ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 045

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch Amelie Schiprowski

October 2021

www.econtribute.de

UNIVERSITÄT BONN
Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1– 390838866 is gratefully acknowledged.

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch[†] Amelie Schiprowski[§]

October 18, 2021

Abstract

Interviewing is a decisive stage of most processes that match candidates to firms or organizations. This paper studies how and why the interview assessment of a candidate depends on the other candidates seen by the same evaluator. We leverage novel administrative data covering about 29,000 one-to-one interviews from a study grant admission process where candidates are quasi-randomly assigned to evaluators and time slots. We find that a candidate's assessment decreases in the quality of the other candidates seen by the same evaluator, and most strikingly in the quality of the previous candidate. This effect is strongest when candidates are similar in terms of their observable characteristics. The reduced-form patterns and the results of a structural estimation suggest that evaluators exhibit a contrast effect which is caused by the interplay between the associative recall of prior candidates and the attention to salient quality differences.

JEL Codes: D91, M51

[†] Institute of Labor Economics (IZA), Schaumburg-Lippe Str, 5-7, D-53113 Bonn. Email: radbruch@iza.org

[§] University of Bonn, Adenauerallee 24-42, 53113 Bonn, Germany. Email: amelie.schiprowski@uni-bonn.de

The main specifications and variable definitions in this project were pre-registered under osf.io/t65zq. We thank Johannes Abeler, Steffen Altmann, Maria Balgova, Pedro Bordalo, Stefano DellaVigna, Markus Dertwinkel-Kalt, Thomas Dohmen, Armin Falk, Andreas Grunewald, Lena Janys, Andreas Klümper, Michael Kosfeld, Danielle Li, Andreas Lichter, George Loewenstein, Sebastian Schaube, Andrei Shleifer, Florian Zimmermann and Ulf Zoelitz for helpful discussions and comments. The paper further benefited from feedback at several seminars and conferences. We thank the study grant organization for the data provision and for numerous fruitful discussions. Julia Wilhelm, Stefanie Steffans, and especially Annica Gehlen provided outstanding research assistance. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1– 390838866 and CRC TR 224.

1 Introduction

Subjective assessments are commonly used to measure quality or performance in high-stakes situations. Examples include the evaluation of employees, the screening of applicants or the grading of students. Given that subjective assessments can have long-lasting consequences for individual life outcomes, it is important to understand their underlying formation.

One context where subjective assessments are prevalent is interviewing, which is a decisive stage of most processes that match candidates to firms and organizations. A core feature of interviews is that the single assessment does not occur in social isolation, as evaluators usually observe several candidates in a sequence. This feature provides the opportunity to learn about the expected quality of the available candidate pool. At the same time, processing sequential information is prone to errors. For instance, evaluators may judge a candidate in light of prior interview experiences. The quality of recently interviewed candidates can thereby have a direct negative spillover on the assessment of the current candidate. This phenomenon, which is commonly referred to as 'contrast effect' (e.g., Pepitone & DiNubile, 1976; Simonson & Tversky, 1992; Bhargava & Fisman, 2014), bears the potential to distort interview assessments. As a result, it might induce firms and organizations to hire or admit the wrong candidates.

In this paper, we provide large-scale evidence on the interdependence of candidate assessments in a real-world interview setting. We study how and why the assessment of a candidate depends on the quality of the other candidates interviewed by the same evaluator. Our analysis focuses on three questions. First, we analyze how the assessment of a candidate changes if the quality of another candidate increases. Second, we ask how the effect depends on that candidate's relative position in the interview sequence. Having identified an over-proportional negative spillover from the previous candidate, we thirdly investigate the behavioral mechanism. Guided by theoretical insights (Bordalo et al., 2020), we study how the interplay between the evaluator's memory and attention generates contrasting against the previous candidate. More specifically, we empirically assess conjectures of the theoretical framework and structurally estimate the underlying model.

The analysis relies on novel administrative data covering about 29,000 interviews from a

study grant admission process with high stakes. The process is organized through assessment center style admission workshops. Every workshop has a committee of eight evaluators, who each conduct about twelve one-to-one interviews over a period of two days. Three main features make this setup ideal to study how candidates influence each others' assessments: first, candidates are quasi-randomly assigned to evaluators and time slots; second, each candidate has a clearly defined reference group, as evaluators observe a closed sequence of candidates; and third, each candidate receives three as-good-as independent assessments, which facilitates the measurement of otherwise unobserved candidate quality.

Exploiting the quasi-random assignment and ordering of candidates, we estimate how the assessment of a candidate changes if the measured quality of another candidate in the same interview sequence increases. We proxy a candidate's unobserved quality through an independent third-party assessment (TPA). More specifically, the TPA is defined as the sum of two independent ratings made by other evaluators. To address issues related to multiple hypothesis testing, selective data-slicing and the arbitrary definition of candidate quality, we pre-registered the main specifications and variable definitions used in the empirical analysis.[1]

The results show that the same candidate is evaluated worse when assigned to an interview sequence with better candidates. Both previously and subsequently observed candidates have a similar negative influence. A striking exception to the overall pattern is the previous candidate, whose influence is about three times stronger than that of the average other candidate. As a consequence of the previous candidate's influence, the evaluators' votes exhibit a strong negative autocorrelation. We find that an evaluator who votes in favor of admitting a candidate observed in period $t-1$ is about 6 percentage points less likely to vote in favor of the candidate observed in period t (16% relative to the mean). The effect also translates into changes in the relative ranking of candidates.

We investigate the channel underlying the previous candidate's striking negative influence. The discussion is guided by theoretical insights from Bordalo et al. (2020), who show how the interplay between associative memory and attention can generate contrast effects. In this

[1] The pre-registration can be found at osf.io/t65zq. Prior to pre-registration, we had access to a pilot dataset, which is not included in the analyses for this paper.

framework, candidates are evaluated against a quality norm. More precisely, the attention of evaluators is attracted to salient differences between the candidate's quality and the norm. The norm is formed through the associative recall of previously experienced candidates. Associative recall retrieves prior interview experiences from memory. The process is associative because it places a higher weight on more similar experiences. As time generates (superficial) similarity between two interviews, the previous candidate's quality receives a strong weight in the norm.

We link the data to the framework in two ways: through reduced-form evidence and through a structural estimation. The reduced-form analysis assesses insights from the framework regarding the incidence and strength of contrast effects under different circumstances. First, we find that breaks, which reduce the similarity in time between two subsequent interviews, decrease the previous candidate's influence. Second, additional dimensions of similarity influence the intensity of sequential contrasting. The previous candidate has a stronger influence if she looks more similar in terms of observable characteristics, such as gender, socio-economic background or study field. More precisely, additional similarity matters in relative terms: the candidate in $t-1$ has the highest (lowest) influence when she looks more (less) similar to the current candidate than the candidate in $t-2$. Third, results show that the previous candidate's influence weakens over the interview sequence, as the evaluator's memory database of experienced candidates expands. Finally, they confirm that only large and salient differences between current and prior candidate quality lead to contrasting.

In a further step, we structurally estimate the model to assess its quantitative fit with the data. In particular, we estimate the evaluators' recall process and the weights that previous candidates receive in the quality norm. Estimation is based on the method of simulated moments. The results show that a simple parameterization of the recall process is able to provide quantitatively meaningful predictions regarding the influence of previous candidates. With respect to the salience of quality differences, the estimates provide overall a good fit with the empirical pattern, but slightly over-predict the effect of very large quality differences. Finally, we use the estimates to calculate counterfactual scenarios. We estimate that about 15% of candidates would get a different rating if the recall process played no role and all evaluators con-

trasted candidates against an objective quality norm. Moreover, about 8% of evaluators would rank a different candidate as the best candidate in their sequence.

While the proposed mechanism is both qualitatively and quantitatively in line with the data, we acknowledge and assess the possibility of alternative explanations. Most importantly, previous evidence by Chen et al. (2016) suggests that a negative autocorrelation in decisions may stem from a gambler's fallacy. Adapted to our setting, evaluators would underestimate the probability that two candidates of similar quality follow each other. For several reasons, it is unlikely that a gambler's fallacy explains the previous candidate's influence. First, we do not find that the negative autocorrelation in votes increases after a 'streak' of more than one yes vote. Second, outcomes are still related to a quality measure of the previous candidate once we condition on the previous binary decision. Both findings are not consistent with a simple gambler's fallacy model in which decision makers expect binary reversals. Moreover, a key distinction between a gambler's fallacy and a contrast effect lies in their timing. The gambler's fallacy changes the prior belief about the next candidate, whereas the contrast effect occurs only when observing the next candidate. We find that the influence of the previous candidate depends on the quality difference to the next candidate. This makes it unlikely that the effect works through prior beliefs.

The results of this paper imply that minor changes in relative candidate ordering can have a major impact on assessments. This has relevant implications for many hiring and admission situations — the economics job market being only one among many settings where candidates are assessed through sequential interviews. Despite the strategic importance of hiring and admission decisions for firms and organizations, only scarce evidence exists on the underlying screening process (Oyer & Schaefer, 2011).[2] In particular, little is known about the formation of subjective assessments through personal interviews. We contribute by documenting dis-

[2] Previous studies on candidate screening have, for example, studied the determinants of callback rates (e.g., Bertrand & Mullainathan, 2004; Kroft et al., 2013), the impact of algorithmic recommendations (Horton, 2017; Bergman et al., 2020), and the influence of job-testing technologies (Autor & Scarborough, 2008; Hoffman et al., 2018; Estrada, 2019). Moreover, Simonsohn and Gino (2013) show that interviewers are prone to narrow bracketing. More broadly related, the literature has documented sources of errors in subjective assessments. For example, Ginsburgh and van Ours (2003) show that a pianist's absolute order of appearance matters for her assessment in a piano competition and Li (2017) estimates the influence of bias versus expertise when evaluators assess grant proposals in their own field.

tortions in candidate assessments and rankings that can arise due to the sequential nature of interviewing.

We also contribute to the literature on path dependence in real-world decision making, and on the incidence of contrast effects in particular. Existing evidence on contrast effects stems from the contexts of renting (Simonsohn & Loewenstein, 2006; Simonsohn, 2006; Bordalo et al., 2019), speed dating (Bhargava & Fisman, 2014) and financial markets (Hartzmark & Shue, 2018).[3] This paper documents the influence of contrast effects at a key stage of labor market matching.[4] Moreover, we provide new evidence on their behavioral foundation. This complements recent evidence on contrast effects, in particular the one by Hartzmark and Shue (2018). Using aggregate price data from financial markets, the study shows that the impact of contrast effects is measurable even in market level outcomes. At the same time, the underlying individual decisions and information histories are unobserved, which makes it difficult to understand the behavioral foundation of the aggregate effect. Our empirical approach provides a complement based on individual level data, allowing for theory-founded insights on the strength of contrast effects under different circumstances. Moreover, we structurally estimate the underlying model and show that a simple model with associative recall provides a good fit of the empirical patterns.

More broadly, this study relates to field evidence on reference-dependent decision making (for an overview, see Donoghue & Sprenger, 2018), and backward-looking, adaptive reference points in particular (e.g., Thakral & Tô, 2020; DellaVigna et al., 2021). Our results show that evaluators use previous candidates as a reference when forming an assessment. Models of

[3] Other studies have also provided field evidence on a negative relationship between a current decision and the characteristics or outcomes of the previous decision. Using hospital data on child deliveries, Singh (2021) finds that complications in their prior delivery makes physicians more likely to switch to the other delivery mode for the subsequent patient, but does not identify the mechanism. As discussed above, Chen et al. (2016) document a negative autocorrelation in the decisions of asylum judges, loan officers and sport judges, which they attribute to a gambler's fallacy. A positive path dependence has been found for jury decision making in criminal courts (Bindler & Hjalmarsson, 2018) and sport judges (Damisch et al., 2006; Kramer, 2017).

[4] An additional difference between the setup of this paper and existing studies lies in the timing of decisions. Existing studies consider sequential decision-making, where decisions occur directly after observing a choice option. In the setup of this paper, final decisions are made at the end of a closed sequence. Our results show that instantaneous errors that occur when observing a choice option persist even when ex-post adjustments are possible.

associative memory (Bordalo et al., 2020; Mullainathan, 2002) provide a foundation for such backward-looking reference dependence. We apply insights that arise from this approach to a relevant labor market context and show that it yields new implications for organizational design. We provide both reduced-form and structural evidence on how associative memory can help understanding economic decision-making in the field. This offers a complement to approaches that conceptualize and test the role of memory for economic decision making in a fully controlled lab environment (e.g., Enke et al., 2020; Bordalo et al., 2021).

The remainder of the paper is structured as follows. Section 2 informs about the institutional setting and background. Section 3 describes and summarizes the data. The empirical analysis is presented in section 4. Section 5 assesses the underlying mechanism and section 6 provides a structural estimation of the underlying model. Section 7 concludes.

2 Institutional Setting

We study the admission workshops of a large, merit-based study grant program for university students in Germany. The workshops have high stakes, as the grant offers a large number of monetary and non-monetary benefits.[5] In the following, we describe the setup of the admission workshops. Additional institutional background on the study grant program is provided in Appendix A.

Background Admission workshops take place over the course of one weekend and resemble the structure of assessment centers. About 48 candidates participate in each workshop.[6] The admission committee is formed by eight evaluators. Moreover, a representative of the study

[5] At the beginning of our sample period, the monetary scholarship ranged from 1,800 to approximately 10,000 euros, depending on parents' earnings. In 2020, the monetary scholarship ranges between 3,600 and about 14,000 euros per year. Given that there are no tuition fees at German universities, the scholarship covers up to the entire living costs of a student. Additional grants can be received for stays abroad. Non-monetary benefits include access to cost-free summer schools and language classes, a strong signal on one's CV, as well as networking opportunities. Students are admitted for the period of their entire university studies, subject to a positive interim evaluation.

[6] The baseline workshop schedule is designed for 48 candidates. Anticipating short-notice cancellations, the program slightly over-books each workshop. If more or fewer than 48 candidates show up, the workshop follows a slightly adjusted schedule. We use the actual schedule with the actual number of participants.

grant organization is permanently present and moderates the workshop.

Candidates are first-year university students. They were pre-selected as the top 2% of their high school's graduation cohort. Prior to the workshop, candidates submit a written CV and their school transcripts. During the workshop, each candidate is assessed through two one-to-one interviews of 35 minutes and a group discussion. Each of the three assessments is made independently by a different evaluator. The final decision is based on the sum of the three equally-weighted assessments.

Evaluators are alumni of the study grant program, now working in diverse professions. They commonly participate in one admission workshop every one or two years. No information about candidates is given to the evaluators before the workshop, and vice versa.

The assignment of candidates to evaluators and the assignment of time slots are quasi-randomized (c.f. randomization checks in section 3).[7] Both candidates and evaluators are assigned an ID. A fixed schedule then matches candidate IDs to evaluator IDs and time slots. Neither evaluators nor candidates know the assignment ex ante.

Workshop Schedule Table 1 sketches an evaluator's schedule during the admission workshop.[8] Upon arrival on Friday night, evaluators receive a short briefing by the representative of the grant organization and prepare the interviews which they conduct on Saturday. For this purpose, they receive each candidate's CV, school records and a letter of recommendation. On Saturday, evaluators each conduct six interviews and watch five group discussions.[9] In the evening, they receive the documents of the candidates whom they interview on Sunday. On Sunday, evaluators conduct six interviews and assess one group discussion. The detailed schedule – including candidate assignments to evaluators and time slots – is shown in Ap-

[7] Randomization occurs conditional on gender, with the aim of ensuring a balanced gender composition in the group discussion (see also randomization checks in section 3.2).

[8] We describe here the schedule for the 2013/14 academic year. In the following years, the schedule was slightly adjusted with respect to the group discussions, but the length and ordering of the interviews were not affected.

[9] Every group discussion includes approximately six candidates and takes place over six time slots. In each time slot, one candidate has to give a short presentation on a self-chosen topic and moderate the following discussion. Evaluators do not interfere in the discussion. Moreover, evaluators do not receive any information about the candidates who they observe in the group discussions, except for their names, study major and visually observable characteristics such as gender. They base their rating on the candidate's presentation and her contributions to the discussion.

Table 1: Stylized Evaluator Schedule

	Friday	**Saturday**	**Sunday**
Morning		**interviews** (≈ 3) + group discussions (≈ 3)	**interviews** (≈ 6) + group discussion (≈ 1)
Afternoon		**interviews** (≈ 3) + group discussions (≈ 2)	committee meeting
Evening	preparation	preparation	

pendix Figure A.1. The schedule also reveals that no evaluator sees the same candidate twice and that there is little overlap in the set of candidates seen by two evaluators.

Assessment and Admission Decision We focus on the formation of assessments in the one-to-one interviews. Evaluators are asked to rate candidates according to their intellectual abilities, ambition and motivation, communication skills, social engagement and broadness of interests, which comprise the program's selection criteria. There are no fixed guidelines regarding the interview questions, but suggestions for suitable types of questions.

Evaluators summarize their assessment on a scale from 1 to 10. A rating of 8 points or above implies a yes vote, i.e., an assessment in favor of accepting the candidate. 9 points are supposed to reflect a strong yes vote and 10 points are reserved for outstanding candidates. A candidate is accepted if she receives at least two yes votes and a total of at least 23 points. There is no admission quota, meaning that the committee can in principle decide to admit all or none of the candidates at the workshop. Evaluators are informed about these rules at the start of the workshop.

Evaluators are asked to determine their individual assessments after having seen all of their assigned candidates. During the workshop, an important rule is that evaluators may not discuss specific candidates with other evaluators before the final committee meeting. There is a common understanding that every candidate should receive the chance of being evaluated independently. In the final meeting on Sunday afternoon, a list with candidate IDs is read out aloud and every evaluator who has assessed the respective candidate states her rating. Subse-

quently, ratings are aggregated and — following a short justification by the responsible evaluators — candidates at or above the cut-off of 23 points are admitted. Ratings for candidates at the margin to admission can be adjusted after a discussion by the committee. According to anecdotal evidence, such adjustments usually concern about two to three out of about 150 votes per workshop. We observe the final ratings of each candidate.[10]

3 Data and Measurement

In this section, we describe the data source, assess the random assignment and ordering of candidates and explain our baseline measure of candidate quality.

3.1 Data Description

Data Source We employ data on the full population of admission workshops for recent high-school graduates that took place during the academic years 2013/14 to 2016/17. The data contain 312 admission workshops, including 29,466 interview ratings on 14,733 candidates.[11] The ratings were made by 2,496 evaluators.[12]

For each candidate, we observe the interview and group presentation slots, as well as the resulting ratings and admission decision. In addition, the data report the candidate's gender, age, study major, high-school grade, an indicator of migration background and an indicator of a first generation status. For the evaluators, we observe gender, study major, age and prior workshop experience.

Summary Statistics Figure 1a plots the sample distribution of interview ratings. Ratings range from 1 to 10, and the average rating in the sample is 6.6 points, with a standard devi-

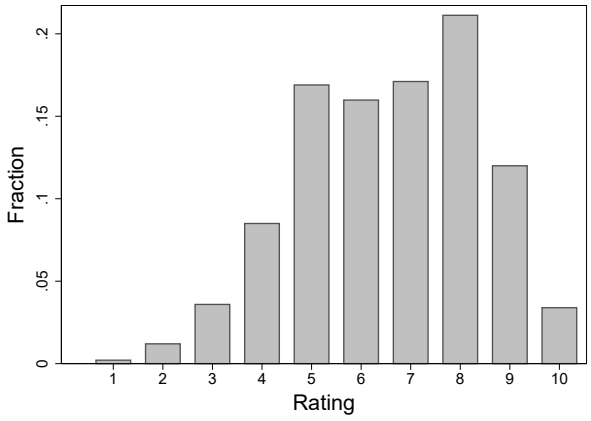
[10] To test whether the adjustment procedure influences our results, we run several robustness checks where marginal candidates are excluded.

[11] We had to exclude 36 workshops because the final assignment of candidate IDs was not documented. Moreover, we dropped 45 individual candidates (0.003%) because their candidate ID is missing, which means that we do not observe their assigned evaluators and time slots.

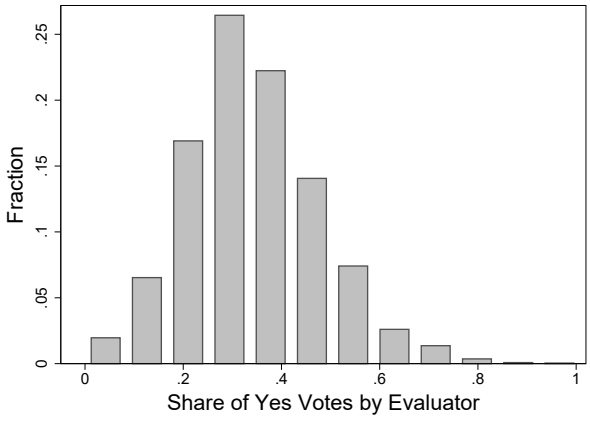
[12] We observe 1,724 unique evaluators. We treat every evaluator-workshop observation as independent, as there is usually a large time lag between two workshops. The average evaluator participates in about 1.8 workshops in the sample. 46% of evaluators participate in only one of the workshops in the sample period.

Figure 1: Distribution of Assessments

(a) Distribution of Interview Ratings



(b) Distribution of Evaluator-Specific Yes Vote Shares



Note: Panel (a) shows the distribution of interview ratings (N=29,466). A rating of ≥ 8 points implies a yes vote. Panel (b) shows the distribution of evaluator-level yes vote shares (N=2,496).

ation of 1.8. For the empirical analysis, we standardize the rating distribution to have a mean of zero and a standard deviation of one at the level of the academic year to account for possible shifts in the overall distribution of ratings over time. A rating of 8 points or more defines a yes vote. As shown in Figure 1b, there is substantial heterogeneity in the share of yes votes per evaluator, which ranges from 0 to 1, with a mean of 0.37 and a standard deviation of 0.14.[13]

Appendix Table B.1 reports summary statistics on candidate and evaluator characteristics. About 55% of candidates are female. The average candidate is 19.6 years old, 16% have a migration background and 26% are first generation students. Close to half of the evaluators are female and the average evaluator is 42 years old. Evaluators and candidates come from various fields of study.

3.2 Randomization Checks

The empirical analysis relies on the assumption that individuals are as-good-as randomly assigned to and ordered within an interview sequence. These conditions should be met by the

[13]This also translates into a wide range of workshop-specific admission rates from about 0.09 to about 0.46 (see Appendix Figure B.1). The average workshop has an admission rate of 0.25, with a standard deviation of 0.07.

