Radbruch, Jonas; Schiprowski, Amelie

Working Paper

Interview Sequences and the Formation of Subjective Assessments

ECONtribute Discussion Paper, No. 045

Provided in Cooperation with:
Reinhard Selten Institute (RSI), University of Bonn and University of Cologne

Suggested Citation: Radbruch, Jonas; Schiprowski, Amelie (2020) : Interview Sequences and the Formation of Subjective Assessments, ECONtribute Discussion Paper, No. 045, University of Bonn and University of Cologne, Reinhard Selten Institute (RSI), Bonn and Cologne

This Version is available at:
https://hdl.handle.net/10419/228848

ECONtribute
Markets & Public Policy

ECONtribute

Discussion Paper

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch Amelie Schiprowski

December 2020

ECONtribute Discussion Paper No. 045

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

Cluster of Excellence

UNIVERSITÄT BONN Universität zu Köln

www.econtribute.de

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch[†] Amelie Schiprowski[§]

December 17, 2020

Abstract

Interviewing is a decisive stage of most processes that match candidates to firms or organizations. This paper studies how the interview assessment of a candidate depends on the other candidates seen by the same evaluator, and their relative timing in particular. We leverage novel administrative data covering about 29,000 one-to-one interviews conducted within the admission process of a prestigious study grant program. Identification relies on the quasi-random assignment of candidates to evaluators and time slots. We find that a candidate's assessment decreases when her evaluator receives a better candidate draw. Moreover, the influence of the previous candidate is about three times stronger than the influence of the average other candidate in the sequence. The empirical pattern suggests that evaluators exhibit a contrast effect caused by the interplay between the associative recall of prior candidates and the attention to salient quality differences.

[†]IZA and University of Bonn, Schaumburg-Lippe Str, 5-7, 53113 Bonn, Germany. Email: radbruch@iza.org

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

This project was pre-registered under osf.io/t65zq. For helpful discussions and comments, we thank Johannes Abeler, Steffen Altmann, 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. The paper further benefited from feedback received at the CESifo Conference on Behavioral Economics 2019, the CRC 224 Conference, the briq/IZA workshop on the Behavioral Economics of Education 2019, the Colloquium on Personnel Economics 2019, the University of Cologne, the EALE-SOLE-AASLE World Conference 2020, DIW Berlin and the University of Lausanne. We thank the study grant organization for the data provision and for numerous fruitful discussions. Julia Wilhelm and Stefanie Steffans provided excellent research assistance. Amelie Schiprowski acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1– 390838866 and CRC TR 224 (Project B04).

1 Introduction

Subjective assessments are commonly used to measure quality and 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 especially prevalent is interviewing, which is a decisive stage of most processes that match candidates to firms or 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, the processing of sequential information is prone to errors. For instance, evaluators may judge a candidate in light of recent 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 (sequential) contrast effect (e.g., Bhargava and Fisman, 2014; Pepitone and DiNubile, 1976; Simonson and Tversky, 1992), bears the potential to distort interview assessments. As a result, it might induce firms and organizations to systematically hire or admit the wrong candidates.

In this paper, we provide causal 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 seen by the same evaluator. Our analysis focuses on three questions. First, we analyze how the assessment of a candidate changes if another candidate's quality 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 then investigate contrast effects as a potential mechanism. Guided by recent theoretical insights (Bordalo, Gennaioli, and Shleifer, 2020), we study how the interplay between the evaluator's memory and attention generates contrasting against the previous candidate. In particular, we empirically assess the conditions under which evaluators over-react more or less to quality differences between candidates.

The analysis relies on novel administrative data 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. Overall, the data cover about 29,000 interviews. Three main features make this setup ideal to study how candidates influence each other's 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 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.[1] 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.[2]

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. An exception to the overall pattern is the previous candidate, whose influence is about three times stronger than that of the average other candidate. Once we condition on the average TPA of candidates in the sequence, only the previous candidate's TPA has a meaningful additional influence. 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 p.p. less likely to vote in favor of the candidate observed in period t (16% relative to the mean). Moreover, the results reveal that not only the absolute probability of a yes vote is significantly distorted, but

[1] Importantly, the other evaluators see the same candidate at different points in time and in different interview sequences.

[2] 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.

also the relative ranking of candidates.

We then investigate the channel underlying the previous candidate's striking negative influence. The discussion is guided by recent theoretical insights from Bordalo, Gennaioli, and Shleifer (2020), who show how the interplay between associative memory and attention can generate contrast effects. Following the framework, candidates are evaluated against a quality norm. More precisely, evaluators attract their attention to salient differences between a candidate's quality and the norm. This norm is based on the associative recall of previously experienced candidates. Associative recall retrieves prior interview experiences from memory. The process is associative because it more heavily weighs more similar experiences. As recency can generate (superficial) similarity, the previous interview experience receives a strong weight when forming the norm. This strong weight can generate contrasting with respect to the previous candidate's quality.

We discuss and empirically assess insights from the framework regarding the incidence and strength of contrast effects. First, we find that breaks, which reduce the similarity between two interviews, decrease the previous candidate's influence. More specifically, similarity in time matters in relative terms: the influence of the previous candidate depends on the relative proximity of the interview in t-1, compared to the interview in t-2. Second, additional dimensions of similarity influence the intensity of sequential contrasting. The autocorrelation is stronger if the previous candidate is more similar in terms of observable characteristics, such as gender, socio-economic background or study field. Again, similarity also matters in relative terms: the candidate in $t-1$ has the highest (lowest) influence when she is more (less) similar than the candidate in $t-2$ in terms of her observable characteristics. Third, we confirm the prediction that only large and salient differences between current and prior candidate quality attract the evaluator's attention, whereas small differences do not lead to contrasting. We also find suggestive evidence that small quality differences can generate assimilation effects for specific groups of candidates. Finally, results show that the previous candidate's influence weakens over the interview sequence, as the evaluator's memory database of experienced candidates expands.

We also assess the relevance of alternative mechanisms. In particular, previous evidence

by Chen, Moskowitz, and Shue (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 our findings. First, we do not find that the negative autocorrelation in yes votes increases after a 'streak' of more than one yes vote. Second, outcomes are still related to a quality measure of the previous candidate after 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. Even if we additionally condition on the previous rating, to capture the evaluator's uncertainty in her yes vote, the influence of quality persists. However, the most important distinction between a gambler's fallacy and a contrast effect lies in the point of time when the bias occurs. 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 strongly suggests that the effect does not occur before observing the next candidate. Overall, the tests that lead Chen, Moskowitz, and Shue (2016) to conclude in favor of a gambler's fallacy do not support its relevance in our setup.

The results of this paper imply that minor changes in relative candidate ordering can have major impacts on the selection outcome. This has relevant implications for many hiring and admission situations, with 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 and Schaefer, 2011).[3] In particular, little is known about the formation of subjective assessments through personal interviews. We contribute by showing that the combination of associative recall and contrasting can distort both

[3] Previous studies on candidate screening have — for example — studied the impact of algorithmic recommendations (Bergman, Li, and Raymond, 2020; Horton, 2017), and the influence of job-testing technologies (e.g., Autor and Scarborough, 2008; Estrada, 2019; Hoffman, Kahn, and Li, 2018). Moreover, Simonsohn and Gino (2013) show that interviewers are prone to narrow bracketing. More broadly related, existing 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.

individual assessments and relative candidate rankings. The resulting errors question the reliance on single subjective assessments and provide a rationale for combining several independent assessments per candidate or (additionally) relying on technology-based screening devices (see, e.g., Hoffman, Kahn, and Li, 2018).

We also contribute to the literature on path dependence in real-world decision making and (sequential) contrast effects in particular. Existing evidence stems from the contexts of renting (Simonsohn, 2006; Simonsohn and Loewenstein, 2006), speed dating (Bhargava and Fisman, 2014) and financial markets (Hartzmark and Shue, 2018).[4] This paper differs from the existing studies in several ways. First, we generalize from settings with sequential decision-making to a setting where decisions are made at the end of a closed sequence. Our results show that instantaneous errors are sufficiently strong to persist even when ex-post adjustments can be made after observing all candidates. Second, we provide evidence of contrast effects in labor markets, and more particularly interviewing, which is a key stage in the job matching process. Third, we apply recent theoretical insights on the behavioral foundation of contrast effects to better understand their underlying nature. We thereby complement other recent evidence on contrast effects, in particular Hartzmark and Shue (2018). Using aggregate price data from financial markets, the study documents that contrast effects distort equilibrium prices. The evidence impressively shows that the impact of contrast effects is measurable even in aggregate market outcomes. At the same time, the underlying individual decisions and information histories are unobserved, which makes it difficult to understand the behavioral foundation behind the aggregate contrast effect. Our empirical approach provides a complement based on individual level data, where the decision maker's database is observed in detail. This allows testing theory-based insights on the strength of contrast effects under different circumstances.

More broadly, this study relates to field evidence on reference dependent decision making (for an overview, see Donoghue and Sprenger, 2018), and backward-looking, adaptive reference points in particular (e.g., DellaVigna et al., 2020; Thakral and Tô, 2020). Our results show

[4]Other studies have provided field evidence on a negative relationship between a current decision and the characteristics or outcomes of prior decisions, due to a gambler's fallacy (e.g., Chen, Moskowitz, and Shue, 2016), as discussed above. A positive path dependence has been found for jury decision making in criminal courts (Bindler and Hjalmarsson, 2018) and sport judges (Damisch, Mussweiler, and Plessner, 2006; Kramer, 2017).

that evaluators use the previous candidate as a reference when forming an assessment. Models of associative memory (Bordalo, Gennaioli, and Shleifer, 2020; Mullainathan, 2002) provide a foundation for such backward-looking reference dependence. In this study, we apply insights that arise from this approach to a relevant labor market context. We thereby provide evidence on how associative memory influences economic decision-making in the field. Most closely related is the study by Bordalo, Gennaioli, and Shleifer (2019), whose findings suggest that memory-based reference points lead to sequential contrasting of rent prices. Moreover, we understand our approach as a complement to studies that conceptualize and test the role of memory for economic decision making in a fully controlled lab environment (e.g., Enke, Schwerter, and Zimmermann, 2020).

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 discusses the underlying mechanism, with a focus on the role of associative memory and attention. Section 6 concludes.

2 Institutional Setting

We study the admission workshops of a large, merit-based study grant program for university students in Germany. Stakes for the candidates are high, as being selected into the program yields 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. The admission committee is formed by eight evaluators.

[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.

About 48 candidates participate in each workshop.[6] An employee of the study grant organization is permanently present.

Candidates are first-year university students. They were pre-selected by their high-school principals, who can nominate about 2% of a graduating cohort. Most commonly, principals nominate the 2% with the highest GPA. Prior to their workshop participation, candidates submit a written CV and their school transcripts. During the workshop, each candidate participates in two one-to-one interviews lasting 35 minutes and one group discussion. Each task is assessed by a different evaluator, implying that every candidate receives three independent assessments: one per interview and one for the group discussion. The final decision is based on the sum of the three equally-weighted assessments.

Evaluators are scholarship alumni 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. Moreover, no interaction between evaluators and candidates takes place afterwards. The workshop therefore constitutes a closed sequence of interaction.

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.[8]

Workshop Schedule Table 1 sketches an evaluator's schedule during the admission workshop.[9] Upon arrival on Friday night, evaluators receive a short briefing by an employee of the scholarship organization and prepare the interviews which they conduct on Saturday. For this

[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.

[7] Randomization occurs conditional on gender, with the aim of ensuring a balanced gender composition in the group discussion.

[8] Evaluators know their ID prior to the workshop, but they do not have any information on any candidate nor their assigned candidates, which renders the knowledge irrelevant.

[9] We describe here the schedule for the 2013/14 academic year. In the following years, the schedule was slightly adjusted such that all group discussions took place on Saturday. The length and ordering of the interviews were not affected. We use the appropriate schedule to calculate breaks and interview ordering.

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	

purpose, they receive each candidate's CV, school records and a letter of recommendation written by the high-school principal. On Saturday, evaluators each conduct six interviews and rate five group discussions. In the evening, they receive the documents of the candidates who they interview on Sunday. On Sunday, evaluators conduct six interviews and rate one group discussion. Every group discussion includes approximately six candidates and takes place over six time slots.[10] The detailed schedule – including candidate assignments to evaluators and time slots – is shown in Appendix 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 (usually one, at maximum two).

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 is no clear guideline regarding the interview structure and the questions asked, but the employee of the organization gives 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

[10] 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.

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. Evaluators are informed about these rules at the start of the workshop. Moreover, the employee of the institution states explicitly that there is no admission quota and that the committee is free to admit all or none of the candidates present at the workshop.

