Radbruch, Jonas; Schiprowski, Amelie

Working Paper

Interview sequences and the formation of subjective assessments

Discussion Paper, No. 497

Provided in Cooperation with:
University of Munich (LMU) and Humboldt University Berlin, Collaborative Research Center Transregio 190: Rationality and Competition

Suggested Citation: Radbruch, Jonas; Schiprowski, Amelie (2024) : Interview sequences and the formation of subjective assessments, Discussion Paper, No. 497, Ludwig-Maximilians-Universität München und Humboldt-Universität zu Berlin, Collaborative Research Center Transregio 190 - Rationality and Competition, München und Berlin

This Version is available at:
https://hdl.handle.net/10419/289827

RATIONALITY & COMPETITION
CRC TRR 190

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch (HU Berlin)
Amelie Schiprowski (University of Bonn)

Discussion Paper No. 497

February 16, 2024

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch[†] Amelie Schiprowski[§]

February 14, 2024

Abstract

Interviewing is a decisive stage of most processes that match candidates to firms and organizations. This paper studies how and why a candidate's interview outcome depends on the other candidates interviewed by the same evaluator. We use large-scale data from high-stakes admission and hiring processes, where candidates are quasi-randomly assigned to evaluators and time slots. We find that the individual assessment decreases as the quality of other candidates assigned to the same evaluator increases. The influence of the previous candidate stands out, leading to a negative autocorrelation in evaluators' votes of up to 40% and distorting final admission and hiring decisions. Our findings are in line with a contrast effect model where evaluators form a benchmark through associative recall. We assess potential changes in the design of interview processes to mitigate contrasting against the previous candidate.

JEL Codes: D91, J20, M51

[†] HU Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email: jonas.radbruch@hu-berlin.de

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

The main specifications and variable definitions in this project were pre-registered under osf.io/t65zq. We thank several seminar and conference participants, and in particular Johannes Abeler, Steffen Altmann, Maria Balgova, Pedro Bordalo, Stefano DellaVigna, Thomas Dohmen, Andreas Grunewald, Lena Janys, Andreas Lichter, Andrei Shleifer, Uri Simonsohn, Florian Zimmermann and Ulf Zoelitz for helpful discussions and comments. Julia Wilhelm, Stefanie Steffans, and especially Annica Gehlen provided outstanding research assistance. Support by the German Research Foundation (DFG) through EXC 2126/1– 390838866, CRC TR 224 (Project A05, Schiprowski) and CRC TR 190 (Project Z, Radbruch) is gratefully acknowledged.

1 Introduction

Subjective assessments are commonly used to measure quality or performance. Examples include the evaluation of employees, the screening of applicants and the grading of students. As subjective assessments can have long-lasting consequences for individuals and organizations, it is important to understand their underlying formation.

The personal interview, which is a decisive stage of most hiring and admission processes, is a context where subjective assessments are particularly prevalent. A core feature of interviewing is its sequential nature, as evaluators encounter one candidate after the other, often at a high frequency. This can have important consequences for the assessment and relative comparison of candidates. The difficulty to process sequential information —for example due to memory limitations— may lead evaluators to assess the current candidate relative to the previous one. The relevance of this phenomenon, commonly known as the sequential contrast effect, has been documented in laboratory experiments (e.g., Pepitone & DiNubile, 1976; Kenrick & Gutierres, 1980; Wexley et al., 1972) and a few real-world applications, such as speed dating (Bhargava & Fisman, 2014), housing choices (Simonsohn & Loewenstein, 2006), and financial markets (Hartzmark & Shue, 2018).[1] In the context of interviewing, contrast effects bear the potential to cause arbitrary spillovers from one candidate's quality to the next candidate's assessment, distorting hiring and admission outcomes.

The main contribution of this paper is to provide large-scale field evidence on the quantitative importance and behavioral nature of contrast effects in high-stakes admission and hiring processes. First, we estimate how the evaluation of a candidate is affected by the quality of the other candidates in the same interview sequence, depending on their relative order. Having identified a striking negative influence of the previous candidate's quality, we analyze how this influence varies with the evaluator's prior experiences and the similarity between subsequent candidates. We then study how a contrast effect model with associative recall can explain our

[1] Additional field studies have documented different types of interdependence in subjective assessments or decisions. In particular, Simonsohn and Gino (2013) show that MBA interview assessments are influenced by the average score of other candidates seen on the same day. They suggest that evaluators engage in narrow bracketing and target a certain number of positive decisions per day. Chen et al. (2016) attribute a negative autocorrelation in decisions by asylum judges, loan officers, and baseball umpires to the influence of a gambler's fallacy.

empirical findings and discuss alternative mechanisms. In a final step, we explore policies to mitigate the influence of contrast effects on hiring and admission decisions.

