How voters use grade scales in evaluative voting

Antoinette Baujard, Frédéric Gavrel, Herrade Igersheim, Jean-François Laslier, Isabelle Lebon

GATE LYON SAINT-ÉTIENNE
UMR 5824

93, chemin des Mouilles
69130 Ecully - France

Maison de l'Université, Bâtiment B
10, rue Tréfilerie
42023 Saint-Etienne cedex 02 - France

http://www.gate.cnrs.fr
gate@gate.cnrs.fr

WP 1729 – October 2017

How voters use grade scales in evaluative voting

Antoinette Baujard, Frédéric Gavrel, Herrade Igersheim, Jean-François Laslier, Isabelle Lebon

Abstract:

During the first round of the 2012 French presidential election, participants in an in situ experiment were invited to vote according to "evaluative voting", which involves rating the candidates using a numerical scale. Various scales were used: (0,1), (-1,0,1), (0,1,2), and (0,1,...,20). The paper studies scale calibration effects, i.e., how individual voters adapt to the scale, leading to possibly different election outcomes. The data show that scales are not linearly equivalent, even if individual ordinal preferences are not inconsistent. Scale matters, notably because of the symbolic power of negative grades, which does not affect all candidates uniformly.

Keywords:
Evaluative Voting, Range voting, Approval voting, In Situ Experiment, Calibration

JEL codes:
D72, C93

Working Paper

How voters use grade scales in evaluative voting*

Antoinette Baujard†, Frédéric Gavrel‡, Herrade Igersheim§, Jean-François Laslier¶, Isabelle Lebon‖

August 22, 2017

Abstract

During the first round of the 2012 French presidential election, participants in an *in situ* experiment were invited to vote according to "evaluative voting", which involves rating the candidates using a numerical scale. Various scales were used: (0,1), (-1,0,1), (0,1,2), and (0,1,...,20). The paper studies scale calibration effects, i.e., how individual voters adapt to the scale, leading to possibly different election outcomes. The data show that scales are not linearly equivalent, even if individual ordinal preferences are not inconsistent. Scale matters, notably because of the symbolic power of negative grades, which does not affect all candidates uniformly.

Classification JEL: D72, C93

Keywords: Evaluative Voting, Range voting, Approval voting, In Situ Experiment, Calibration

*A previous version of the paper has been presented at the Social Choice and Welfare Conference held in Boston in 2014. We extend grateful thanks to all the members of the Community Councils, all the participants, and all the volunteers who helped us on April 22, 2012. (See http://www.gate.cnrs.fr/spip.php?rubrique94#Merci). This research was made possible thanks to the support of our research centers (CREM, PSJE, BETA and GATE L-SE), the grant ANR13-BSH1-0010, as well as fundings of the Chaire d'Excellence "Welfare economics" (CNRS and Université Jean Monnet), and the Foundation of the University of Strasbourg. We also thank Ahmad Fliti, Clemens Puppe and two referees for insightful comments.

†Univ Lyon, UJM Saint-Etienne, GATE Lyon Saint-Etienne UMR 5824, F-42 023 Saint-Etienne; antoinette.baujard@univ-st-etienne.fr

‡CREM (UMR CNRS 6211), University of Caen Normandie and Condorcet Center; frederic.gavrel@unicaen.fr

§CNRS and Beta (UMR 7522), University of Strasbourg; igersheim@unistra.fr

¶CNRS and Paris School of Economics (UMR 8545); jean-francois.laslier@ens.fr

‖CREM (UMR CNRS 6211), University of Caen Normandie, and Condorcet Center; isabelle.lebon@unicaen.fr; corresponding author.

1 Introduction

When using "evaluative voting" (also called range voting, grade voting or utilitarian voting) the voter rates candidates on a predetermined numerical scale, and the winner is the candidate whose score, obtained by summing the grades of all voters, is the highest. Using "approval voting" the voter can support as many candidates as she wishes and the winner is the candidate who collects the highest number of supporters. Notice that approval voting is a special case of evaluative voting, with only two possible grades $(0, 1)$.

The study of evaluative voting is important for several reasons. Firstly, this summation mechanism is common to many instances of multi-criteria decision; for instance, when the requirement for passing an exam is that the average grade be higher than a given trigger, or when teams or sportsmen are ranked according to some average score.

Secondly, evaluative voting is used for some political elections, and many current electoral systems embody some elements of it. A key feature of evaluative voting is a form of independence: the voter can evaluate all the candidates in turn. This independence cannot exist under uni-nominal voting rules as in one round or two-round plurality, nor under the rules that require the voter to submit a ranking of candidates like the Borda rule; but this key feature is however present in many European open-list systems (Farrell 2001). Approval type, i.e. two-level grade voting, is used for the election of the municipal council in villages in France, where voter are allowed to cross out the names of unwanted candidates. It is also close to the Swiss case where voters can approve up to a fixed number of candidates, if this number is large enough (Lachat et al. 2017). Apart from independence, another feature of evaluative voting, also encountered in several political systems, is that voters can express some degree of preference. Three-level grade-voting is possible in Latvia, where voters can cross out, leave as it is, or mark a "plus" for each candidate of her chosen party list (Laslier et al. 2015), under "cumulative voting", in Germany and Luxembourg, voters have the option of giving several points to the same candidate.

Thirdly, previous experiments in the field and on the web have confirmed that citizens appreciate these rules[1]. They appreciate grade voting more than rules allowing less expression; they tend to prefer longer scales (but not too long ones, as we will see), and they prefer scales with negative grades to scales with only positive grades. There exists in different countries some activist associations supporting evaluating voting rules and lobbying for their use in political elections (e.g. rangevoting.org, votedevaleur.org). One of the various debates in this non-academic community concerns the scope of

[1]This point may be important in a context where elections are not so attractive and turnout decreases, see Garman 2017.

the scale. It is often argued that longer scales require more political commitments from citizens and convey more information, so that such scales are better in principle. This argument is debatable. Among other reasons, if individuals were to vote strategically, longer scales may enhance more strategic behaviors, and voters who vote strategically will have even more power than voters who vote honestly. Although we have observed that strategic behavior might not be such a central issue in grade-voting in political elections (Igersheim et al. 2016)[2], we still face a lack of scientific and pragmatic arguments about the use of different grade scales for evaluative voting.

Although important in practice, the properties of evaluative voting have seldom been studied in theory, at least until recently (exceptions include Yilmaz 1999; Gaertner and Xu 2012; Smaoui and Lepelley 2013; Pivato 2013). Voting by grading is similar to a simple utilitarian calculus, and is indeed sometimes called "Utilitarian voting" (Hillinger 2004a,b). Given individual utilities, which are supposed to be of a quantitative nature and interpersonally comparable, the axiomatic approach to utilitarianism has lent justification to the idea that social aggregation can be realized through sums, or generalized sums (Debreu 1960, Gorman 1968, D'Aspremont and Gevers 1977, Maskin 1978, Wakker 1989). This mathematical theory is useful for understanding how we can, or should, aggregate grades (Pivato 2013, Macé 2015), but it has not much to say about how voters would use grading scales for voting purposes.

As will be seen, the data show that varying the scale length or the scale labels induces non-equivalent voting results, because voters vote differently. To study the properties of evaluative voting, including the special case of approval voting, one must understand how voters use the numerical grades in the various scales. But we do not know very much about this. There are very few references in the academic literature comparing different forms of grade voting in practice (Baujard et al. 2014, Igersheim et al. 2016).

This paper is not an activist stance to defend evaluative voting, it is a contribution to the scientific debate on the relative qualities and shortcomings of grade voting systems. It aims to show why and how variations of scales matter. We call this the calibration problem: how the choice of one individual translates her preference into a vote, using different rules.

Although the real-life examples mentioned above involve elections in multi-member districts, the present paper will consider a simpler case: the election of a single candidate. We designed an experiment in which voters were asked to vote in the same election using two different scales. As we can show on the basis of these experimental data, a change in grade labels suffices to

[2]For theoretical and/or experimental works on strategic behavior under approval voting, see Brams and Fishburn 1978; Laslier 2009; Lethinen 2008; Van der Straeten et al. 2010.

change the way in which individuals calibrate their preference, so that we refute assumptions involving the equivalence of scales. Examination of our experimental data in a political context will help understand the way in which different candidates are affected by changing scales.

The paper is organized as follows. Section 2 describes the design of the experiment, links to the related theory and presents the specific hypotheses we want to test. Section 3 compares the information provided by approval voting and evaluative voting, enabling us to confirm the hypothesis that voters have consistent preference orderings. Section 4 studies the assumption that grade labels are neutral and highlight the importance of labeling effects in the presence of negative grades. Section 5 examines the assumption of invariance with scale length, that seems to hold when negative grades are excluded. Section 6 concludes.