Evaluators are asked to determine their individual assessments after having seen all of their assigned candidates. Moreover, they are not allowed to exchange opinions with other evaluators about candidates before the final committee meeting. This rule is strictly enforced by the employee of the scholarship organization, who wants to ensure that every candidate receives the chance of being evaluated independently. Moreover, evaluators have a high intrinsic motivation for compliance, as they are alumni who have received many benefits from the program. 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.[11] Subsequently, ratings are aggregated and — following a short justification by the responsible evaluators — candidates at or above the cut-off of 23 points are accepted for the scholarship. Ratings for candidates at the margin to admission can be adjusted after a discussion by the committee. Such adjustments usually happen for about two to three out of about 150 votes per workshop. We observe the final ratings of each candidate.[12]

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.

[11] In this process, it is not easily possible to trace the behavior of other evaluators, as assessments are collected with high frequency and not ordered with respect to evaluator IDs.

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

3.1 Data Description

Data Source & Sampling We employ data on the full population of admission workshops for recent high-school graduates that took place during the 2013/14 to 2016/17 academic years. The data contain 312 admission workshops, including 29,466 interview ratings on 14,733 candidates.[13] The ratings were made by 2,496 evaluators.[14]

For each candidate, we observe her 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 non-academic parental background. We further observe basic characteristics of the evaluator, namely 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 largest mass of ratings lies between 5 and 8 points. The average rating in the sample is 6.6 points, with a standard deviation 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 across evaluators, which reflects that evaluators do not face a quota. The evaluator-specific share of yes votes ranges from 0 to 1, with a mean of 0.37 and a standard deviation of 0.14.[15]

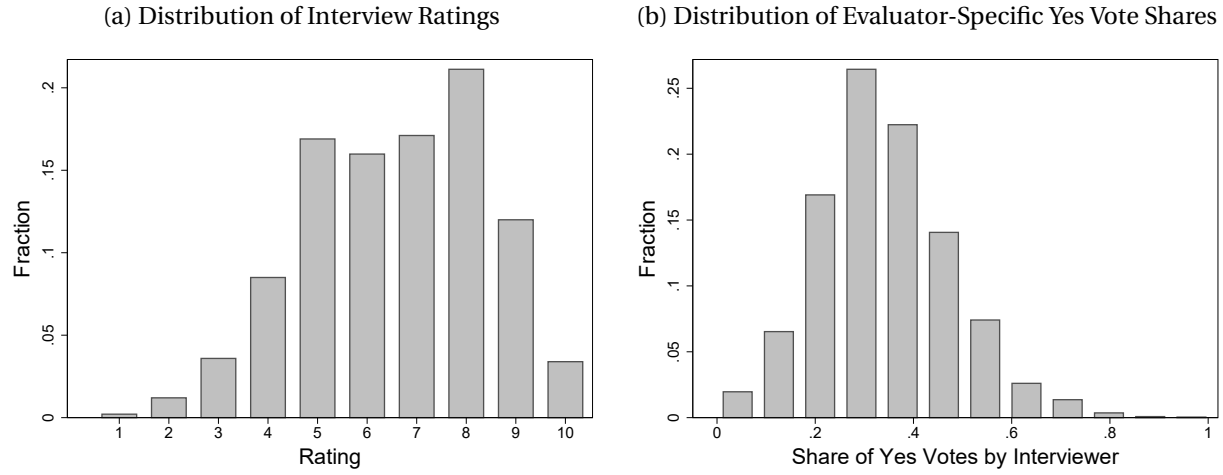
Appendix Table B.1 reports summary statistics on evaluator and candidate characteristics. Close to half of the evaluators are female and the average evaluator is 42 years old. Evaluators come from various study backgrounds, with the most dominant one being humanities (45%), followed by STEM (36%). About 60% of evaluators participate in their first workshop,

[13] 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.

[14] 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.

[15] 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.

Figure 1: Distribution of Assessments at the Individual and Aggregate Level



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).

about 20% have two prior workshop participations, and about 20% have previously participated three or more times. The average evaluator conducts twelve interviews per workshop. Among candidates, about 55% are female. The average candidate is 19.6 years old, 16% of candidates have a migration background and 26% a non-academic parental background. The average applicant achieved 92% of the maximum possible high-school GPA. The most frequent study field is STEM (37%), followed by medicine (24%), social sciences (20%) and humanities (18%).

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 institutional setup as described in section 2. The only candidate characteristic taken into account for the assignment of candidate IDs is gender, because the scholarship organization aims to have 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 as-

signed to evaluator i are not systematically related to the characteristics of the other candidates assigned to i. We test this implication by regressing a given observable 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.[16] 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 for sorting of candidates to evaluators. Appendix Table B.2 additionally provides evidence that candidate and evaluator characteristics are not systematically related.

To assess the assumption of quasi-random ordering, we test for the presence of an autocorrelation in candidate characteristics, conditional on own gender. Panel B of Table 2 presents the results from a regression of the current candidate's characteristics on the previous candidate's characteristics, conditional on own gender and workshop fixed effects.[17] It shows no indication of systematic ordering by observed candidate characteristics.

Table 2: Assessment of Quasi-Random Assignment & Ordering

	GPA	Age	Migrant	1st Generation	STEM	Social Sciences
	(1)	(2)	(3)	(4)	(5)	(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, we only use the same sample as in the estimations below, i.e., we only use candidates who are not the first in an interview sequence. Following Guryan, Kroft, and Notowidigdo (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).

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

[17] Again, following Guryan, Kroft, and Notowidigdo (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.

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. Given the institutional context, quality is a description of how well a candidate meets the study grant's selection criteria (see section 2 for details). True candidate quality is unobserved by design, as the assessment process would not need to take place otherwise. Any measurement of quality therefore needs to be thought of as an approximation.

Our preferred measure of a candidate's quality is based on third party assessments (TPA) made independently by other evaluators. 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. More precisely, we use the sum of the two within-year standardized ratings. One of the other ratings is based on the candidate's second interview and the other on her performance in the group discussion.[18] The main idea behind this approach is twofold: first, evaluators use the same criteria when rating quality; and second, they all measure these criteria with noise, but their noise terms are independent of each other. Below, we discuss these two advantages in more detail.

The first advantage is that all evaluators are supposed to rate the same dimensions of quality and ability. The correlation between the individual rating and the sum of the other two evaluators' ratings is 0.38.[19] Given that evaluators differ in their leniency and see the same candidate under different circumstances, we interpret this correlation as strong.[20]

The second advantage is that the other two evaluators' ratings are as-good-as independent of the evaluator's own assessment behavior. This is a result of the workshop schedule. Evaluators see the same candidate at very different points in time and the sets of candidates seen by two evaluators hardly overlap (see workshop schedule in Appendix Figure A.1). Im-

[18] 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.

[19] The two interview ratings are correlated by a factor of 0.36. As expected, the correlation with the group discussion rating — which is based on a different task — is smaller and amounts to about 0.23.

[20] 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.

portantly, this implies that evaluator A's assessment of a given candidate is influenced by a different candidate pool than evaluator B's assessment of that candidate. Moreover, two evaluators never see the same two candidates in the same relative order. Finally, evaluators are not allowed to discuss candidates before ratings are joined in the final committee meeting. This rule is enforced by the employee of the scholarship organization who is present throughout the workshop (see section 2 for details).[21] In Appendix Table B.3, we assess one implication of the independence assumption. The idea is that an evaluator's characteristics are likely to correlate with her candidate ratings; for instance, female evaluators are on average more lenient. On the contrary, evaluator characteristics should not influence the candidate's TPA, i.e., the sum of ratings the candidate was given by the other two evaluators. 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.

An alternative way to measure candidate quality is through pre-determined characteristics, in particular high-school GPA. However, GPA is likely to be a poor predictor of fit with the scholarship criteria, which extend beyond grade performance. Indeed, Appendix Table B.4 shows that individual assessments increase in high-school GPA, but the power of observed candidate characteristics to predict interview ratings is low (R-Squared ≈ 0.02). Nevertheless, we construct an alternative quality measure based on pre-determined candidate characteristics to check the robustness of the qualitative results pattern.

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 assess-

[21] One context 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.

ments.[22]

4.1 Influence of the Other Candidates and the Role of Relative Timing

Econometric Specification

We estimate how the assessment of a candidate interviewed in period t is affected by the measured quality of the candidate interviewed in period $t+k$. As described in section 3.3, we use the sum of the other two evaluators' independent assessments as our preferred quality measure. We refer to this measure as the third-party assessment (TPA). For each value of $k \in \{-11, ..., -1, 1, ..., 11\}$, we perform a separate estimation of the following regression model:[23]

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

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, as well as dummies for 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).

[22] 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.

[23] 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.

In a supplementary regression, we study the influence of another candidate conditional on the average quality of candidates in the evaluator's interview sequence (excluding t). For this purpose, we replace $\overline{TPA}_{i,-\{t,t-k\}}$ by the leave-one-out mean $\overline{TPA}_{i,-t}$. Thereby, β_k measures the additional effect that the candidate in $t+k$ has beyond contributing to the average quality of the sequence.

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

Results

Panel (a) of Figure 2 plots the coefficients β_k from equation 1, resulting from separate regressions for each value of $k = \{-11, ..., -1, 1, ..., 11\}$. The outcome is the candidate's standardized interview rating. The corresponding coefficients and p-values are shown in Appendix Table C.1.

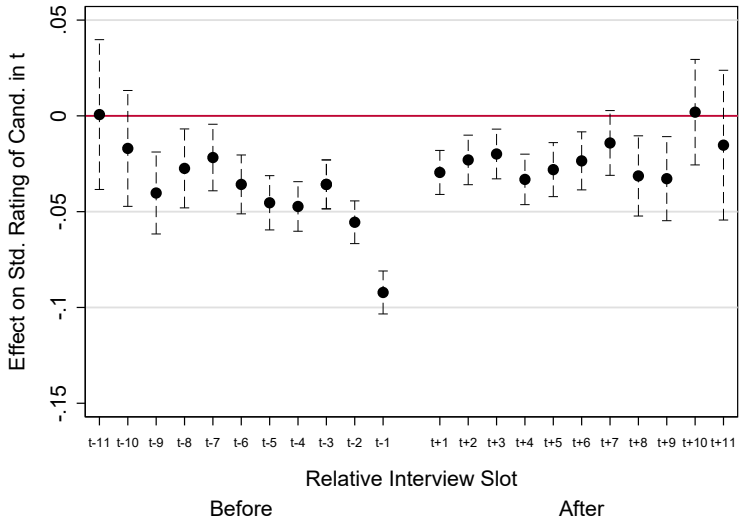
The figure documents three main results. First, the rating of a candidate decreases in the quality (measured through TPA) of any other candidate seen by the same evaluator. If another candidate's TPA increases by one standard deviation, the candidate's rating decreases by about 2 to 5% of a standard deviation. Second, candidates interviewed before t ($k < 0$) as well as 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.[24] 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 only the previous candidate has a meaningful additional in-

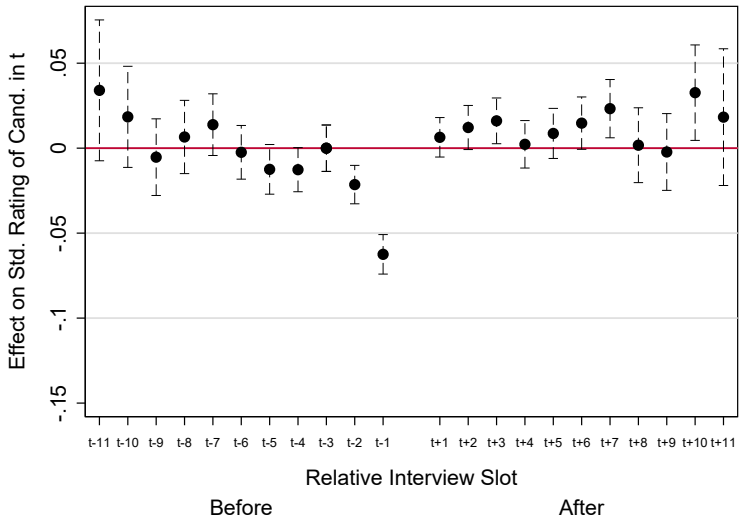
[24] The average of the coefficients with $k < -1$ amounts to 3.1 p.p. and is significantly larger than the average of the coefficients with $k > 0$, which is 2.1 p.p. (c.f. Appendix Table C.1 for the corresponding p-values).

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

(a) Unconditional Effect



(b) Conditional on Leave-One-Out Mean



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.

fluence on the rating. This suggests the existence of two separate effects: an influence of the other candidates' average quality and an additional influence of the previous candidate's quality.