Table 2: Assessment of Quasi-Random Assignment & Ordering

	GPA (1)	Age (2)	Migrant (3)	1st Generation (4)	STEM (5)	Social Sciences (6)
Panel A						
Leave-One-Out Mean	-0.000	0.000	0.002*	0.000	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Panel B						
Lag	0.003	0.002	-0.003	0.001	0.009	0.009
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
N	26970	26970	26970	26970	26970	26970

Note: In Panel A, "Leave-one-Out Mean" is the average value of the respective variable at the evaluator level, excluding the candidate in t. In Panel B, "Lag" refers to the previous candidate's value of the respective outcome variable. Regressions control for own gender and include workshop fixed effects. In both panels, the first candidate in the sequence, for whom no previous candidate exists, is excluded. Following Guryan et al. (2009), we control for the fact that an individual cannot be assigned to herself by including the workshop leave-one-out mean of the respective variable in Panel A, and the evaluator leave-one-out mean of the respective variable in Panel B. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

institutional setup as described in section 2. The only candidate characteristic taken into account for the assignment of candidate IDs is gender, because the organization aims at gender-balanced group discussions. We thus assume the random assignment and ordering conditional on own gender. In the following, we assess this central assumption.

Quasi-random assignment to evaluators implies that the characteristics of a candidate assigned to evaluator i are not systematically related to the characteristics of the other candidates assigned to i. To test this implication, we regress a given characteristic of a candidate on the leave-out mean characteristic of the other candidates assigned to the same evaluator, conditional on own gender and workshop fixed effects.[14] The results of this exercise are reported in Panel A of Table 2. In line with quasi-random assignment at the workshop level, we find no evidence of candidate sorting. Appendix Table B.2 additionally shows that candidate and evaluator characteristics are not systematically related to each other.

To assess the assumption of quasi-random ordering, we test for the presence of an autocor-

[14] Following Guryan et al. (2009), we control for the fact that an individual cannot be assigned to herself using the workshop leave-out mean of the respective variable.

relation in candidate characteristics, conditional on own gender and workshop fixed effects.[15] Panel B of Table 2 reports the estimates, which show no indication of systematic ordering by candidate characteristics.

3.3 Third-Party Assessment as a Measure of Candidate Quality

Our aim is to analyze how a candidate's assessment changes when the quality of another candidate in the same interview sequence increases. In our context, quality describes how well a candidate meets the study grant's selection criteria (see section 2). True candidate quality is unobserved by design — the assessment process would not need to take place otherwise. Therefore, any measurement of quality needs to be thought of as an approximation.

Our preferred approximation is based on an independent third party assessment (TPA) of a candidate's quality.[16] Given our setup, we define TPA as the sum of the candidate's other two ratings, which are made independently by two of the other seven evaluators at the workshop. One of the other ratings is based on the candidate's second interview and the other on her performance in the group discussion.[17] The main idea behind this approach is twofold: first, evaluators use the same criteria when rating quality; second, they all measure these criteria with noise, but their noise terms are independent of each other. Below, we discuss these two features in more detail.

First, all evaluators are supposed to rate the same dimensions of quality. The correlation between the individual rating and the sum of the other two evaluators' ratings is 0.38. Given that evaluators differ in their leniency and see the same candidate under different circumstances, we interpret this correlation as strong.[18]

[15] Again, following Guryan et al. (2009), we control for the fact that an individual cannot follow herself using the leave-one-out mean characteristics of the other candidates assigned to the same evaluator.

[16] An alternative way to measure candidate quality is through pre-determined characteristics, in particular high-school GPA. However, GPA is a poor predictor of fit with the scholarship criteria (see Appendix Table B.4), which extend beyond grade performance. Nevertheless, we construct an alternative quality measure based on pre-determined candidate characteristics to check the robustness of the main results pattern.

[17] Combining both ratings for the quality measure has the advantage of reducing noise. As a robustness check, we also run analyses using either only the other interview rating or only the group discussion rating as a measure of quality.

[18] As one point of comparison, Card et al. (2019) find a correlation of about 0.25 between two referee reports of the same paper in four leading journals in economics.

Second, the other two evaluators' ratings are as-good-as independent of the evaluator's own assessment behavior. Due to the design of the process, evaluators see the same candidate at very different points in time and the sets of candidates seen by two evaluators hardly overlap (see schedule in Appendix Figure A.1). Crucially, two evaluators never see the same two candidates in the same relative order. Moreover, evaluators are not allowed to discuss candidates before the final committee meeting (see section 2 for details).[19] In Appendix Table B.3, we empirically assess a key implication of the independence assumption. The idea is that we expect an evaluator's characteristics to correlate with her rating of a candidate. For instance, female evaluators are on average more lenient. On the contrary, evaluator characteristics should not correlate with the candidate's TPA, i.e., the other two evaluators' assessments. In line with this intuition, the results show that a candidate's rating — but not her TPA — correlates with the characteristics of the evaluator who made the rating.

4 Empirical Analysis

In this section, we analyze how the assessment of a candidate changes if another candidate's quality increases. In particular, we study how the influence of another candidate varies with the relative timing of her interview. In section 4.1, we relate a candidate's assessment to the other candidates' measured quality. In section 4.2, we estimate the autocorrelation in assessments.[20]

[19] One incidence where the independence assumption is potentially violated is the discussion of marginal candidates in the final committee meeting (c.f. section 2). Here, it can occur that an evaluator changes her rating following the arguments of another evaluator. We run robustness checks where we exclude marginal candidates from the estimation sample, and the estimates are unaffected.

[20] The analyses in this section are pre-registered. We uploaded the pre-registration before accessing the dataset used for this paper, including the main hypothesis and the econometric specifications. Prior to pre-registration, we had access to a data for the 2012/13 academic year. This "pilot" dataset is not contained in the estimation sample used for this paper.

4.1 Influence of the Interview Sequence

4.1.1 Econometric Specification

We estimate how the assessment of a candidate interviewed in period t is affected by the quality of the candidate interviewed in period $t+k$, as measured through her third-party assessment (TPA, see section 3.3). For each value of k, $k \in \{-11, ..., -1, 1, ..., 11\}$, we perform a separate estimation of the following regression model:[21]

$$(1) \qquad Y_{i,t} = \beta_k \, TPA_{i,t+k} + \gamma_k \, \overline{TPA}_{i,-\{t,t+k\}} + \pi \, TPA_{i,t} + X'_{i,t} \, \sigma + \eta_w + \epsilon_{i,t}$$

The outcome variable $Y_{i,t}$ is the standardized rating made by evaluator i of the candidate interviewed in period t. $TPA_{i,t+k}$ is the standardized TPA of the candidate interviewed by evaluator i at time $t+k$. The coefficient of interest, β_k, measures the influence of $TPA_{i,t+k}$ on the rating of the candidate interviewed in t.

The standardized leave-two-out mean $\overline{TPA}_{i,-\{t,t+k\}}$ controls for the average TPA of the other candidates in the interview sequence, excluding both the candidate in t and the candidate in $t+k$. $TPA_{i,t}$ denotes the candidate's own standardized TPA. The vector $X_{i,t}$ includes observed characteristics of candidates and evaluators as reported in Table B.1 and an indicator of the candidate's absolute order in the sequence. η_w controls for workshop fixed effects, corresponding to the level of randomization. Standard errors are clustered at the workshop level (N=312).

For each value of $k, k \in \{-11, ..., -1, 1, ..., 11\}$, we run a separate regression including the largest possible set of candidates, i.e., all candidates for whom period $t+k$ exists.

[21] Recall that the institutional setting allows every other candidate within an interview sequence to matter equally, as final ratings are set after the last interview took place. Therefore, both previously and subsequently observed candidates can potentially influence a candidate's evaluation.

4.1.2 Results

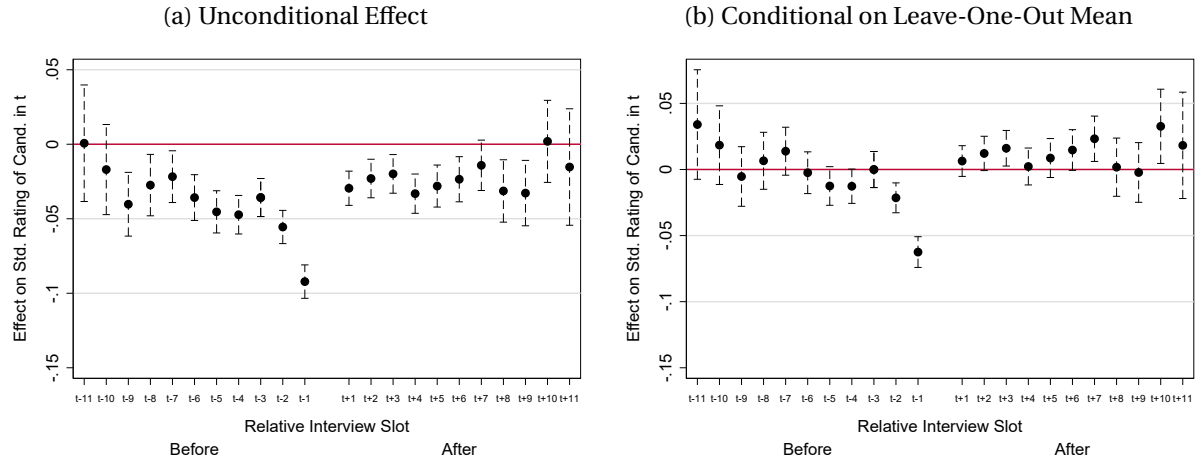
Panel (a) of Figure 2 plots the estimates of β_k from equation 1. The corresponding estimates are reported in Appendix Table C.1. The figure documents three main results. First, the rating of a candidate decreases in the measured quality of any other candidate seen by the same evaluator. If another candidate's TPA increases by one standard deviation, the candidate's rating decreases by about 2 to 5% of a standard deviation. Second, both candidates interviewed before t ($k < 0$) and candidates interviewed afterwards ($k > 0$) have an influence, suggesting that evaluators adjust their ratings after having seen everyone. However, candidates interviewed before have on average a slightly stronger negative influence.[22] Third, the influence of the previous candidate strikingly stands out: if the previous candidate's TPA increases by one standard deviation, the individual rating decreases by about 10% of a standard deviation. Appendix Figure C.1 shows a similar pattern when considering the probability of receiving a yes vote (rating ≥ 8 points) as an outcome.

Panel (b) provides evidence that the overall negative influence of the other candidates can be captured by controlling for the average quality of the sequence (leave-one-out mean TPA, $\overline{TPA}_{i,-t}$). The figure reveals that especially the previous candidate has a meaningful additional influence on the rating. This suggests the existence of two separate effects: an influence of the other candidates' average quality and an additional influence of recently observed quality.

Appendix Tables C.2 and C.3 report the estimated coefficients of own, previous and leave-one-out mean TPA and provide several robustness checks. Panel A of Table C.2 shows that the influence of the previous candidate's TPA amounts to about 17% of the influence of the candidate's own TPA and to about 55% of the influence of the sequence's leave-one-out mean TPA. Panels B to D show that these estimates are robust to the exclusion of marginal candidates (panel B), the estimation with evaluator fixed effects (panel C) and the estimation with candidate fixed effects (panel D). Table C.3 documents the robustness of the results for different measures of candidate quality, including a prediction based on observable characteristics. The

[22] The average of the coefficients with $k < -1$ amounts to 3.1% of a std. deviation and is significantly larger than the average of the coefficients with $k > 0$, which is 2.1% of a std. deviation (see Appendix Table C.1 for the corresponding p-values).

Figure 2: Effect of Candidate Quality in $t+k$ on Std. Rating of Candidate in t



Note: Panel (a) shows the estimated coefficients β_k from equation 1, resulting from separate regressions for each value of $k = \{-11, ..., -1, 1, ..., 11\}$. The coefficients measure how the standardized TPA of the candidate interviewed in $t+k$ affects the standardized rating of the candidate in t. TPA = third-party assessment of candidate quality (see section 3.3 for details). Panel (b) estimates the additional effect of the candidate interviewed in $t+k$, beyond her contribution to the average quality of the sequence (leave-one-out mean, excluding the candidate in t). Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level. Appendix Table C.1 reports the corresponding coefficients and p-values.

overall pattern, as well as the relative importance of own, previous and leave-one-out mean quality is very robust.

4.2 Autocorrelation in Assessments

The presented estimates have shown that the previous candidate's quality has a strong negative spillover on the current candidate's assessment. We now complement this causal evidence by an estimate of the autocorrelation in assessments.

The appeal of the autocorrelation is that it directly reflects the evaluator's own perception of candidates. A potential drawback is that the autocorrelation may in principle also contain the current candidate's influence on the previous candidate, which would prevent a one-directional interpretation. However, the previous analysis revealed that only the previous and not the next candidate has an influence beyond her contribution to the average quality of candidates in the sequence. This provides a justification for interpreting the autocorrelation as a

measure of the previous candidate's influence, once we condition on the average strength of the interview sequence.

4.2.1 Econometric Specification

We estimate the autocorrelation using the following specification:

$$Y_{i,t} = \delta\, Y_{i,t-1} + \theta\, \overline{Y}_{i,-t} + X'_{i,t}\,\mu + \omega_w + \zeta_{it} \tag{2}$$

$Y_{i,t}$ and $Y_{i,t-1}$ denote evaluator i's assessment of the candidates in t and $t-1$, respectively. The parameter of interest δ measures the autocorrelation between $Y_{i,t}$ and $Y_{i,t-1}$. To condition on the other candidates' average influence, we include the evaluator's mean assessment of candidates in the interview sequence, excluding the candidate in t (leave-one-out mean, $\overline{Y}_{i,-t}$). $\overline{Y}_{i,-t}$ always contains both the leave-one-out mean rating and the leave-one-out mean share of yes votes, to control for differences in both the average rating on the 1-10 scale and the propensity to give a yes vote. Note that the leave-one-out mean assessment also controls for differences in evaluator leniency.[23]

The specification controls for workshop fixed effects (ω_w),[24] as well as evaluator and candidate covariates $X_{i,t}$ (including interview order and the candidate's TPA).

4.2.2 Results

Table 3 reports the linear autocorrelation in evaluator assessments, based on an estimation of equation 2.[25] Columns 1 (without controls) and 2 (with controls) show that a one standard deviation increase in the rating of the previous candidate is associated with a 7% of a standard

[23] An alternative strategy is the use of evaluator fixed effects. However, as first noted by Nickell (1981), fixed effects introduce a downward bias when auto-regressive models are estimated on finite panels (here: $\overline{T} = 12$). They are therefore not suited in our context.

[24] Note that the use of workshop fixed effects in the context of an auto-regressive model also creates the potential for a 'Nickell bias'. However, T now amounts to $\approx 8 \times 12$ (the number of evaluator assessments per workshop), which makes the bias negligible.

[25] In Appendix Figures D.1 and Figure D.2, we additionally allow for a non-linear relationship.

Table 3: Autocorrelation in Assessments

	Rating (Std.)		P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)	(6)
Rating (t-1) (std.)	-0.066***	-0.070***				
	(0.007)	(0.006)				
Yes (t-1)			-0.057***	-0.406***	-0.027***	-0.023***
			(0.006)	(0.042)	(0.004)	(0.004)
Leave-one-out Mean Rating	0.256***	0.289***	0.076***	-0.691***	-0.017***	0.037***
	(0.018)	(0.018)	(0.009)	(0.054)	(0.006)	(0.005)
Leave-one-out Share Yes	-0.975***	-0.886***	-0.398***	-3.510***	-0.284***	-0.201***
	(0.078)	(0.072)	(0.048)	(0.238)	(0.028)	(0.027)
Controls	No	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.00	0.37	6.43	0.15	0.25
R-Squared	0.01	0.16	0.11	0.22	0.08	0.42
N	26970	26970	26970	26970	26970	26970

Note: All regressions include workshop fixed effects. Controls include candidate characteristics, evaluator characteristics and interview order. The leave-one-out means are computed at the level of the evaluator's interview sequence. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=312).

deviation decrease in the rating of the current candidate. Column 3 quantifies the autocorrelation in binary yes votes (rating ≥ 8 points). The probability of receiving a yes vote decreases by about 6 percentage points (16% relative to the mean) if the previous candidate received a yes instead of a no vote.

Columns 4 and 5 report the relationship between a yes vote in $t-1$ and a candidate's relative rank in the evaluator's rating distribution. Candidates who follow a candidate with a yes vote move down about 0.4 ranks on average (column 4) and are about 3 percentage points less likely to receive the best rating given by the evaluator (column 5). Note that, due to the leave-one-out mean controls, this result is not mechanical. The estimates thus reveal that the previous candidate's additional influence also distorts relative rankings. Finally, column 6 shows that the probability of admission — which is based on the sum of the three independent ratings — decreases by about 2.3 percentage points (10% relative to the mean) if the previous candidate received a yes vote in one of the interviews.[26]

[26] Appendix Table D.1 shows that the results on relative ranking and admission also hold when using the previous candidate's TPA as the regressor.

In all columns (except for the ranking outcomes in columns 4 and 5), the evaluator's leave-one-out mean rating shows a positive coefficient, which reflects the role of evaluator leniency. Conditional on the leave-one-out mean rating, the leave-one-out share of yes votes shows a negative coefficient. The individual likelihood of receiving a yes vote thus decreases if the evaluator gives more yes votes to the other candidates.

Appendix Table D.2 shows that the estimated autocorrelation is robust to the inclusion of candidate fixed effects. In line with the prediction of a downward bias that arises when estimating auto-regressive models on a finite panel (Nickell, 1981), coefficients become more negative when we control for evaluator leniency using evaluator fixed effects instead of leave-out means (Table D.3). Appendix Figure D.3 documents that there is no significant autocorrelation in assessments beyond t-2. Finally, Appendix Tables D.4 to D.5 show that the size of the autocorrelation exhibits little heterogeneity with respect to evaluator and candidate characteristics. Notably, the autocorrelation does not differ if evaluators have more prior interview experience, nor if they participated in a training for interviewing skills.

5 Behavioral Mechanism

The previous section provided evidence of two distinct influences on the formation of assessments: first, the average quality of the other candidates in the sequence decreases the individual assessment; and second, the previous candidate's quality has a strong additional influence. There are several straight-forward ways to explain the influence of the other candidates' average quality, such as learning about an uncertain admission threshold or an implicit target for the number of yes votes. However, it is difficult to reconcile the additional influence of more recent candidates with standard arguments.

This section discusses the behavioral mechanism underlying the influence of recently interviewed candidates. One intuitive mechanism is a contrast effect, where current assessments are negatively influenced by previous impressions. In the following, we discuss how the notion of a sequential contrast effect, caused by the interplay between the evaluator's memory and attention (Bordalo et al., 2020), can explain the previous candidate's influence. We then assess

the empirical relevance of conjectures that arise from this mechanism and that can help to further understand the nature of the effect.[27] In a final step, we discuss alternative mechanisms, notably a gambler's fallacy.

5.1 Contrast Effects and the Role of Associative Memory

5.1.1 Theoretical Intuition

Evaluators exhibit a contrast effect if they evaluate a current candidate against a (background) reference or norm. The notion of contrast effects is well known in the economics and psychology literature (see, for example, Pepitone & DiNubile, 1976; Simonson & Tversky, 1992; Bhargava & Fisman, 2014). In a recent contribution, Bordalo et al. (2020) propose a theoretical foundation for the emergence of contrast effects. Based on the concept of associative recall, the framework studies how reference norms are formed and how choice options are evaluated against them. In the following, we discuss the main intuition of how the framework explains the influence of the previous candidate. Appendix E.1 provides a more formal discussion of this intuition.

Valuation of Candidate Quality An evaluator decides on the rating of a candidate interviewed in period t, based on her valuation of that candidate.[28] We focus on the instantaneous valuation of the candidate formed at the time of the interview t, thereby abstracting from any ex-post adjustments which can occur after seeing all candidates. We define the instantaneous valuation as:

$$V_t = \underbrace{\tilde{q}_t}_{\substack{(perceived)\\ quality}} + \underbrace{\sigma(\tilde{q}_t, q_t^n)}_{salience} \times \underbrace{(\tilde{q}_t - q_t^n)}_{surprise} \tag{3}$$

[27] Most of the analyses in this section were not pre-registered, as they are based on insights from a recent theoretical framework.

[28] At this stage, we remain agnostic about how valuations translate into ratings and yes votes. We will come back to this step in the structural estimation (section 6).

The valuation V_t not only depends on the candidate's own quality as perceived by the evaluator ($\tilde{q}_t$), but also on its difference to the reference norm (q_t^n), i.e., the 'surprise'. The extent to which this surprise affects the valuation is determined by the salience function $\sigma(\tilde{q}_t, q_t^n)$.[29] Importantly, the salience of a given surprise varies with its size, which renders the impact of the quality norm non-linear. Small surprises do not capture the evaluator's attention and are therefore not salient, i.e., $\sigma(\tilde{q}_t, q_t^n)$ is low. Larger surprises are more salient, yet with diminishing sensitivity.[30] A change in the quality norm therefore affects not only the difference between a candidate's quality and the norm, but also the degree of attention that is directed towards it. When a difference is sufficiently large to attract the evaluator's attention, contrast effects arise because the evaluator reacts to the observed difference.

Formation of Quality Norm through Recall Bordalo et al. (2020) use the notion of associative recall to understand how the reference norm is formed in a given choice situation. Adapted to our setup, the idea is that evaluators form a reference norm through the recall of their prior interview experiences. The recall process is associative: an experience is weighted more heavily if it is similar to the current one. The norm is thus a similarity-weighted average of previously observed candidate quality.

Similarity can be defined along different dimensions. An obvious contextual stimulus when interviewing sequentially is time.[31] In this case, the quality of the most recent candidate strongly influences the norm, even though the time dimension does not have any normative relevance for the evaluation of candidates. But similarity can also include other dimensions, such as a candidate's observable characteristics (e.g., gender or study field). Importantly, similarity is of relative nature: increasing the similarity with one interview experience reduces the extent to

[29] We abstract from anchoring, present in the original model of (Bordalo et al., 2020). Anchoring adds a second layer to the valuation, where the valuation of a candidate is not only contrasted against the quality norm, but also anchored towards it. We formally discuss this extension in Appendix E.1 and provide an empirical assessment later in the section.

[30] Formally, $\sigma(\tilde{q}_t, q_t^n)$ is a salience function that is symmetric, homogeneous of degree zero, increasing in $\frac{x}{y}$ for $x \geq y > 0$ and $\sigma(y, y) = 0$; bounded by $\lim_{x/y \to \infty} \sigma(x/y, 1) = \sigma$.

[31] Bordalo et al. (2020) argue that "critically contextual stimuli such as location and time, act as cues that trigger recall of similar past experiences" (p. 1401). The location dimension is constant in our setting. Moreover, it is a well-established finding in psychology that recency is a key determinant of how well a prior experience is remembered (see, e.g., Kahana, 2012).

which another interview experience is recalled.