2 Theory, design, and hypotheses

2.1 Theoretical questions

Individuals have political preferences, but we cannot observe them directly. The main way by which we can access political preferences is through their translation into vote choices (Schnellenbach 2015), constrained by voting rules. Studying different calibrations, i.e. how preferences are expressed through different evaluative scales, provides a way of studying the properties of political preferences by comparison, as well as the calibration effects themselves.

First, it is standard in the economic literature to capture an individual preference by a pre-order, *i.e.* a reflexive and transitive binary relation (Arrow 1951). Arrovian preferences are ordinal, they do not contain quantitative statements, they are binary qualitative statements of the form "I prefer a to b". This assumption could however be challenged by a richer conception of preferences, taking into account cardinality and multidimensionality (*e.g.*, Sen 1977 and the debate on welfarism). While the nature of preferences has been extensively debated and tested (taking into account risk, time, and market context), and despite significant debates in the political domain, there is little empirical or experimental work aimed at confirming or refuting the assumption of preference consistency in a real political context[3]. By observing and analyzing the low number of inconsistent pre-orderings induced by both approval and evaluation rankings, this paper confirms the assumption that pre-orders do capture part of political preferences.

[3]Regenwetter et al. 2011 tackles related points from the statistical point of view.

Second, the expression of such preferences raises a calibration problem. If a utility function captures a preference pre-ordering, any monotone increasing transformation of this utility function will also capture it. Since evaluative voting proceeds by simply summing grades, the outcome of the election — that is, the identity of the elected candidate as well as the relative scores — is preserved by linear transformation of the grades.

Therefore, when translating Arrovian preferences into a graded scale, each voter has to calibrate her preference. The precise individual calibration process depends on the grade scale, both with respect to the scale length (the number of available grades), and with respect to the grades labels (compare, for instance the scale $(0, 1, 2)$ with the scale $(-1, 0, 1)$). The refutation of the invariance assumption implies that two linearly equivalent rules may provide non-equivalent, even if consistent, information on voters' preferences and evaluations of candidates, and potentially non-equivalent collective rankings.

Identifying precisely these effects and their aggregate consequences remains, however, an open issue, to which this paper will contribute. In the context of opinion surveys (rather than votes), Schwarz et al. (1991) established that numeric values may change the meaning of scale labels and thus the result of the evaluation. They also provide an intuitive explanation for this phenomenon. People are reluctant to use minimal negative grades in a questionnaire concerning their success in life: this would be interpreted as an explicit failure, whereas the use of the minimal positive grade could be interpreted as a mere absence of success. As reluctance to use negative grades may be due to the particular context, this paper offers some possible interpretations of the specific use of negative grades by voters in a political context.

The existence of some label effects will indeed be confirmed by our data. More importantly, we highlight that the variations depend on the type of candidates, a point that could not be shown in the context selected by Schwarz et al. (1991). This allows us to understand which symbolic contents of grades create bias, and how they do so.

2.2 Design of the experiment

We were able to conduct an *in situ* experiment during the first round of the French presidential election on April 22, 2012. Among the $4,319$ voters who were invited to participate in the experiment, $2,340$ accepted. Each participant tested two alternative rules for deciding on the ten candidates in this election: approval voting and a variant of evaluative voting. We refer the reader to Baujard et al. (2013) for a complete presentation of the protocol,

and to Grofman et al. (2011) for more details of the *in situ* methodology.[4]

The experiment was performed in five polling stations. In one station (in the city of Saint-Etienne), voters used approval voting (henceforth: AV) and evaluative voting (henceforth EV) with the $(0, 1, 2)$ scale. In two stations (in Louvigny, a small town near Caen), voters used AV and the $(-1, 0, 1)$ scale. In two stations (in the city of Strasbourg), voters used AV and the twenty-one-level scale $(0, ..., 20)$. Notice that the 21-scale is routinely used in the French education system. Hence all voters who participated in the experiment made use of two scales of differing lengths, and one of them (Approval) was common to all voters. When voters did not mark any box to grade a candidate, this lack of response was considered as a zero-grade; this rule was written on the voting ballot and explained to each voter.

A characteristic feature of the *In Situ* method is that we take care not to ask the voters their "true underlying preferences" or "true opinion" but we distinguished what we do from an opinion survey and explicitly asked them to consider our experimental vote as they would consider a real one. Even if we cannot prove it, we have no reason to think that the participants did not do the job seriously so that those inclined to strategic thinking might be strategic and others not. We thus assume that the mix of sincere and of various strategic considerations that might be prevailing in reality is captured by our methodology, and has the consequences described in the paper.

2.3 Hypotheses to be tested

This paper addresses the calibration issue by assessing three hypotheses regarding voters' preferences and vote results.

The first hypothesis (H1), which deals with ordinal consistency, is required in order to study calibration. Comparing two scales, ordinal preferences are preserved if a voter never gives a strictly higher grade to candidate A than to candidate B under one system (showing that she strictly prefers A to B), and a lower grade under the other system (showing that she strictly prefers B to A).

Our protocol made it possible to track inconsistencies at the individual level between approval voting and one type of evaluative voting. Although the literature does not define what a "small proportion" of mistakes is for such experiments, we conclude that the observed inconsistencies do not call into question the entire idea that approval voting and evaluative voting define, for each voter, a pre-ordering of the candidates.

[4] In 2012 we used the phrases *Vote par évaluation*, *Vote par approbation*, *Vote par note*, that are well understood by the voters.

We also stress that inconsistencies are, a priori, more likely to appear with mere fine-grained scales (such as our 21-level scale). This is indeed what we observe, when testing H1 :

Hypothesis (H1– Identical proportion of inconsistencies) *The proportion of ranking inversions is equivalent for the different scales of evaluation.*

We refute hypothesis H1 because a long-scale evaluation led to a number of errors significantly higher than an evaluation using a short scale.

The second hypothesis (H2) deals with scales with different labels but of the same length (in our experiment: $(0, 1, 2)$ and $(-1, 0, 1)$). It corresponds to a test of invariance under translation, i.e. adding a constant to each numerical label should not change the way voters vote, nor the overall collective result.

Hypothesis (H2– Invariance with respect to labels) *Numerical scales of the same length but with different labels are linearly equivalent.*

We refute H2: the statistical distribution of the lowest, median and highest grades depends on the scale. We also refute the related hypotheses that (i) the thresholds of approvals are equivalent, and that (ii) the labeling effects are uniform across candidates.

The third hypothesis (H3) deals with scales of different lengths. It corresponds to a test of invariance under linear transformation, i.e., differences in lengths of grade scales should not modify rankings or relative scores of candidates. Observations of lengths 2 (approval voting), 3, and 21 are available, but we here focus on the comparisons of $(0, 1, 2)$ with $(0, 1, ..., 20)$.

Hypothesis (H3– Invariance with respect to length) *Numerical scales of different lengths are equivalent.*

This hypothesis is not refuted, nor do we refute the related hypothesis concerning approval thresholds and candidate-specific effects.

The following sections are based on the data collected. As previously explained, approval voting was used at all polling stations, whereas for evaluative voting, each city tested only one evaluative grade scale. This solution made it possible to present to the voters of each site a unique experimental protocol, in the letters sent to them and at the public information meetings. Voters, knowing in advance the exact alternative voting rules proposed to them, would be more likely to think ahead before voting, to better understand the aims and stakes of this scientific experiment, and to eventually take part in it.

By design, the *In Situ* method indeed leaves the voters free to participate

in the survey or not. The resulting self-selected samples have a significant important ideological bias, with conservative voters being under-represented. Beyond these trends, our results are also affected by the disparity of voters' opinions from one polling station to another. To make experimental voting of different sites directly comparable, we have to correct them from these disparities. We weight the individual observations to make the distributions of official votes on the different sites match the national results. Such weighting is usually done through various socio-economic variables that are correlated with political standing. In our case, we do not have such variables but we directly have access to the individual political standing through the official vote, at least for those participants who effectively stated it. We take as the reference distribution the actual distribution of the votes in the 2012 first round of the official election so that both the site representation bias and the voter participation bias should be offset (see Baujard et al. 2013 for more details). The rankings of the candidates from the different grade scales are thus comparable both to each other and to the official national results.[5] The following uses the weighted data.

3 Consistency of ordinal preferences

Approval voting, tested in all three locations, is a two-level scale. Each ballot divides the candidates into only two classes: those who are approved, and those who are not. Supposedly, approved candidates are preferred to non-approved candidates. Within each class the candidates cannot be disentangled. In three-level evaluation voting, as tested both in Louvigny and Saint-Etienne, the candidates are divided into three classes. Supposedly the set of the least preferred candidates gets the lowest grade G_L, then the set of more preferred the middle grade G_M, and the favorite candidates (including strategically selected ones) should get the highest grade G_H. For a more extensive scale of grades, the 21-level scale as tested in Strasbourg, voters can rank each of the 10 candidates in 21 different classes.