The analysis relies on register data from two high-stakes interview processes. Our primary data source covers about 29,000 interviews from the admission process of a prestigious study grant program funded by the German government. The program yields several monetary and non-monetary benefits, including a generous stipend, mentoring and the access to an active network. We complement the analysis with data on about 8,000 interviews from the hiring process of a large consulting company that selects employees for high-paying internships and permanent positions. The study grant's admission process is organized through two-day workshops, where evaluators conduct twelve one-to-one interviews. In the hiring process, evaluators conduct three one-to-one interviews on each assessment day. The following features of the two setups are key for our analysis: first, candidates are quasi-randomly assigned to evaluators and time slots; second, each candidate has a clearly defined reference group, as evaluators observe closed sequences of candidates; third, evaluators do not face an explicit quota, as admissions and job offers occur on a rolling basis; and fourth, each candidate receives three independent assessments, facilitating the measurement of unobserved candidate quality.

Exploiting the quasi-random assignment and ordering of candidates, we estimate how the assessment of a candidate changes when the measured quality of another candidate in the same interview sequence increases. As a proxy for unobserved candidate quality, we rely on an independent third-party assessment (TPA). Specifically, the TPA is defined as the sum of two independent ratings made by different evaluators. To address issues related to multiple hypothesis testing, selective data-slicing and discretion in the definition of candidate quality, we pre-registered the main specifications and variable definitions.[2]

The results show that the same candidate is evaluated worse when assigned to an interview sequence with better candidates. However, the impact of other candidates strongly depends on their position in the sequence. In particular, the influence of the immediately preceding

[2] The pre-registration can be accessed at osf.io/t65zq. It refers to the study grant admission process. Prior to pre-registration, we had access to a pilot dataset, which is excluded from the analyses in this paper. When analyzing the hiring data, we stick to the same pre-registered specifications unless we need to adapt them to the slightly different institutional setup.

candidate is about three times stronger than the influence of the average other candidate in the sequence. A one standard deviation increase in the previous candidate's quality measure is about 25% (admission process) to 45% (hiring process) as influential as a one standard deviation decrease in a candidate's own quality measure. This leads to a strong negative autocorrelation in evaluators' binary decisions. In the admission (hiring) process, candidates who follow a candidate with a yes vote are about 15% (40%) less likely to receive a yes vote themselves. The magnitude of this autocorrelation is substantial compared to other factors that affect evaluator decisions. For instance, it is comparable in size to the effect of a one (two) standard deviation change in evaluator leniency in the admission (hiring) process.[3] The previous candidate's influence persists beyond the single interview and leads to large changes in the final decisions taken by the respective admission and hiring committees. Specifically, an additional yes vote given to the previous candidate in one out of two interviews reduces the probability of being admitted or hired by about 20% relative to the average.

We proceed by investigating how the influence of the previous candidate depends on the decision environment of the evaluator. We first document that the influence decreases over the interview sequence, as evaluators encounter more candidates. This can also explain the stronger average effect in the hiring process, where sequences are shorter. Conversely, experiences from past interview sequences do not mitigate the influence. Second, longer breaks between interviews are associated with a lower autocorrelation. Third, the previous candidate exerts a stronger influence when being more similar to the current candidate; for example, in terms of gender and study background.

Based on the empirical findings, we discuss the behavioral mechanism behind the previous candidate's strong influence. An intuitive mechanism is a contrast effect, where evaluators assess candidates relative to a quality benchmark or norm. To fix ideas, we consider a contrast effect model where the norm is formed through associative recall, based on the framework by Bordalo et al. (2020).[4] Applied to our setting, associative recall suggests that evaluators retrieve

[3] Decision-maker leniency has been shown to have large effects on individual outcomes (see, e.g., Bhuller et al., 2020, for evidence on differences in judge leniency).

[4] The notion of associative recall is a guiding principle in psychological research on memory (see, e.g., Kahana, 2012; Kahana et al., 2022). We summarize the relevant literature in Appendix A.

previous interview experiences from memory based on their contextual similarity to the current interview. Thereby, more recent and similar candidates receive a stronger weight in the quality norm, which can explain the previous candidate's influence and its heterogeneity. Additional reduced-form results show that distinctive features of the framework are in line with the data. Specifically, a key implication of associative recall is interference, whereby relatively more recent and similar interviews disrupt the recall of older and less similar interviews. In line with this notion, we find that the strength of contrasting depends on the relative — rather than absolute — recency and similarity between interviews. As further evidence favoring a contrast effects explanation, we find that the previous candidate's influence is stronger within than between sub-dimensions of candidate quality. To complement the reduced-form analysis, we evaluate the framework's quantitative plausibility with a simple structural estimation. The results indicate that the framework can capture essential moments of the data.