In summary, the framework predicts the incidence of contrast effects through the interplay of associative recall, which forms the reference norm, and the attention to salient quality differences. The notion of a 'sequential contrast effect' — i.e., contrasting with respect to the previous candidate — is incorporated in a natural way: recent interview experiences receive a strong weight in the quality norm because time generates a (potentially misleading) similarity. In the following, we assess the empirical relevance of insights that arise from this framework regarding the strength of contrast effects under different circumstances.

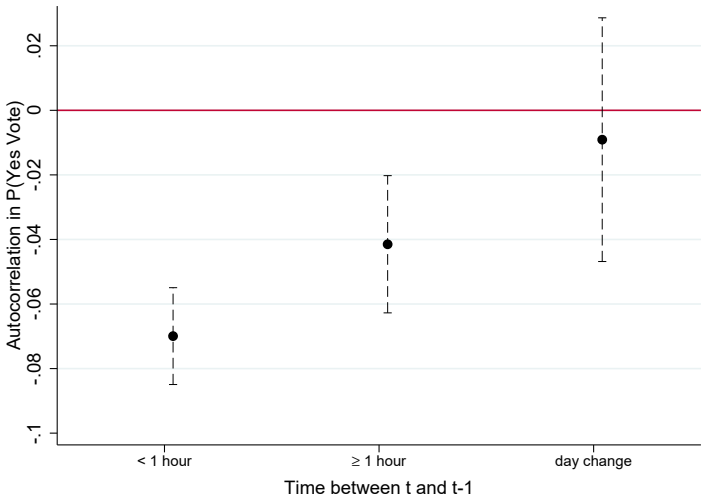
5.1.2 Insights Regarding the Previous Candidate's Influence

The Role of Breaks Associative recall predicts a strong influence of the previous candidate through similarity in the time dimension. A natural conjecture is that the strength of the influence should decrease when there is a larger break between two interviews. Figure 3 shows how the estimated autocorrelation — as a measure for the average influence of the previous candidate — varies with the time gap between two interviews. The pattern is in line with the intuition that longer breaks weaken the autocorrelation between yes votes in t and t-1. If there is an hour or more between two interviews, the autocorrelation amounts to only -4.2, instead of -7.0 percentage points in cases where there is less than an hour break. If interviews are separated by a day change, the autocorrelation approaches zero.

Additional Dimensions of Similarity Associative memory implies that another candidate's influence on the current candidate's valuation increases with her relative similarity. So far, we have made the presumption that similarity is driven by the time dimension. If we allow for observable candidate characteristics as a second dimension, the weight of the previous candidate will also depend on similarity with respect to these characteristics. A natural conjecture is that the previous candidate's influence increases if she shares more characteristics with the current candidate.

To assess the empirical relevance of this conjecture, we analyze how the autocorrelation in

Figure 3: Breaks and the Size of the Autocorrelation

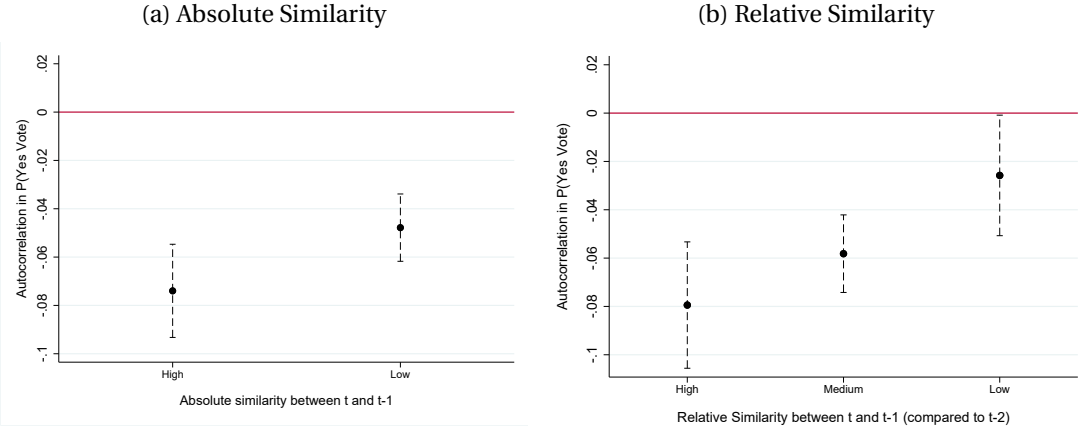


Note: The black dots plot estimates of the autocorrelation based on equation 2, where the previous candidate's yes vote is interacted with the time gap between the end of the interview in t-1 and the start of the interview in t. N=26,970. The dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level.

yes votes differs if two subsequent candidates are more or less similar in terms of their observable characteristics. We construct a simple "similarity index", which is defined as the number of observed characteristics shared between the current and previous candidate (including gender, migration status, first generation status and study field). We interact a median split of the index with the vote of the previous candidate. Panel (a) of Figure 4 shows the result. In line with theoretical intuition, the autocorrelation is significantly stronger if the observed similarity between two subsequent candidates is higher.

A distinctive feature of the framework is the notion of relative similarity. According to this notion, it matters how similar the previous candidate is compared to other preceding candidates. To assess this conjecture, we allow the influence of the previous candidate to depend on how similar the candidate in t-1 is compared to the candidate in t-2, who is still recent and provides a possible point of comparison in case the candidate in t-1 lacks similarity. More precisely, we compare three cases: (i) the candidate in t-1 is similar to the candidate in t, and the candidate in t-2 is not (high rel. similarity); (ii) the candidates in t-1 and t-2 are equally similar to the candidate in t (medium rel. similarity); (iii) the candidate in t-1 is not similar to the

Figure 4: Similarity of Candidate Characteristics and the Size of the Autocorrelation



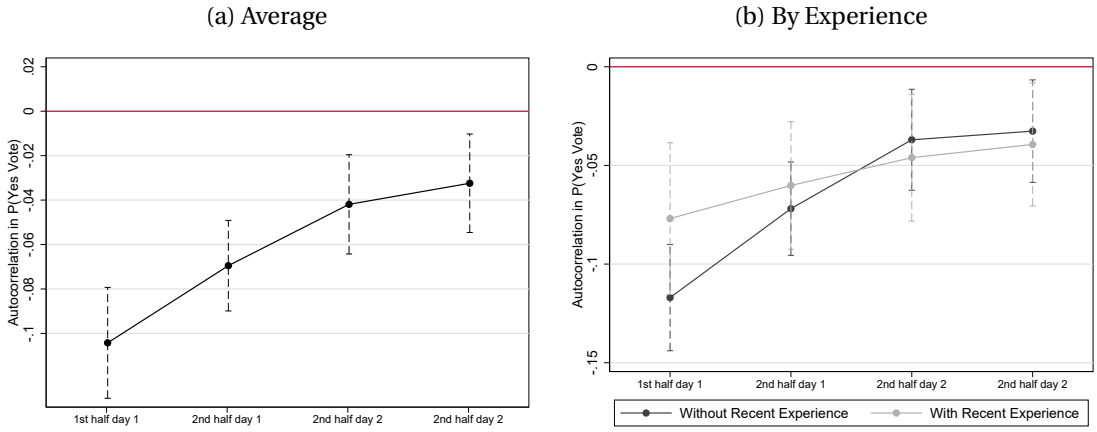
Note: Panel (a) shows estimates of the autocorrelation based on equation 2, where the previous candidate's yes vote is interacted with a median split of the similarity index, defined as the number of observable characteristics (gender, migration status, first generation status and study field) which the candidate in t and the candidate in t-1 have in common. In panel (b), similarity of t and t-1 is defined relative to the similarity between t and t-2. N=26,970 (panel a) & N=24,474 (panel b). The dashed lines show 95% confidence intervals.

candidate in t-1, and the candidate in t-2 is similar (low rel. similarity). Panel (b) of Figure 4 shows that, as relative similarity decreases, the strength of the autocorrelation decreases from about -8 to about -3 percentage points.

In Appendix E.2.1, we perform the same exercise considering every characteristic separately. The overall pattern is consistent, although the single characteristics yield a less powerful variation than the joint index. The strongest pattern is visible for gender, which is both a very salient characteristic and yields high statistical power due to roughly equal gender shares.

Based on our pre-registration, we discuss in Appendix E.2.2 whether the influence of similarity is symmetric with respect to gender. Models of memory consider similarity to work symmetrically (see, e.g., Kahana, 2012). Symmetric similarity states that the perceived similarity between two subsequent candidates does not depend on who is compared to whom. Put differently, the perceived similarity of candidate A following candidate B equals the perceived similarity of candidate B following candidate A. However, there also exist findings suggesting that similarity can sometimes be one-directional (Tversky, 1977), and our pilot data also pointed into that direction. Table E.1 shows that the evidence is overall more in line with the

Figure 5: Adjustment over the Interview Sequence



Note: Panel (a) shows estimates of the autocorrelation based on equation 2, where the previous candidate's yes vote is interacted with the absolute time of the current interview. In panel (b), there is an additional interaction with the evaluator's background experience. Recent experience is defined as having interviewed for the program in the previous academic year (36%). N=26,970. The dashed lines show 95% confidence intervals.

notion of symmetric similarity, as both male and female candidates are more strongly influenced by previous candidates of the same gender.

Size of the Memory Database Over the course of the interview sequence, evaluators continuously experience more candidates and thereby expand their memory of candidate quality. Given that similarity is based on relative weights, this expansion should lead to a reduced weight of the previous candidate in the quality norm as long as other preceding candidates are similar to the current candidate to at least some extent. In addition, having seen more candidates increases the probability of having observed a candidate who is very similar to the candidate in t and interferes with the previous candidate's influence.[32] We therefore expect the influence of the previous candidate to decrease over the course of the sequence. Figure 5 (a) is in line with this conjecture. It shows that the autocorrelation in votes weakens from about -10 percentage points for interviews conducted during the first half of the first interview day to

[32]With our non-experimental data, we cannot distinguish these two mechanisms. However, the general idea is common to both of them: the evaluator has a larger database with more variance from which she can retrieve experiences. These other experiences reduce the relative weight of the previous candidate and therefore reduce her influence.

about -3 percentage points for interviews conducted during the second half of the second day. Panel (b) shows that there is little difference between the patterns of experienced versus non-experienced evaluators, although experience appears to reduce contrasting at the beginning of the sequence.

Attention and Salience of the Surprise The estimates presented in section 4 revealed that the previous candidate's quality has on average a negative influence. However, the theoretical framework predicts a more nuanced pattern, where the effect depends on the size of the surprise as defined by the difference between a candidate's own quality and the quality norm. We therefore expect to observe contrasting only for 'large' differences, as small differences are not salient. What constitutes small and large differences in our context is a priori unclear.

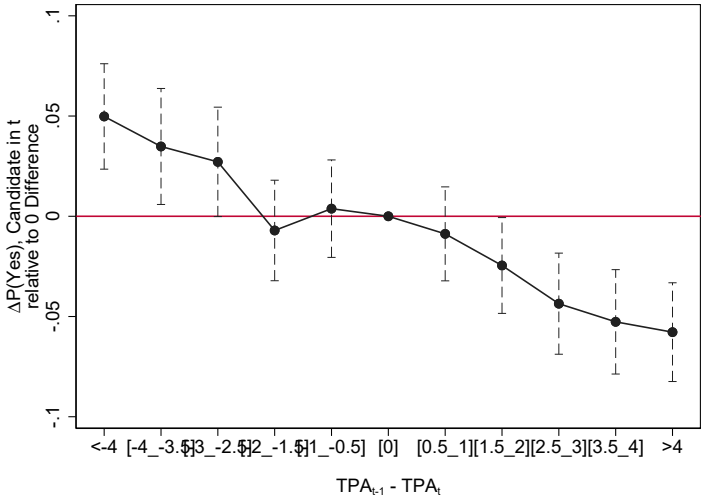
As our preferred proxy of the difference between the norm and the quality of the current candidate, we use the difference in the measured quality (TPA score) between the current and the previous candidate. This approach hinges on the assumption that the previous candidate indeed constitutes an important part of the norm. We then relate the current candidate's probability of a yes vote to categories of this difference, while flexibly controlling for the current candidate's TPA. The identifying variation thereby stems from changes in the previous candidate's TPA.

Figure 6 reports the estimates. Moving from the left (positive surprise) to the right (negative surprise) coincides with increasing the quality of the previous candidate.[33] In line with the previous results, the average slope is negative. However, we also observe a flat relationship around zero. This is in line with the notion that small quality differences do not attract the evaluator's attention. At larger absolute TPA differences, contrasting kicks in and leads to meaningful changes in the current candidate's probability to obtain a yes vote.

The original framework by Bordalo et al. (2020) suggests that evaluators not only contrast, but also anchor candidates in the case of a small quality difference, which leads to assimilation effects (see Appendix E.1 for more details). The evidence presented in Figure 6 does not

[33] To interpret differences in TPA: a one point increase in own TPA is associated with a 5 p.p. increase in the probability of a yes vote. Therefore, differences in TPA very quickly represent significant differences in quality of candidates.

Figure 6: Influence of Quality Differences



Note: The x-axis shows the difference in TPA between the candidate in t and the candidate $t-1$. The y-axis shows estimated coefficients on the probability of receiving a yes vote for the candidate in t. The underlying regression includes dummies for the candidate's own TPA. Further controls are the leave-one-out mean TPA, candidate characteristics, evaluator characteristics and interview order. N=26,970. 95% confidence intervals, with standard errors clustered at the workshop level.

provide evidence of assimilation. This may be due to the specific setup at hand: it is in the nature of candidate selection to differentiate between candidates and, therefore, pay a lot of attention to quality differences. Yet, we cannot reject the presence of assimilation assimilation, as the analysis might simply be unable to detect small assimilation effects. We revisit this question in the structural estimation (section 6).

In Appendix E.3, we explore the robustness of the presented pattern with respect to different proxies of the norm. In panel (a) of Figure E.4, we use the average TPA of the two previous candidates, and in panel (b) the average TPA of all previous candidates. In both specifications, we observe that evaluators show little reaction to small quality differences.

5.2 Alternative Mechanisms

We now discuss two alternative mechanisms that could also explain the previous candidate's influence: sequential updating about candidate quality and a gambler's fallacy.

Sequential (Bayesian) Updating Under sequential updating, prior candidates of high quality increase the belief about the average quality and therefore decrease the assessment of subsequent candidates. While this mechanism could produce a negative autocorrelation in assessments, it cannot explain why more recent candidates should have a stronger influence.

Gambler's Fallacy The belief in the law of small numbers states that individuals erroneously believe small samples to be representative of the population. It is — for example — modeled via the belief that signals are not i.i.d., but drawn from an urn without replacement (see, e.g., Rabin, 2002; Benjamin, 2019). An immediate implication is the gambler's fallacy, which expresses the mistaken belief that a 'good draw' should follow a 'bad draw' and vice versa. Under the gambler's fallacy, evaluators hold downward (upward) biased priors about the next candidate's quality after observing a strong (weak) candidate, which can produce a negative autocorrelation in assessments.

Three empirical arguments speak against a major role of the gambler's fallacy in explaining the previous candidate's influence. First, the gambler's fallacy works through the prior belief about the candidate in t and thus occurs before seeing that candidate. This implies that the influence of the candidate in t-1 should not depend on the size of the surprise, i.e., the quality difference, as it is the case in our setup (see Figure 6).

Second, the gambler's fallacy predicts streaks of votes to matter: two yes votes in a row should decrease the prior about the upcoming candidate more than one no vote followed by one yes vote. A direct conjecture is that the influence of two prior yes votes should be stronger than the influence of only one prior yes vote. Appendix Table E.2 shows that this is not the case in our data.

Finally, we follow Chen et al. (2016) and test for the influence of the previous candidate's continuous quality conditional on the previous binary decision. In a simple gambler's fallacy model, evaluators expect binary reversals, implying a negative autocorrelation in binary votes. As a result, once we condition on the previous vote, a simple gambler's fallacy does not predict any further correlation with a measure of the previous candidate's quality. Columns (1) and (2) of Appendix Table E.3 show that the influence of the previous candidate's measured quality

persists after controlling for the previous candidate's yes vote. This contradicts the prediction of a simple gambler's fallacy. However, as pointed out by Chen et al. (2016), the result could still be in line with a more complicated version, where the conditional influence of previous quality reflects the evaluator's uncertainty about the previous yes vote. For this purpose, we leverage the rating, which expresses the strength of the vote. Again, the influence of the previous candidate's quality measure persists (columns 3), further pointing towards contrasting as the predominant mechanism.

6 Structural Estimation

The previous section provided consistent evidence that the evaluators' behavior is in line with a combination of contrasting and associative recall. To further strengthen the link between the theoretical framework and the empirical evidence, this section structurally estimates the underlying model. While the reduced-form results have shown that the framework yields empirically relevant conjectures, the structural estimation can assess its quantitative plausibility. Moreover, we can use the structural estimates to calculate counterfactual scenarios regarding the formation of quality norms.

6.1 Parameterization

For the estimation, we parameterize the final valuation of a candidate interviewed in period t. The final valuation is composed of the instantaneous valuation as modeled in section 5.1 and an ex-post adjustment term:

$$V_t^{final} = \alpha \times \tilde{q}_t + \sigma(\tilde{q}_t, q_t^n) \times (\tilde{q}_t - q_t^n) - \beta \times \bar{q}_{-t} + u_t \tag{4}$$

In this expression, the instantaneous valuation depends on the candidate's own measured quality $\tilde{q}_t$ and on the difference between $\tilde{q}_t$ and the quality norm q_t^n, multiplied by its salience $\sigma(\tilde{q}_t, q_t^n)$. These terms correspond to the model presented in section 5.1. To capture the possi-

bility of ex-post adjustments, we add the auxiliary term $\beta \times \bar{q}_{-t}$, where $\bar{q}_{-t}$ is the leave-one-out mean measured quality of the other candidates seen by the evaluator. We thereby account for the reduced-form finding that evaluators adjust their assessments ex-post to the average quality of all candidates. u_t denotes a normally distributed implementation error (see Appendix F for details).

We parameterize the quality norm q_t^n as a weighted average of the preceding candidates' qualities: $q_t^n(c_t) = \sum_{l=1}^{t-1} \tilde{q}_{t-l}\omega_{t-l}$. The weight ω_{t-l} of the candidate observed in period $t-l$, for $l = 1, 2, \ldots, t-1$, is determined by her relative similarity to the current candidate: $\omega_{t-l} = \frac{S_{t-l}}{\sum_{m=1}^{t-1} S_{t-m}}$, where S_{t-l} denotes the similarity between t and $t-l$. In line with the reduced-form evidence, we assume that $\omega_{t+l} = 0$, i.e., subsequently observed candidates do not influence the norm and the recall process is only backward looking.

In the full specification, similarity depends on two contextual attributes: relative time and observable candidate characteristics. These two dimensions of similarity are multiplicatively separable and jointly define absolute similarity of the candidate in $t-l$:

$$S_{t-l} = S_{t-l}^{time} \times S_{t-l}^{char} = e^{-\delta_1(l-1)} \times e^{-\delta_2(\mathbb{1}_{diff})_{t-l}}$$

Similarity in time, S_{t-l}^{time}, exponentially decreases in the time lag l between two interviews. The speed of this process is determined by δ_1. For simplicity, similarity in characteristics, S_{t-l}^{char}, enters only in binary terms: the indicator $(\mathbb{1}_{diff})_{t-l}$ equals one if the candidate in $t-l$ differs from the current candidate in terms of her observable characteristics.[34] If this is the case, similarity in time gets multiplied by the factor $e^{-\delta_2} < 1$ and, thus, absolute similarity is lower.

The salience function $\sigma(\tilde{q}_t, q_t^n)$ defines how much attention is attracted to a given quality difference. We follow Bordalo et al., 2020 and assume salience to follow the functional form $\sigma(\tilde{q}_t, q_t^n) = \sigma \frac{e^{\theta(x-1)^2}}{1+e^{\theta(x-1)^2}} - \frac{\sigma}{2}$, $x = \frac{\tilde{q}_t}{q_t^n}$. This function evaluates to zero for zero quality differences and is bounded by $\frac{\sigma}{2}$. The parameter σ describes how strongly quality differences influence

[34] In line with the reduced-form analysis in section 5.1, we construct an index counting the number of shared observable characteristics. $(\mathbb{1}_{diff})_{t-l}$ equals one if two candidates share not more than the median number (i.e., two) of characteristics.

the valuation, whereas θ determines how quickly differences become salient.

The data do not report the evaluators' valuations, but their ratings on a discrete 1-10 scale. We therefore add a transformation process that maps the latent valuations to the observed ratings. To this end, we bin the ordered (simulated) valuations into groups corresponding to the share of candidates that receive a given rating in the observed distribution.[35]

In the appendix, we also present estimates from three variants of the presented model. First, we estimate the original framework postulated by Bordalo et al. (2020), which features anchoring to the quality norm. Moreover, we estimate two variants without associative recall, which serve as benchmarks for our main estimates. In the first benchmark model, we estimate a model with $\delta_1 = 0$ and $\delta_2 = 0$. This eradicates associative recall, such that all previous candidates receive the same weight in the norm. In the second benchmark model, we replace the quality norm with the expected quality (i.e., the sample average). This eradicates not only associative recall, but recall in general.

6.2 Estimation & Identification

Estimation We estimate the model parameters using the method of simulated moments. Let $m(\xi)$ denote the vector of simulated moments as a function of the model parameters, and $\hat{m}$ the vector of empirical moments. The estimator chooses the parameter vector $\hat{\xi}$ that minimizes the distance $(m(\xi) - \hat{m})' W (m(\xi) - \hat{m})$. As a weighting matrix W, we use the diagonal of the inverse of the variance-covariance matrix.[36,37] In every simulation step, we simulate a population of 10,000 evaluators, who each interview 12 candidates. We use the DFO-LS algorithm to solve the minimization problem.[38] To address concerns regarding local minima in

[35] For example, 3.4% of candidates receive a rating of ten. Therefore, the 3.4% of candidates with the highest valuation are assigned a rating of ten in the simulation process. As a result of this procedure, the estimated distribution of ratings is mechanically fitted to the observed distribution. Note that the unconditional moments of the ratings distribution are not targeted otherwise in the estimation procedure.

[36] Altonji and Segal, 1996 show that using the full inverse of the variance-covariance matrix can lead to numerical instability of the estimator.

[37] We multiply the weights of moments that describe the previous candidates' influence (depending on similarity) by a factor of five, as they are key moments describing the recall process and we want to ensure that the estimator targets them. See DellaVigna et al., 2021 for a similar approach. Appendix F.6.1 presents a robustness checks where we use the identify matrix to weight the moments.