Appendix Tables C.2 and C.3 report the estimated coefficients of own, previous and leave-one-out mean TPA for $k = -1$ and provide several robustness checks. To put the influence of the previous candidate into perspective, Panel A of Table C.2 shows that the influence of the previous candidate's TPA is 17% as large as the influence of the candidate's own TPA and about 55% as large as 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 proxies of candidate quality, including a prediction based on observable characteristics. The 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 a measure of 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 — compared with the previous analysis — 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 being as-good-as one-directional, once we condition on the average strength of the interview sequence.

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, $\overline{Y}_{i,-t}$). Note that the leave-one-out mean assessment also controls for differences in evaluator leniency.[25] $\overline{Y}_{i,-t}$ 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.

The specification controls for workshop fixed effects (ω_w),[26] as well as evaluator and candidate covariates $X_{i,t}$ (including interview order and the candidate's TPA).

Results

Table 3 reports the linear autocorrelation in evaluator assessments, based on an estimation of equation 2.[27] In columns 1 (without controls) and 2 (with controls) of Table 3, the outcome is the standardized rating of the candidate interviewed at time t. Irrespective of the inclusion of controls, a one standard deviation increase in the rating of the previous candidate is associated with a 7% of a standard deviation decrease in the rating of the current candidate.

[25] 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.

[26] 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.

[27] In Appendix Figure D.1, we additionally allow for a non-linear relationship. Moreover, Appendix Figure D.2 shows the non-linear autocorrelation for candidates below and above the median TPA level, respectively.

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)			-0.058***	-0.407***	-0.027***	-0.023***
			(0.006)	(0.042)	(0.004)	(0.004)
Leave-one-out Mean Rating	0.256***	0.298***	0.080***	-0.663***	-0.014***	0.039***
	(0.018)	(0.018)	(0.009)	(0.053)	(0.006)	(0.005)
Leave-one-out Share Yes	-0.975***	-0.874***	-0.394***	-3.473***	-0.281***	-0.199***
	(0.078)	(0.071)	(0.049)	(0.237)	(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).

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

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 the result is not mechanical due to the leave-one-out mean controls. The estimates thus reveal that the previous candidate's additional influence also distorts relative rankings. It therefore carries potential relevance for contexts in which the aim is to find the relatively best candidate(s). Finally, column 6 shows that the probability of admission — which is based on the sum of the three independent ratings — decreases by 2.3 percentage points if the previous candidate received a yes vote in one of the interviews (10% relative to the mean).[28]

[28] 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 training for interviewing skills. Finally, we provide a back-of-the-envelope quantification on the reversal of admission outcomes induced by the binary autocorrelation in Appendix E.

5 Potential Mechanisms

In the previous section, we 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, which produces a strong negative autocorrelation in an evaluator's assessments. This section discusses potential mechanisms underlying the findings.

As a potential explanation for the influence of the other candidates' average quality, Appendix F.1 presents an illustrative theoretical framework where evaluators learn about the admission threshold through the other candidates' quality. The framework yields the straightforward prediction that the likelihood of a positive assessment decreases in the quality of the other candidates observed by the same evaluator. However, it is difficult to reconcile the pre-

vious candidate's additional influence with standard arguments.

One intuitive behavioral 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, Gennaioli, and Shleifer, 2020) can explain the previous candidate's influence. We then assess the empirical relevance of additional predictions that arise from this mechanism and can help to further understand the nature of the influence. In a final step, we discuss alternative mechanisms — notably a gambler's fallacy — and argue that they are not in line with our empirical findings.[29]

5.1 Contrast Effects and the Role of Associative Memory

Evaluators exhibit contrast effects 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, Bhargava and Fisman, 2014; Pepitone and DiNubile, 1976; Simonson and Tversky, 1992). In a recent contribution, Bordalo, Gennaioli, and Shleifer (2020) propose a theoretical framework that includes a formulation of contrast effects. Based on the concept of associative recall, the model directly addresses the question of what constitutes the norm for the evaluation of choice options, i.e., against what reference a choice is contrasted. In the following, we discuss the main intuition of how the framework explains the influence of the previous candidate. Appendix F.2 provides a more formal discussion of this intuition.

Valuation of Candidate Quality An evaluator votes on the admission of a candidate. She votes in favor of admission if her valuation of the candidate exceeds an evaluator-specific threshold.[30] The valuation is formed upon interviewing the candidate in period t and is defined as:

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

[30] The threshold can depend, e.g., on the evaluator's leniency and the average quality of the other observed candidates.

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

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 $\sigma(\tilde{q}_t, q_t^n)$.[31] 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 therefore are not salient, i.e., $\sigma(\tilde{q}_t, q_t^n)$ is low. Larger surprises are more salient, yet with diminishing sensitivity.[32] 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.

Experience-based quality norm Bordalo, Gennaioli, and Shleifer (2020) use the notion of associative recall to address the question of what constitutes the reference norm 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. Recall of these experiences 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 that

[31] We abstract from anchoring, present in the original model of (Bordalo, Gennaioli, and Shleifer, 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 F.2 and provide an empirical assessment later in the section

[32] 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 $\lim_{x/y \to \infty} \sigma(x/y, 1) = \sigma > 1$.

triggers the recall of past experiences in the context of sequential interviewing is time.[33] When defining similarity based on time, associative recall triggers the recall of recently observed candidates. Thereby, 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. Note that similarity might 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 which another interview experience is recalled.

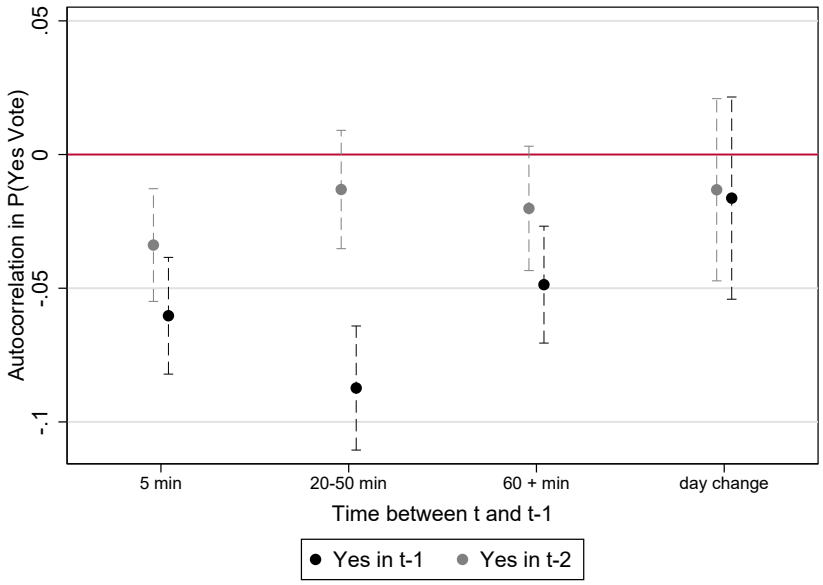
In summary, the framework by Bordalo, Gennaioli, and Shleifer (2020) predicts the incidence of contrast effects through the interplay of associative recall, which forms the background norm, and the attention to salient quality differences. The notion of 'sequential contrast effect' — i.e., contrasting with respect to the previous candidate — is incorporated in a natural way: due to associative recall, recent interview experiences receive a strong weight in the quality norm because recency leads to a (potentially misleading) similarity. In the following, we discuss and assess the empirical relevance of predictions that arise from this framework regarding the incidence and strength of contrast effects under different circumstances.

Additional 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. In the following, we test how the estimated autocorrelation — as a measure for the average influence of the previous candidate — changes with relative similarity between two interviews. Based on associative recall, we expect that the strength of the influence decreases when there is a larger break between two interviews, as long as we keep the time difference to other interviews constant. In real-world contexts as studied in this paper, a change in the time gap between two interviews implies a simultaneous

[33] Bordalo, Gennaioli, and Shleifer (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).

Figure 3: Autocorrelation by Time Lag between t and t-1



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. The gray dots repeat the exercise, but replace the yes vote of the candidate in t-1 with the yes vote of the candidate in t-2. N=26,970 (black dots); N= 24,474 (gray dots). The dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level.

change in the time gap to other previous interviews. Therefore, the prediction on the role of breaks is less clear cut.

Figure 3 confirms this ambiguity. On the one hand, it documents the tendency that longer breaks weaken the autocorrelation between yes votes in t and t-1. If interviews are separated by a day change, the autocorrelation approaches zero. On the other hand, the relationship between the strength of the autocorrelation and elapsed time is not monotonic: short breaks of five minutes are associated with a weaker autocorrelation than longer breaks of 20 to 50 minutes. A potential explanation is that interviews with a five-minute break involve stronger recall of the interview in t-2, which is now mechanically also closer in time. Therefore, the relative similarity of the previous candidate may be smaller, even though its own absolute similarity increased. The gray dots — which show the autocorrelation in votes between t and t-2 — are in line with this intuition: votes for candidates who are interviewed after a five-minute break

show a relatively strong autocorrelation with votes in t-2. For slots with longer breaks, this correlation is close to zero.[34] Furthermore, note that the two estimated autocorrelations are identical if t follows a change in day: in case of an extremely large time gap between t and t-1, the marginally higher recency of t-1 plays no role. Overall, these results are in line with the conjecture that relative similarity in time matters for the size of the autocorrelation.

Additional dimensions of similarity The previous results have shown that time is an important dimension of similarity from the evaluator's perspective. We now assess the relevance of 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 associative recall is driven by similarity in time. If we allow recall to be based on additional (socio-demographic) candidate characteristics, 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 yes votes differs if two subsequent candidates are more or less similar in terms of their observable characteristics. More precisely, 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 background, parental background and study field).[35] 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 the theoretical conjecture, the autocorrelation is stronger in cases where the observed similarity between two subsequent candidates is high. If the candidate in t shares more than the median number of characteristics (2) with the candidate in t-1, the autocorre-

[34] Appendix Figure F.2 shows the same figure, replacing the x-axis with time between t and t-2. It shows that if the interview in t-2 ends only 45 to 60 minutes before the interview in t, the autocorrelation with t-2 is almost as strong as the autocorrelation with t-1.

[35] We abstract from age and GPA as two additional observable characteristics because candidates differ little along these dimensions. All candidates are pre-selected based on having a high GPA and all candidates are in the first year of their undergraduate program.

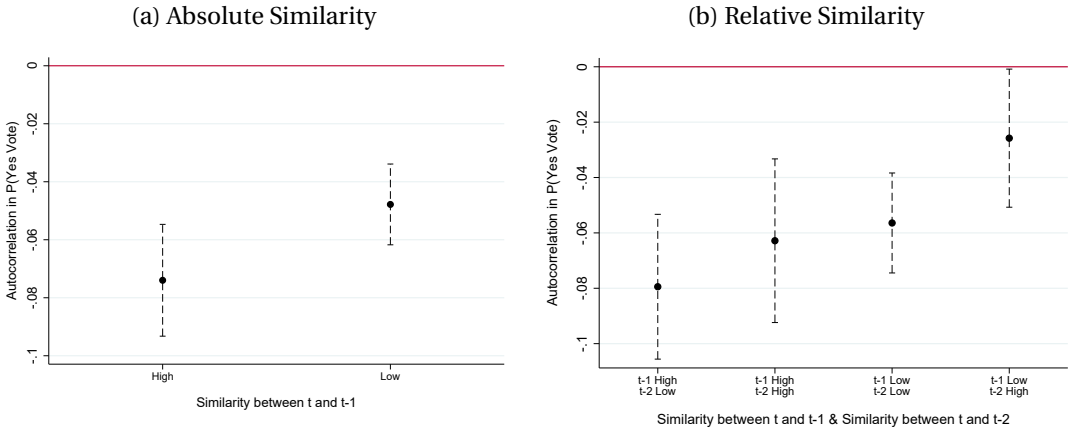
lation amounts to about 7.4 percentage points, while it lowers to about 4.8 percentage points when similarity is at or below the median.

Having established that candidate characteristics define similarity in addition to the time dimension, we now account for the notion of relative similarity. Relative similarity implies that it matters how similar the previous candidate is compared to other preceding candidates. To assess this conjecture empirically, we allow the influence of the previous candidate to depend on the similarity of the candidate in t to the candidate in t-1 as well as to the candidate in t-2. The idea is that if the candidate in t-2 is rather similar to the candidate in t, she may be recalled strongly and (partially) block the influence of the candidate in t-1.[36] The results strongly support the idea that relative similarity matters for the previous candidate's influence. The strength of the autocorrelation increases in the relative similarity of the previous candidate, up to 8 percentage points. It strikingly reduces to 2.6 percentage points in the case where not only similarity to the candidate in t-1 is low, but also similarity to the the candidate in t-2 is high. Moreover, the pattern also reveals that it makes no difference if the candidates in t-1 and t-2 both have a high or both a low similarity to the candidate in t. For the influence of the previous candidate, it rather seems to matter whether she is relatively more similar to the candidate in t than the candidate in t-2. In Appendix Figure F.3, we perform a similar 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.