Although a contrast effect model with associative recall offers a qualitatively and quantitatively plausible way to explain the findings, other behavioral mechanisms can also lead to a negative autocorrelation in decisions. We assess the potential relevance of sequential learning about a quality threshold and other belief-based explanations as the gambler's fallacy. Our main findings and additional empirical tests rule out simple versions of these alternative mechanisms. While more complicated versions could be used to explain parts of the results, it is difficult to align them with all patterns in the data.

Irrespective of its behavioral mechanism, the influence of the previous candidate significantly distorts assessments within professional selection processes. We explore different policy interventions designed to counteract this distortion. We first document that an information treatment implemented by the study grant program turned out to be ineffective. We then simulate and discuss the potential of alternative solutions, such as the implementation of a reordering algorithm, the collection of additional independent evaluations, and the flagging of specific interview assessments for final committee discussions. Although these approaches cannot easily reduce contrast effects to zero, they hold the potential to reduce the magnitude of the resulting distortion.

The results of this paper demonstrate that decisions by professional interviewers can be distorted by the evaluation of candidates against an arbitrary benchmark. Despite the critical importance of interviews in labor market matching, the underlying decision process largely remains a 'black box.' Most related, Simonsohn and Gino (2013) find that the likelihood of admission into an MBA program decreases with the proportion of candidates admitted by the interviewer on the same day, attributing this to daily narrow bracketing. Conversely, our analysis focuses on comparisons between candidates based on their exact position in the interview sequence. Our findings reveal quantitatively important contrast effects, which imply that even minor changes in candidate ordering can have a major impact on the selection outcome.[5] This result also complements the study by Hoffman et al. (2018), indicating that job-testing technologies outperform HR managers in selecting candidates for low-skilled jobs.[6] While many organizations have begun to implement job-testing technologies, interviews remain central to most candidate selection processes. Therefore, an empirical understanding of human assessments is key to enhance the validity of hiring and admission decisions.

Our findings also contribute to the literature on negative path dependence in decision-making (see Appendix A for a detailed overview). Initial evidence of contrast effects comes from laboratory experiments (e.g., Pepitone & DiNubile, 1976; Wexley et al., 1972; Kenrick & Gutierres, 1980). Existing field studies have used data on rental choices (Simonsohn & Loewenstein, 2006; Simonsohn, 2006; Bordalo et al., 2019), a speed dating field experiment (Bhargava & Fisman, 2014) and financial market prices (Hartzmark & Shue, 2018). Chen et al. (2016) document a negative autocorrelation in the decisions of asylum judges, loan officers and baseball umpires, which they attribute to a gambler's fallacy while remaining open towards contrast effects as an alternative explanation. More generally, there is increasing evidence that individuals overreact to recent experiences. Singh (2021) finds that physicians change the mode of delivery in response to complications in the previous case, Jin et al. (2023) document a posi-

[5] Another key distinction between our study and Simonsohn and Gino (2013) is the scale and structure of the data sources. While their data encompass 31 evaluators conducting ≈ 9,000 interviews, our two datasets include ≈3,000 evaluators from two distinct processes, conducting a total of ≈ 37,000 interviews.

[6] Additional studies on the effect of technology-based candidate screening include Autor and Scarborough (2008), Horton (2017), Estrada (2019), and Bergman et al. (2020).

tive autocorrelation in physician decisions, and Bhuller and Sigstad (2023) show that judges change their sentencing behavior in response to recent reversals of their decisions. In this study, we provide evidence that sequential contrast effects produce significant distortions in labor market decisions with high stakes, even when individuals have the opportunity to correct their initial assessments ex post. Moreover, our findings offer new insights into the influence of the decision environment, the role of memory and the potential for policy interventions by firms and organizations.

More broadly, this paper relates to field evidence on reference-dependent decision-making (for an overview, see Donoghue & Sprenger, 2018), and backward-looking, adaptive reference points in particular (e.g., Thakral & Tô, 2021; DellaVigna et al., 2022). Our results provide evidence that evaluators use recent and similar candidates as a reference when forming an assessment. Memory-based models of economic decision-making conceptualize how past experiences influence economic decisions (e.g., Mullainathan, 2002; Bordalo et al., 2020; Wachter & Kahana, 2023).[7] We provide field evidence that this concept helps to understand real-world decision-making and the formation of backward-looking reference points in particular.