[38] We use a Python implementation of this algorithm (Gabler, 2021).

the criterion function, we estimate the model several times, using ten randomly-chosen initial values from a uniform distribution over the parameter space. We use the estimates with the minimum weighted distance as the final parameter estimate. Moreover, we conducted Monte Carlo exercises and confirmed that the estimation method is able to back out the true parameters of a simulated dataset. Appendix F.1 provides further estimation details.

Identification To identify the model parameters, we use moments that describe how a candidate's rating reacts to her own and the other candidates' measured quality. One key set of moments, which serves primarily the identification of the associative recall parameters δ_1 and δ_2, describes how the influence of preceding candidates varies with similarity in time and candidate characteristics. The first salience parameter σ is primarily identified from the relationship between ratings and the size of the difference between current and previous candidate quality. As only large changes in the second salience parameter θ lead to measurable changes in this relationship, we abstain from estimating θ and calibrate it to different values.[39] Moments that identify the incidental parameters α and β capture how ratings respond to the measured own and leave-one-out mean quality, respectively. Appendix F.2 provides details about all moments and their link to the structural parameters. It also documents that the criterion function indeed reacts to changes in the model parameters.

6.3 Results

Main Parameter Estimates Table 4 presents parameter estimates for two versions of the model: a reduced version which only includes similarity in the time dimension (columns 1 and 3) and the full model with candidate characteristics as a second dimension of similarity (columns 2 and 4). We estimate both models once with a high and once with a low calibrated value of θ, thereby varying how quickly quality differences become salient.

The structural estimates lead to the following observations: first, our simple parameter-

[39] In the main estimation, we use $\theta = 30$ and $\theta = 100$. Appendix Figure F.6 illustrates the influence of θ on the salience of quality differences and their valuation. Importantly, we will find that the other parameter estimates are as good as invariant to the value of θ.

Table 4: Structural Estimates

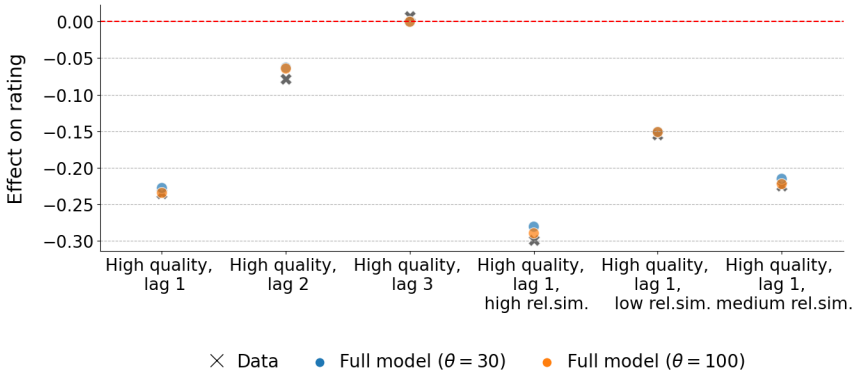
	Only similarity in time ($\theta = 100$) (1)	Full model ($\theta = 100$) (2)	Only similarity in time ($\theta = 30$) (3)	Full model ($\theta = 30$) (4)
Similarity Parameters				
δ_1	0.943	1.22	1.048	1.207
	(0.251)	(0.178)	(0.139)	(0.117)
$e^{-\delta_2}$		0.219		0.235
	.	(0.076)	.	(0.122)
Salience Parameters				
σ	0.173	0.17	0.18	0.181
	(0.027)	(0.018)	(0.020)	(0.022)
$\theta^{\dagger}$	100.0	100.0	30.0	30.0
	.	.	.	.
Incidental Parameters				
α	0.155	0.155	0.162	0.158
	(0.015)	(0.009)	(0.011)	(0.012)
β	0.251	0.249	0.25	0.244
	(0.022)	(0.017)	(0.020)	(0.025)
Weights				
ω_{t-1}	0.662	0.696	0.694	0.697
ω_{t-2}	0.244	0.239	0.232	0.239
ω_{t-3}	0.094	0.077	0.08	0.077
ω_{t-1} | high rel. sim		0.894		0.888
ω_{t-1} | medium rel. sim		0.686		0.684
ω_{t-1} | low rel. sim		0.386		0.399
Weighted SSE	124.746	128.012	146.924	146.066
Number of moments	21	24	21	24

Note: The table shows estimates of the parameters in equation 4, with standard errors in brackets. Columns (1) and (3) report estimates from a reduced model where similarity is only based on the time dimension. Columns (2) and (4) report estimates from the full model, including similarity in terms of candidate characteristics. The second salience parameter θ is calibrated to 100 in columns (1) and (2) and to 30 in columns (2) and (4). The weights describe the weight that a previously interviewed candidate receives in the quality norm. “Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). “High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. Estimation is based on the method of simulated moments (see Appendix F.1 for details). † = calibrated. SSE= Sum of Squared Errors.

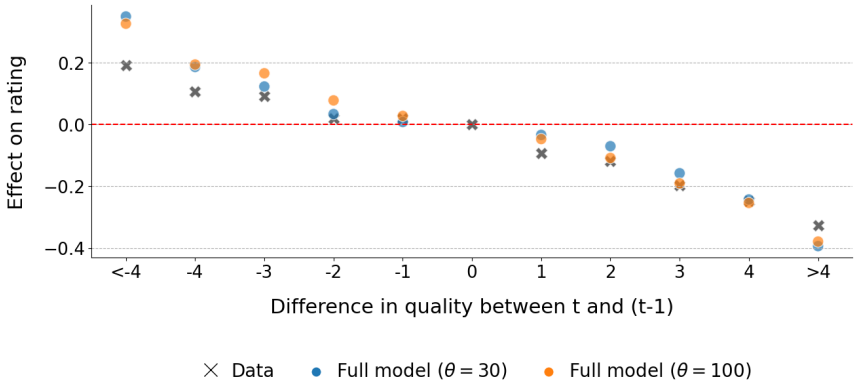
ization of the evaluators' recall process is able to provide quantitatively reasonable predictions regarding the influence of previous candidates. The estimate of δ_1 amounts to about 1 (columns 1 and 3), implying a strong decline in similarity with increasing time lag, which is in line with the reduced-form findings. This results in a high weight of the previous candidate in the norm of roughly 65-70%. The candidate interviewed in period t-2 still receives a meaningful weight of about 25%, while the weight of the candidate in t-3 is still positive, but small. When adding the second dimension of similarity (columns 2 and 4), the interpretation of δ_1 changes. It now describes how absolute similarity in time evolves for candidates of high similarity in characteristics. The estimate of $e^{-\delta_2}$, i.e., the factor by which absolute similarity in

Figure 7: Empirical Moments and Model Fit: Influence of Previous Candidates

(a) Role of similarity (time & characteristics)



(b) Role of quality difference



Note: This figure documents the model fit for the estimates reported in columns (2) and (4) of Table 4. In panel (a), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. “rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). “High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. In panel (b), the moments describe the effect of a given quality difference between the candidates in t and t-1. The fit with additional moments is illustrated in Appendix Figure F.7.

time is multiplied if two candidates are observationally more different, amounts to about 0.2. Jointly, the estimates of δ_1 and δ_2 imply that the previous candidate's weight varies strongly with her relative similarity: the previous candidate has a weight of almost 90% if she is more similar to the current candidate than the candidate in t-2, but only about 40% if her relative similarity is low (i.e., the candidate in t-2 is more similar). Panel (a) of Figure 7 reveals that the parameterization of the recall process can match the corresponding empirical moments

almost perfectly. To further validate this claim, Appendix Figure F.8 shows that this also holds when empirical moments are computed based on the pilot dataset, i.e., in a 'hold-out' sample. Moreover, Appendix Figure F.9 shows that the model is able to provide a good fit of moments that also describe the recall process but were not targeted in the estimation.

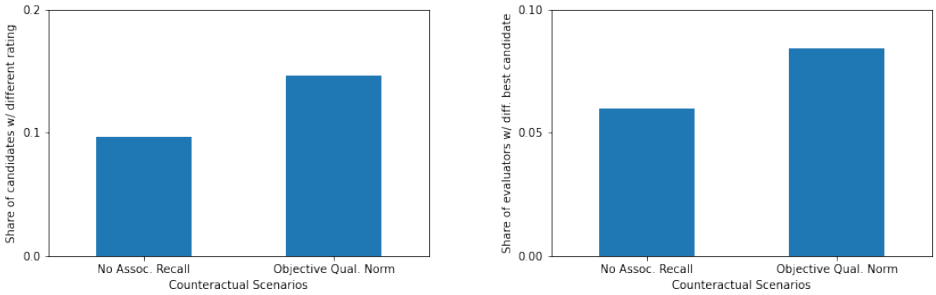
The second key component of the model is the relationship between the differences in quality and the valuation, as described by the salience function. As discussed above, the proposed functional form is characterized by the parameters σ and θ. We estimate σ and calibrate θ to the values 30 and 100, where the latter implies a higher salience of small differences (see Appendix Figure F.6 for an illustration). The estimate of σ is almost invariant to the value of θ and amounts to about 0.17-0.18. Panel (b) of Figure 7 shows that the structural estimates with both values of θ provide overall a good fit with the empirical evidence. However, we also note that they over-predict the effect of large quality differences at the boundaries. One possible explanation is the nature of the evaluation process: while the model predicts a strong impact of large quality differences, interviewers might be reluctant to implement this in their ratings. More generally, the figure illustrates the difficulty in pinning down an exact functional form for the salience of differences. While the overall criterion value is slightly smaller for a higher value of θ, the model with a lower θ provides a better fit of the flat part for small negative quality differences. Acknowledging additionally that quality differences are estimated with noise, we refrain from drawing any strong conclusions about the exact form of the salience function in our setup.

Model Variant & Benchmarks In Appendix F.6.2, we provide estimates of a model variant with anchoring to the norm as present in the original framework (c.p., Bordalo et al., 2020). The model captures the relative importance of previous candidates about equally well, but the pattern of quality differences slightly worse. This is mostly due to the fact that the model predicts small assimilation effects, which are not found in the empirical relationship.

Appendix F.6.3 presents estimates from two benchmark models. In the first benchmark (columns 1 and 3 of Table F.6), we eradicate associative recall by setting $\delta_1 = \delta_2 = 0$ and thus assume that the quality norm is the average quality of all prior candidates. This assumption

Figure 8: Simulation of Counterfactual Quality Norms

(a) Change in rating (b) Change in choice of best candidate



Note: The simulations are based on the estimates from column (2) of Table 4. The left bar corresponds to a counterfactual where the norm is based on an unweighted average of the previous candidates' quality ($\delta_1 = \delta_2 = 0$). The right bar corresponds to a counterfactual where the norm is the sample average of measured quality. In panel (a), the y-axis shows the share of candidates who would get a different rating under the given counterfactual. In panel (b), it shows the share of evaluators who would give their highest valuation to a different candidate under the given counterfactual.

results in a considerably worse overall fit with the empirical moments (SSE=228 vs. 125 in the main specification). The second benchmark (columns 2 and 4) assumes that the norm is not formed through recall at all, but simply consists of the expected quality. This model has no chance of predicting the influence of previous candidates, which further reduces the overall fit (SSE=741). Taken together, the estimation of the two benchmark models provides evidence that associative recall is key to explain the empirical pattern.

6.4 Counterfactual Experiments

In a final step, we use the structural estimates to conduct two counterfactual experiments regarding the formation of quality norms.

The first counterfactual corresponds to an intervention that reduces associative recall. We consider the extreme case where similarity plays no role for recall, implying that all previous candidates have an equal weight in the norm. We use the estimates from column (2) of Table 4 and simulate ratings once with the estimated values of δ_1 and δ_2 and once with δ_1 and δ_2 set to 0. Results show that about 9% of all ratings change under this counterfactual scenario (left bar

in Panel a of Figure 8). To move away from our specific context, where valuations translate into ratings on a 1-10 scale, we also calculate how often the candidate with the highest valuation per evaluator changes. In other words: which share of evaluators would choose a different candidate as the best candidate in the sequence? Results show about 6% of evaluators would select a different candidate as the best one (panel b).

In the second counterfactual, quality norms are not formed through recall at all. Instead, all candidates are compared against the same “objective" quality norm, namely the expected quality (sample average). Intuitively, such an intervention would have a slightly larger impact, as the draw of the previous candidates has no longer any influence on the norm. The simulations suggest that about 15% of all ratings would change (panel a) and about 8% of evaluators would rank a different candidate as the best candidate (panel b).

All in all, the simulations illustrate that the nature of quality norms matters for the assessment result in a quantitatively meaningful way. Making the formation of norms less prone to (associative) recall could therefore yield a significant improvement in the precision of evaluations and the selection of candidates.

7 Conclusion

Using large-scale data on real-world interviews, this paper shows that the quality of a candidate has a strong negative spillover on the assessment of the next candidate. This spillover extends far beyond the influence of any other candidate observed by the same evaluator. We conduct an empirical investigation and structural estimation of the underlying mechanism and argue that the previous candidate's strong influence is in line with a sequential contrast effect that is rooted in associative memory. The evaluator's attention is attracted to salient differences between the current and the previous candidate, who is strongly recalled due to similarity in the time dimension. This effect is further strengthened if the candidates are similar in terms of their observable characteristics.

The findings in this paper help to understand how people make subjective assessments in the light of prior experiences. They show that minor changes in candidate sorting and or-

dering can have major consequences on the assessment outcome. This carries implications for the organizational design of processes through which assessments are reached. First, the influence of individual biases can be mitigated by combining several assessments of a candidate and ensuring their independence. More precisely, it is key to minimize the overlap in the set and ordering of candidates seen by different evaluators. However, the collection of many subjective assessments will usually come at non-negligible costs. An alternative answer to the influence of human errors may lie in the combination of subjective assessments with more objective screening devices, such as algorithm-based job-testing technologies (e.g., Autor & Scarborough, 2008; Hoffman et al., 2018). At present, it remains unclear how well these technologies perform when selecting from a high-ability candidate segment.

Moreover, the paper shows that understanding the behavioral foundation behind evaluation errors yields additional design implications. A key insight is that evaluators focus on time as a superficial dimension of similarity, which is normatively irrelevant for the assessment. This calls for the design of interview processes that reduce the evaluators' focus on superficial similarity. If the aim is only to reduce the previous candidate's weight in the evaluation norm, a possible solution might be to avoid the aggregation of different dimensions of similarity. For example, organizations may want to order candidates in a way that subsequent candidates are not observationally similar to each other.

While such small organizational changes may successfully reduce the previous candidate's influence, they do not ensure that candidates are always evaluated against a relevant reference norm. An alternative approach lies in the design of interventions which target the retrieval of evaluation norms more directly. Such interventions require first of all a clearly defined and objective norm that is of relevance for the assessment process – such as precisely specified evaluation criteria or a clear profile of the ideal candidate. Crucially, organizations then need to nudge evaluators to actually retrieve the relevant norm upon interviewing. If and how organizational interventions can alter the formation of norms for evaluation remains a question for future research.

References

Altonji, J. G., & Segal, L. M. (1996). Small-sample bias in GMM estimation of covariance structures. *Journal of Business & Economic Statistics*, *14*(3), 353–366.

Autor, D. H., & Scarborough, D. (2008). Does job testing harm minority workers? evidence from retail establishments. *The Quarterly Journal of Economics*, *123(1)*, 219–277.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics*, 69–186.

Bergman, P., Li, D., & Raymond, L. (2020). Hiring as exploration. *mimeo*.

Bertrand, M., & Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, *94*(4).

Bhargava, S., & Fisman, R. (2014). Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economics and Statistics*, *96*(3), 444–457.

Bindler, A., & Hjalmarsson, R. (2018). *Path dependency in jury decision making*.

Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., & Shleifer, A. (2021). Memory and representativeness. *Psychological Review*, *128*(1), 71–85.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2019). Memory and reference prices: An application to rental choice. *AEA Papers and Proceedings*, *109*, 572–76.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, *135*(3), 1399–1442.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2019). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, *135*(1), 269–327.

Chen, D., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, *131*(3), 1181–1242.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, *12*(3), 166–178.

DellaVigna, S., Heining, J., Schmieder, J. F., & Trenkle, S. (2021). Evidence on job search models from a survey of unemployed workers in Germany. *Quarterly Journal of Economics*, *forthcoming*.

Donoghue, T., & Sprenger, C. (2018). Chapter 1 - reference-dependent preferences. In S. D. B. Douglas Bernheim & D. Laibson (Eds.), *Handbook of behavioral economics: Applications and foundations 1* (pp. 1–77). North-Holland.

Enke, B., Schwerter, F., & Zimmermann, F. (2020). Associative memory and belief formation. *mimeo*.

Estrada, R. (2019). Rules versus discretion in public service: Teacher hiring in mexico. *Journal of Labor Economics*, *37*(2), 545–579.

Gabler, J. (2021). A Python tool for the estimation of (structural) econometric models. https://github.com/OpenSourceEconomics/estimagic

Galiani, S., & Pantano, J. (2021). Structural models: Inception and frontier. *NBER working paper*, (w28698).

Ginsburgh, V. A., & Ours, J. C. van. (2003). Expert opinion and compensation: Evidence from a musical competition. *The American Economic Review*, *93*(1), 289–296.

Guryan, J., Kroft, K., & Notowidigdo, M. J. (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *AEJ:Applied Economics*, *1*(4), 34–68.

Hartzmark, S. M., & Shue, K. (2018). A tough act to follow: Contrast effects in financial markets. *Journal of Finance, 73(4)*, 1567–1613.

Hoffman, M., Kahn, L., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, *133*(2), 765–800.

Horton, J. J. (2017). The effects of algorithmic labor market recommendations: Evidence from a field experiment. *Journal of Labor Economics*, *35*(2), 345–385.

Kahana, M. (2012). Foundation of human memory. *Oxford University Press*.

Kramer, R. S. S. (2017). Sequential effects in olympic synchronized diving scores. *Royal Society Open Science*, *4*, 1–9.

Kroft, K., Lange, F., & Notowidigdo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics*, *128*(3), 1123–1167.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the nih. *American Economic Journal: Applied Economics*, *9*(2), 60–92.

Mullainathan, S. (2002). Memory-based model of bounded rationality. *Quarterly Journal of Economics*, *117*, 735–774.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*(6), 1417–1426.

Oyer, P., & Schaefer, S. (2011). Personnel economics: Hiring and incentives. *Handbook of Labor Economics, Vol 4b*, 1769–1823.

Pepitone, A., & DiNubile, M. (1976). Contrast effects in judgments of crime severity and the punishment of criminal violators. *Journal of Personality and Social Psychology*, *33*(4), 448–459.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, (Vol. 117, No. 3), 775–816.

Simonsohn, U. (2006). New Yorkers commute more everywhere: Contrast effects in the field. *The Review of Economics and Statistics*, *88*(1), 1–9.

Simonsohn, U., & Gino, F. (2013). Daily horizons: Evidence of narrow bracketing in judgment from 10 years of MBA-admission interviews. *Pyschological Science*, *24*(2), 219–224.

Simonsohn, U., & Loewenstein, G. (2006). Mistake #37: The effect of previously encountered prices on current housing demand. *The Economic Journal*, *116*(508), 175–199.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, *29 (3)*, 281–95.

Singh, M. (2021). Heuristics in the delivery room. *Science*, *374*(6565), 324–329.

Thakral, N., & Tô, L. T. (2020). Daily labor supply and adaptive reference points. *American Economic Review*, *forthcoming*.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84(4)*, 327–352.

Appendix

A Additional Material: Institutional Setting and Descriptives

A.1 Study Grant Program

Candidates at the admission workshops apply for admission into a large merit-based study grant program in Germany. The program is prestigious and has a strong reputation for being highly competitive. It is mostly financed by the German Ministry of Education and administered by a foundation. Students in the program receive (in 2020) a lump-sum payment of at least 300 euros per month. Recipients can additionally receive up to 861 euros per month, depending on their parents' earnings.[1] Additional financial support is offered when spending a semester abroad. In addition, the program offers a large, cost-free course program including language classes abroad, summer schools and academic workshops. Finally, its benefits include many networking opportunities and a high signaling value. As a consequence of these financial and career-related benefits, the stakes for being accepted into the program are high.

The program offers several admission channels. Apart from being nominated by a high-school principal, candidates can qualify for participation in an admission workshop by passing a written test or being nominated by their university during the course of their studies. In this paper, we concentrate on nominations by high-school principals, for two reasons: first, they constitute the most important admission channel (around 55% of all candidates); and second, candidates who participate at later stages of their university studies are no longer randomly matched to evaluators, but rather assigned according to their field of study.

[1] All German students are eligible for financial aid up to 861 euros per month, dependent on their parents' earnings. However, payments have to be repaid after graduation by students who do not receive a merit-based scholarship. The lump-sum payment was increased during our sample period from 150 to 300 euros and the additional monetary benefits are adjusted every year.

A.2 Workshop Schedule

Figure A.1: Illustration of Schedule

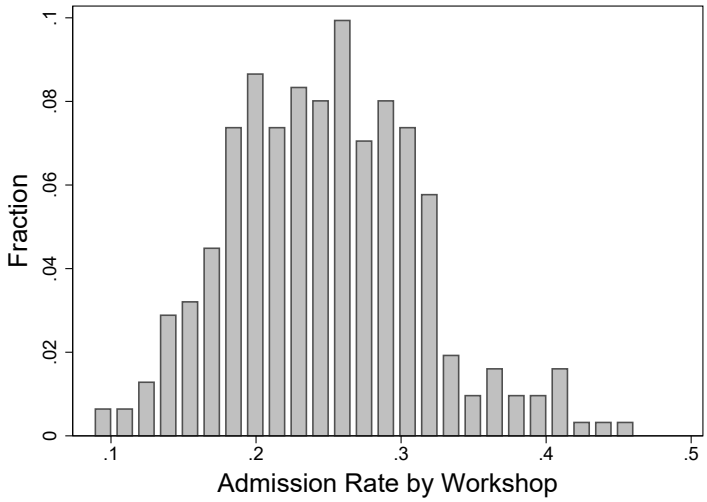
	Duration (minutes)	Type	Interviewer							
			A	B	C	D	E	F	G	H
Day 1	30	Group	1	7	13	19	25	31	37	43
	35	Interview 1	9	15	21	27	33	39	45	3
	35	Interview 1	46	4	10	16	22	28	34	40
	20	**Break**								
	30	Group	2	8	14	20	26	32	38	44
	35	Interview 1	35	41	47	5	11	17	23	29
	35	Interview 1	24	30	36	42	48	6	12	18
	60	**Lunch**								
	30	Group	3	9	15	21	27	33	39	45
	35	Interview 1	31	37	43	1	7	13	19	25
	30	Group	4	10	16	22	28	34	40	46
	20	**Break**								
	35	Interview 1	20	26	32	38	44	2	8	14
	30	Group	5	11	17	23	29	35	41	47
Day 2	35	Interview 2	43	1	7	13	19	25	31	37
	35	Interview 2	38	44	2	8	14	20	26	32
	20	**Break**								
	35	Interview 2	33	39	45	3	9	15	21	27
	30	Group	6	12	18	24	30	36	42	48
	35	Interview 2	28	34	40	46	4	10	16	22
	60	**Lunch**								
	35	Interview 2	23	29	35	41	47	5	11	17
	35	Interview 2	18	24	30	36	42	48	6	12

Note: The time table illustrates the assignment of candidates to evaluators and time slots. Candidates are identified by an ID between 1 and 48. Evaluators are identified by an ID between A and H at the respective time slot. When a candidate ID appears in a slot denoted “Group", this means that the candidate presents in front of her group and moderates a discussion. Interviews are 35 minutes + 5 minutes break.