In Appendix F.4, we explore for the case of gender whether the influence of similarity is symmetric. 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. Symmetric similarity is a common assumption in models of memory (see, e.g., Kahana, 2012). Table F.1 shows that the data is in line with the notion of symmetric simi-

[36] We concentrate on the candidate in t-2 as this candidate is still recent and provides a possible point of comparison in case the candidate in t-1 lacks similarity.

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 background, parental background and study field) which the candidate in t and the candidate in t-1 have in common. In panel (b), there is an additional interaction with the similarity between t and t-2. N=26,970 (panel a) & N=24,474 (panel b). The dashed lines show 95% confidence intervals.

larity with respect to gender. For example, both male and female candidates are more strongly influenced by previous candidates of the same gender. This result contradicts the hypothesis that the previous candidate's influence is asymmetric with respect to gender, which we registered in our pre-analysis plan. The hypothesis was based on the notion by Tversky (1977) that similarity can be directional and asymmetric. The pilot data that was used prior to the preregistration pointed towards gender asymmetric contrasting. We found consistent evidence that women were equally harmed by strong male and female candidates, while male candidates were harmed by strong male but not by strong female candidates. This would have been in line with asymmetric similarity where female candidates are contrasted against males, but not vice versa. As we this pattern does not consistently replicate in the present data, we rather conclude in favor of the idea that similarity has a symmetric influence on recall. A detailed analysis and discussion of the hypothesis is provided in Appendix F.4.

Attention and Size of the Surprise The estimates presented in section 4 revealed that the previous candidate's quality has on average a negative influence on the current candidate's

valuation. However, the presented 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 perceived quality and the quality norm.[37] We therefore expect to observe contrasting only for 'large' differences, as small differences do not attract the evaluator's attention, i.e., they 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 TPA score (based on exact number of points) 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. In particular, we relate the current candidate's probability of a yes vote to categories of this difference, while flexibly controlling for the current candidate's TPA in points. The identifying variation thereby stems from changes in the previous candidate's TPA.

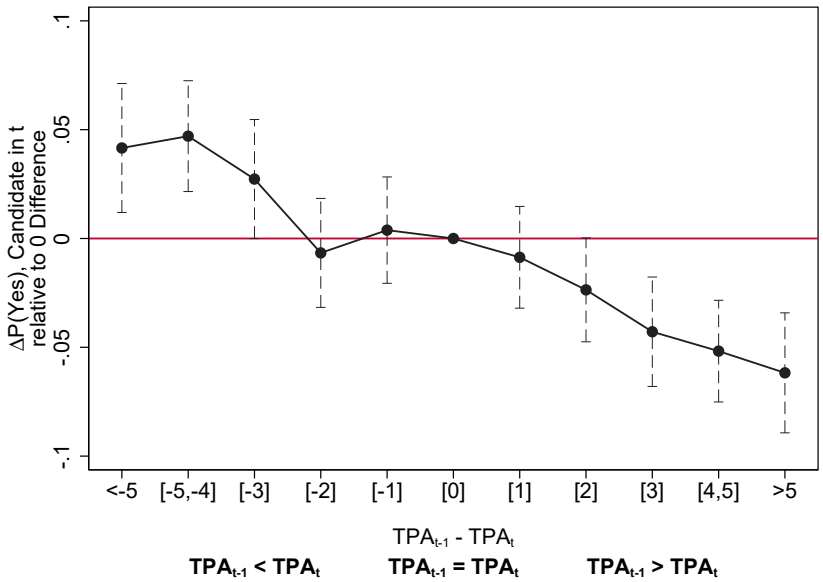
Figure 5 shows the corresponding estimates. Moving from the left (positive surprise) to the right (negative surprise) coincides with increasing the quality of the previous candidate.[38] In line with the previous results, the average slope is negative. However, we also observe a flat relationship in the case of small TPA differences. This is directly in line with the theoretical prediction that small quality differences do not attract the evaluator's attention. At larger absolute TPA differences, contrasting kicks in and leads to economically meaningful changes in the current candidate's probability of a yes vote.

In Appendix F.5, we explore the robustness of the presented pattern. Most importantly, we test the robustness with respect to different approximations of the norm. In panel (a) of Figure F.6, we study the difference between current TPA and the average TPA of the two previous candidates. Given that the second-most-recent candidate is still relatively recent, her quality may also be a meaningful part of the norm. In panel (b) we use the average TPA of all previous

[37] In Appendix F.2, we also discuss the role of anchoring and how it affects this relationship formally. For an intuitive discussion see below. Moreover, Appendix Figure F.1 plots the theoretical prediction for the relationship between the quality norm and the valuation (including anchoring).

[38] 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 (see section 4.1). Therefore, differences in TPA very quickly represent significant differences in quality of candidates; for example, a difference of 3 points is associated with a 15 p.p. difference in the likelihood of receiving a yes vote.

Figure 5: 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.

candidates as an approximation of the norm. Appendix Figure F.7 replicates Figure 5 with more fine-grained categories of TPA difference. In all specifications, we observe that evaluators do not react to small quality differences, but strongly react to larger differences.

Assimilation versus Contrasting The original framework by Bordalo, Gennaioli, and Shleifer (2020) suggests that evaluators not only contrast, but also assimilate candidates (see Appendix F.2 for more details). Whether contrasting or assimilation dominates depends on the (salience of) differences between two candidates. In particular, the framework predicts that assimilation dominates if two subsequent candidates have a small and non-salient quality difference.

The evidence presented in Figure 5 does not provide evidence of assimilation effects. 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 rule out the presence of assimilation assimilation, as the analysis might simply be

unable to detect small assimilation effects. In particular, the average pattern could hide potentially heterogeneous effects of assimilation on different types of candidates. Assimilation might, for instance, be only one-directional: evaluators follow the aim of separating candidates into high- and low-quality candidates. As a result, they might assimilate good to slightly better and bad to slightly worse candidates, but not vice versa.

To explore this possibility, we study in Figure 6 the impact of quality differences on candidates of high-quality and low-quality candidates, respectively.[39] Panel a provides suggestive evidence that low-quality candidates are assimilated downward when following a slightly worse candidate. In turn, when following a better candidate, they are contrasted immediately. A potential explanation is that evaluators use a slightly worse candidate in t-1 to realize that the candidate in t also belongs to the category of candidates who do not fulfill the criteria for a yes vote. In turn, a slightly better candidate in t-1 is used to differentiate the candidate in t from potentially suitable candidates. Analogously, high quality candidates (panel b) benefit from following a candidate who is slightly better. This suggests upward-assimilation into the category of suitable candidates. Again, contrasting immediately occurs when the previous candidate is of lower quality.[40]

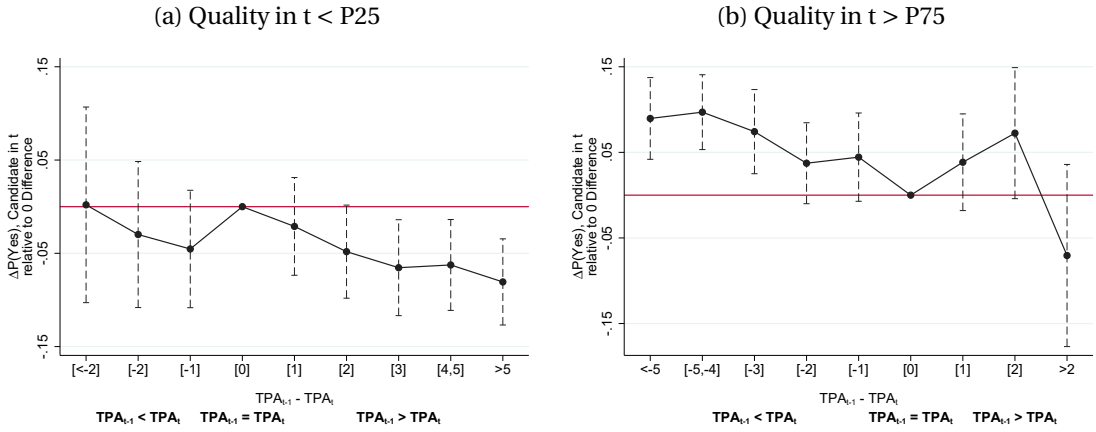
Wile the pattern provides interesting insights regarding the possible existence of assimilation effects and their direction, it is based on reduced statistical power and therefore needs to be interpreted with caution. It remains that contrasting is in this setting on average the dominating force, while assimilation has a comparatively minor influence.

Size and variability of the memory database Over the course of the admission workshop, evaluators continuously experience more candidates and thereby expand their memory database 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 pre-

[39] Note that, by construction, the difference in quality is only partially supported in the two panels. For high (low) quality candidates, it is not possible to follow a much better(worse) candidate.

[40] Appendix Figure F.8 shows the pattern for candidates whose quality is in the second or third quartile.

Figure 6: Influence of Quality Differences by Quality of Candidate in t



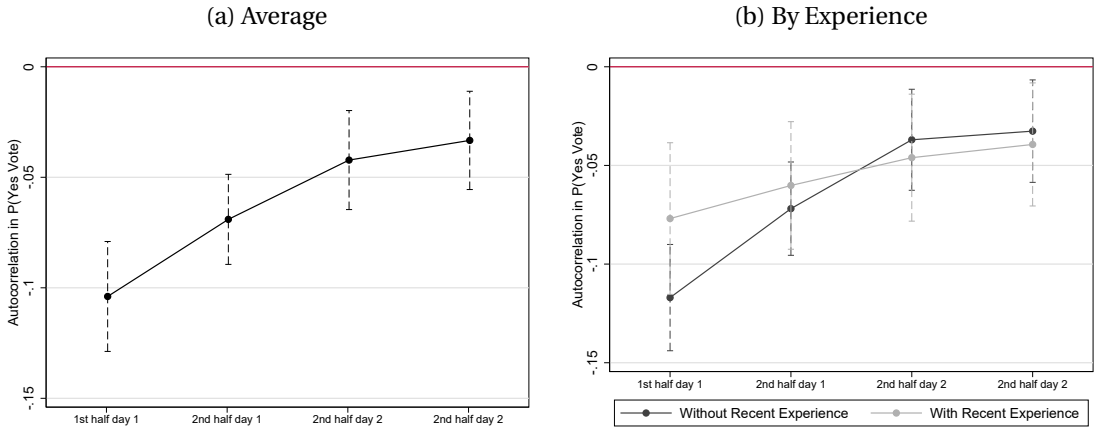
Note: Panel (a) includes candidates whose quality (measured by TPA) is below the 25th percentile. Panel (b) includes candidates whose quality (measured by TPA) is above the 75th percentile. The x-axis shows the difference in TPA between the candidate in t and the candidate $t-1$. The y-axis shows estimated 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. The dashed lines show 95% confidence intervals.

ceding candidates are similar to the current candidate to at least some extent.[41] We therefore expect that the negative autocorrelation in yes votes weakens over the course of the interview sequence. Figure 7 (a) is in line with this intuition. It shows that the autocorrelation weakens from about -0.1 for interviews conducted during the first half of the first interview day to about -0.03 for interviews conducted during the second half of the second day. Nonetheless, even at the end of the interview sequence, the autocorrelation remains statistically significant.

In Figure 7 (b), we check whether the pattern is steeper for inexperienced compared to experienced evaluators. More precisely, we compare evaluators who have interviewed for the program in the previous academic year (36%) to those who have not (64%). The results suggest that relatively recent background experience weakens contrasting at the beginning of the

[41] Additionally, the probability of observing a candidate who has similar characteristics as the current one increases. Potentially, the variance in the database also increases. These effects can block the recall of the previous candidate, beyond the pure effect of enlarging the database. With our non-experimental data where several aspects vary at the same time, we cannot distinguish these mechanisms. However, the general idea is common to all of them: the evaluator has a larger database with more variance from which he or she can retrieve experiences. These other experiences reduce the relative weight of the previous candidate and therefore reduce her influence.

Figure 7: 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. N=26,970. The dashed lines show 95% confidence intervals.

interview sequence. However, the two profiles converge over the remaining sequence. The average of the autocorrelation over the two days therefore does not significantly differ between evaluators with and without recent experience (see also the more detailed results from Appendix Table D.4).