As is standard in voting theory, we assume that voters' preferences are pre-orders, hence complete and transitive. If a voter v approves candidate i and disapproves candidate j, she reveals that she prefers i to j. There is an induced pre-order with AV, let us say $\succeq^v_{AV}$; here $i \succeq^v_{AV} j$. If the same voter v gives a better grade to i than she gives to j, she reveals she prefers i to j. There is thus also an induced pre-order through EV; here $i \succeq^v_{EV} j$.

Voter v's two ballots are inconsistent if there is a pair of candidates for which there is an inversion of ranking between AV and EV: for instance $i \succ^v_{AV} j$ but $j \succ^v_{EV} i$, where $\succ$ denotes strict preference.

[5] The weighting scheme is described in the Appendix (A.1).

Logically, these induced strict rankings should not be inconsistent. In practice there may be some inconsistencies whose presence reflects the complexity for individuals in codifying their opinions, as well as purely material errors in filling the forms in.

Table 1 reports information on this issue.

Table 1 – Frequency of inconsistent ballots

	Inconsistent ballots
Saint-Etienne: $(0,1)$ and $(0,1,2)$	3.77%
Louvigny: $(0,1)$ and $(-1,0,1)$	3.41%
Strasbourg: $(0,1)$ and $(0,...,20)$	8.9%

We have assumed an identical frequency of inconsistencies regardless of the rating scale:

Hypothesis (H1– Identical proportion of inconsistencies) *The proportion of ranking inversions is equivalent across all evaluation scales.*

We find a frequency of 3.77% in Saint-Etienne, and 3.41% in Louvigny. These frequencies are not significantly different at the 5% threshold. In Strasbourg, the proportion is equal to 8.9%, which is significantly higher than in the two other sites (at the 5% threshold). This result refutes hypothesis H1. The 21-level EV scale multiplies by almost three the number of AV/EV combinations which can induce inversions. The complexity of this broad scale for individuals is likely to be the reason for the higher share of inconsistent ballots.

Should these figures be considered large or small? First notice that, unlike the usual practice in Experimental Economics, and the usual practice for surveys, the *in situ* method involves no incentive for participants. Besides, the social pressure in acting consistently is weak because ballots and questionnaires are completed anonymously in a voting booth and behind a curtain, and not facing an interviewer. Given the context, these inconsistency figures are rather low, especially for EV3. Notice that, among these cases, there can be different subsets of inconsistencies. A thorough scrutiny of these various cases, shown in Appendix A.2, leads us to confirm that the phenomenon of genuine inconsistency is insignificant. The experiment was conducted *in situ*, implying imperfect material conditions, and a lack of incentive for consistency. Nevertheless, inconsistent ballots are remarkably rare.

Voters' preferences can then be seen as consistent. The remaining analysis focuses on consistent ballots only.

4 Comparing scales of same length

This section focuses on scales of equal length. The theory says:

Hypothesis (H2 – Invariance with respect to labels) *Numerical scales of the same length but with different labels are linearly equivalent.*

An implication of H2 is that the grade k attributed to candidate C under the scale $(-1, 0, 1)$ should exactly correspond to the grade $k + 1$ attributed to this candidate under the scale $(0, 1, 2)$. Testing this correspondence requires that the scales be normalized: we add $+1$ to each grade given with the scale $(-1, 0, 1)$ in order to compare it with the $(0, 1, 2)$ scale. Differences between voting stations are controlled by our data weighting, so that the observed effects should only be due to differences in the way participants use the grade scales.

The normalized grades appear to differ markedly from one scale to another. The average normalized grade, equal to 0.82 with scale $(-1, 0, 1)$, drops to 0.56 with scale $(0, 1, 2)$. In other words, switching from a scale with positive values to a 0-centered scale leads to an increase in the global average grade of 46% (see table (2)). This very large difference cannot be attributed to mere random variation.

Let us consider the two distributions of grades more closely. If the two scales were linearly equivalent, the following would hold: all lower grades in one scale (respectively middle and higher) would exactly correspond to the lower grades of the other scale (respectively middle and higher). Is this hypothesis correct? To compare the observed (normalized) distribution, we applied a χ^2 test. The answer is that the two distributions are significantly different at the 5% threshold. Hence the data reveal that the two scales are in practice not linearly equivalent, as held by hypothesis H2.

However, there is a straightforward explanation for the significant differences between the two distributions, and for the increase of the normalized mean up to 0.82 under $(-1, 0, 1)$: the candidates who are not evaluated receive grade 0, the median grade under this rule; whereas they receive the lower grade (0) under $(0, 1, 2)$. The (normalized) scores of EV(-1,0,1) are thus raised mechanically compared to EV(0,1,2)'s when voters do not evaluate all candidates. Unfortunately, the collected data do permit the required correction. When using grade zero, many participants did not indicate whether this grade translated an evaluation or just no response. But we shall here show there is more to it than that.

4.1 Different approval thresholds

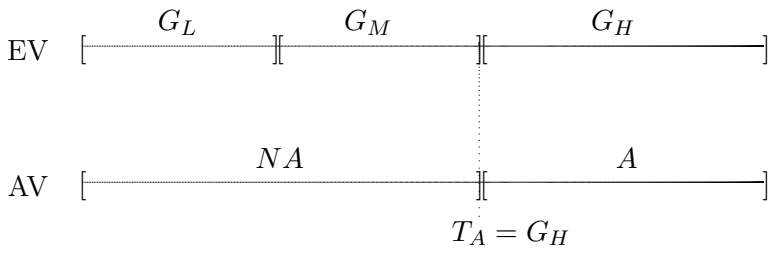
To compensate for the lack of information we compare the AV ballot with the EV ballot of the same voter. We restrict the EV data to the grades that voters gave to the candidates they approved under AV. Assuming that approved candidates are always evaluated under the two EV rules, which sounds very reasonable, the lacking information no longer affects the comparison. This implies that the (normalized) grades attributed to the approved candidates should not be too different under the distinct three-level scales.

For each voter, call "approval threshold" the lowest of the grades given to an approved candidate. The structure of the grades assigned to approved candidates may be controlled by studying the approval threshold. However, two situations must be distinguished: (i) when non-approved candidates have lower grades than approved ones, (ii) when some non-approved candidates have the same grade as approved ones[6].

To distinguish between these, we represent the preference pre-orderings of approbation and evaluation for each observed ballot by two consistent segments, whose left (right) extremity corresponds to the worse (best) candidate. We define the approval threshold as a grade on the evaluation segment. The grades assigned to a non-approved candidate are therefore always lower than (or equal to) the approval thresholds.

Let us here illustrate two categories among the observed ballots (respectively Categories 2 and 1 in the Appendix A.3 list). Consider first the simple case (Category 2) where all approved candidates receive the highest grade (G_H) while all non-approved candidates receive middle or lowest grades (G_M or G_L). Hence the threshold of approval (T_A) is equal to the highest grade. This category of ballots is referred to as $T_A = G_H$, and illustrated by Figure 1.

Figure 1 – Approval threshold $T_A = G_H$



[6]Remember that inconsistent ballots were ruled out. This means that approved candidates cannot have lower grades than non-approved ones.

Suppose now that all approved candidates, and at least one non-approved candidate, receive the highest grade (G_H) (Category 1). At first glance such a configuration might look paradoxical. Our explanation is that, in the subset of candidates who obtain the highest grade, some of them are better than others: the approved candidates. All of them receive the maximum grade because, according to voter's preferences, the distance between approved candidates and the best non-approved candidates was too small (relative to other non-approved candidates). This category of ballots is referred to as $T_A > G_H$ (see Appendix A.3 for a representation).

Following this line of reasoning, we distinguish 5 categories of ballots and corresponding approval thresholds:

- Threshold $T_A > G_H$ describes the state in which one part of the candidates who attracted the highest grade are non-approved (Cat. 1).
- Threshold $T_A = G_H$ describes the state in which all approved candidates and them only, attracted the highest grade (Cat. 2).
- Threshold $G_M < T_A < G_H$ describes the state in which both approved and non-approved candidates received the median grade (Cat. 3).
- Threshold $T_A = G_M$ describes the state in which only approved candidates received at least the median grade, and the non-approved candidates only received the lowest grade (Cat. 4).
- Threshold $G_L < T_A < G_M$ describes the state in which some of the approved candidates received the lowest grade (Cat. 5).

These categories are displayed in Appendix A.3.