B Additional Material: Data and Measurement

In the following, we provide additional material on the data sources, randomization checks and the measurement of candidate quality. Figure B.1 shows the distribution of workshop-level admission rates. Table B.1 provides summary statistics on candidate and evaluator characteristics. Table B.2 provides evidence that there is no indication of systematic sorting of candidates to evaluators. Table B.3 shows the relationship between evaluator characteristics and a candidate's rating (column 1) as well as her third-party assessment (column 2). It shows that an evaluator's characteristics only influence her own rating of a candidate, and does not have any spillover on the TPA made by the other two evaluators. Table B.4 presents results from a regression of individual ratings on candidate characteristics.

Figure B.1: Distribution of Workshop-Specific Admission Rates



Note: The figure shows the distribution of workshop-level admission rates (N=312).

Table B.1: Summary Statistics on Evaluator and Candidate Characteristics

	Evaluators		
	N	Mean	SD
Female	2496	0.48	0.50
Age	2496	42.02	11.58
Field: Humanities	2496	0.45	0.50
Field: Social Sciences	2496	0.10	0.31
Field: STEM	2496	0.36	0.48
Field: Medicine	2496	0.08	0.28
Field: Others	2496	0.01	0.09
Experience: 0	2496	0.62	0.48
Experience: 1	2496	0.11	0.31
Experience: 2	2496	0.08	0.28
Experience: 3+	2496	0.18	0.39
Number of interviews	2496	11.81	0.71

	Candidates		
	N	Mean	SD
Female	14733	0.55	0.50
Age	14733	19.62	1.41
Migration Background	14733	0.16	0.37
1st Generation Student	14733	0.26	0.44
High School GPA (in %)	14733	92.07	7.78
Field: Humanities	14733	0.18	0.39
Field: Social Sciences	14733	0.20	0.40
Field: STEM	14733	0.37	0.48
Field: Medicine	14733	0.24	0.43
Field: Others	14733	0.01	0.10

Table B.2: Randomization Check: Relation between Candidate and Evaluator Characteristics

	Candidate Characteristic			
	(1) Female	(2) Age	(3) Field: STEM	(4) Field: Soc. Sciences
Female Evaluator	0.003			
	(0.004)			
Evaluator Age		0.000		
		(0.001)		
Evaluator Field: STEM			-0.008	
			(0.005)	
Evaluator Field: Soc.Sc.				-0.005
				(0.007)
Outcome Mean	0.55	19.62	0.37	0.20
N	29466	29466	29466	29466

Note: Soc.Sc.=Social Sciences. Regressions include workshop fixed effects. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Table B.3: Influence of Evaluator Characteristics on Rating and TPA

	Rating (Std.)	TPA (Std.)
	(1)	(2)
Female	0.036***	0.004
	(0.013)	(0.012)
Age	0.005***	0.000
	(0.001)	(0.001)
Field: Social Sciences	0.026	0.008
	(0.022)	(0.019)
Field: STEM	0.027*	0.004
	(0.015)	(0.012)
Field: Medicine	0.018	-0.010
	(0.025)	(0.024)
Field: Others	0.012	-0.009
	(0.069)	(0.061)
Experience	-0.025***	0.003
	(0.002)	(0.002)
p-value (joint significance)	0.00	0.78
N	29466	29466

Note: Humanities is the omitted study field. Experience is a continuous variable of prior workshop participations by an evaluator. All regressions include workshop fixed effects. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Table B.4: Influence of Candidate Covariates on Ratings and Admission

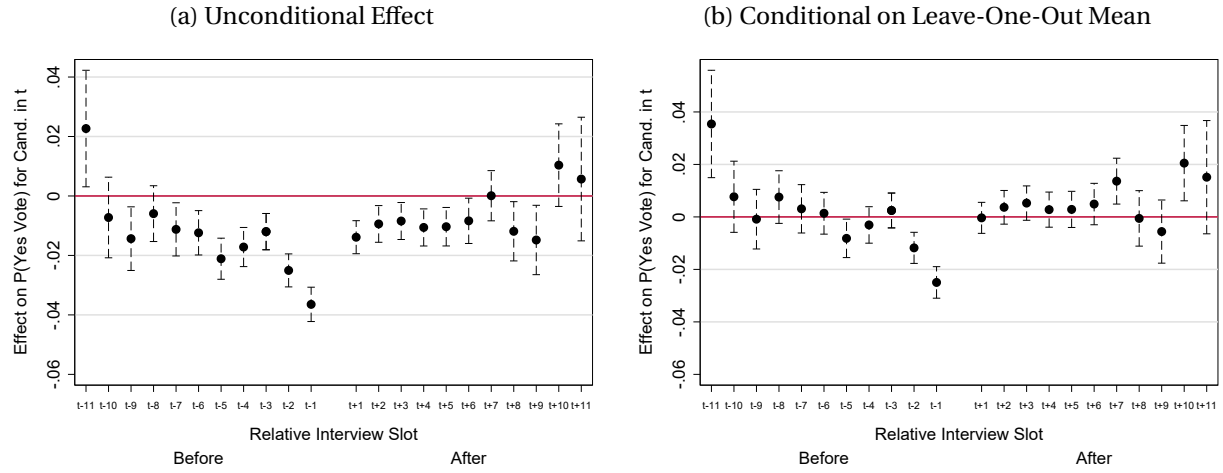
	Rating (Std.) (1)	Admission (2)
GPA Decile: 1	-0.117***	-0.039***
	(0.028)	(0.015)
GPA Decile: 2	-0.132***	-0.061***
	(0.029)	(0.013)
GPA Decile: 3	-0.054*	-0.036**
	(0.031)	(0.015)
GPA Decile: 4	0.009	0.008
	(0.029)	(0.016)
GPA Decile: 6	0.006	-0.004
	(0.033)	(0.017)
GPA Decile: 7	0.084***	0.028*
	(0.029)	(0.015)
GPA Decile: 8	0.089***	0.037**
	(0.028)	(0.015)
GPA Decile: 9	0.141***	0.041**
	(0.031)	(0.016)
GPA Decile: 10	0.208***	0.074***
	(0.027)	(0.015)
Female	-0.070***	-0.052***
	(0.014)	(0.008)
Age	0.059***	0.022***
	(0.007)	(0.003)
Migration Background	0.205***	0.097***
	(0.018)	(0.010)
1st Generation Student	-0.005	0.019**
	(0.016)	(0.009)
Field: Social Sciences	0.007	-0.006
	(0.022)	(0.012)
Field: STEM	-0.107***	-0.071***
	(0.019)	(0.010)
Field: Medicine	-0.013	-0.019*
	(0.021)	(0.011)
Field: Others	-0.117*	-0.056*
	(0.069)	(0.033)
Outcome Mean	-0.00	0.25
R-Squared (Within)	0.02	0.02
N	29466	14733

Note: Humanities is the omitted study field. All regressions include workshop fixed effects. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

C Additional Material: Influence of the Other Candidates and the Role of Relative Timing

Figure C.1 is analogous to Figure 2 (Figure 2b), using the probability of a yes vote as an alternative outcome. Table C.1 reports the coefficients and corresponding p-values illustrated in Figures 2, 2b, and C.1. Tables C.2 and C.3 report coefficients for $k = -1$ and show their robustness to alternative specifications (Table C.3) and to the use of alternative quality measures (Table C.2).

Figure C.1: Effect of Candidate Quality in $t+k$ on the Yes Vote Probability of Candidate in t



Note: Panel (a) shows the estimated coefficients β_k from equation 1, resulting from separate regressions for each value of $k = \{-11, ..., -1, 1, ..., 11\}$. The coefficients measure how the standardized TPA of the candidate interviewed in $t+k$ affects the probability of the candidate in t receiving a yes vote. TPA = third-party assessment of candidate quality (see section 3.3 for details). Panel (b) estimates the additional effect of the candidate interviewed in $t+k$, beyond her contribution to the leave-one-out mean. Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level. Table C.1 reports the corresponding coefficients and p-values.

Table C.1: Coefficients and p-Values Corresponding to Figures 2 and C.1

	Std. Rating, Unconditional			Std. Rating, Conditional			P(Yes), Unconditional			P(Yes), Conditional		
	Coeff.	p-value	p-value (adj.)	Coeff.	p-value	p-value (adj.)	Coeff.	p-value	p-value (adj.)	Coeff.	p-value	p-value (adj.)
t-11	0.001	0.972	1.000	0.034	0.104	1.000	0.023	0.022	0.479	0.035	0.001	0.013
t-10	-0.0170	0.2673	1.0000	0.0184	0.2212	1.0000	-0.0072	0.2917	1.0000	0.0077	0.2635	1.0000
t-9	-0.0403	0.0002	0.0045	-0.0053	0.6436	1.0000	-0.0144	0.0082	0.1796	-0.0009	0.8812	1.0000
t-8	-0.027	0.009	0.192	0.007	0.547	1.000	-0.006	0.212	1.000	0.008	0.137	1.000
t-7	-0.0217	0.0136	0.3000	0.0138	0.1331	1.0000	-0.0112	0.0133	0.2932	0.0031	0.5096	1.0000
t-6	-0.0358	0.0000	0.0001	-0.0024	0.7615	1.0000	-0.0124	0.0011	0.0234	0.0014	0.7300	1.0000
t-5	-0.045	0.000	0.000	-0.012	0.093	1.000	-0.021	0.000	0.000	-0.008	0.028	0.615
t-4	-0.0473	0.0000	0.0000	-0.0127	0.0546	1.0000	-0.0172	0.0000	0.0000	-0.0031	0.3851	1.0000
t-3	-0.0358	0.0000	0.0000	-0.0000	0.9975	1.0000	-0.0120	0.0001	0.0024	0.0025	0.4670	1.0000
t-2	-0.056	0.000	0.000	-0.021	0.000	0.004	-0.025	0.000	0.000	-0.012	0.000	0.002
t-1	-0.0922	0.0000	0.0000	-0.0625	0.0000	0.0000	-0.0365	0.0000	0.0000	-0.0250	0.0000	0.0000
t+1	-0.0295	0.0000	0.0000	0.0064	0.2783	1.0000	-0.0139	0.0000	0.0000	-0.0004	0.9045	1.0000
t+2	-0.0230	0.0005	0.0102	0.0122	0.0650	1.0000	-0.0094	0.0027	0.0593	0.0037	0.2613	1.0000
t+3	-0.0199	0.0025	0.0555	0.0161	0.0187	0.4120	-0.0084	0.0076	0.1679	0.0053	0.1135	1.0000
t+4	-0.0332	0.0000	0.0000	0.0023	0.7483	1.0000	-0.0106	0.0008	0.0183	0.0028	0.4141	1.0000
t+5	-0.0281	0.0001	0.0020	0.0087	0.2462	1.0000	-0.0103	0.0017	0.0366	0.0029	0.4126	1.0000
t+6	-0.0235	0.0022	0.0495	0.0147	0.0598	1.0000	-0.0084	0.0306	0.6725	0.0049	0.2211	1.0000
t+7	-0.0141	0.0993	1.0000	0.0233	0.0075	0.1658	0.0001	0.9847	1.0000	0.0136	0.0021	0.0466
t+8	-0.0313	0.0032	0.0695	0.0017	0.8757	1.0000	-0.0119	0.0188	0.4130	-0.0006	0.9161	1.0000
t+9	-0.0328	0.0032	0.0708	-0.0022	0.8446	1.0000	-0.0148	0.0123	0.2711	-0.0056	0.3595	1.0000
t+10	0.0019	0.8898	1.0000	0.0327	0.0217	0.4765	0.0104	0.1406	1.0000	0.0205	0.0048	0.1061
t+11	-0.0153	0.4382	1.0000	0.0183	0.3682	1.0000	0.0057	0.5874	1.0000	0.0151	0.1644	1.0000
Joint test		0.00			0.00			0.00			0.00	
$(t-1=t+1)$		0.00			0.00			0.00			0.00	
Before vs. after		0.04			0.01			0.08			0.22	

Note: The table shows the coefficients and p-values corresponding to Figures 2 and C.1. P-values are adjusted using Bonferroni. At the bottom, we report tests on the equality of $t-1$ and $t+1$. The last line reports a test on the equality of the average coefficient for $k<-1$ and the average coefficient for $k\geq 1$.

Table C.2: Additional Influence of the Previous Candidate: Robustness to Sample and Specification

	Std. Rating		P(Yes Vote)
	(1)	(2)	(3)
Panel A: Baseline			
TPA (std.), t-1	-0.063***	-0.062***	-0.025***
	(0.006)	(0.006)	(0.003)
Leave-one-out Mean TPA (std.)	-0.108***	-0.110***	-0.043***
	(0.009)	(0.008)	(0.003)
TPA (std.), t	0.360***	0.348***	0.144***
	(0.006)	(0.006)	(0.003)
Panel B: Exclusion of marginal candidates			
TPA (std.), t-1	-0.065***	-0.064***	-0.025***
	(0.007)	(0.007)	(0.004)
Leave-one-out Mean TPA (std.)	-0.108***	-0.111***	-0.037***
	(0.010)	(0.010)	(0.004)
TPA (std.), t	0.346***	0.332***	0.139***
	(0.008)	(0.008)	(0.004)
Panel C: Estimation with Interviewer FE			
TPA (std.), t-1	-0.062***	-0.061***	-0.024***
	(0.006)	(0.006)	(0.003)
TPA (std.), t	0.389***	0.377***	0.155***
	(0.006)	(0.006)	(0.003)
Panel D: Estimation with Candidate FE			
TPA (std.), t-1	-0.064***	-0.062***	-0.022***
	(0.008)	(0.008)	(0.004)
Leave-one-out Mean TPA (std.)	-0.074***	-0.076***	-0.029***
	(0.010)	(0.010)	(0.004)
Controls	No	Yes	Yes
Outcome Mean	0.00	0.00	0.37
N	26970	26970	26970

Note: TPA = third-party assessment of candidate quality (see section 3.3 for details). The leave-one-out mean is computed at the level of the evaluator's interview sequence. All regressions include workshop fixed effects. Controls include candidate characteristics, evaluator characteristics and interview order. In Panel B, marginal candidates are candidates whose sum of ratings is at or one point below the admission cut-off (22 or 23 points). It is possible that individual ratings of these candidates were adjusted during the final committee meeting. In Panel C, the leave-one-out mean TPA is omitted due to collinearity with interviewer fixed effects. In Panel C, the candidate's own TPA is omitted due to collinearity with candidate fixed effects. $^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$. Standard errors are clustered at the workshop level (N=312).

Table C.3: Additional Influence of the Previous Candidate: Robustness to Alternative Quality Measures

	Std. Rating		P(Yes Vote)
	(1)	(2)	(3)
Panel A: Baseline			
TPA (std.), t-1	-0.063***	-0.062***	-0.025***
	(0.006)	(0.006)	(0.003)
Leave-one-out Mean TPA (std.)	-0.108***	-0.110***	-0.043***
	(0.009)	(0.008)	(0.003)
TPA (std.), t	0.360***	0.348***	0.144***
	(0.006)	(0.006)	(0.003)
Panel B: TPA includes group discussion rating only			
TPA (std.), t-1	-0.036***	-0.036***	-0.014***
	(0.006)	(0.006)	(0.003)
Leave-one-out Mean Rating group (std.)	-0.076***	-0.076***	-0.029***
	(0.009)	(0.009)	(0.004)
Rating Group (std.)	0.212***	0.201***	0.083***
	(0.007)	(0.007)	(0.003)
Panel C: TPA includes other interview rating only			
TPA (std.), t-1	-0.059***	-0.058***	-0.024***
	(0.006)	(0.006)	(0.003)
Leave-one-out Mean Rating oth. int. (std.)	-0.087***	-0.090***	-0.035***
	(0.008)	(0.008)	(0.003)
Rating other int. (std.)	0.344***	0.330***	0.136***
	(0.007)	(0.007)	(0.003)
Panel D: Predicted quality based on GPA, age and major			
Predicted Rating (std.), t-1	-0.025***	-0.024***	-0.005
	(0.007)	(0.007)	(0.003)
Leave-one-out Mean TPA (std.)	-0.055***	-0.058***	-0.028***
	(0.011)	(0.011)	(0.005)
Predicted Rating (std.)	0.203***	0.170***	0.077***
	(0.008)	(0.009)	(0.004)
Controls	No	Yes	Yes
Outcome Mean	0.00	0.00	0.37
N	26970	26970	26970

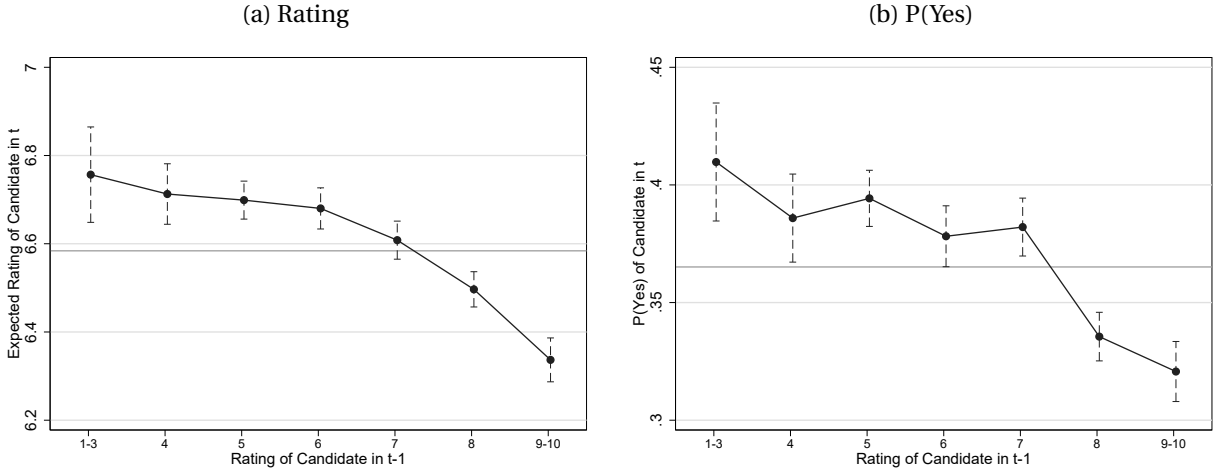
Note: TPA = third-party assessment of candidate quality (see section 3.3 for details). The leave-one-out mean is computed at the level of the evaluator's interview sequence. All regressions include workshop fixed effects. Controls include candidate characteristics, evaluator characteristics and interview order. In Panel D, we predict ratings by regressing the rating on characteristics of the candidates, while leaving out the workshop itself. In addition to candidate controls, the prediction is based on indicators of the candidate's home and university federal state. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

D Additional Material: Autocorrelation in Assessments

Figure D.1 plots the non-linear autocorrelation. It shows that the autocorrelation in ratings is more pronounced at the higher end of the previous candidate's rating distribution, while being rather flat for ratings at the lower end of the distribution.

Moreover, we provide several robustness checks for the estimated autocorrelation presented in Table 3 (section 4.2). Table D.1 shows that the effects on ranking and admission outcomes replicate when using the previous candidate's TPA instead of her rating as the regressor. Tables D.2 and D.3 report results from regressions with candidate fixed effects and evaluator fixed effects, respectively. In Figure D.2, we interact the non-linear influence of the previous candidate's quality and the non-linear autocorrelation with a median split in own quality. Figure D.3 shows the autocorrelation beyond t-1. Tables D.4 to D.5 test for heterogeneity in the autocorrelation with respect to evaluator and candidate characteristics.

Figure D.1: Non-Linear Autocorrelation in Ratings



Note: The figures plot margins based on estimates of equation 2, controlling for workshop fixed effects, the evaluator's leave-one-out mean assessment of candidates in the sequence, evaluator and candidate characteristics and interview order. Ratings of 8 points and above imply a yes vote. The gray vertical line shows the outcome average. N=26,970. Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level.

Table D.1: Influence of the Previous Candidate's TPA on Ranking and Admission Outcomes

	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)
TPA (std.), t-1	-0.208***	-0.013***	-0.011***
	(0.020)	(0.002)	(0.002)
Leave-one-out Mean TPA (std.)	-0.367***	-0.021***	-0.017***
	(0.012)	(0.002)	(0.003)
TPA (std.), t	1.214***	0.082***	0.279***
	(0.020)	(0.002)	(0.003)
Controls	Yes	Yes	Yes
Outcome Mean	6.43	0.15	0.25
R-Squared	0.17	0.07	0.42
N	26970	26970	26970

Note: TPA = third-party assessment of candidate quality (see section 3.3 for details). The leave-one-out mean is computed at the level of the evaluator's interview sequence. All regressions include workshop fixed effects. Controls include candidate characteristics, evaluator characteristics and interview order. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Table D.2: Robustness Checks: Autocorrelation Estimated with Candidate Fixed Effects

	Rating (Std.)	P(Yes Vote)	Rank	P(Best)
	(1)	(2)	(3)	(4)
Rating (t-1) (std.)	-0.072***			
	(0.009)			
Yes (t-1)		-0.061***	-0.419***	-0.031***
		(0.008)	(0.052)	(0.006)
Leave-one-out Mean Rating	0.313***	0.064***	-0.644***	-0.014*
	(0.020)	(0.010)	(0.061)	(0.008)
Leave-one-out Share Yes	-0.549***	-0.108**	-2.275***	-0.178***
	(0.078)	(0.050)	(0.259)	(0.036)
Controls	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.37	6.43	0.15
N	26970	26970	26970	26970

Note: All regressions include candidate fixed effects. The leave-one-out mean is computed at the level of the evaluator's interview sequence. As the admission outcome does not vary at the candidate level, this outcome is omitted from the table. Further controls include evaluator characteristics and interview order. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=312).