5.2 Alternative Mechanisms

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

Sequential (Bayesian) Updating We first consider sequential updating about candidate quality, where evaluators form ratings immediately after observing each candidate. 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 ratings, two main reasons speak against its plausibility and relevance in this setting. First, candidates observed before and candidates observed afterwards matter similarly, which is not in line with immediate sequential updating (see Fig-

ure 2). Second, the ordering of prior candidates should be irrelevant for sequential updating. The quality of the previous candidate should not matter more than the quality of candidates observed in other preceding periods.

Gambler's Fallacy The belief in the law of small numbers (or representativeness heuristic) 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 (c.f. Benjamin, 2019; Rabin, 2002). 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 underestimate the probability that two candidates of similar quality follow each other. Therefore, they 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 and most importantly, the gambler's fallacy works through the prior belief about the upcoming candidate's quality. Therefore, under a gambler's fallacy, the influence of the previous candidate should not depend on the size of the surprise, i.e., the difference between the two candidates' quality. In opposition to this prediction, Figure 5 revealed that the influence of the previous candidate is a function of the difference between current and previous candidate quality.

Second, the gambler's fallacy predicts streaks of assessments to matter: two yes votes in a row should decrease the prior about the upcoming candidate stronger than one no vote followed by one yes vote. As a direct conjecture, the influence of two prior yes votes should be stronger, compared to one prior yes vote. As Appendix Table F.2 shows that we find no evidence in this direction. The probability of a yes vote does not decrease any further if not only the candidate in t-1, but also the candidate in t-2 receives a yes vote. This contradicts the evidence in Chen, Moskowitz, and Shue (2016), who find that the autocorrelation is stronger after a streak of two decisions, therefore concluding in favor of a gambler's fallacy.

Finally, we follow Chen, Moskowitz, and Shue (2016) and test for the influence of the pre-

vious 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 the previous candidate's quality. Under a contrast effect, evaluators instead react negatively to the continuous quality of the previous candidate. Columns (1) and (2) of Appendix Table F.3 show that the influence of the previous candidate's quality, measured through the third-party assessment (TPA), persists after controlling for the previous candidate's yes vote. This rejects the prediction of simple gambler's fallacy. However, as pointed out by Chen, Moskowitz, and Shue (2016), the result could still be in line with a more complicated version of the gambler's fallacy, 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 can express the strength of the vote, i.e., if it is a clear yes or no vote. Therefore, we control for the evaluator's uncertainty in column (3) by including the previous rating. The influence of the previous candidate's quality measure is unaffected, further pointing towards a contrast effect as the predominant mechanism.

6 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 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 large differences between the current and the previous candidate, who is strongly recalled due to similarity in the time dimension.

The findings in this paper help to understand how people make subjective assessments of individuals in the presence of others. They show that minor changes in candidate sorting and ordering can have major consequences in terms of who is selected. This carries implications for the organizational design of processes through which assessments are reached. First, our

results illustrate that it is crucial to mitigate the influence of individual biases by combining several assessments of a candidate and ensuring their independence. More precisely, it is key to minimize the overlap in the set and — importantly — the 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 thus lie in the combination of subjective assessments with more objective screening devices, such as algorithm-based job-testing technologies (e.g., Autor and Scarborough, 2008; Hoffman, Kahn, and Li, 2018). At present, it remains unclear how well these technologies perform when selecting from a high-ability segment of candidates.

Moreover, the paper shows that understanding the behavioral foundations behind evaluation errors yields additional implications for organizational design. Due to the interplay between memory and attention, the strength of contrast effects differs between circumstances. For instance, a candidate is contrasted less against the previous candidate when the two candidates share little similarity in terms of their observable characteristics. Organizations and firms can exploit this knowledge; for example, to strategically order candidates in a way that minimizes sequential contrasting.

Alternatively, organizations and firms can leverage our results to tackle the bias directly by raising awareness among evaluators. To date, it is unclear whether such interventions lead to an actual improvement in the validity of subjective assessments, as their effectiveness remains an open question. More generally, the efficiency of measures that target the validity of interview assessments will always depend on their benefits and costs. Yet, an understanding of the evaluation process — as provided by this paper — is a necessary prelude to analyzing this trade-off.

References

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*.

Bernheim, B. D., & Rangel, A. (2004). Addiction and cue-triggered decision processes. *American Economic Review*, *94*(5), 1558–1590.

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., 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 [qjaa007]. *The Quarterly Journal of Economics*.

Caeyers, B., & Fafchamps, M. (2020). Exclusion bias in the estimation of peer effects, (DP14386).

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

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.

DeGroot, M. H. (2005). *Optimal statistical decisions* (Vol. 82). John Wiley & Sons.

DellaVigna, S., Heining, J., Schmieder, J. F., & Trenkle, S. (2020). Evidence on job search models from a survey of unemployed workers in germany. *mimeo*.

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.

Gennaioli, N., & Shleifer, A. (2010). What comes to mind. *The Quarterly Journal of Economics*, *125*(4), 1399–1433.

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.

Laibson, D. (2001). A cue-theory of consumption. *The Quarterly Journal of Economics*, *116*(1), 81–119.

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.

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

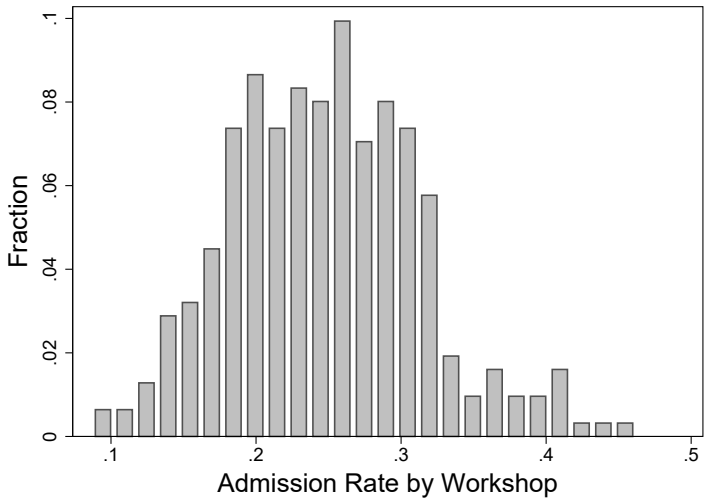
	Duration (minutes)	Type	Interviewer							
			A	B	C	D	E	F	G	H
Day 1	30	Group	1	7	13	19	25	31	37	43
	35	Interview 1	9	15	21	27	33	39	45	3
	35	Interview 1	46	4	10	16	22	28	34	40
	20	**Break**								
	30	Group	2	8	14	20	26	32	38	44
	35	Interview 1	35	41	47	5	11	17	23	29
	35	Interview 1	24	30	36	42	48	6	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 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: 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	0.031**	-0.000
	(0.014)	(0.012)
Age	0.004***	0.001
	(0.001)	(0.001)
Field: Social Sciences	0.033	0.013
	(0.022)	(0.020)
Field: STEM	0.028*	0.006
	(0.016)	(0.013)
Field: Medicine	0.018	-0.013
	(0.027)	(0.026)
Field: Others	0.004	-0.009
	(0.068)	(0.064)
Experience	-0.021***	0.000
	(0.003)	(0.003)
p-value (joint significance)	0.00	0.97
N	26970	26970

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

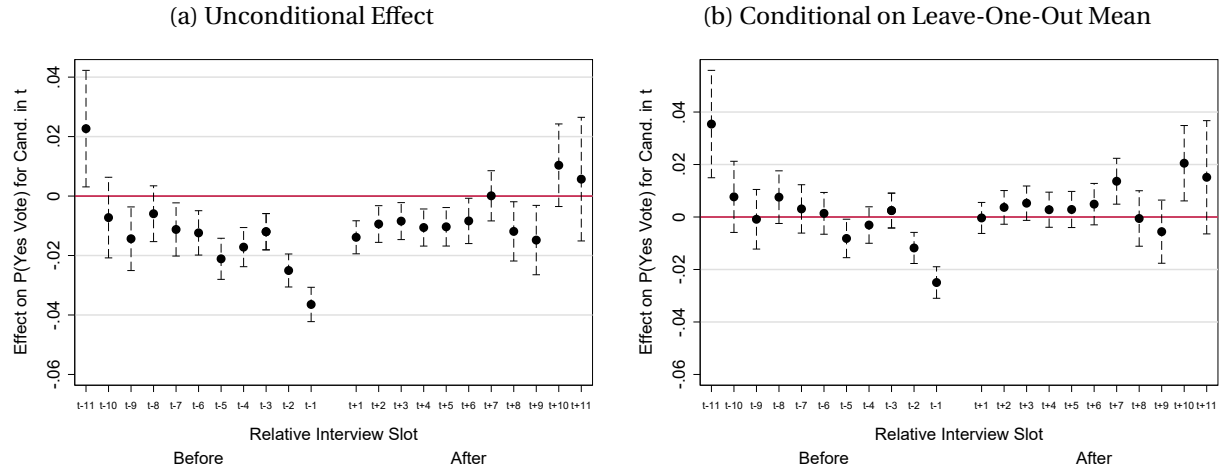
	Rating (Std.) (1)	Admission (2)
GPA Decile: 1	-0.117***	-0.039***
	(0.028)	(0.015)
GPA Decile: 2	-0.132***	-0.061***
	(0.029)	(0.013)
GPA Decile: 3	-0.054*	-0.036**
	(0.031)	(0.015)
GPA Decile: 4	0.009	0.008
	(0.029)	(0.016)
GPA Decile: 6	0.006	-0.004
	(0.033)	(0.017)
GPA Decile: 7	0.084***	0.028*
	(0.029)	(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

Panel (a) (Panel (b)) of 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 $k = -1$ (influence of the previous candidate's TPA) and show their robustness to alternative specifications (Table C.3) as well as 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

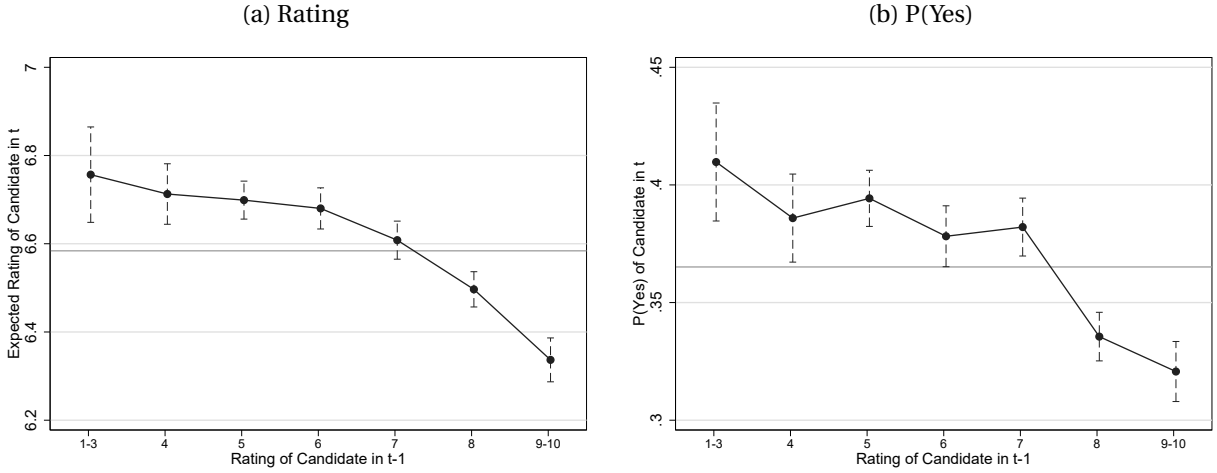
Note: TPA = third-party assessment of candidate quality (see section 3.3 for details). The leave-one-out mean is computed at the level of the evaluator's interview sequence. All regressions include workshop fixed effects. Controls include candidate characteristics, evaluator characteristics and interview 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

In Figure D.1, we additionally allow for a non-linear relationship. Panel (a) of Figure D.1 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. Overall, candidates receive on average 0.4 fewer points when following a candidate with a very high (9-10 points) instead of a very low (1-3 points) rating.

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. Figure D.2 reports how the non-linearity in the autocorrelation differs between candidates above and below the median TPA. 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).