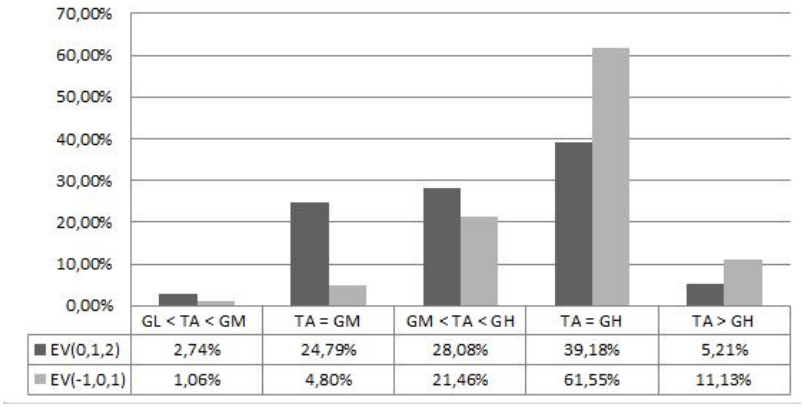
It is worth noting that, according to our interpretation, categories 1, 3 and 5 reflect the fact that the 3-level scale may be too short. We will see below that the results obtained under the broad scale EV21 do not contradict this (reasonable) explanation.

If all voters perceive two evaluative scales of equal length in the same manner, approval thresholds should have the same structure, as in the following hypothesis.

Hypothesis (H2.1 – No labeling effects for scales with same length) *Up to a 1-point translation, the two 3-level scales offered generate the same structure of approval thresholds.*

Figure 2 reports the empirical structure of approval thresholds as previously defined under scales $(0, 1, 2)$ and $(-1, 0, 1)$, i.e., the frequencies of each categories. The two distributions of approval thresholds look extremely different. Applying a χ^2 test to them, we can easily show that the distributions are

Figure 2 – Approval threshold in Saint-Etienne (EV(0,1,2)) and Louvigny (EV(-1,0,1))



	GL < TA < GM	TA = GM	GM < TA < GH	TA = GH	TA > GH
EV(0,1,2)	2,74%	24,79%	28,08%	39,18%	5,21%
EV(-1,0,1)	1,06%	4,80%	21,46%	61,55%	11,13%

significantly different at the 5% threshold. Hence we reject hypothesis H2.1: the structure of approval thresholds changes from one scale to the other.

Under scale $(-1, 0, 1)$, for a large majority of voters (72.68%, sum of $T_A > G_H$ and $T_A = G_H$ cases), the approval threshold coincides with the highest grade, which is here the only (strictly) positive grade. Under $(0, 1, 2)$, voters more frequently retain the middle grade (the lowest positive grade) as the approval threshold (52.87%, sum of $G_M < T_A < G_H$ and $T_A = G_M$ cases). We observe that voters behave as if they can only give a positive grade to approved candidates. When the scale reduces the range of positive grades, this automatically raises the scores. This focus on positive grades may also be seen as a reluctance to award negative and even zero grades to approved candidates. This implies that the grade scale strongly influences preference calibration.

4.2 Candidate-specific label effects

We now accept that, at a general level, the grade scale influences the behavior of the voters. The centered scale $(-1, 0, 1)$ tends to move the attributed grades up by comparison with the scale $(0, 1, 2)$. If this upward bias were uniform among candidates, this would not be much of a problem; although not linear, the transformation might even leave rankings unaffected. In the converse case, the voting results could be different. One may wonder whether the phenomenon affects all candidates in the same way. Hence the following hypothesis:

Hypothesis (H2.2 – No candidate-specific labeling effects) *Labeling bias affects uniformly the evaluation of all candidates, .*

Clearly, acceptance or refutation of this assumption presumes a comparison of the grade structure assigned to each candidate in $(0, 1, 2)$ and $(-1, 0, 1)$. In doing so, we must keep in mind that changes in the grade distribution result not only from the evolution of voter behavior, but also from the moving of the neutral grade from $G_L = 0$ with scale $(0, 1, 2)$ to $G_M = 0$ with the scale $(-1, 0, 1)$. In fact, the use of the same "zero" label to signify abstention with the two scales only affects the relative share of the low and the median grades. By contrast, a change in the frequency of the highest grade must result from voters ticking the box corresponding to the best grade in front of their favorite candidates' names. The consistency of the votes we have previously emphasized leads us to believe that this change in behavior does not owe anything to chance, but comes from a genuine reflection on the part of the voters.

Start by considering the evolution of average (normalized) scores obtained with the two 3-level scales (see table 2 and appendix for a presentation of average scores).

Table 2 – Evolution of average scores and of proportion of highest grade, and comparison of grade structure per candidate, from EV(0,1,2) to EV(-1,0,1)

Candidates	Score evolution	Significance at 5% threshold	Evolution of G_H prop.	Significance at 5% threshold	Significance of grade distribution difference (χ^2 test)
Hollande	+21 %	Significant	+10%	Not signif.	Significant
Sarkozy	+6.0 %	Not signif.	-3%	Not signif.	Not signif.
Le Pen	-2.5 %	Not signif.	-14%	Not signif.	Not signif.
Mélenchon	+35 %	Significant	+41%	Significant	Significant
Bayrou	+21 %	Significant	+15%	Not signif.	Significant
Joly	+80 %	Significant	+155%	Significant	Significant
Dupont -Aignan	+109 %	Significant	+113%	Significant	Significant
Poutou	+120 %	Significant	+67%	Significant	Significant
Arthaud	+130 %	Significant	+125%	Significant	Significant
Cheminade	+321 %	Significant	+631%	Significant	Significant

At the candidate level, Table 2 shows significant differences between candidates: some candidates' scores are multiplied by 4 whereas other candidates are endowed with equal (not significantly modified) normalized scores. We note that the two candidates whose score has not changed significantly, namely M. Le Pen and N. Sarkozy, have neither significant change in the proportion of the highest grade. This is also the case of the two candidates whose average score increased the least, namely F. Bayrou and F. Hollande.

We now need to look at the grade distribution for each candidate under the two 3-level scales. Figure 3 illustrates how the distribution of grades for each candidate varies from scale $(0, 1, 2)$ to scale $(-1, 0, 1)$. Table 8, which reports the numerical adjusted distributions of grades per candidate in these two 3-level scales, is presented in Appendix (A.4). In addition, Table 2 shows the results of χ^2 tests for the observed differences in structure.

Figure 3 – From (0,1,2) to (-1,0,1) grade distributions

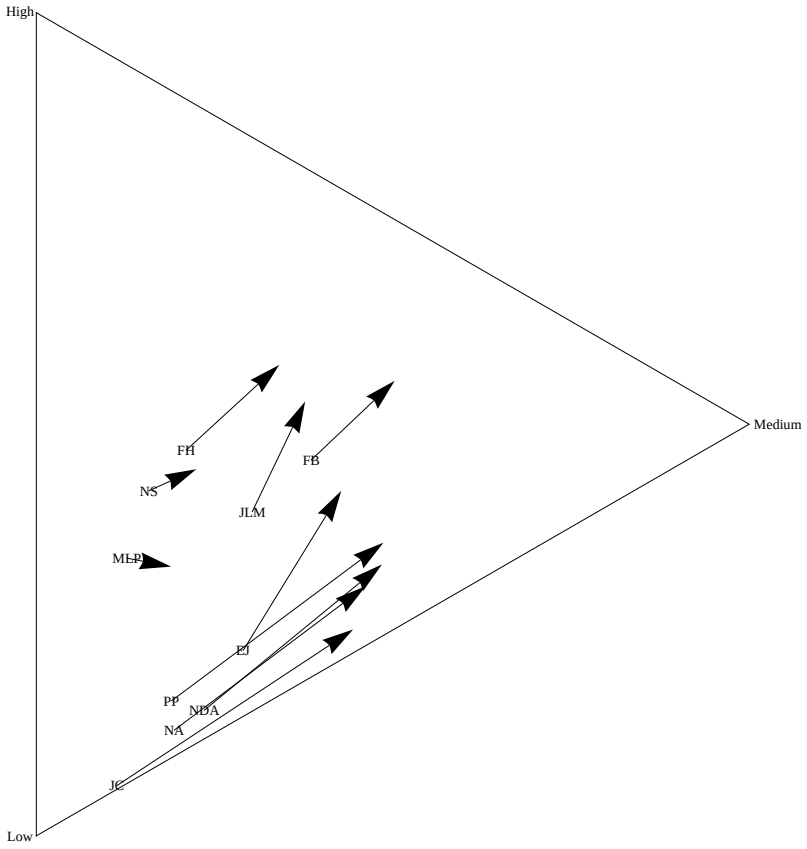


Figure 3 reads as follows. The statistical distribution of the grades received by a given candidate, in a 3-level scale is a vector of three positive numbers that sum to 1, hence a point in the two-dimensional simplex. The three apexes of the triangle are the degenerated grade profiles and respectively correspond to the Lowest (G_L), Middle (G_M) and Highest (G_H) grades. Notice, as a mathematical nicety, that with the triangle shown as it is, the height of a point on the Figure corresponds to the candidate aggregate score (for instance, 100% of middle grades yield the same score as a 50-50 mix of Highest and Lowest grades). The arrows in Figure 3 show how the grade profile of each candidate changes when going from the $(0, 1, 2)$ scale to the $(-1, 0, 1)$ scale.