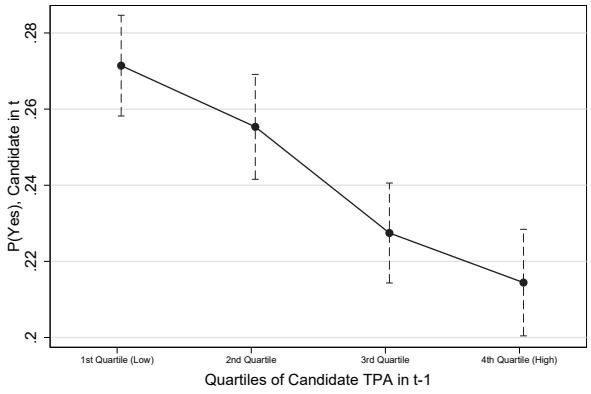
Table D.3: Robustness Checks: Autocorrelation Estimated with Evaluator Fixed Effects

	Rating (Std.)	P(Yes Vote)	Rank	P(Best)	P(Admission)
	(1)	(2)	(3)	(4)	(5)
Rating (t-1) (std.)	-0.145***				
	(0.006)				
Yes (t-1)		-0.139***	-0.973***	-0.067***	-0.052***
		(0.006)	(0.046)	(0.005)	(0.004)
Controls	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.00	0.37	6.43	0.15	0.25
N	26970	26970	26970	26970	26970

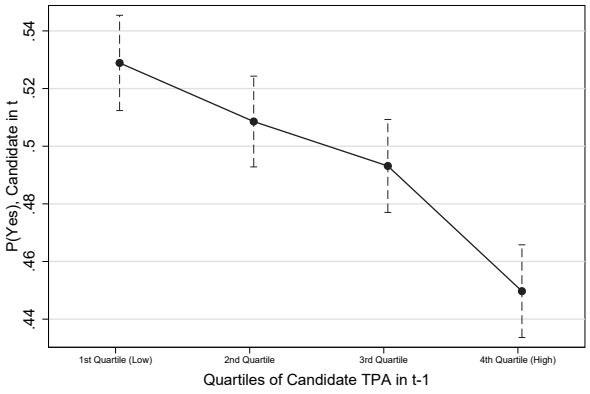
Note: All regressions include evaluator fixed effects. Due to collinearity, the evaluator's leave-one-out mean assessments are omitted. Further controls include candidate characteristics and interview order. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Figure D.2: Influence of the Previous Candidate, by Current Candidate's TPA

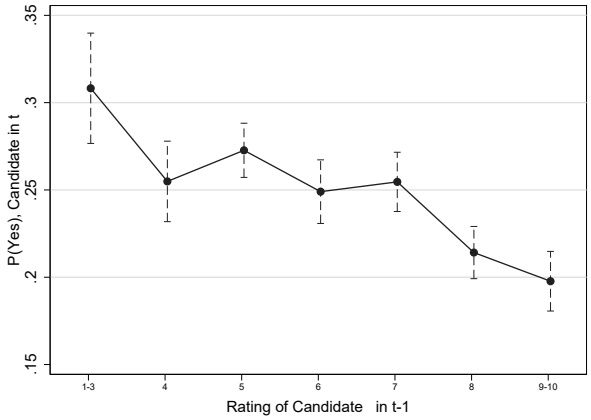
(a) Effect of Previous TPA for Candidates of Low TPA



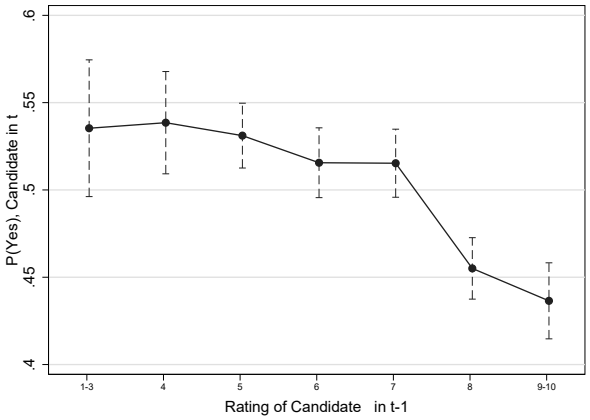
(b) Effect of Previous TPA for Candidates of High TPA



(c) Autocorrelation for Candidates of Low TPA



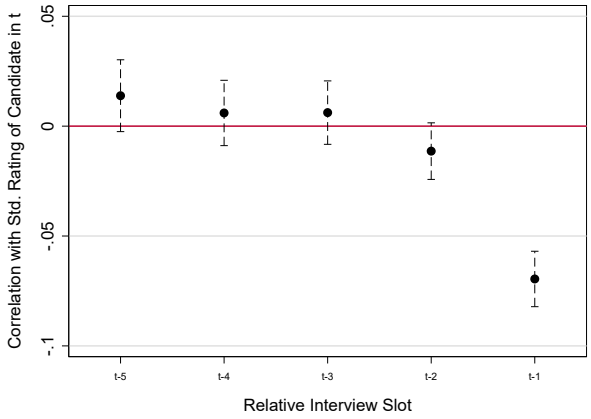
(d) Autocorrelation for Candidates of High TPA



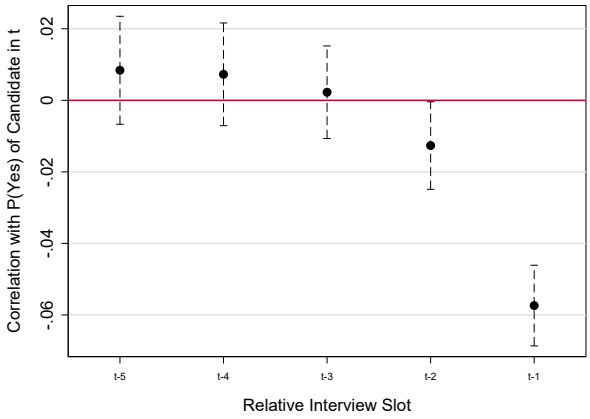
Note: "Low TPA": third-party assessment of quality ≤ median. "High TPA": third-party assessment of quality > median. Estimates result from two-way-interacted regression models. The regression underlying panels (a) and (b) controls for workshop fixed effects, the leave-one-out mean TPA at the evaluator level, candidate characteristics (including TPA), evaluator characteristics and interview order. The regression underlying panels (c) and (d) controls for workshop fixed effects, the evaluator's leave-one-out mean assessments, candidate characteristics (including TPA), evaluator characteristics and interview order. N=26,970. 95% confidence intervals, with standard errors clustered at the workshop level.

Figure D.3: Autocorrelation Beyond t-1

(a) Rating



(b) P (Yes Vote)



Note: Each coefficient results from a separate regression, where the assessment of the candidate in t is related to the assessment of the candidate in $t+k$, $k \in \{-5,...,-1\}$. All regressions include workshop fixed effects and the evaluator's leave-one-out mean in ratings and yes votes. Further controls include candidate characteristics (including TPA), evaluator characteristics and interview order. 95% confidence intervals, with standard errors clustered at the workshop level.

Table D.4: Heterogeneity in the Autocorrelation: Evaluator Characteristics

	P(Yes Vote)			
	(1)	(2)	(3)	(4)
Yes (t-1)	-0.057***	-0.066***	-0.053***	-0.058***
	(0.007)	(0.008)	(0.008)	(0.007)
Experience: 1 x Yes (t-1)	0.032			
	(0.021)			
Experience: 2 x Yes (t-1)	-0.014			
	(0.022)			
Experience: 3+ x Yes (t-1)	-0.014			
	(0.014)			
Age > Median x Yes (t-1)		0.018		
		(0.012)		
Female x Yes (t-1)			-0.008	
			(0.012)	
Training x Yes (t-1)				0.002
				(0.015)
Controls	Yes	Yes	Yes	Yes
Outcome Mean	0.37	0.37	0.37	0.37
N	26970	26970	26970	26970

Note: All regressions include workshop fixed effects and control for the evaluator's leave-out mean of ratings and yes votes. Experience denotes the number of prior workshop participations. Training equals one if the evaluator participated in an interviewer training before the workshop. Controls are candidate and evaluator characteristics and interview order. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Table D.5: Heterogeneity in the Autocorrelation: Candidate Characteristics

	P(Yes Vote)					
	(1)	(2)	(3)	(4)	(5)	(6)
Yes (t-1)	-0.060***	-0.055***	-0.050***	-0.058***	-0.057***	-0.064***
	(0.009)	(0.006)	(0.008)	(0.006)	(0.006)	(0.007)
Female x Yes (t-1)	0.004					
	(0.011)					
Age > Median x Yes (t-1)		-0.016				
		(0.015)				
GPA > Median x Yes (t-1)			-0.016			
			(0.011)			
Migration Background x Yes (t-1)				0.001		
				(0.016)		
Parents w/out Univ. Degree x Yes (t-1)					-0.000	
					(0.013)	
Field: STEM=1 x Yes (t-1)						0.017
						(0.012)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.37	0.37	0.37	0.37	0.37	0.37
N	26970	26970	26970	26970	26970	26970

Note: All regressions include workshop fixed effects and control for the evaluator's leave-out mean of ratings and yes votes. Controls are candidate and evaluator characteristics and interview order. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=312).

E Additional Material: Behavioral Mechanism

E.1 Additional Material: Theoretical Intuition

In the following, we provide a more formal framework for the intuition discussed in section 5. The framework adapts the model by Bordalo et al. (2020) to our setting.

Setup An evaluator decides on the rating of a candidate interviewed in period t, based on her valuation of that candidate. Observing a candidate with perceived quality $\tilde{q}_t$ in a context c_t defines an interview experience. This experience cues the recall of past interview experiences, which are used to form the reference norm on which the evaluation is based.

Experience-based quality norm The norm for a candidate interviewed in period t is formed by recalling and weighting past experiences of other candidates. In this process, interviews that are similar in terms of context c_t receive a stronger weight, where the similarity of an interview that took place in period $t-l$ is measured by the function $S(c_{t-l})$. Observed context variables that vary in our setup are the time of interview as well as candidate characteristics (e.g., gender or study field). Similarity decreases in the distance between two interview contexts. In the most simple case where only the Euclidean distance in time between two interviews is considered, $S(c_{t-l}) = S(|t-(t-l)|)$ for l=1,..,11, where t indicate the point in time of one interview and $t-l$ indicates the point in time of another interview. $S : \mathbb{R}_+ \to \mathbb{R}_+$ is decreasing.

Intuitively, the quality norm is a similarity-weighted average of observed past quality, formally written as:

$$q_t^n(c_t) = \sum_{l=1}^{t-1} \tilde{q}_{t-l}\omega_{t-1},$$

where the weight of a prior interview experience e_{t-l} is determined by her relative similarity to the current interview experience:

$$\omega_{t-1} = \frac{S(c_{t-l})}{\sum_{l=1}^{t-1} S(c_{t-l})},$$

Importantly, the notion of relative similarity implies that an increase in similarity of one candidate decreases the weight of any other observed candidate.

Valuation The valuation is a function of the quality norm and the quality of the candidate herself. We focus on the instantaneous valuation formed at the time of the interview t, thereby abstracting from any ex-post adjustment which can occur after seeing all candidates. We define the instantaneous valuation as:

$$V_t = \tilde{q}_t + \sigma(\tilde{q}_t, q_t^n(c_t)) \times (\tilde{q}_t - q_t^n(c_t)) \tag{5}$$

The valuation V_t not only depends on the candidate's own quality as perceived by the evaluator ($\tilde{q}_t$), but also on its difference to the reference norm (q_t^n). The salience function σ determines how much this difference — i.e., the surprise relative to the norm — attracts the evaluator's attention. Large surprises are more salient, with diminishing sensitivity. More formally, $\sigma(\tilde{q}_t, q_t^n)$ is a salience function that is symmetric, homogeneous of degree zero, increasing in $\frac{x}{y}$ for $x \geq y > 0$ and $\sigma(y, y) = 0$; bounded by $\lim_{x/y \to \infty} \sigma(x/y, 1) = \sigma$.

Alternative approach: Valuation with anchoring to the norm The original framework by Bordalo et al. (2020) proposes a slightly different expression for the valuation, which includes the notion of anchoring to the norm. Adapted to our setup, the valuation is then defined as:

$$V_t = q_t^n(c_t) + \sigma(\tilde{q}_t, q_t^n(c_t)) \times (\tilde{q}_t - q_t^n(c_t)) \tag{6}$$

According to this model, the quality norm q_t^n affects valuation V_t in two ways. First, the valuation is anchored to the norm. Second, it increases in the difference between the candidate's own quality and the norm, as described above. In this framework, the valuation of a candidate reacts to a change in the (perceived) quality of the previous candidate as follows:

$$\frac{\partial V_t}{\partial \tilde{q}_{t-1}} = \omega_{t-1} + \frac{\partial \sigma(\tilde{q}_t, q_t^n)}{\partial q_t^n} \omega_{t-1}(\tilde{q}_t - q_t^n) - \sigma(\tilde{q}_t, q_t^n)\omega_{t-1}$$

The first term describes the anchoring of the current valuation to the norm. Anchoring leads to a positive influence of the previous candidate's quality on the current candidate's valuation, i.e., assimilation. The second and third terms describe contrasting: an increase in the previous candidate's quality makes the current candidate look 'surprisingly' weak(er), thereby reducing her valuation.

It is straightforward to see that the strength of both anchoring and contrasting depends on ω_{t-1}, the weight of the previous candidate in the norm. However, which of the two counteracting mechanisms dominates depends on the size of the surprise as described by $q_t - q_t^n$. If the surprise is small, it does not capture the evaluator's attention. Anchoring is thus relatively important and can lead to assimilation of two subsequent candidates. For larger surprises, contrasting as described by the second and third parts dominates.

Our empirical results — as reported in Figure 6 of section 5 — neither reject nor confirm the presence of assimilation effects due to anchoring. On the one hand, the observed pattern would be in line with a valuation without the assimilation component. In the absence of assimilation, a flat relationship for small differences is predicted based on the low salience of such. On the other hand, we cannot reject the presence of assimilation. Our estimates are not sufficiently precise to exclude the notion that small positive (negative) differences have a positive (negative) effect, which would hint at assimilation. In particular, the model parameters can be specified to yield the prediction of very small assimilation effects, which we might simply be unable to detect.[1,2]

[1] One example is a specification where differences in quality attract attention very quickly and strongly. Small surprises already attract the attention of the evaluator, the salience is large and it strongly increases with TPA differences.

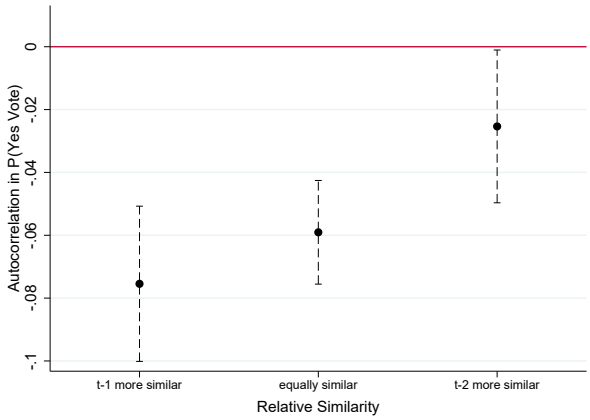
[2] Another potential explanation would be a small symmetric measurement error. A measurement error in the differences could lead to a relatively flat relationship. Intuitively, a (locally) u-shaped curve would (due to the measurement error) take values from the left and right, thereby producing a 'flat' curve. However, note that such a symmetric measurement error would not induce a linear effect to flatten around zero.

E.2 Additional Material: Additional Dimensions of Similarity

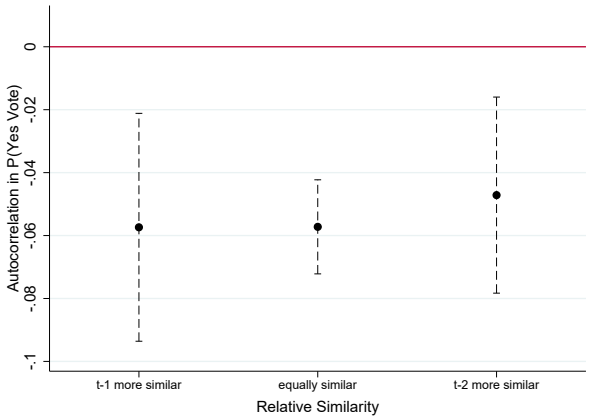
E.2.1 The Role of Additional Similarity

Figure E.1: The role of relative similarity in characteristics

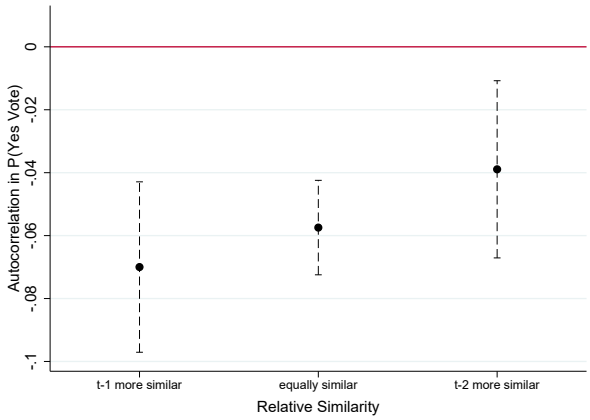
(a) Gender



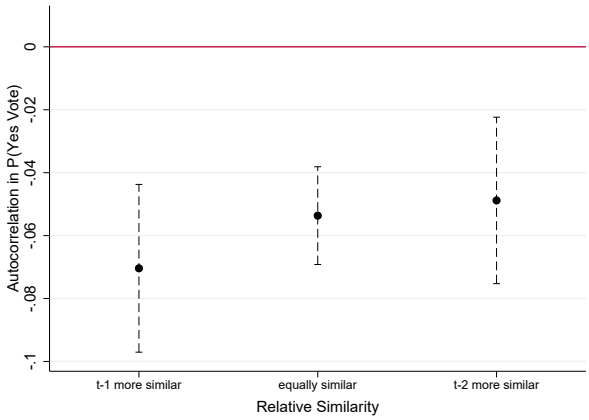
(b) Migration Background



(c) 1st Generation Student



(d) Study Field



Note: The figure presents estimates of the autocorrelation based on equation 2, where the previous candidate's yes vote is interacted with her relative similarity to the candidate in t in a given observable characteristics. "t-1 more similar" = the candidate in t-1, but not the candidate in t-2 shares a given characteristic with the candidate in t. "Equally similar" = both t-1 and t-2 either do or do not share a given characteristic with the candidate in t. "t-2 more similar" = the candidate in t-2, but not the candidate in t-1 shares a given characteristic with the candidate in t. 95% confidence intervals, with standard errors clustered at the workshop level.

E.2.2 Symmetric Similarity and the Role of Gender

We test whether both females and males are more strongly influenced by subsequent candidates of their own gender. We pre-registered the hypothesis that the influence varies with respect to the sequencing of gender. In particular, the results based on our pilot dataset showed an asymmetry: while the gender of the previous candidate did not matter for female candidates, male candidates were not harmed by following a strong female candidate. This asymmetry is reported in columns 1 and 2 of Table E.1 and Figure E.2. Both show that male candidates are as-good-as unaffected by the measured quality and rating of previous candidates who are female. In turn, the gender of the previous candidate does not significantly matter for female candidates. This finding pointed towards asymmetric similarity, where female candidates are compared with previous candidates of both genders, but male candidates are not compared with female candidates (Tversky, 1977). Moreover, the asymmetry in the previous candidate's influence had relevant implications for the 'gender assessment gap': in the pilot data, males who follow a male candidate are 5% more likely to receive a yes vote than females, compared with 20% for males who follow a female candidate.

Columns 3 and 4 of Table E.1 and Figure E.3 show the results from our replication exercise based on the main data. They reject the hypothesis that female candidates have no influence on male candidates. While panel (b) of Figure E.3 provides suggestive evidence that male candidates are more affected by previous strong male than by previous strong female candidates, the pattern is not clearly distinguishable from the one for female candidates (panel a). This also implies that the size of the gender gap is not significantly affected by the previous candidate's gender.

Table E.1: Gender Sequence and the Influence of the Previous Candidate

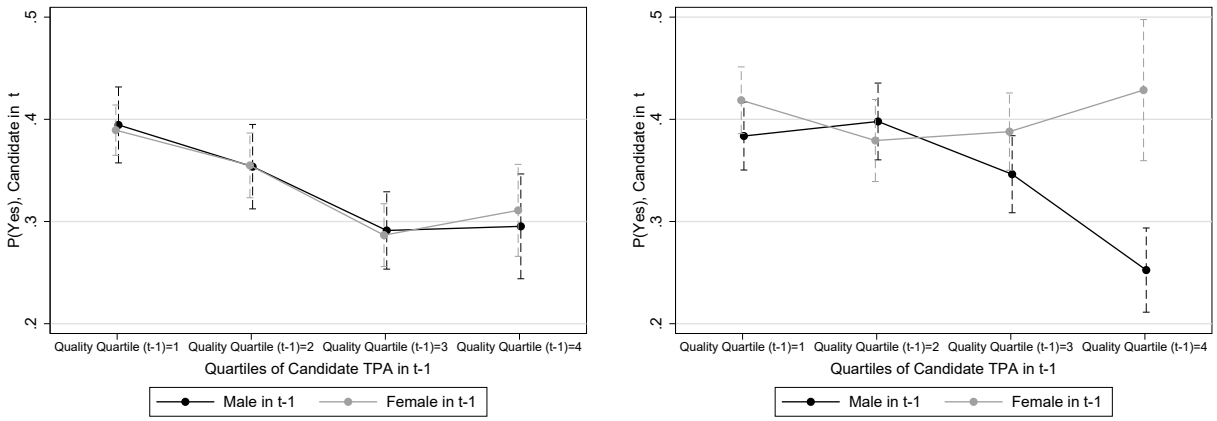
	Pilot Data		Main Data	
	(1) Rating (Std.)	(2) P(Yes Vote)	(3) Rating (Std.)	(4) P(Yes Vote)
Male × Male (t-1) × TPA (std.), t-1	-0.079***		-0.064***	
	(0.019)		(0.012)	
Male × Female (t-1) × TPA (std.), t-1	-0.025		-0.055***	
	(0.024)		(0.011)	
Female × Male (t-1) × TPA (std.), t-1	-0.059**		-0.070***	
	(0.023)		(0.013)	
Female × Female (t-1) × TPA (std.), t-1	-0.084***		-0.063***	
	(0.017)		(0.010)	
Male × Male (t-1) × Yes (t-1)		-0.071***		-0.069***
		(0.019)		(0.012)
Male × Female (t-1) × Yes (t-1)		-0.015		-0.050***
		(0.023)		(0.012)
Female × Male (t-1) × Yes (t-1)		-0.087***		-0.052***
		(0.023)		(0.012)
Female × Female (t-1) × Yes (t-1)		-0.106***		-0.058***
		(0.016)		(0.009)
Controls	Yes	Yes	Yes	Yes
p-value: Male (t) coeffs equal	0.08	0.08	0.57	0.26
p-value: Female (t) coeffs equal	0.37	0.41	0.66	0.71
N	8522	8522	26970	26970

Note: All regressions include workshop fixed effects and control variables. Columns 1 and 3 also control for the leave-one-out mean TPA of the interview sequence. Columns 2 and 4 also control for the evaluator's leave-one-out mean of ratings and yes votes. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Figure E.2: Interaction between Prior Candidate Quality and the Gender Sequence: Pilot Data

(a) Female Candidate in t

(b) Male Candidate in t

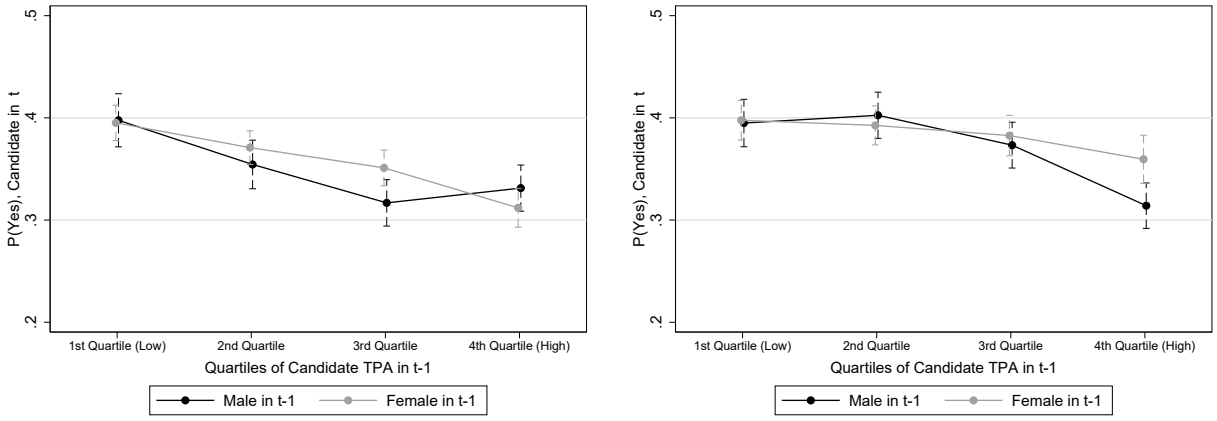


Note: The "pilot data" include the academic year 2012/13 (N=8,522). Estimates in panels (a) and (b) result from the same two-way-interacted regression model. Controls include the leave-one-out mean TPA of the interview sequence, candidate and evaluator characteristics, interview order and workshop fixed effects. 95% confidence intervals, with standard errors clustered at the workshop level.