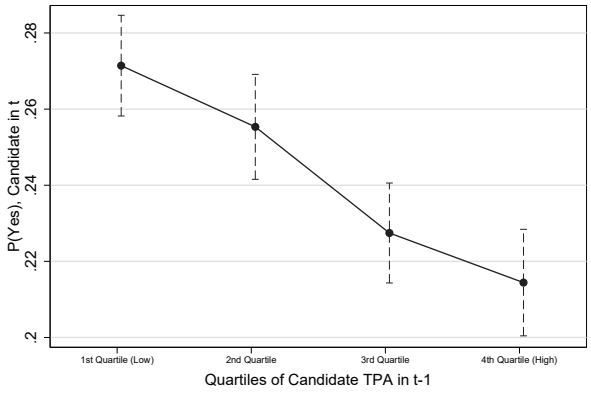
Table D.3: Robustness Checks: Autocorrelation Estimated with Evaluator Fixed Effects

	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

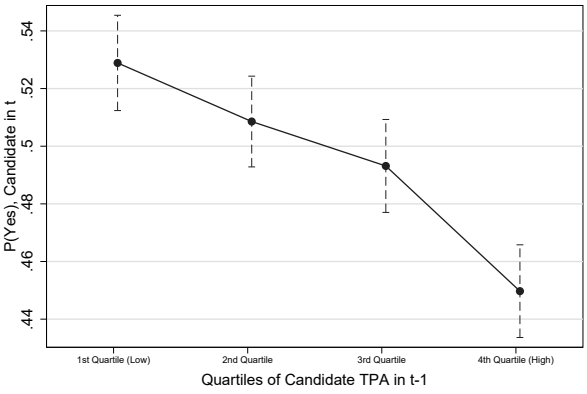
Note: All regressions include evaluator fixed effects. Due to collinearity, the evaluator's leave-one-out mean assessments are omitted. Further controls include candidate characteristics and interview order. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$. Standard errors are clustered at the workshop level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

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

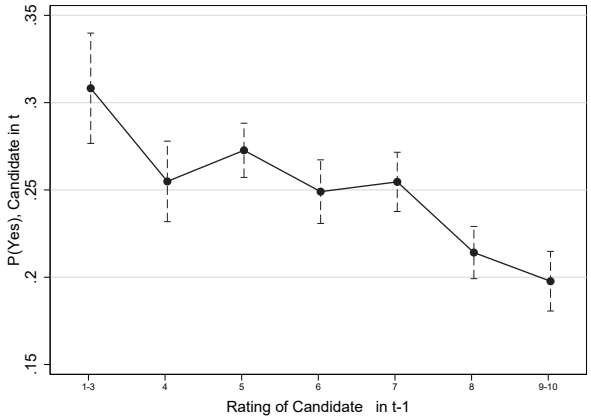
(a) Causal Effect for Candidates of Low TPA



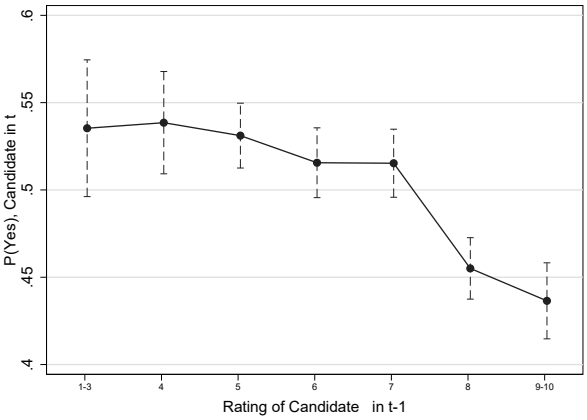
(b) Causal Effect for Candidates of High TPA



(c) Autocorrelation for Candidates of Low TPA



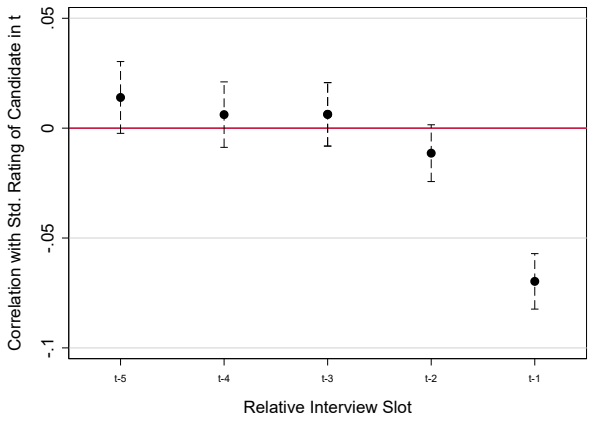
(d) Autocorrelation for Candidates of High TPA



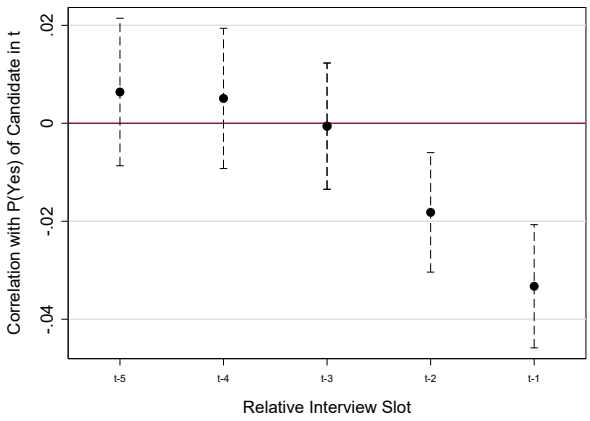
Note: "Low TPA": third-party assessment of quality ≤ median. "High TPA": third-party assessment of quality > median. Estimates result from two-way-interacted regression models. The regression underlying panels (a) and (b) controls for workshop fixed effects, the leave-one-out mean TPA at the evaluator 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: Reversal of Admission Outcomes

In a final step, we provide a back-of-the-envelope quantification on the reversal of admission outcomes induced by the autocorrelation. The reversal rate tries to capture the number of votes and admission decisions that are reversed due to the negative autocorrelation in votes.

To compute reversals, we follow the approach by Chen, Moskowitz, and Shue (2016). Using their approach, we derive the share of reverted decisions from a simple regression $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$. Taking expectations, $E(Y) = \frac{\beta_0}{1-\beta_1}$. Assuming that the rate of positive decisions, $P(Y = 1)$, would be equal in the absence of the autocorrelation, reversal can be due to two situations. If the previous candidate received a no vote, the negative autocorrelation increases the current candidate's probability of a yes vote by $\beta_0 - P(Y = 1)$, i.e., her empirical probability of receiving a yes vote minus her (assumed) counterfactual probability. If the previous candidate received a yes vote, the current candidate is not sufficiently likely to receive a yes vote by $P(Y = 1) - (\beta_0 + \beta_1)$, i.e., the counterfactual probability of a yes vote minus the empirical probability. The expected number of reversals is the weighted instance of the two cases $(\beta_0 - P(Y = 1))P(Y_{t-1} = 0) + (P(Y = 1) - (\beta_0 + \beta_1))P(Y_{t-1} = 1)$. Substituting $P(Y = 1) = \frac{\beta_0}{1-\beta_1}$, the rate of affirmative decisions becomes $R = -2\beta_1 P(Y = 1)(1 - P(Y = 1))$.

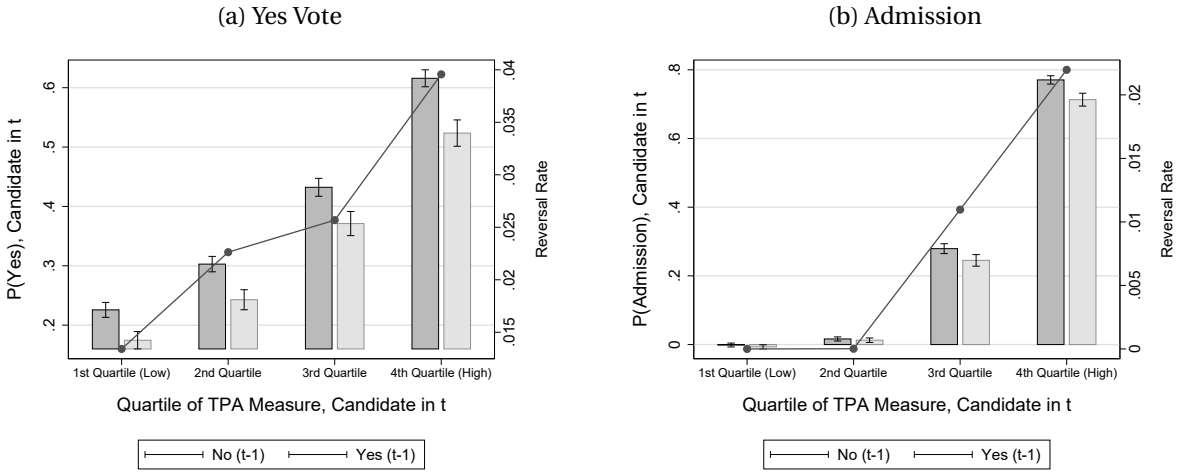
Intuitively, the reversal rate is expected to vary over the distribution of candidate quality. The outcomes of candidates with very weak admission prospects are less likely to be reverted than those of candidates who are more likely to reach the threshold for a yes vote. We therefore calculate a separate reversal rate for each quartile of candidate quality, as measured by the third party assessment (TPA).

Figure E.4 illustrates the reversal pattern. In Panel (a), the outcome is the individual yes vote. The share of reversals is about 1.5% for candidates from the lowest quartile and 2.5% for candidates from the second or third quartile. Candidates from the fourth quartile — who have in expectation the best prospects of receiving a yes vote — have a reversal rate of about 4%.

When we consider the admission outcome (Panel b), the pattern looks similar. Here, the reversal rate is 0 for candidates from the two lowest quartiles. This is partly mechanical, given that candidates who receive low ratings from the other two evaluators have close-to-zero chances

of being admitted, irrespective of the outcome of the third assessment. However, candidates in the upper two quartiles are at the margin of admission. About 1.2% to 2.2% of these candidates would obtain a different admission decision if one of the three evaluators did not have autocorrelated votes.

Figure E.4: Influence on Admission Outcomes by Candidate Quality



Note: The bars plot margins based on the estimation of equation 2, where the previous candidate's yes vote is interacted with quartiles of the current candidate's quality as measured by TPA (=third-party assessment). The connected line plots the fraction of votes (panel a) and admission decisions (panel b) that are reverted due to the autocorrelation. The computation of reversal rates is described in Appendix E. N=26,970. 95% confidence intervals, with standard errors clustered at the workshop level.

F Additional Material: Potential Mechanisms

This section first provides a model of evaluator learning that is able to explain the overall influence of the other candidates seen by the same evaluator (F.1). It then outlines the model of associative memory and attention proposed by Bordalo, Gennaioli, and Shleifer, 2020 adapted to our setting (F.2).

F.1 A Framework of Evaluator Learning

We lay out a framework where a rational risk-neutral evaluator votes on the admission of a closed sequence of candidates. The evaluator's aim is to accept candidates whose quality exceeds a threshold. The evaluator forms beliefs about each candidate's quality based on noisy signals. Moreover, she infers the average of the quality distribution through the observed signals. This average determines the evaluator's beliefs about the quality threshold. Signals are received sequentially, but decisions are made at the end of the sequence. This is a key difference compared with sequential decision-making models, where updating and decisions occur after each period. Finally, Appendix F provides additional empirical results related to the discussion in the main text.

Setup Suppose that a candidate observed by evaluator i at time t has quality $q_{i,t} \sim \mathcal{N}(\theta_0, \sigma_0^2)$. The risk-neutral evaluator observes a noisy signal of quality, $\tilde{q}_{i,t} = q_{i,t} + \epsilon_{i,t}$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

The evaluator votes on the admission decision of each observed candidate. Each decision is supposedly independent, as the evaluator does not face a quota. Therefore, the evaluator evaluates the two alternatives of voting in favor of or against admitting a candidate. Admission yields a value $V_{accept} = \mathbb{E}(q_{i,t})$, while the value of a rejection is $V_{reject} = \underline{q}_i$.

In the expression of V_{reject}, $\underline{q}_i$ is a predefined quality threshold. It can be expressed as $\underline{q}_i = \alpha_i * \mathbb{E}(q)$, where $\alpha_i > 1$ is an evaluator-specific term capturing — for example — differences in leniency between evaluators. The candidate receives a yes vote if $V_{accept} > V_{reject}$, i.e.

$$\mathbb{E}(q_{i,t}) > \alpha_i \mathbb{E}(q) \tag{4}$$

This decision rule implies that an evaluator votes in favor of a candidate if her posterior belief about the candidate's quality exceeds the threshold. The threshold depends on the expected quality of all candidates. Therefore, the evaluator has to form posterior beliefs about the quality of the candidate and the average quality of all candidates.