An invariant grade structure from one scale to another requires arrows to be mere points. Without actually being reduced to a point, the arrows corresponding to two candidates, M. Le Pen and N. Sarkozy, are very short; the χ^2 test confirms that the distribution of their grades is not significantly different from one scale to the other. For all the other candidates, by contrast, the structure is significantly different.

Figure 3 conveys more information: the orientation and the length of an arrow indicate for each candidate the precise modification his or her grade profile. The arrows associated with the eight candidates other than M. Le Pen and N. Sarkozy show a shift of votes from the lowest grade to the median one, but also an increase in the share of the highest grade, except for F. Bayrou and F. Hollande, whose arrows are nearly perpendicular to the apex G_H.

At this point, we have sufficient evidence to refute the hypothesis H2.2. We have indeed shown that, for some candidates, the structure of grades was not significantly altered by the move from scale $(0, 1, 2)$ to scale $(-1, 0, 1)$, this not only being due to the "label zero" effect, since the share of the highest grade increases significantly.

To understand the origin of these differences in voter behavior we must refer to the characteristics of candidates. Baujard et al. (2014) showed that candidates fall into three categories according to how they are perceived by the electorate.

- "Exclusive" candidates, such as M. Le Pen and N. Sarkozy, are highly appreciated by a segment of voters, but they are clearly rejected by the rest of them.
- "Inclusive" candidates, like F. Bayrou, F. Hollande and J.-L. Mélenchon, are appreciated — yet not necessarily very much appreciated — by a large part of the electorate.
- "Small" candidates, such N. Arthaud, N. Dupont-Aignan, E. Joly, P. Poutou, J. Cheminade, are little known to voters.

As the χ^2 test confirms, exclusive candidates have the same grade distribution with both rules, hence unchanged average scores. Conversely, distributions are significantly different for inclusive and for small candidates. The average score increases slightly (from 21% to 35%) for inclusive candidates, and much more (by 80% to more than 300%) for small candidates (see Table 8).

Because they are not known to everyone, small candidates are often not rated by voters; the change in the structure of their rating may thus reflect the displacement of the label zero. But simultaneously their support is more

often characterized by the highest grade which voters see as the only way of awarding them "points".

In this election the inclusive candidates belong to big parties or are individually well known, and it is hard to imagine that very many participants would hold no opinion about them. However, we also see a shift from the lowest grade (0) to the median grade (0) for these candidates. Since these candidates rarely prompt a strong impulse for rejection among voters, we should probably treat this as a reluctance to use a negative grade. This result confirms the finding of Schwarz et al. (1991) in a different context: individuals are much less inclined to give negative grades because of a reluctance related to the strong symbolic content of such grades. However, the political context on which we here focus allows us to develop this general observation, showing that this reluctance varies greatly from candidate to candidate: some of them attract negative grades, others do not.

Labeling effects generate non-uniform transformations of electoral results over candidates when scales vary. These transformations typically depend on the kinds of candidate, implying that the presence of certain labels in the scale favor some candidates and disfavor others. Independently of the effect of the change in the normalized neutral grade, the presence of a negative grade (and of a unique positive grade) has two consequences. First, it favors inclusive candidates relative to exclusive ones. Second, it favors unknown candidates, or candidates with respect to whom voters are indifferent. This second effect means that a scale with negative grades could in extreme cases lead to an unknown candidate being elected, if voters would indeed continue to give the zero-grade, rather than the negative one, to the unknown candidates, simply because they have no negative feeling towards them. In our case this effect is not that extreme: indeed, “small” candidates remain in any case far from the top as their scores remain quite modest in absolute terms (See Table 8 in the Appendix for a comparison of normalized scores)..

However, two evolutions of the results following this change of scale are symptomatic of this risk, the first concerns the relative scores and the second the ranking of the candidates. As shown in Table 3, the candidates placed in the first and last positions are the same with both three-level scales. But if we refer to the standardized scores of Table 8, the score of the last candidate increases from 12.76% of the score of the first with scale (0,1,2) to 44.74% with scale (-1,0,1), which is a very substantial raise in his relative score. In addition, M. Le Pen who in an exclusive candidate, goes from 5th to 8th place, which means that three of the smaller candidates score better than her, as shown in Table 3. This exclusive candidate gets the highest grade (1) more often than any of these small candidates, but also gets much more often the lowest grade (-1).

Table 3 – Average scores and ranking, for the different voting rules.

Scale	AV		EV(0,1,2)		EV(-1,0,1)		EV21	
	Ave.	Rank	Ave.	Rank	Ave.	Rank	Ave.	Rank
Hollande	.49	1	.94	1	.14	1	9.45	1
Sarkozy	.40	2	.85	3	-.10	4	7.72	4
Bayrou	.39	3	.91	2	.11	2	7.94	2
Mélenchon	.39	4	.78	4	.06	3	7.91	3
Le Pen	.27	5	.67	5	-.34	8	4.94	6
Joly	.27	6	.46	6	-.17	5	6.8	5
Poutou	.13	7	.32	7	-.28	6	4.07	7
Dupont-Aignan	.11	8	.32	8	-.33	7	3.37	9
Arthaud	.08	9	.26	9	-.39	9	3.52	8
Cheminade	.03	10	.12	10	-.49	10	2.21	10

In these conditions, it is relevant to consider whether a higher value associated with the highest grade could override the distortion caused by the existence of the negative minimum grade and, if so, what this value would be.[7] In other words, is there a x such that the grade distribution observed on the scale (-1,0,1) would give with the scale (-1,0, x) the same ranking as the one obtained with the scale (0,1,2)? The answer is the following. It is necessary that x is at least equal to 7.5 for M. Le Pen to reach a score higher than that of all the small candidates. But the ranking thus obtained (F. Hollande; J.-L. Mélenchon; N. Sarkozy; F. Bayrou; M. Le Pen; E. Joly; Ph. Poutou; N. Dupont-Aignan; N. Arthaud; J. Cheminade) is not identical to that resulting from the scale (0,1,2). Notably, F. Bayrou, who is in 2nd position with both 3-level scales, is downgraded by two places for $x = 7.5$. With this value of x, this inclusive candidate benefits less from the feeling of low rejection but moderate support that voters have towards him. The impossibility to realign the rankings resulting from the two scales is a further proof of the existence of a candidate-specific labeling effect.

5 Comparing scales of various length

To focus on the effects induced by their length, we first compare the scales $(0, 1, 2)$ and $(0, ..., 20)$, and do not here consider the zero-centered scale $(-1, 0, 1)$, in order to avoid the distortion due to 0 being the default grade.

Table 3, which reports candidates' scores under the different rules, shows significant differences in observed rankings. Candidates located at ranks 1, 2, 7 and 10 are the same. By contrast, the candidates of ranks 3, 4 are

[7] We thak a referee for this suggestion

permutated, and the same holds for ranks 5 and 6 and for ranks 8 and 9. These changes seem to capture a scale length effect upon the expression of voters' preferences. The differences are however too weak to accept without further investigation of whether they are statistically significant. To that end, we reduce the 21-level scale to a 3-level scale by merging them into the three evenly-spaced categories $(0, ..., 6)$, $(7, ..., 13)$, and $(14, ..., 20)$. On the basis of the new reconstructed distribution, we are now able to test the third following hypothesis:

Hypothesis (H3 – Invariance with length) *Numerical scales of different lengths are equivalent.*

5.1 No difference at the aggregate level

We first consider aggregate differences. Two approaches may be considered to test how a larger scale may affect aggregate results. After reducing the 21-level scale to three levels, the first approach directly compares the new grade distribution with the results obtained with the actual $(0, 1, 2)$-scale. The second approach considers the approval thresholds determined after bundling the 21-level scale into three levels only, and compares them with the approval thresholds obtained with the actual 3-level scale. This second approach allows us to extend the analysis to the zero-centered scale.

Following the first approach, the grade distribution attributed by voters for all candidates according to both scales is presented in Table 4.