Figure E.3: Interaction between Prior Candidate Quality and the Gender Sequence: Main Data

(a) Female Candidate in t

(b) Male Candidate in t

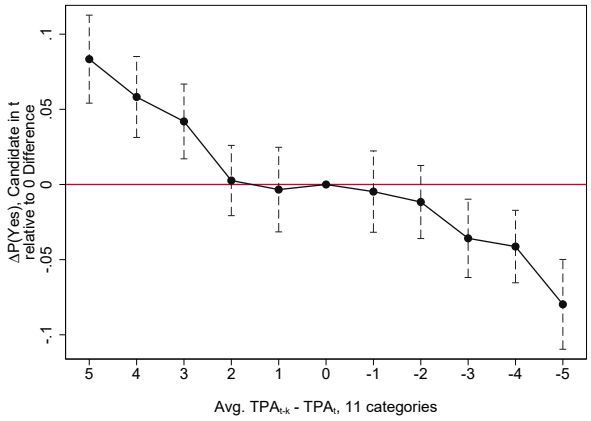


Note: Estimates in panels (a) and (b) result from the same two-way-interacted regression model. Controls include the leave-one-out mean TPA of the interview sequence, candidate and evaluator characteristics, interview order and workshop fixed effects. 95% confidence intervals, with standard errors clustered at the workshop level. 95% confidence intervals.

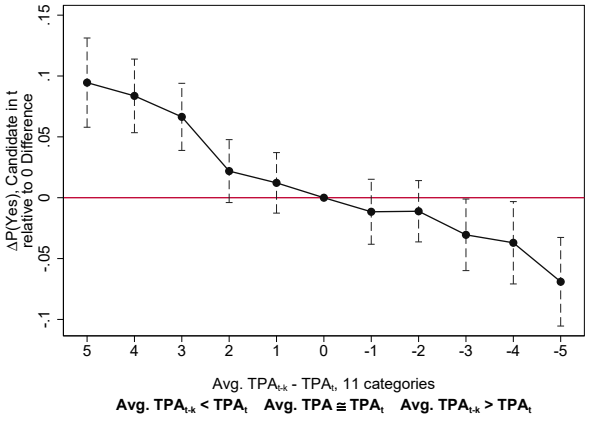
E.3 Additional Material: Attention and Size of the Surprise

Figure E.4: Alternative Proxies of the Quality Norm

(a) Norm=Average TPA of Previous Two Candidates



(b) Norm=Average TPA of All Previous Candidates



Note: In both panels, the x-axis denotes the difference between current candidate's TPA and a proxy for the quality norm in eleven equally-sized categories. In panel (a), the norm is approximated by the average TPA of the two previous candidates. In panel (b), the norm is approximated by the average TPA of all previous candidates. The y-axis shows estimated coefficients on the probability of receiving a yes vote for the candidate in t. The underlying regression includes dummies for the candidate's own TPA. Further controls are the leave-one-out mean TPA, candidate characteristics, evaluator characteristics and interview order. N=26,970. 95% confidence intervals, with standard errors clustered at the workshop level.

E.4 Additional Material: Alternative Mechanisms

Table E.2: Test for Additional Influence of Streaks

	Rating (Std.)	P(Yes Vote)
	(1)	(2)
Yes (t-1)=1	-0.130***	-0.058***
	(0.016)	(0.007)
Yes (t-1) and (t-2)	0.012	0.006
	(0.024)	(0.012)
Controls	Yes	Yes
N	24474	24474

Note: The table tests whether the rating (column 1) and the probability of a yes vote (column 2) changes when the evaluator gives the two preceding — instead of the one preceding — candidates a yes vote. All regressions include workshop fixed effects, the evaluator's leave-one-out mean rating and re of yes votes, candidate characteristics (including TPA), evaluator characteristics and interview order dummies. The regressions are based on candidates with at least two preceding candidates, explaining why the number of observations is smaller than in the main analyses. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are clustered at the workshop level (N=312).

Table E.3: Previous Decisions and Previous Quality

	P(Yes Vote)		
	(1)	(2)	(3)
TPA (std.), t-1	-0.025***	-0.018***	-0.017***
	(0.003)	(0.003)	(0.003)
Yes (t-1)		-0.046***	-0.036***
		(0.006)	(0.010)
Rating (t-1) (std.)			-0.006
			(0.005)
Controls	Yes	Yes	Yes
Outcome Mean	0.37	0.37	0.37
R-Squared	0.12	0.12	0.12
N	26970	26970	26970

Note: All regressions include workshop fixed effects, the evaluator's leave-one-out mean rating, share of yes votes and leave-one-out mean of TPA, candidate characteristics (including TPA), evaluator characteristics and interview order dummies. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level (N=312).

F Additional Material: Structural Estimation

F.1 Estimation

We use a minimum-distance estimator to estimate the model described in section 6. Let $m(\xi)$ denote the the vector of simulated moments as a function of the model parameters, and $\hat{m}$ the vector of moments observed in the data. The estimator chooses the parameter vector $\hat{\xi}$ that minimizes the distance $(m(\xi)-\hat{m})'W(m(\xi)-\hat{m})$. As a weighting matrix W, we use the diagonal of the inverse of the variance-covariance matrix.[3]

In particular, we estimate the variance of the parameters using

$$(\hat{G}'W\hat{G})^{-1}(\hat{G}'W(1+J_m/J_s)\hat{\Sigma}W\hat{G})(\hat{G}'W\hat{G})^{-1}/N$$

where $\hat{G} \equiv \nabla_\xi m(\xi)$, $\hat{\Sigma} \equiv Var(m(\xi))$,[4] J_m is the number of empirical observations used to calculate the moment, and J_s is the corresponding number of simulated observations.

We calibrate the standard deviation of the error term to a value of 1.68, which corresponds to the standard deviation of the residual rating (conditional on own, previous and leave-one-out mean measured quality). We fix the draw of errors across all estimations.

In every simulation step, we simulate a population of 10,000 evaluators, who each interview 12 candidates. We solve the minimization problem using a Python implementation of the DFO-LS algorithm (Gabler, 2021). We impose the following box constraints on the parameters: $\alpha \geq 0$ (own quality), $\beta \leq 0$ (average quality), $\sigma > 0$, $\delta_1 > 0$ and $e^{-\delta_2} \in (0,1]$. We calibrate θ to different values.

To increase our confidence in the identification and estimation of the structural parameters, we simulate a set of 3,000 evaluators with given model parameters, which corresponds roughly to our actual sample size. We start our estimation procedure (as described above, with

[3] Altonji and Segal, 1996 show that using the full inverse of the variance-covariance matrix can lead to numerical instability of the estimator.

[4] We assume a zero covariance across the following sets of moments: (i) moments that describe the relation between a candidate's own quality measure and the assessment; (ii) moments that capture adjustment of assessments to the average quality measure of the other candidates (leave-one-out mean); (iii) moments that capture the additional influence of the previous candidates' measured quality.

10,000 draws) at a perturbed initial value and check that the estimator is able to back out the original parameters that were used to simulate the data.

For each model whose estimates are reported in the paper or appendix, we picked 10 random starting points from the parameter space and reported the model with the lowest criterion value. As some parameters are only bounded from above or below, we use a targeted parameter space, from which we randomly pick start values. We allow $\alpha \in (0,2]$, $\beta \in [-2,0)$, $\sigma(0,2]$, $\delta_1 \in [0,10]$, and $e^{-\delta_2} \in (0,1)$. As this parameter space is still rather large, we pick one vector of starting values by hand to make sure that at least one sensible combination is considered.

F.2 Identification

To identify the model parameters, we use moments that correspond to distinct aspects of the data. They describe how a candidate's rating reacts to her own and the other candidates' quality, as measured through the third-party assessment (TPA).

The variation in the preceding candidates' influence identifies the importance of relative time for similarity-based recall as described by δ_1. More precisely, the moments describe how ratings respond to the quality of candidates in the three preceding interview slots. We capture this through the coefficients of separate OLS regressions that link a current candidate's rating to a dummy indicating whether the candidate in t-1/ t-2/t-3 was of high measured quality (i.e., in the highest quality quartile).

The role of similarity in candidate characteristics, as captured by δ_2, is identified through moments that describe the interaction between the previous candidate's influence and her relative similarity in terms of observed characteristics. A distinctive feature of the model is that similarity enters in relative terms. To capture this notion, we allow the influence of the previous candidate to depend on how similar the candidate in t-1 is compared to the candidate in t-2, who is still recent and provides a possible point of comparison in case the candidate in t-1 lacks similarity. We construct three cases (in analogy with Figure E.1): (i) the candidate in t-1 is similar to the candidate in t, and the candidate in t-2 is not (high relative similarity); (ii)

the candidates in t-1 and t-2 have the same similarity to the candidate in t (medium relative similarity); (iii) the candidate in t-1 is not similar to the candidate in t-1, and the candidate in t-2 is similar (low relative similarity). Again, we use the coefficients of an OLS regression that links the current candidate's rating to the previous candidate's quality, interacted with her relative similarity. The model parameter δ_2 determines how the previous candidate's influence differs between these three cases.

To identify the parameters of the salience function, we use the regression coefficients in analogy to those reported in Figure 6. The coefficients measure how assessments react to the difference between the current candidate's measured quality and the measured quality of the previous candidate, whom we expect to have an important weight in the norm. As discussed above, the parameters σ and θ determine the shape of this relationship. To ease the identification of σ, we calibrate θ, thereby fixing the range where differences are not fully salient.[5]

Finally, we identify the incidental parameters α and β through moments that describe how a candidate's rating varies with her her own quality measure and with the other candidates' average quality measure. More specifically, we use the average rating per quartile of the current candidate's measured quality to identify α. Identification of the adjustment parameter β relies on the average rating conditional on the quartile of the average measured quality of the other candidates seen by the same evaluator (leave-one-out mean).

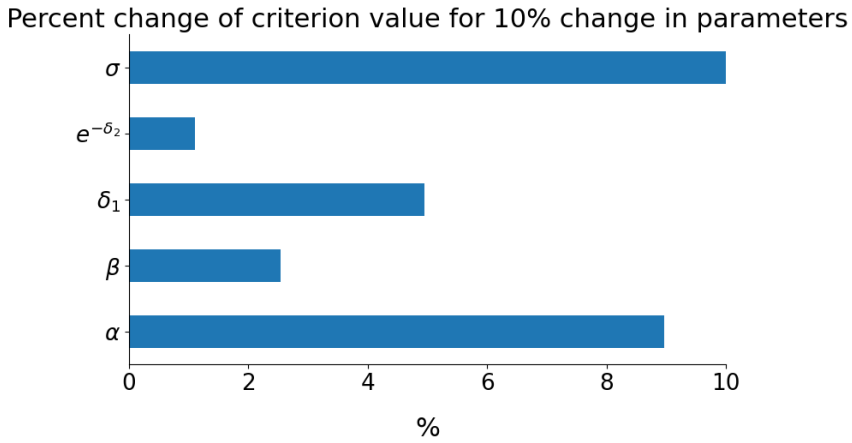
The empirical moments are calculated using the full set of available data. To account for the level of randomization and in line with the reduced-form analysis, all moments are computed conditional on workshop fixed effects and candidate characteristics. Moreover, moments that target δ_1, δ_2 and σ control for own and leave-one-out mean measured quality.

To check whether the moments indeed provide meaningful information on the model parameters, we test whether the value of the criterion function is reactive to changes in a given parameter. The idea is that the criterion function should increase when a parameter moves away from its estimated value — unless none of the moments is sensitive to that parameter.

[5] In the main estimation, we use $\theta = 30$ and $\theta = 100$. Figure F.6 illustrates the influence of θ on the salience of quality differences and their valuation.

For that purpose, we impose a 10% change away from the estimated value of a given parameter, leaving the others unchanged (see, e.g., Galiani & Pantano, 2021). Figure F.5 shows the result of this exercise. The change in the objective function is always positive, showing that the moment fit is reactive to the model parameters.

Figure F.5: Sensitivity of Criterion Function to Changes in Model Parameters



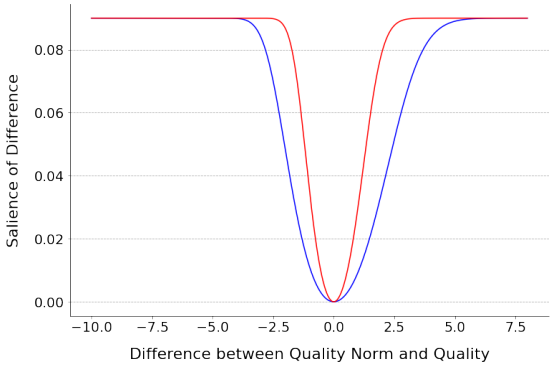
Note: The bars in this figure illustrate the percentage change in the criterion function to a ten percent change in a given model (holding the other parameters constant). The figure is truncated at 10%.

F.3 Influence of θ on Salience and Valuation

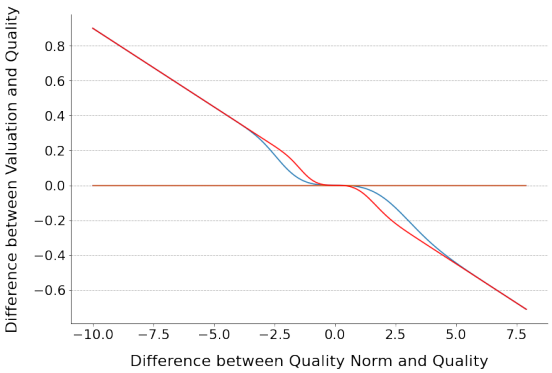
The following figures plot the salience of quality differences and the reaction of the valuation to quality differences for candidates of average measured quality (TPA=12), varying the norm. We plot both relationships for $\theta = 30$ and $\theta = 100$. A higher θ implies that the salience of differences increases faster and thus contrasting kicks in earlier.

Figure F.6: Salience and Valuation for $\theta = 30$ and $\theta = 100$

(a) Salience of Quality Differences



(b) Influence of Quality Differences on Valuation

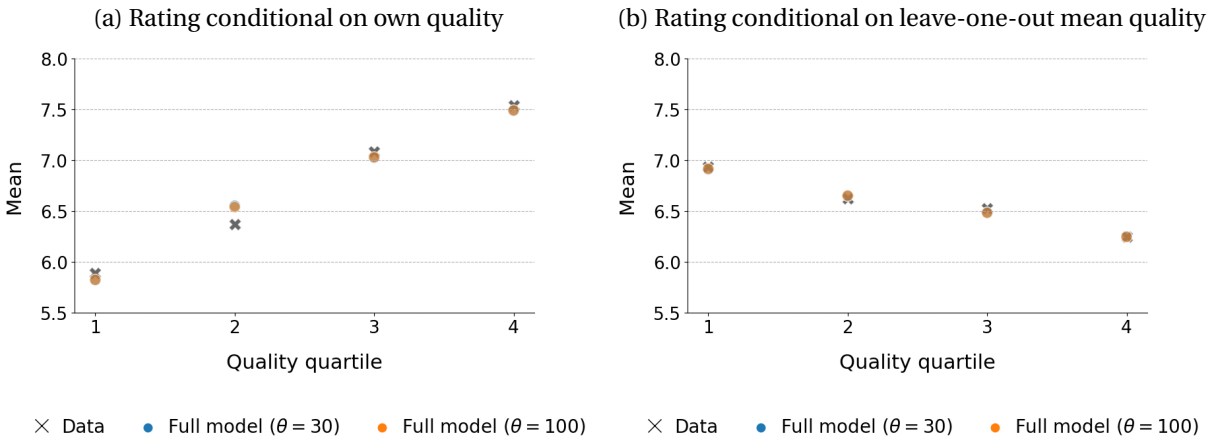


Note: This figure illustrates the influence of θ on the salience of quality differences and their valuation. In panel (a), the y-axis denotes the salience of quality differences and in panel (b) the candidate's valuation. We hold the quality of the candidate constant to 12 and vary the quality norm. The salience function is defined as $\sigma(\tilde{q}_t, q_t^n) = \sigma \frac{e^{\theta(x-1)^2}}{1+e^{\theta(x-1)^2}} - \frac{\sigma}{2}$, $x = \frac{\tilde{q}_t}{q_t^n}$ with $\sigma = 0.18$ and $\theta = 30$ (blue line) as well as $\theta = 100$ (red line).

F.4 Fit with Additional Moments

This Figure plots the fit with the additional moments that were used to obtain the estimates in Table 4.

Figure F.7: Fit of Simulated Moments and Empirical Moments



Note: This figure documents the model fit for the estimates reported in columns (2) and (4) of Table 4. Panels (a) and (b) describe the average rating conditional on the quartile of own and leave-one-out mean quality, respectively.

F.5 Validation Checks

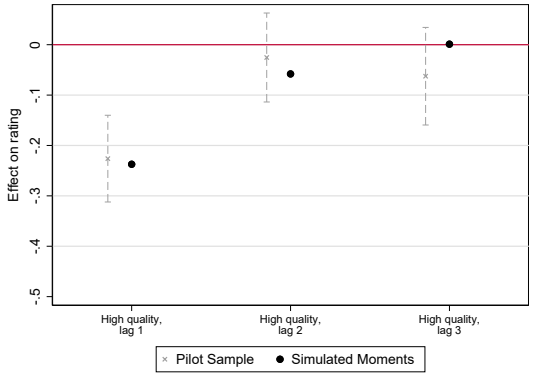
We provide evidence from two validation checks regarding the parameterization of the recall process and the role of similarity therein.

First, Figure F.8 documents that the model is able to predict the role of similarity beyond the estimation sample. In the Figure, empirical moments are computed based on the pilot dataset. As we did not use the pilot data to compute the empirical moments for the structural estimation, the approach resembles the one of a “hold-out" sample. Results show that the simulated moments closely fit the moments in the pilot data, suggesting that the model estimation does not merely fit some artefact of the estimation sample.

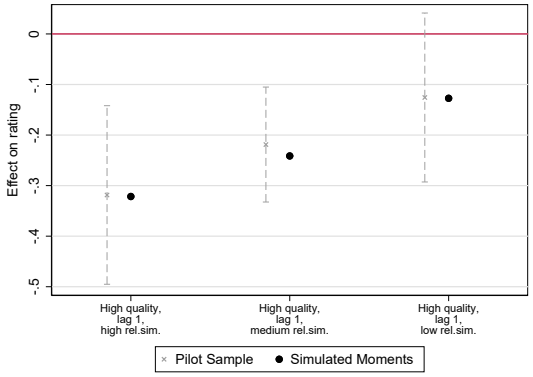
Second, Figure F.9 provides evidence on the model fit with moments that were not targeted in the estimation, but also describe the role of similarity. We compute these moments based on the simulated data and compare them to the corresponding empirical moments in both the estimation sample and the pilot sample. In Panel (a), we consider the influence of the previous candidate conditional on absolute similarity with the current candidate (as opposed to relative similarity in the targeted moments). The predicted role of absolute similarity is close to the one observed empirically in both samples. In panel (b), we consider the previous candidate's influence over the interview sequence, i.e., from the first half of the first day (first three interviews) to the second half of the second day (last three interviews). The model predicts a reduction of the previous candidate's influence which is quantitatively in line with the moments from both datasets.

Figure F.8: Role of Similarity: Simulated Moments and Empirical Moments in Pilot Sample

(a) Relative time



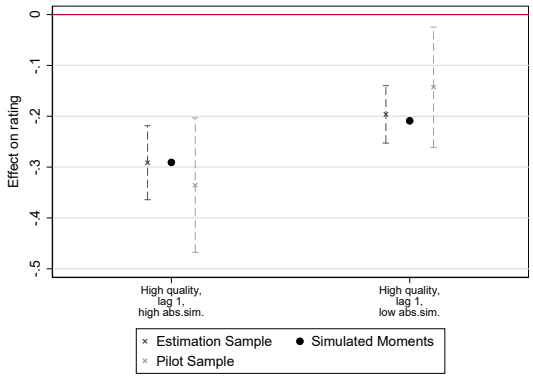
(b) Relative similarity in characteristics



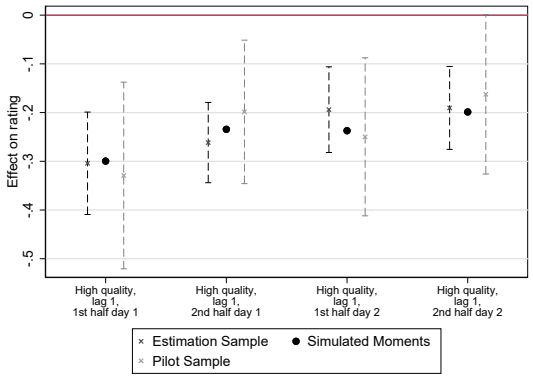
Note: This figure documents the fit of the model (parameters from column (2) of Table 4) with empirical moments of the pilot data. The pilot data were not used to construct the empirical moments for the estimation. In panel (a), the empirical moments describe the effect of following a high quality candidate, depending on the time lag of that candidate's interview. In panel (b), the effect of following a high quality candidate is interacted with her relative similarity in terms of observable characteristics (gender, migration status, study field, 1st generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. Dashed lines describe 95% confidence intervals of the empirical moments.