As postulated above and in line with the institutional framework, we assume that the evaluator updates her belief about the mean quality after observing all (uncorrelated) signals $\tilde{q}_{i1}, \tilde{q}_{i2}, ..., \tilde{q}_{iT}$. Let $\overline{\tilde{q}_i}$ be the average of all signals. Following Bayes' rule, we can express the updated belief about $\mathbb{E}(q)$ (c.f. DeGroot, 2005):

$$\theta_1 = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2 T}\theta_0 + \frac{T\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2 T}\overline{\tilde{q}_i} \tag{5}$$

In a second step, the evaluator forms posterior beliefs about the quality of each individual candidate, given her posterior belief about the average. The belief about the quality of a candidate is thus a precision weighted average of the signal and the posterior belief about the average:

$$q_{posterior,i,t} = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2}\theta_1 + \frac{\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2}\tilde{q}_{i,t} \tag{6}$$

Decision Rule Plugging the two posterior beliefs into equation 4 yields the following decision rule:

$$\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2}\theta_1 + \frac{\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2}\tilde{q}_{i,t} > \alpha_i \theta_1 \tag{7}$$

The decision rule shows that an evaluator votes in favor of a candidate if her posterior belief about the quality of this candidate exceeds the threshold, which depends on the posterior of the mean quality. In this rule, a candidate's signal acts in two counteracting ways: on the one hand, it affects the posterior belief about the candidate's individual quality; and on the other hand, it affects the threshold, as it increases the posterior belief about the average quality. To solve the model, we assume that the first effect dominates, i.e. the threshold reacts less than the posterior belief about individual quality. Formally, this assumption is satisfied iff $\frac{2\sigma_\epsilon^2+\sigma_0^2 T}{\sigma_\epsilon^2+\sigma_0^2} > \alpha_i$. For $T \geq 2$, it is sufficient that $\alpha_i < 2$. While we cannot test this condition formally, it is plausible that this condition is fulfilled in our data, as the left-hand side of the condition is increasing in T and T is relatively large in our data. We can then derive a threshold for the signal of each candidate:

$$\tilde{q}_{i,t} > \Big[\alpha_i - \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2+\sigma_0^2}\Big]\Big[\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2+\sigma_0^2 T}\theta_0 + \frac{(T-1)\sigma_0^2}{\sigma_\epsilon^2+\sigma_0^2 T}\overline{\tilde{q}_{i,-t}}\Big] \frac{\sigma_\epsilon^2+\sigma_0^2}{\sigma_0^2}\Big[1 - \frac{\alpha_i(\sigma_\epsilon^2+\sigma_0^2)}{\sigma_\epsilon^2+\sigma_0^2 T} + \frac{\sigma_\epsilon^2}{(\sigma_\epsilon^2+\sigma_0^2 T)}\Big]^{-1}$$

A candidate is therefore accepted if her signal $\tilde{q}_{i,t}$ exceeds the threshold $\underline{q_{i,t}}(\alpha_i, \sigma_\epsilon^2, \sigma_0^2, \overline{\tilde{q}_{i,-t}}, T, \theta_0)$. The threshold increases in α_i, reflecting the notion that evaluators with a higher leniency have lower thresholds. It further increases in the average signal, which implies that the individual probability of a yes vote decreases in the (average) signals of the other candidates observed by the same evaluator.

As a direct consequence of the threshold rule, the average quality of the other candidates observed by the same evaluator can affect the rating of the single candidate. Besides, it is easy to see that the partial derivatives of the threshold with respect to the signals of any other candidate do not depend on the timing of a particular candidate's interview. Therefore, the threshold rule cannot rationalize why the previous candidate has a stronger impact than the other candidates.

F.2 A Framework of Contrast Effects Based on Associative Memory and Attention

In the following, we provide a more formal framework for the intuition discussed in section 5. The framework adapts the model by Bordalo, Gennaioli, and Shleifer, 2020 to our setting.

We consider an evaluator who has to vote on the admission of a candidate interviewed in period t. The evaluator decides on her votes after interviewing all candidates. She votes in favor of admitting the candidate interviewed in t if her valuation V_t of that candidate exceeds an admission threshold, which depends on the quality of all other observed candidates.[1] Valuation V_t depends on the candidate's quality as perceived by the evaluator, $\tilde{q}_t$.[2] Observing a candidates with perceived quality $\tilde{q}_t$ in a context c_t defines an interview experience e_t. 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. Observed context variables that vary in our setup are the time of interview as well as candidate characteristics (e.g., gender or study field). Given an interview experience e_t, the evaluator recalls experienced candidates and uses their perceived quality to form the norm $q_t^n(c_t)$. More similar interview experiences are overweighted, where the similarity of an interview that took place in period $t-l$ is measured by the function $S(e_{t-l}, c_{t-l})$. 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, $S(e_{t-l}, c_{t-l}) = S(|t-(t-l)|)$ for l=1,..,11, where t indicate the point in time of one interview and $t-l$ indicates the point in time of other interviews. $S : \mathbb{R}_+ \to \mathbb{R}_+$ is decreasing.

[1] A simple example of a threshold is $\alpha\mathbb{E}(q)$, where $\alpha >$ is an evaluator-specific constant (capturing, e.g., leniency) and $\mathbb{E}(q)$ denotes the expected average candidate quality.

[2] We model the valuation as being based on a candidate's posterior quality signal. We are agnostic about the exact point in time at which the valuation takes place. To some degree, our results suggests that it takes place at the time of the interview itself, as only the previous candidate matters beyond the group effect.

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

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

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

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

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 evaluator evaluates the candidate in t given her recall-based quality norm q_t^n and given the candidate's perceived quality $\tilde{q}_t$. Following Bordalo, Gennaioli, and Shleifer (2020), we write the valuation as:

$$V_t = q_t^n(c_t) + \sigma(\tilde{q}_t, q_t^n(c_t)) \times (\tilde{q}_t - q_t^n(c_t)) \tag{8}$$

Valuation V_t is composed of two terms. First, it is anchored to the evaluator's quality norm q_t^n, which depends on context c_t (note that the anchoring term is omitted for simplicity in section 5). Second, it increases in the difference between the candidate's own quality as perceived by the evaluator and the norm. 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 $\lim_{x/y \to \infty} \sigma(x/y, 1) = \sigma > 1$.

Assimilation vs. Contrasts We can use the postulated framework to study how the valuation of a candidate reacts to a change in the (perceived) quality of the previous candidate. Formally, the reaction is described by:

$$\frac{\partial V_t}{\partial \tilde{q}_{t-1}} = w_{t-1} + \frac{\partial \sigma(\tilde{q}_t, q_t^n)}{\partial q_t^n} w_{t-1}(\tilde{q}_t - q_t^n) - \sigma(\tilde{q}_t, q_t^n) w_{t-1} \tag{9}$$

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. 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.

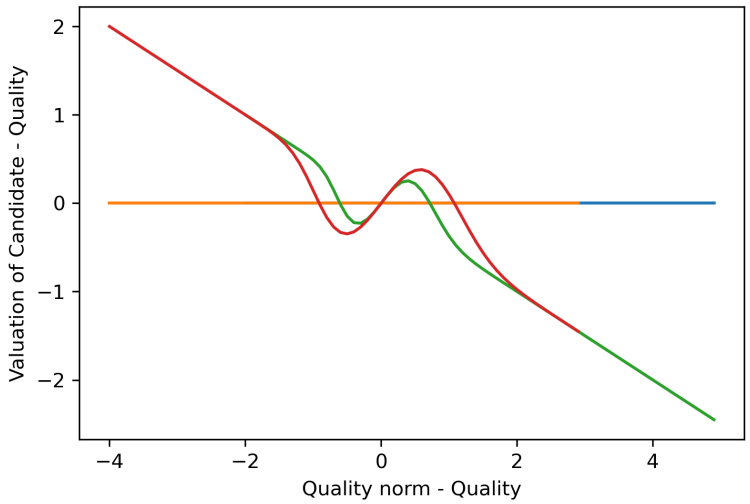
It is straightforward to see that the strength of both anchoring and contrasting depends on w_{t-1}, the weight of the previous candidate in the norm.[3] However, which of the two counteracting mechanisms dominates depends on the size of the surprise as described by $q_t - q_t^n$. If the surprise is small, it does not capture the evaluator's attention. Anchoring is thus relatively important and can lead to assimilation in the valuation of two subsequent candidates. For larger surprises, contrasting as described by the second and third parts dominates.

Figure F.1 illustrates this pattern by showing how the valuation of a candidate is predicted to deviate from her own quality as a function of the difference between the candidate's quality and the quality $(q_t^n(c_t) + \sigma(\tilde{q}_t, q_t^n(c_t)) \times (\tilde{q}_t - q_t^n(c_t)) - q_t)$.

Our empirical results — as reported in Figure 5 of section 5 — neither reject nor confirm the presence of assimilation effects. 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,

[3] This also explains why another candidate's influence depends on the relative timing of her interview. As similarity depends on relative timing, the weight w_{t-l} depends on l. In the strongest version, it is close to zero for any $l \neq 1$.

Figure F.1: Relationship between Norm and Valuation



Note: The green line displays the values for a candidate with quality=4 and the red line with quality= 6. The norm varies from 2 points to 9 points. The salience function is of the form $\sigma(x,1) = \sigma\frac{e^{\theta(x-1)^2}}{1+e^{\theta(x-1)^2}} - \sigma_0$ for $x \geq 1$, where $\sigma > 1+\sigma_0, \theta > 0$ and $\sigma(1,1) = 0$. In the figure, $\sigma = 3, \sigma_0 = 3/2$, and $\theta = 50$.

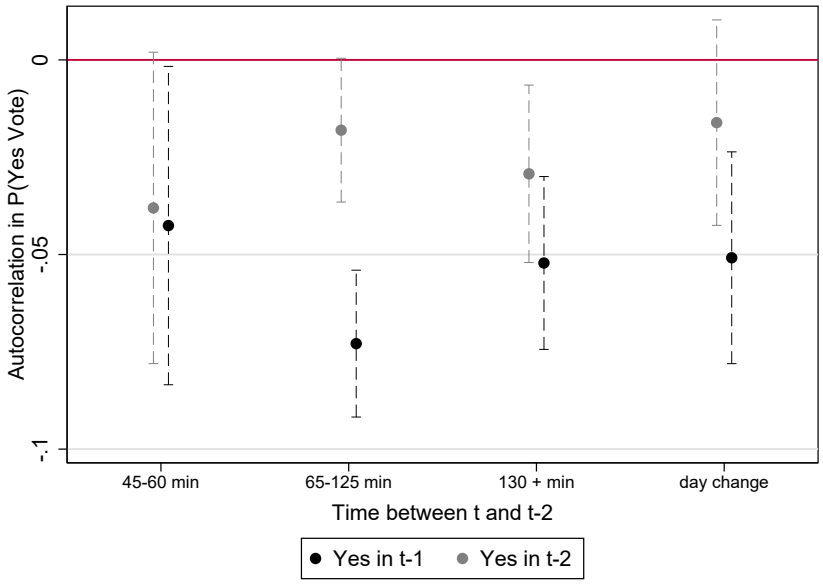
which would hint at assimilation. In particular, the model parameters can be specified to yield the prediction of very small assimilation effects, which we might simply be unable to detect.[4,5]

[4] 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.

[5] 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.

F.3 Additional Material: The Role of Breaks

Figure F.2: Autocorrelation and the Time Lag between t and t-2



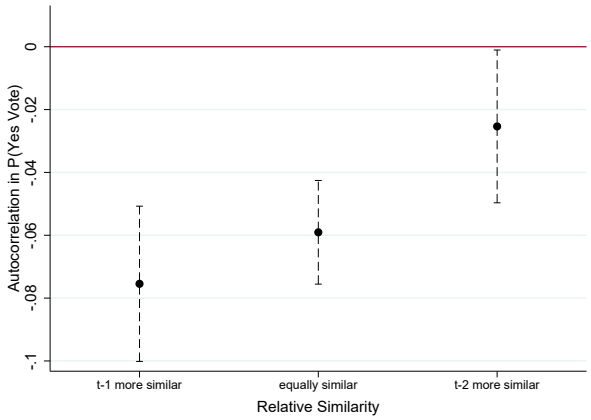
Note: The black dots plot estimates of the autocorrelation based on equation 2, where the previous candidate's yes vote is interacted with the time gap between the end of the interview in t-2 and the start of the interview in t. The gray dots repeat the exercise, but replace the yes vote of the candidate in t-1 with the yes vote of the candidate in t-2. N=26,970 (black dots); N= 24,474 (gray dots). The dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level.