Table 4 – Aggregate grade distribution (%)

Voting rules	G_L	G_M	G_H
EV(0,1,2)	60.7	22.25	17.05
Reduced EV21	59.72	23.25	17.03

A first version of hypothesis H3 may be formulated as follows:

Hypothesis (H3.1 – Invariance of grade distributions with length) *Once reduced to the same length, the grade distributions in scales of different lengths are equivalent.*

At the 5% threshold, the distribution obtained with both grade scales are not significantly different, so that we do not refute hypothesis H3.1. Although there are some ranking inversions, the first approach seems to show that scale length does not matter much. Notice that this does not imply that the grades given by a voter with scale (0, ..., 20) are uniformly distributed within each of the three grade intervals. The data clearly show that some grades are more

frequently used, probably because of their particular symbolic significance. As shown in Table 5, the most frequent ones are: grade 0 (41.77%), which is the default grade in both scales; grade 5 (5.94%); grade 10 (9.91%); and grade 15 (5.61%). These modal frequencies do not however really distort the distribution reduced to three levels.

Table 5 provides further insight regarding the usefulness of such a large scale. Indeed, 66.71% of the grades were symbolic grades (0, 5, 10, 15 or 20). At the 5% threshold, the distribution obtained under EV21 is significantly non-homogeneous. But one can wonder whether beyond these five symbolic grades the distribution of the 16 remaining ones becomes homogeneous or not. Indeed, participants might want to use intermediate grades in order to stress the difference from the symbolic grades they use for most of the candidates. In this case, the grades 1,4,6,9,11,14,16 and 19 should be less used than the grades 2,3,7,8,12,13,17 and 18 which are farther from the five symbolic grades. Even if this scenario is not validated at the 5% threshold, one can point out that between each symbolic grade there seems to be one intermediate grade which attracts a slightly higher number of participants such as grades 2, 8, 12 and 18. Compared to a 21-level scale, one could thus claim that a 9-level scale would be more satisfying for most voters.

Let us now turn to the second approach. Regarding EV21, we compute all the lowest grades given to approved candidates before we map the approval thresholds to three levels.

For instance, the highest threshold G_H is assigned to ballots for which the lowest grade given to an approved candidate is between 14 and 20. Regarding scale $(0, 1, 2)$, this threshold is given to ballots for which all approved candidates received grade 2, whether some non approved candidates received or not the same grade. As explained previously, the distribution of approval thresholds is not affected by the shifting of the neutral grade; we can reintroduce the comparison with the scale $(-1, 0, 1)$.

Before comparing approval thresholds, notice that, with both 3-level scales $(-1, 0, 1)$ and $(0, 1, 2)$, almost a third of the consistent ballots award the same grade to some approved and non-approved candidates. This phenomenon may result from the fact that the scale length was not fine grained enough for voters to express precisely their political opinion on all candidates, and the corresponding ranking. In this perspective, this behavior should disappear under the 21-level scale. However, although these cases become much rarer, they still represent 9.06% of the consistent ballots under the 21-level scale.

The approval thresholds are displayed in figure 4. Hypothesis 3.2 can be tested by considering the threshold distributions:

Hypothesis (H3.2 – No length effects for approval thresholds) *Reduced*

Table 5 – Average number of grades per ballot – EV21 (%)

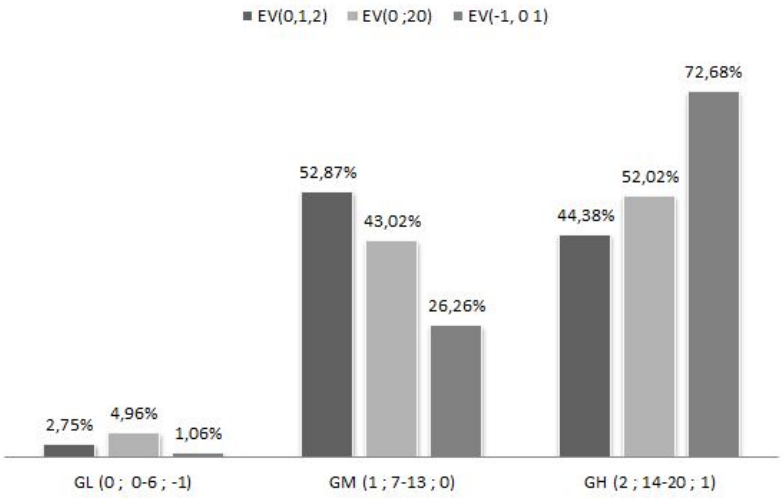
Grades	Av. number (%)
0	41.77
1	2.86
2	3.65
3	1.61
4	1.30
5	5.94
6	1.78
7	1.65
8	3.00
9	2.03
10	9.91
11	1.82
12	3.87
13	1.85
14	2.55
15	5.61
16	1.61
17	1.52
18	2.13
19	0.63
20	2.94

to the same length, both distributions of approval thresholds are equivalent.

At the 5% statistical threshold, the distributions of approval thresholds for $(0, 1, 2)$ and $(0, ..., 20)$ are not significantly different the one from the another, although they are significantly different from the distribution obtained under scale $(-1, 0, 1)$. We have seen above that the presence of a negative grade with scale $(-1, 0, 1)$ distorts the distribution of approval thresholds relative to scale $(0, 1, 2)$. On the other hand, we now know that in terms of approval thresholds, $(0, 1, 2)$ and $(0, ..., 20)$ do not generate significantly different structures. So it is not surprising that scale $(-1, 0, 1)$ and $(0, ..., 20)$ do not generate the same distribution of approval thresholds. Hence we accept hypothesis H3.2: the increase of scale length does not affect evaluation voting at the aggregate level.

We now test the absence of length effect by candidates.

Figure 4 – Approval thresholds in 3 classes, a comparison of EV(0,1,2), EV21 and EV(-1,0,1)



5.2 Few differences at the candidate level

Table 6 displays for each candidate the grade distribution observed for scale $(0, 1, 2)$ and the reduced grade distribution for scale $(0, ..., 20)$.

Regarding the absence of length effect for each candidate we state the following hypothesis:

Hypothesis (H3.3 – No candidate-specific length effects) *Scale length does not affect the evaluation of any candidate.*

A first observation is that for most candidates, the grade distributions are not significantly different from one scale to another. At the 5% threshold, the grade distributions are significantly different from one scale to another for only two candidates. These candidates are N. Sarkozy and E. Joly. N. Sarkozy is given a similar proportion of lowest grades; however he is rewarded less often with the highest grades. E. Joly is given fewer lowest grades and more highest grades in the reduced version of $(0, ..., 20)$ than with $(0, 1, 2)$. The characteristics of these two candidates may fully explain the observed differences.

The interpretation of the results obtained with scale $(0, ..., 20)$ might be linked to their use in schools, since this is the scale used to evaluate pupils in the French school system: 10 often is the minimum grade to pass an exam. Rewarding a candidate with any grade above 10 in this perspective means supporting this candidate, but the scale from 10 to 20 also allows more fine grained evaluation to be captured. As N. Sarkozy was the incumbent

Table 6 – Grade distribution per candidates (%)

Candidates	Voting rules	G_L	G_M	G_H
F. Hollande	EV(0,1,2)	42.21	21.51	36.28
	Reduced EV21	35.42	27.51	37.07
F. Bayrou	EV(0,1,2)	34.86	39.01	26.13
	Reduced EV21	41.38	35.9	22.72
J.-L. Mélenchon	EV(0,1,2)	45.89	30.11	24.00
	Reduced EV21	45.73	27.62	26.65
N. Sarkozy	EV(0,1,2)	49.92	15.6	34.48
	Reduced EV21	48.37	24.57	27.06
E. Joly	EV(0,1,2)	62.38	29.65	7.97
	Reduced EV21	51.54	30.09	17.55
M. Le Pen	EV(0,1,2)	59.67	13.32	27.01
	Reduced EV21	68.61	12.68	18.71
P. Poutou	EV(0,1,2)	74.35	18.83	6.82
	Reduced EV21	71.16	20.91	7.93
N. Arthaud	EV(0,1,2)	76.88	19.97	3.15
	Reduced EV21	74.71	20.43	4.86
N. Dupont-Aignan	EV(0,1,2)	72.53	23.25	4.22
	Reduced EV21	76.93	16.43	6.64
J. Cheminade	EV(0,1,2)	88.33	11.22	0.45
	Reduced EV21	83.33	15.52	1.15

president, grades higher but close to 10 are able to capture some of the dissatisfaction usually associated with governing fatigue, and mapped into 1 after scale reduction. Conversely, E. Joly, as a green, candidate seems to represent ideas that are considered as interesting by many voters who would however not convert their interest into a vote. Regarding the relevance of the school grading system, this led to many grades being below, but close to 10. The reason for this is that her middle grades are often translated into 1 (rather than 0) in the reduced scale. For similar reasons, E. Joly was not often rewarded with grades close to 20, but her highest grades have been be transformed into 2 rather than 1.

