic voting. Springer, Berlin

Laslier J-F (2019) Voter autrement: le recours à l'évaluation. No. 51 in Collection du Cepremap. Presses de l'ENS, Paris

Laslier J-F, van der Straeten K (2002) Approval voting in the French 2002 presidential election: a live experiment. Exp Econ 11(2008):97–195

Laslier J-F, Blais A, Bol D, Golder SN, Harfst P, Stephenson LB, Van der Straeten K (2015) The euro vote plus experiment. Eur Union Politics 16(4):601–615

Macé A (2018) Voting with evaluations: characterizations of evaluative voting and range voting. J Math Econ 79:10–17

Myerson RB (2002) Comparison of scoring rules in Poisson voting games. J Econ Theory 103:219–251

Narens L (1985) Abstract measurement theory. MIT Press, Cambridge

Núñez M, Laslier J-F (2014) Preference intensity representation: strategic overstating in large elections. Soc Choice Welf 42(42):313–340

Pivato M (2013) Formal utilitarianism and range voting. Math Soc Sci 67:50–56

Schram A, Sonnemans J (1996a) Voter turnout as a participation game: an experimental investigation. Int J Game Theory 25:385–406

Schram A, Sonnemans J (1996b) Why people vote: experimental evidence. J Econ Psychol 17:417–442

Smaoui H, Lepelley D (2013) Le système de note à trois niveaux: étude d'un nouveau mode de scrutin. Revue d'Economie Politique 123(6):827–850

Affiliations

Antoinette Baujard[1] · **Herrade Igersheim**[2] · **Isabelle Lebon**[3]

Herrade Igersheim
igersheim@unistra.fr

Isabelle Lebon
isabelle.lebon@unicaen.fr

[1] University Jean Monnet, University of Lyon, GATE Lyon Saint-Etienne, CNRS,
42023 Saint-Etienne, France

[2] Université de Strasbourg, Université de Lorraine, CNRS, BETA, 67085 Strasbourg, France

[3] CREM-CNRS, University of Normandy, UNICAEN, Esplanade de la Paix, 14032 Caen Cedex,
France