F.4 Additional Material: Additional Dimensions of Similarity

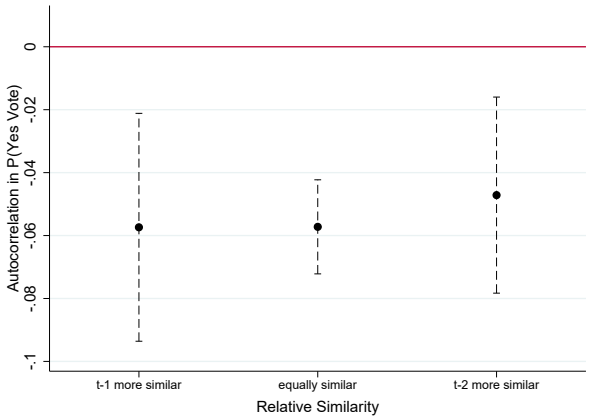
The Role of Additional Similarity

Figure F.3: Interaction between Prior Candidate Quality and the Gender Sequence: Pilot Data

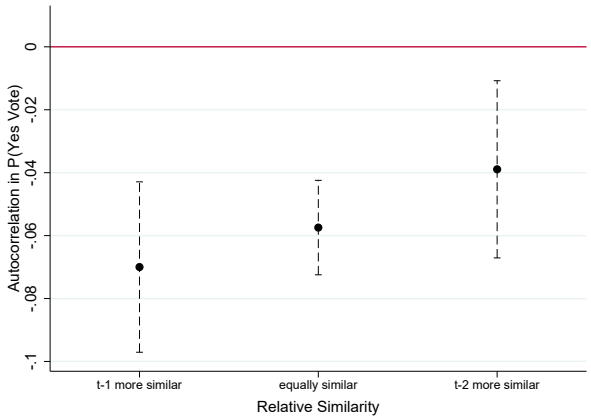
(a) Gender



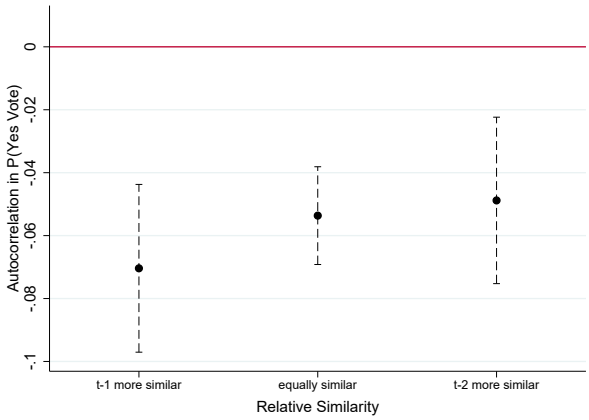
(b) Migration Background



(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.

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 F.1 and Figure F.4. 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 asymmetry 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 F.1 and Figure F.5 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 F.5 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 by the previous candidate's gender

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

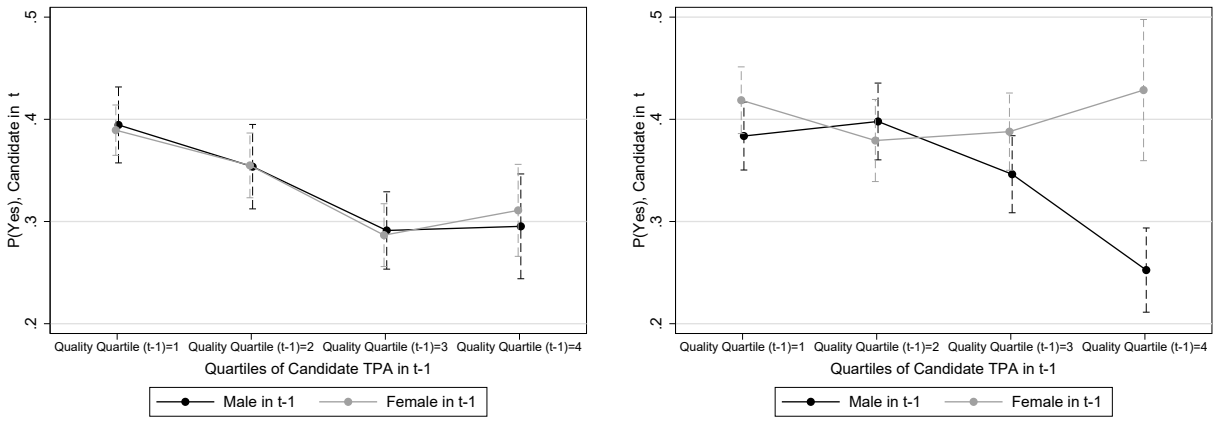
	Pilot Data		Main Data	
	(1) Rating (Std.)	(2) P(Yes Vote)	(3) Rating (Std.)	(4) P(Yes Vote)
Male × Male (t-1) × TPA (std.), t-1	-0.079***		-0.064***	
	(0.019)		(0.012)	
Male × Female (t-1) × TPA (std.), t-1	-0.025		-0.055***	
	(0.024)		(0.011)	
Female × Male (t-1) × TPA (std.), t-1	-0.059**		-0.070***	
	(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 F.4: Interaction between Prior Candidate Quality and the Gender Sequence: Pilot Data

(a) Female Candidate in t

(b) Male Candidate in t

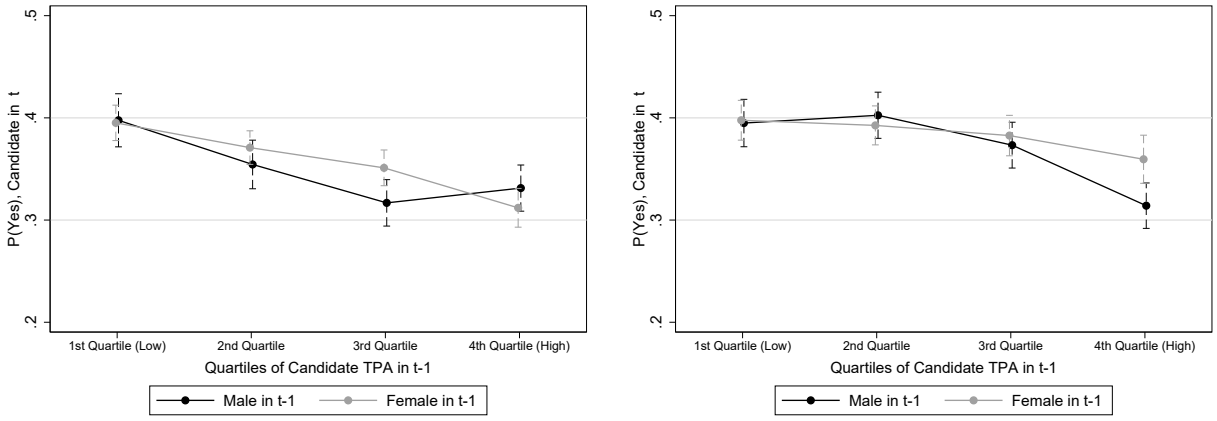


Note: The "pilot data" include the academic year 2012/13 (N=8,522). Estimates in panels (a) and (b) result from the same two-way-interacted regression model. Controls include the leave-one-out mean TPA of the interview sequence, candidate and evaluator characteristics, interview order and workshop fixed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

Figure F.5: Interaction between Prior Candidate Quality and the Gender Sequence: Main Data

(a) Female Candidate in t

(b) Male Candidate in t

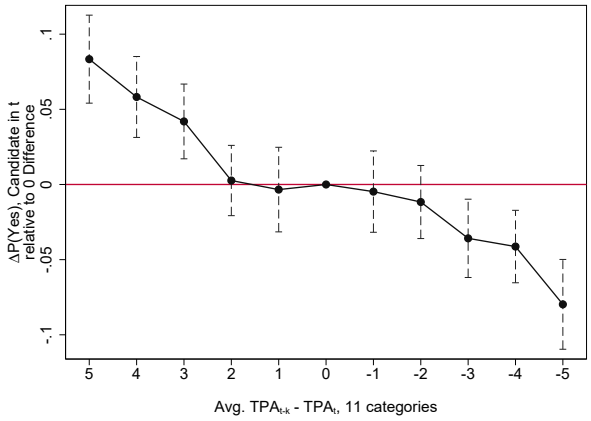


Note: Estimates in panels (a) and (b) result from the same two-way-interacted regression model. Controls include the leave-one-out mean TPA of the interview sequence, candidate and evaluator characteristics, interview order and workshop fixed effects. 95% confidence intervals, with standard errors clustered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

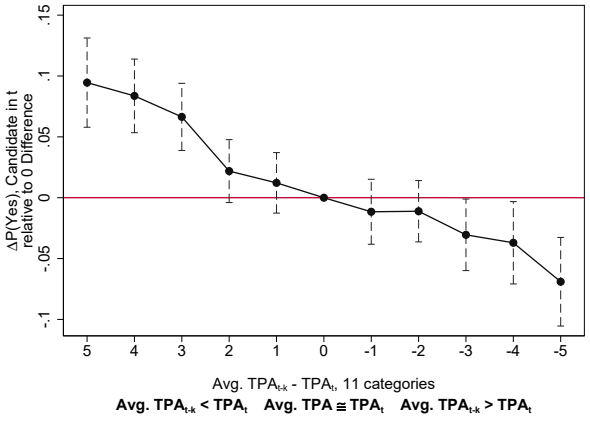
F.5 Additional Material: Attention and Size of the Surprise

Figure F.6: Alternative Proxies of the Quality Norm

(a) Norm=Average TPA of Previous Two Candidates



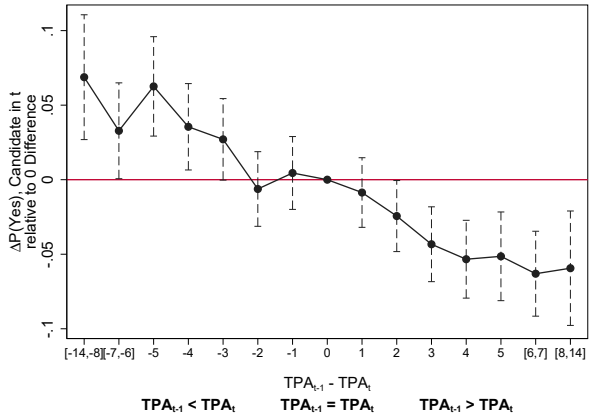
(b) Norm=Average TPA of All Previous Candidates



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.

F.6 Additional Material: Assimilation versus Contrasting

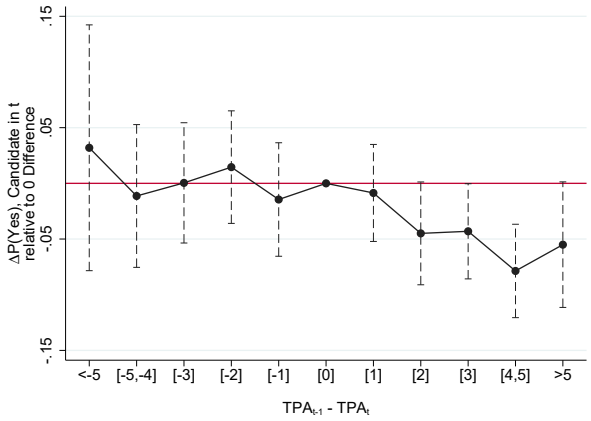
Figure F.7: More Categories for Difference in TPA between Current and Previous Candidate



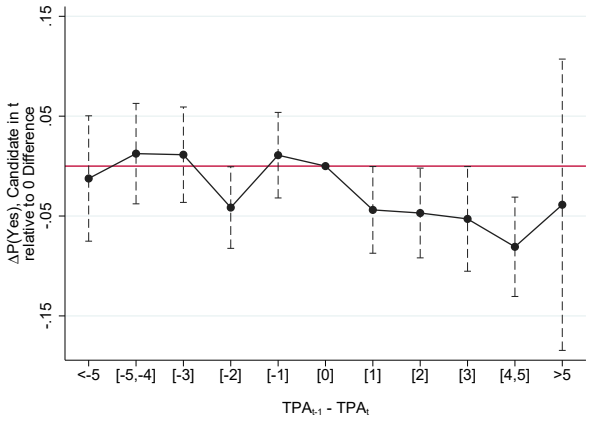
Note: The x-axis denotes the difference in points between the third-party assessment (TPA) of the current and the TPA of the previous candidate. The y-axis shows estimated coefficients on the probability of receiving a yes vote for the candidate in t. The underlying regression includes dummies for the 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.

Figure F.8: Influence of Quality Differences by Quality of Candidate in t

(a) Quality in t P25-P50



(b) Quality in t P50-P75



Note: Panel (a) includes candidates whose quality (measured by TPA) is between the 25th and 50th percentile. Panel (b) includes candidates whose quality (measured by TPA) is is between the 50th and 75th percentile. 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. The dashed lines show 95% confidence intervals.

F.7 Additional Material: Alternative Mechanisms

Table F.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 F.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).