Differences between these two candidates should not however be over-interpreted, since for both of them, they become insignificant at the 1% threshold. We conclude we can hardly reject H3.3. A change in the scale length does not create a major bias in the grade distribution.

6 Concluding remarks

The data from the *in situ* experiment conducted in the context of the French Presidential election in 2012 allow us to establish that, when confronted with different grade scales in evaluative voting, voters rank the candidates almost without inconsistencies; but that they grade them in different manners, and this may generate different electoral results. Electoral preferences are indeed calibrated - i.e. translated into numerical grades - differently under the different grade scales. This paper disentangles these calibration effects and shows how they work. It discusses when this is worrying or not; it shows how different scales may be associated with different properties.

We first observed the small number of inconsistent ballots: ballots for which the underlying approval ranking and the underlying evaluation ranking are inconsistent. This observation is in accordance with choice theory where underlying preferences are standardly represented as pre-orders. In the context of a field experiment, where mistakes due to lack of time or other factors could be expected, this striking feature lends confidence, especially for the 3-level scales.

Secondly, focusing on the two observed 3-level scales, we established that voters use $(-1, 0, 1)$ and $(0, 1, 2)$ in significantly different ways. The $(0, 1, 2)$ scale leads voters to assign a zero either to candidates they know well but reject, and to "unknown" candidates; the scale $(-1, 0, 1)$ often prompts them to award -1 to rejected candidates, and the zero grade to unknown candidates who do not inspire the same feeling. Likewise, the two positive grades 1 and 2 in the scale $(0, 1, 2)$ offer an opportunity that voters seize: to distinguish strength of support. And as they stated in our questionnaire, participants in the experiment appreciated these additional expressive opportunities. Conversely, in $(-1, 0, 1)$, voters often feel the strictly positive grade as the only way to support a candidate. We have shown that the bias created by negative grades is non-uniform across candidates, but is well identified: this confirms a previous result (Baujard et al. 2014), and explains why the scale $(-1, 0, 1)$ disfavors exclusive candidates, and favors unknown candidates and inclusive ones.

A third set of observations is based on the comparison between the scales $(0, 1, 2)$ and $(0, ..., 20)$. The proportion of inconsistent ballots (with respect to approvals) is significantly higher for the long scale than for the 3-level scale. This reveals that the longer scale is more sensitive for voters. We then compared the two scales by merging into just three grades the 21 grades $0, ..., 20$, and we found relatively similar results. Changing the scale length does not generate significant calibration bias.

One uncontroversial discovery of this research is that, as far as scales matter,

evaluation voting should not be used without further tests. A tentative conclusion is that the main difference between variants of evaluative voting is the availability, or not, of negative grades, and the treatment of the neutral grade. Further research should test more precisely the hypothesis that we offer as a close: evaluative voting with non-negative grades is robust to variations in the precise scale in use.

References

Arrow, K. J. *Social Choice and Individual Values.* John Wiley & Sons, New York, 1951.

Baujard, A., Gavrel, F., Igersheim, H., Laslier, J.-F., and Lebon, I. Approval voting, evaluation voting. an experiment during the 2012 French presidential election. *Revue Economique 64*, 2 (2013), 345–356.

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

Brams, S. J., and Fishburn, P. C. Approval voting. *American Political Science Review 72* (1978), 831–847.

D'Aspremont, C., and Gevers, L. Equity and the informational basis of social choice. *Review of Economic Studies 44* (1977), 199–209.

Debreu, G. Topological methods in cardinal utility theory. In *Mathematical Methods in the Social Sciences.* Stanford University Press, 1960, pp. 16–26.

Farrell, D. *Comparing Electoral Systems.* Palgrave, Besingtoke, 2001.

Gaertner, W., and Xu, Y. A general scoring rule. *Mathematical Social Sciences 63*, 3 (2012), 193–196.

Garman, S. Election frequency, choice fatigue, and voter turnout. *European Journal of Political Economy 47* (2017), 19–36.

Gorman, W. M. The structure of utility functions. *The Review of Economic Studies 35* (1968), 367–390.

Grofman, B., Dolez, B., and Laurent, A. *In Situ and Laboratory Experiments on Electoral Law Reform: French Presidential Elections.* Springer, Heidelberg, 2011.

Hillinger, C. On the possibility of democracy and rational collective choice. Discussion paper 2004-21, Department of Economics, University of Munich, 2004.

Hillinger, C. Voting and the cardinal aggregation of judgments. Discussion paper 2004-09, Department of Economics, University of Munich, 2004.

Igersheim, H., Baujard, A., Lebon, I., Laslier, J.-F., and Gavrel, F. Individual behavior under evaluative voting. a comparison between laboratory and in situ experiment. In *Voting Experiments*, A. Blais, J.-F. Laslier, and K. van der Straeten, Eds. Springer, 2016, pp. 257–269.

Lachat, R., Laslier, J.-F., and Van der Straeten, K. Strategic voting in multi-winner elections with approval balloting: An application to the 2011 regional government election in Zurich. In *Strategic voting*, A. Blais, Ed. Springer, 2017, ch. 6.

Laslier, J.-F. The leader rule: a model of strategic approval voting in a large electorate. *Journal of Theoretical Politics 21*, 1 (2009), 113–136.

Laslier, J.-F., Blais, A., Bol, D., Golder, S. N., Harfst, P., Stephenson, L. B., and der Straeten, K. V. The Euro Vote Plus experiment. *European Union Politics 16*, 4 (2015), 601–615.

Lethinen, A. The welfare consequences of strategic behaviour under approval and plurality voting. *European Journal of Political Economy 24*, 3 (2008), 688–704.

Macé, A. Voting with evaluations: When should we sum, what should we sum? Working paper, Ecole Polytechnique, Paris, 2015.

Maskin, E. A theorem on utilitarianism. *The Review of Economic Studies 45* (1978), 93–96.

Pivato, M. Formal utilitarianism and range voting. *Mathematical Social Sciences 67* (2013), 50–56.

Regenwetter, M., Dana, J., and Davis-Stober, C. P. Transitivity of preferences. *Psychological Review 118*, 1 (2011), 42–56.

Schnellenbach, J. Behavioral political economy: A survey. *European Journal of Political Economy 40* (2015), 395–417.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L. Rating scales: Numeric values may change the meaning of scale labels. *The Public Opinion Quarterly 55*, 4 (1991), 570–582.

Sen, A. K. Social choice theory: A re-examination. *Econometrica 45*, 1 (1977), 53–89.

Smaoui, H., and Lepelley, D. Le système de note à trois niveaux: Etude d'un nouveau mode de scrutin. *Revue d'économie politique 123*, 6 (2013), 827–850.

Van der Straeten, K., Laslier, J.-F., Sauger, N., and Blais, A. Strategic, sincere, and heuristic voting under four election rules: an experimental study? *Social Choice and Welfare 35*, 3 (2010), 435–472.

Wakker, P. P. *Additive Representation of Preferences: A New Foundation of Decision Analysis*. Kluwer, Dordrecht, 1989.

Yilmaz, M. R. Can we improve upon approval voting? *European Journal of Political Economy 15*, 1 (1999), 89–100.

A Appendix

A.1 Weighting of data

The five 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. We obtained 1294 usable answers to this question. The analysis is restricted to these voters; in addition, 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. An extensive explanation of this treatment can be found in Baujard et al. (2013).

Table 7 – Weights used for data adjustment

Candidates	Louvigny EV(-1, 0, 1)	Saint-Etienne EV(0, 1, 2)	Strasbourg EV(0, ... , 20)
E. Joly	0.42	1.43	0.29
M. Le Pen	3.62	1.23	4.67
N. Sarkozy	1.88	1.44	2.12
J.-L. Mélenchon	0.77	0.63	0.6
P. Poutou	1.12	0.71	1.5
N. Arthaud	3.31	1.05	2.95
J. Cheminade	0.73	-	0.43
F. Bayrou	0.57	0.89	0.72
N. Dupont-Aignan	1.32	1.67	1.87
F. Hollande	0.69	0.85	0.69

A.2 Scrutiny of all cases of inconsistencies

Some ballots, for instance, just select one candidate for approval and/or evaluation. This can still involve inconsistencies when one candidate is approved in AV, while another is given a better grade. This is the case for three ballots in Saint-Etienne and two ballots in Louvigny. There may be some palpable mistakes due to difficulties in completing the forms: line shifts or inversions of columns, and inversions between the higher grade column and the lower grade under EV3. This is the case for one ballot in Saint-Etienne and two in Louvigny. There are also cases of apparent misunderstandings of the rules,

and particularly of the concept of approval. In three ballots in Louvigny, all candidates were approved, but the three grades are still used to assess a subset of these candidates. A harder case concerns 6 ballots in Louvigny which approve a subset of candidates and assess those exclusively with the whole range of grades. The 21-level EV scale multiplies the number of AV/EV combinations which can induce inversions. Unsurprisingly, we find more inconsistencies with this rule. The complexity of the $(0, ..., 20)$ scale for individuals may also explain why the share of inconsistent ballots, almost 9 %, is three times higher than under EV3. A more careful analysis of these inconsistent pairs of ballots, however, shows mild inconsistencies, i.e., when voters give a slightly better evaluation to a non-approved candidate than to an approved one. The difficulty in grading candidates according to such a fine scale as $(0, ..., 20)$ was clearly expressed by participants in oral or written remarks.