Figure F.9: Untargeted Moments

(a) Influence of t-1, by absolute similarity



(b) Influence of t-1 over the interview sequence



Note: This figure documents the fit of the model (parameters from column (2) of Table 4) with empirical moments that were not targeted in the estimation process. In panel (a), the moments describe the effect of following a high quality candidate, interacted with a median split of the similarity index, defined as the number of observable characteristics (gender, migration status, first generation status and study field). "High abs.sim." = the candidates in t and t-1 is are of above-median similarity. In panel (b), the moments describe the effect of following a high quality candidate, interacted with the time of the current interview. Dashed lines describe 95% confidence intervals of the empirical moments.

F.6 Robustness of Estimates and Additional Specifications

In the following, we provide estimates from alternative specifications and robustness checks. In section F.6.1, we use the identity matrix to weight the moments. In section F.6.2, we estimate the model proposed by Bordalo et al. (2020), which features anchoring to the norm. In section F.6.3, we estimate two model variants without associative recall, which serve as benchmarks for our main estimates.

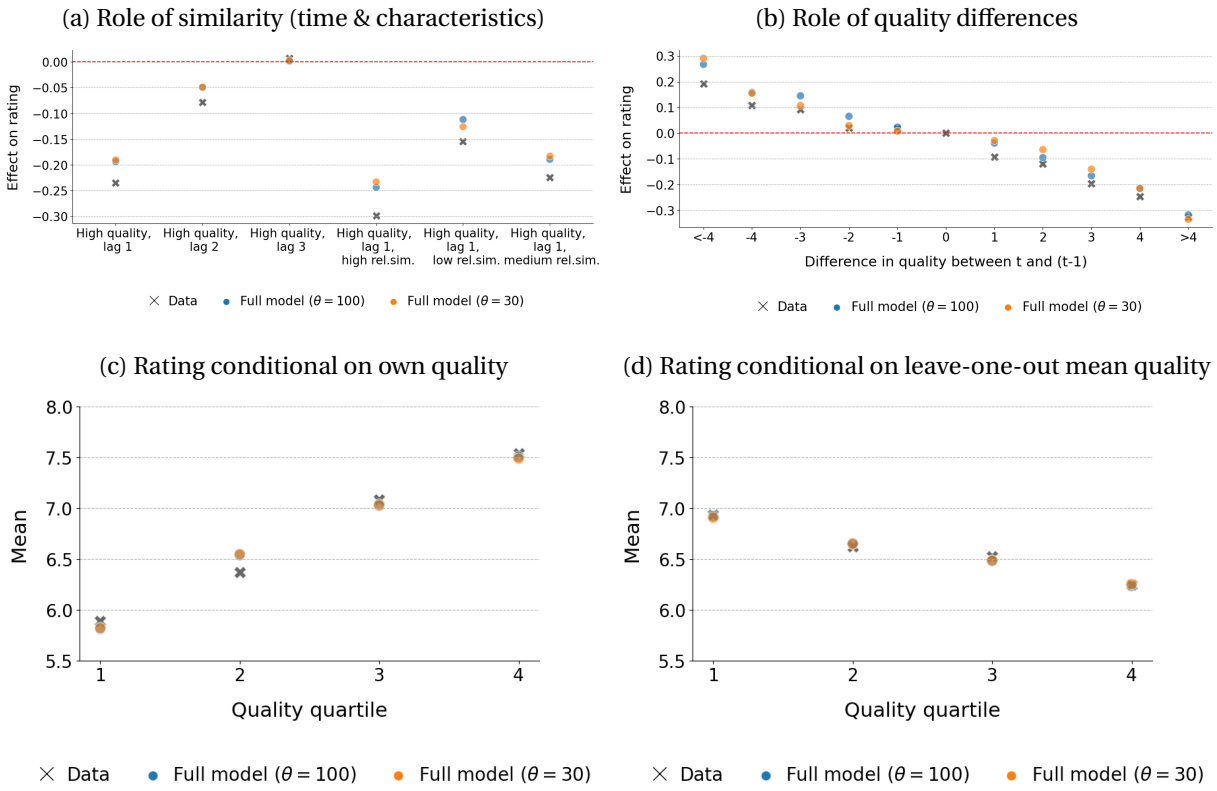
F.6.1 Identity Matrix as Weighting Matrix

Table F.4: Structural Estimates: Identity Matrix

	Only similarity in time ($\theta = 100$) (1)	Full model ($\theta = 100$) (2)	Only similarity in time ($\theta = 30$) (3)	Full model ($\theta = 30$) (4)
Similarity Parameters				
δ_1	1.246	1.444	1.434	1.393
	(0.322)	(0.193)	(0.261)	(0.169)
$e^{-\delta_2}$		0.127		0.178
	.	(0.079)	.	(0.090)
Salience Parameters				
σ	0.127	0.136	0.123	0.145
	(0.025)	(0.015)	(0.019)	(0.023)
$\theta^{\dagger}$	100.0	100.0	30.0	30.0
	.	.	.	.
Incidental Parameters				
α	0.184	0.176	0.187	0.173
	(0.019)	(0.011)	(0.009)	(0.017)
β	0.274	0.274	0.281	0.257
	(0.033)	(0.027)	(0.026)	(0.035)
Weights				
ω_{t-1}	0.747	0.718	0.789	0.726
ω_{t-2}	0.207	0.233	0.183	0.227
ω_{t-3}	0.059	0.067	0.043	0.064
ω_{t-1} | high rel. sim		0.945		0.927
ω_{t-1} | medium rel. sim		0.73		0.729
ω_{t-1} | low rel. sim		0.325		0.387
Weighted SSE	0.059	0.067	0.071	0.078
Number of moments	21	24	21	24

Note: The table shows estimates of the parameters in equation 4, with standard errors in brackets. Columns (1) and (3) report estimates from a reduced model where similarity is only based on the time dimension. Columns (2) and (4) report estimates from the full model, including similarity in terms of candidate characteristics. The second salience parameter θ is calibrated to 100 in columns (1) and (2) and to 30 in columns (2) and (4). "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. Estimation is based on the method of simulated moments (see Appendix F.1 for details). † = calibrated. The weighting matrix is the identity matrix. SSE= Sum of Squared Errors.

Figure F.10: Empirical Moments and Model Fit: Identity Matrix



Note: This figure documents the model fit for the estimates reported in columns (2) and (4) of Table F.4. In panel (a), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. “Rel.sim." describes relative similarity in terms of observable characteristics. “High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. In panel (b), the moments describe the effect of a given quality difference between the candidates in t and t-1. In panels (c) and (d), they describe the average rating conditional on the quartile of own and leave-one-out mean quality, respectively.

F.6.2 Model with Anchoring to the Norm

As an alternative model specification, we estimate the original a model by Bordalo et al. (2020), which features anchoring to the norm. The model is formally described in Appendix E.1. The corresponding estimation equation writes:

$$V_t^{final} = \alpha \times \tilde{q}_t^n + \sigma(\tilde{q}_t, q_t^n) \times (\tilde{q}_t - q_t^n) - \beta \times \bar{q}_{-t} + u_t$$

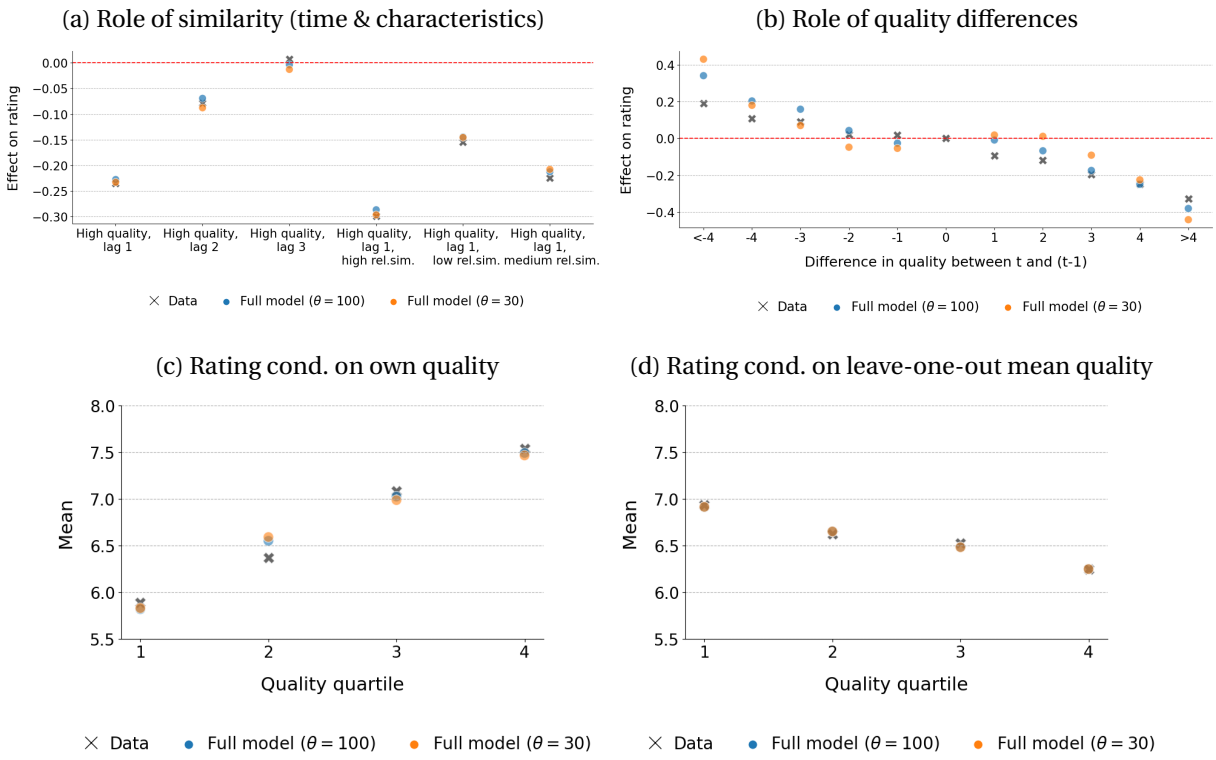
Table F.5 provides the resulting parameter estimates and model fit. The estimates of the similarity parameters are close to those in the baseline model. Note that in this model, the estimate of σ is naturally higher, as it now is the only parameter that captures the influence of own quality. Overall, the model with anchoring leads to a similar, but slightly worse fit with the data than our baseline model. This mostly due to a worse fit with the effect of quality differences (see Panel b of Figure F.11).

Table F.5: Structural Estimates: Model with Anchoring to the Norm

	Only similarity in time ($\theta = 100$) (1)	Full model ($\theta = 100$) (2)	Only similarity in time ($\theta = 30$) (3)	Full model ($\theta = 30$) (4)
Similarity Parameters				
δ_1	0.93	1.128	0.812	0.864
	(0.079)	(0.204)	(0.132)	(0.121)
$e^{-\delta_2}$		0.226		0.268
	.	(0.087)	.	(0.061)
Salience Parameters				
σ	0.492	0.492	0.54	0.532
	(0.012)	(0.013)	(0.016)	(0.015)
$\theta^{\dagger}$	100.0	100.0	30.0	30.0
	.	.	.	.
Incidental Parameters				
α	0.154	0.156	0.147	0.138
	(0.011)	(0.010)	(0.010)	(0.008)
β	0.242	0.244	0.232	0.227
	(0.023)	(0.023)	(0.023)	(0.030)
Weights				
ω_{t-1}	0.657	0.674	0.617	0.605
ω_{t-2}	0.246	0.248	0.257	0.266
ω_{t-3}	0.095	0.087	0.112	0.116
ω_{t-1} | high rel. sim		0.877		0.802
ω_{t-1} | medium rel. sim		0.657		0.568
ω_{t-1} | low rel. sim		0.367		0.327
Weighted SSE	147.726	148.85	246.479	241.45
Number of moments	21	24	21	24

Note: The table shows parameter estimates of the parameters of a model variant with anchoring to the norm (see Appendix E.1 for a description). Columns (1) and (3) report estimates from a reduced model where similarity is only based on the time dimension. Columns (2) and (4) report estimates from the full model, including similarity in terms of candidate characteristics. The second salience parameter θ is calibrated to 100 in columns (1) and (2) and to 30 in columns (2) and (4). "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. Estimation is based on the method of simulated moments (see Appendix F.1 for details). † = calibrated. SSE= Sum of Squared Errors.

Figure F.11: Fit of Simulated Moments and Empirical Moments: With Assimilation



Note: This figure documents the model fit for the estimates reported in columns (2) and (4) of Table F.5. In panel (a), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. “Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). “High/medium/low rel.sim." = the candidate in t-1 is more/less/equally similar to the candidate in t than the candidate in t-2. In panel (b), the moments describe the effect of a given quality difference between the candidates in t and t-1. In panels (c) and (d), they describe the average rating conditional on the quartile of own and leave-one-out mean quality, respectively.

F.6.3 Benchmark Models

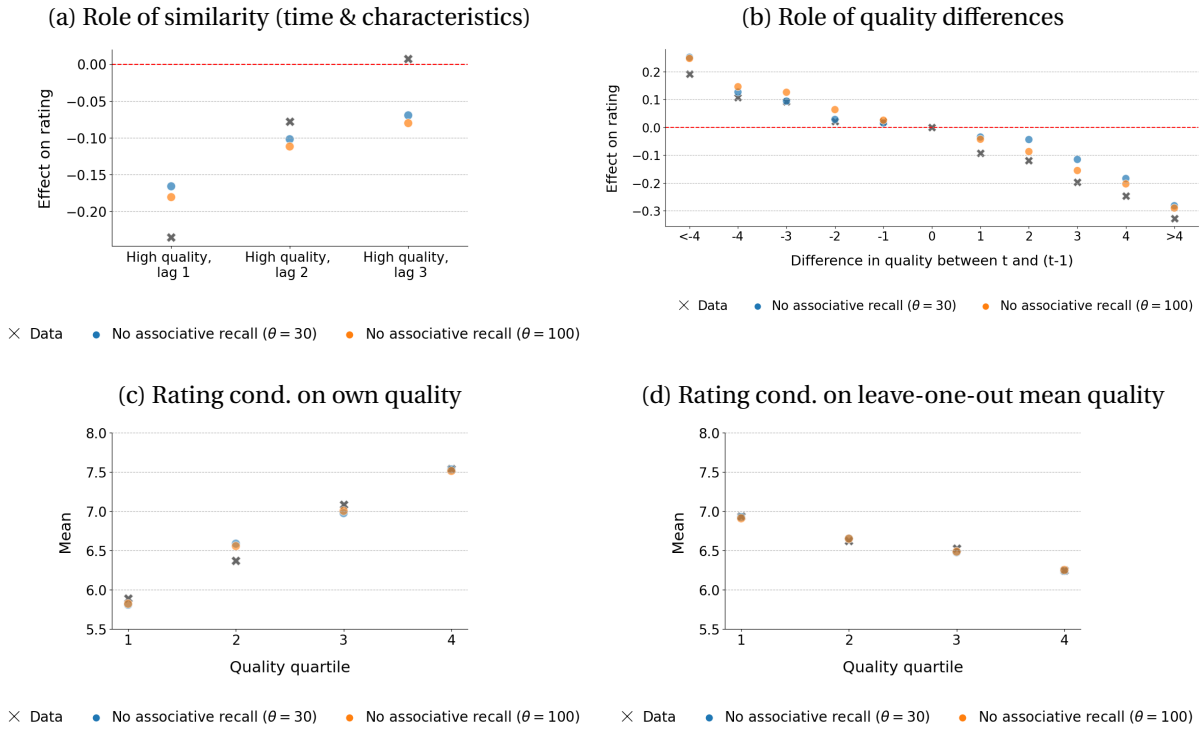
To benchmark our main estimates, we estimate two model variants without associative recall. First, we estimate a model with $\delta_1 = 0$ and $\delta_2 = 0$ (columns 1 and 3 of Table F.6). This eradicates associative recall, such that all previous candidates receive the same weight in the norm. Note that we still expect this model to predict a slightly higher average influence of more recent candidates due to the first two interview slots, where the norm can only consist of the previous or the second two previous candidates. Second, we estimate a model where we replace the quality norm with the expected quality (i.e., the sample average). This eradicates not only associative recall, but recall in general (columns 2 and 4 of Table F.6).

Table F.6: Structural Estimates: Benchmark Models

	No associative recall ($\theta = 100$) (1)	No recall ($\theta = 100$) (2)	No associative recall ($\theta = 30$) (3)	No recall ($\theta = 30$) (4)
Salience Parameters				
σ	0.437	0.0	0.443	0.0
	(0.042)	(0.035)	(0.037)	(0.024)
$\theta^{\dagger}$	100.0	100.0	30.0	30.0
	.	.	.	.
Incidental Parameters				
α	0.032	0.241	0.06	0.241
	(0.022)	(0.020)	(0.016)	(0.013)
β	0.113	0.327	0.145	0.327
	(0.029)	(0.020)	(0.031)	(0.021)
Weighted SSE	227.743	741.358	285.671	741.595
Number of moments	21	21	21	21

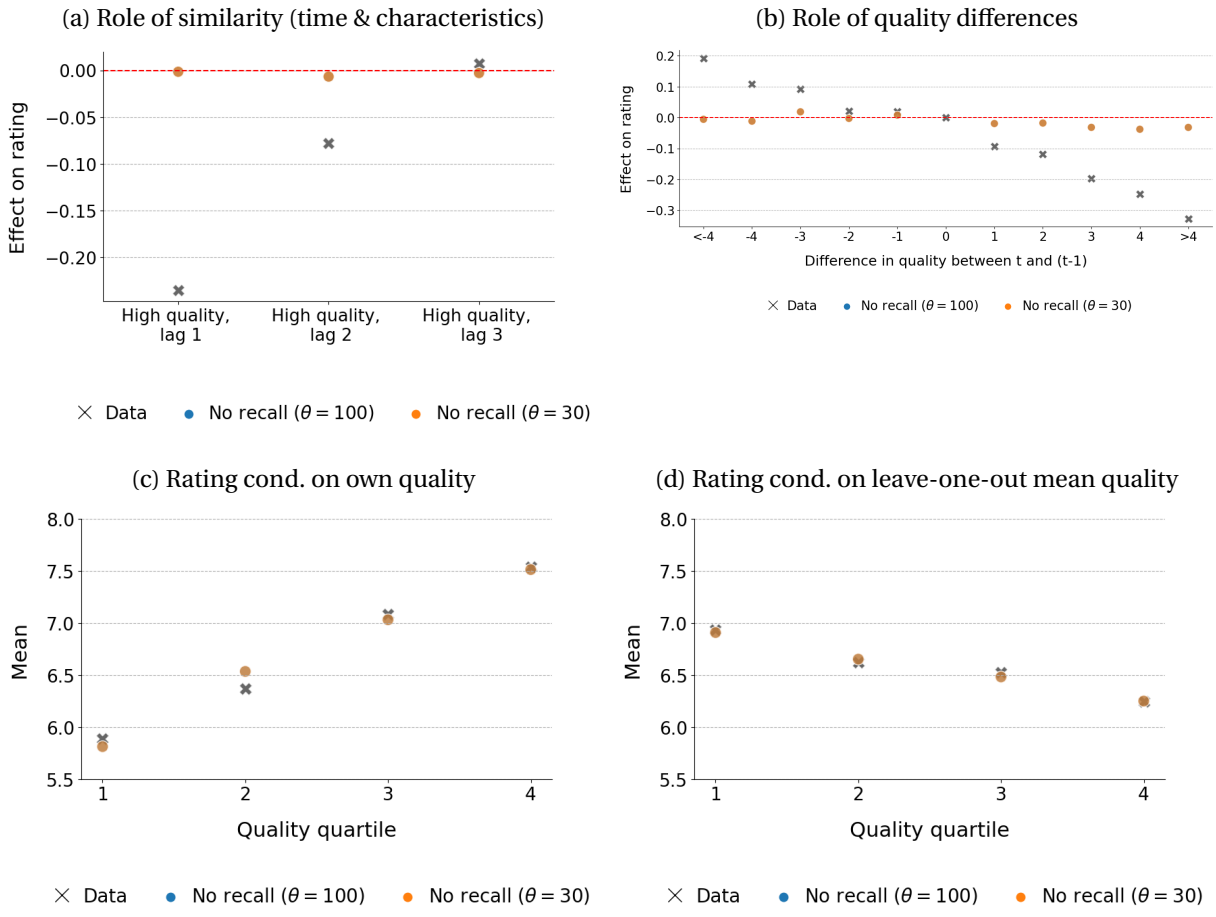
Note: The table shows parameter estimates for two benchmark models. Columns (1) and (3) report estimates from a model where similarity plays no role for recall. Columns (2) and (4) report estimates from a model where the norm is not formed through recall at all, but consists of the expected quality (sample average). Estimation is based on the method of simulated moments (see Appendix F.1 for details). The second salience parameter θ is calibrated to 100 in columns (1) and (2) and to 30 in columns (2) and (4). Estimation is based on the method of simulated moments (see Appendix F.1 for details). † = calibrated. SSE= Sum of Squared Errors.

Figure F.12: Fit of Simulated Moments and Empirical Moments: Benchmark without Associative Recall



Note: This figure documents the model fit for the estimates reported in columns (1) and (3) of Table F.6. In panel (a), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. In panel (b), they describe the effect of a given quality difference between the candidates in t and t-1. Panels (c) and (d) describe the average rating conditional on the quartile of own and leave-one-out mean quality, respectively.

Figure F.13: Fit of Simulated Moments and Empirical Moments: Benchmark without Recall



Note: This figure documents the model fit for the estimates reported in columns (2) and (4) of Table F.6. In panel (a), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. In panel (b), they describe the effect of a given quality difference between the candidates in t and t-1. Panels (c) and (d) describe the average rating conditional on the quartile of own and leave-one-out mean quality, respectively.