A.3 Classification of approval thresholds

When identifying the structure of approval thresholds T_A under a 3-level scale, we are led to distinguish among (at least) five categories of consistent ballots. Among these categories, we distinguish two meta-categories. In some ballots, a grade G is a threshold, in the sense that all non-approved candidates have a lower grade, and all approved candidates are given this grade or higher. In such a case, we write $T_A = G$. In other ballots, some non-approved candidates may be given the same grade G than some approved candidates. These ballots are still consistent in the sense we defined above. Yet the approval threshold is not exactly a grade, but somewhere between two successive grades. In this category of ballots we retain the following convention. Among the grades assigned to approved candidates we shall specify the approval threshold by the lowest relevant bound on the grade scale, such that $T_A > G$.

The five categories can be defined and represented as follows:

1. In the first category of ballots, all approved candidates receive the higher grade while at least one non-approved candidate also receives this grade. Notice that the approval threshold is G_H, but the voter approves only some of the candidates who obtain G_H. This category of ballot is referred to as $T_A > G_H$, and illustrated by Figure 5.

2. In the second category of ballots, all approved candidates receive the higher grade while all non-approved candidates receive lower grades. In other words, the threshold of approval is equal to the higher grade. This category of ballots (2) is referred to as $T_A = G_H$, and illustrated by Figure 6.

Figure 5 – Category (1) $T_A > G_H$

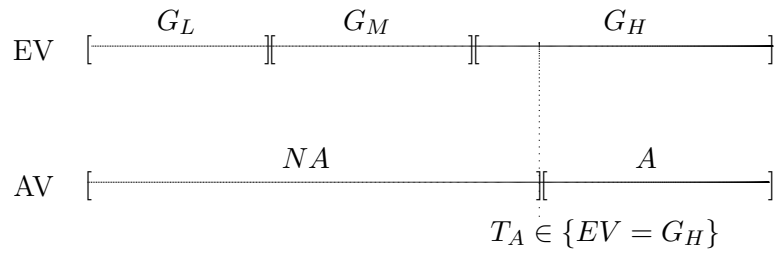
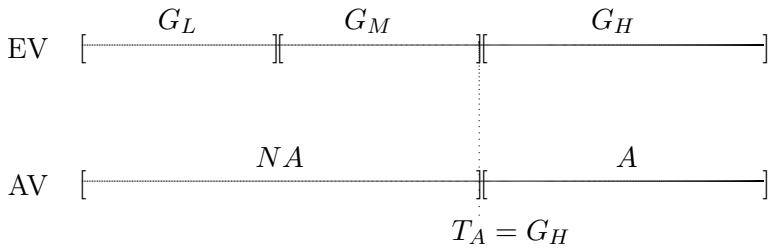


Figure 6 – Category (2) $T_A = G_H$



3. In the third category of ballots, all approved candidates receive either the middle grade or the higher grade, while at least one non-approved candidate also receives the middle grade. The approval threshold is equal to G_M, but some candidates who obtain this grade are not approved. This category of ballots is referred to as $G_M < T_A < G_H$, and illustrated by Figure 7.

4. In the fourth category of ballots, all approved candidates receive either the middle grade or the higher grade, while all non-approved candidates receive the lower grade. This category of ballots is referred to as $T_A = G_M$ and illustrated by Figure 8.

5. In the fifth category of ballots, non-approved candidates all obtain the lower grade. At least one approved candidate obtains the lower grade. Because the approval threshold coincides with the lower grade (G_L) but some non-approved candidates also obtain this grade, this category (5) of ballot is referred to as $G_L < T_A < G_M$, and illustrated by Figure 9.

Notice that the case $T_A = G_L$ was not observed. As illustrated by Figures 5 to 9, in the first two categories, approval is restricted to candidates who

Figure 7 – Category (3) $G_M < T_A < G_H$

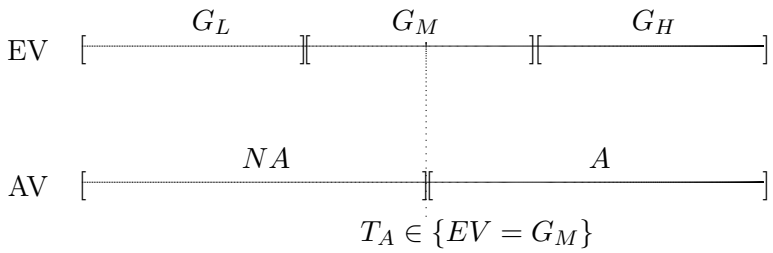
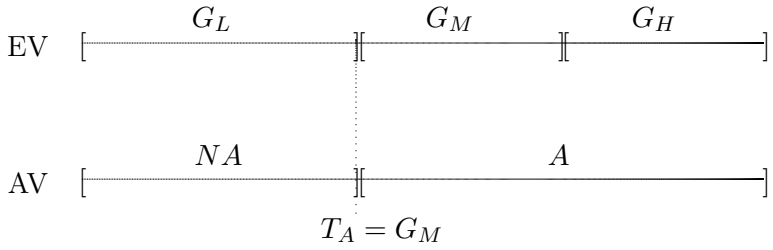


Figure 8 – Category (4) $T_A = G_M$



receive the maximum grade. In the two following cases, some approved candidates obtain a middle grade, while in case 5, some of lower rated candidates are approved.

A.4 Adjusted distribution of grades in 3-level scales

Table 8 describes the adjusted distributions of grades per candidates in each 3-level scale. In each scale, G_L, G_M, G_H is respectively the lowest, the middle and the highest grade. It also presents the average score of each candidate, which is normalized to the scale $(0, 1, 2)$.

Figure 9 – Category (5) $G_L < T_A < G_M$

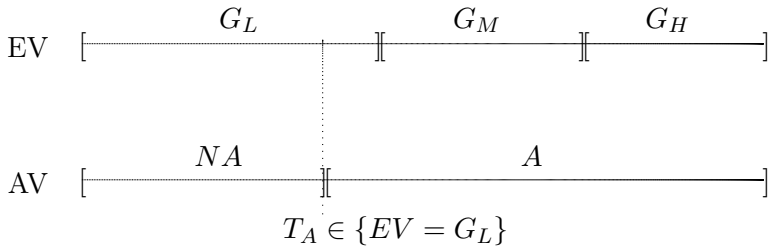


Table 8 – Grade distribution per candidates and average normalized score (%). Comparing 3-level scales.

Candidates	Voting rules	G_L	G_M	G_H	Average normalized score
F. Hollande	EV(0,1,2)	42.21	21.51	36.28	0.94
	EV(-1,0,1)	25.71	34.53	39.76	1.14
N. Sarkozy	EV(0,1,2)	49.92	15.6	34.48	0.85
	EV(-1,0,1)	43.75	22.88	33.37	0.90
M. Le Pen	EV(0,1,2)	59.67	13.32	27.01	0.67
	EV(-1,0,1)	57.44	19.45	23.11	0.66
J.-L. Mélenchon	EV(0,1,2)	45.89	30.11	24.00	0.78
	EV(-1,0,1)	27.99	38.24	33.77	1.06
F. Bayrou	EV(0,1,2)	34.86	39.01	26.13	0.91
	EV(-1,0,1)	17.38	50.62	30.00	1.11
E. Joly	EV(0,1,2)	62.38	29.65	7.97	0.46
	EV(-1,0,1)	36.91	42.71	20.35	0.83
N. Dupont-Aignan	EV(0,1,2)	72.53	23.25	4.22	0.32
	EV(-1,0,1)	42.62	43.38	9.00	0.66
P. Poutou	EV(0,1,2)	74.35	18.83	6.82	0.32
	EV(-1,0,1)	39.75	48.84	11.41	0.72
N. Arthaud	EV(0,1,2)	76.88	19.97	3.15	0.26
	EV(-1,0,1)	46.58	46.32	7.1	0.61
J. Cheminade	EV(0,1,2)	88.33	11.22	0.45	0.12
	EV(-1,0,1)	52.27	44.44	3.29	0.51