2 Institutional Settings

Our analysis is based on data from two distinct interview processes with high stakes. In the following, we provide information on these processes and the corresponding data sources. Table 1 provides an overview of their main features.

2.1 Setting 1: Study Grant Admission Process

Our primary data source stems from the admission process of a large, merit-based study grant program for university students in Germany.

[7] Several lab studies conceptualize and test the role of memory for beliefs and expectations (e.g., Enke et al., 2020; Bordalo et al., 2021; Afrouzi et al., 2023).

Background The grant is government-funded and has the reputation of being highly competitive. It offers a variety of monetary and non-monetary benefits. Specifically, recipients receive a generous monthly stipend and have the opportunity to participate in a large, cost-free course program that includes language classes, summer schools and career workshops. Additional benefits include a high signaling value and access to a network of high-ability peers and alumni. Appendix B provides further information on the program.

The admission process is organized through two-day workshops. Each workshop comprises about 48 candidates, all of whom are first-year university students pre-selected as the top 2.5% of their high school's graduation cohort. There are eight evaluators per workshop, who are mostly alumni of the study grant program. They work in different professions and typically participate in an admission workshop every one or two years. About half of the evaluators have undergone a two-day interviewer training program. A workshop organizer from the study grant foundation is constantly present to lead and moderate the workshop.

Interview Process Candidates undergo two one-to-one interviews and participate in a group discussion round. Each of these three assessments is made independently by a different evaluator. The assignment of candidates to evaluators and the assignment of time slots are quasi-randomized within workshops, conditional on gender.[8] Both candidates and evaluators are quasi-randomly assigned an ID. A fixed schedule then matches candidate IDs to evaluator IDs and time slots (see Appendix Figure B.2).

Evaluators arrive at the workshop on Friday evening and first receive a briefing by the workshop organizer. The briefing informs about the workshop procedures and reminds evaluators of the admission criteria. On Saturday and Sunday, evaluators conduct six one-to-one interviews per day, which they prepare the evening before based on the candidates' CV, school records and letters of recommendation. Between interviews, evaluators also assess six group discussions. In these discussions, a candidate gives a brief presentation on a self-chosen topic and moderates the subsequent discussion, while evaluators serve as passive observers.

[8] The randomization conditional on gender aims to gender-balance the group discussions.

Table 1: Comparison of Settings and Datasets

	Admission Process	**Hiring Process**
Sample size	29,466	8,423
Interviews per sequence	12	3
Assessment	Rating (Scale 1-10)	Rating (Scale 1-3) + sub-scores
Assessments to decision	Cut-off rule (+discussion)	Committee discussion

Assessment and Admission Decision Our study focuses on one-to-one interviews. Evaluators assess candidates according to their intellectual ability, ambition and motivation, communication skills, social engagement, and breadth of interests. The assessment is summarized on a rating scale from one to ten. A rating of eight or higher is considered a 'yes' vote for the candidate's admission. A candidate is accepted upon a minimum of two yes votes and a total of 23 points. There is no admission quota at the workshop level, giving the committee the flexibility to admit any number of candidates. Evaluators are instructed to finalize their assessments after interviewing all assigned candidates. A common practice is to make provisional ratings after each interview and potentially adjust them ex post. To maintain the independence of each candidate's three assessments, evaluators do not discuss individual candidates prior to the final committee meeting. In this meeting, held on Sunday afternoon, the individual ratings are aggregated.[9] Candidates above the threshold are admitted after a brief justification from the evaluators involved. Ratings of candidates at the margin of admission can be adjusted following a committee discussion.[10]

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

[9] A list of candidate IDs is read out aloud and the three evaluators who have assessed the respective candidate report their ratings. In this process, it is not easily possible to trace the behavior of other evaluators, as the assessments are collected at high frequency and not ordered by the evaluator's IDs.

[10] Such adjustments typically affect about two to three out of around 150 votes per workshop. We observe the final ratings of each candidate. To test whether the adjustment procedure influences our results, we perform robustness checks that exclude marginal candidates from the sample.

[11] There are 1,724 unique evaluators. In the main analysis, we treat every evaluator-workshop observation as independent. The average evaluator participates in about 1.8 workshops in the sample.

slots, as well as the resulting ratings and admission decision. In addition, the data report the candidate's gender, age, study major, high-school GPA, an indicator of migration background, and an indicator of being a first-generation student. Observed evaluator characteristics include gender, study major, age, and prior workshop experience.

2.2 Setting 2: Hiring Process

The second data set covers interviews conducted within the hiring process of a large consulting company.

Background Candidates in the data apply for permanent positions ($\approx$ 65%) or internships ($\approx$ 35%) at the German-speaking branch of the consultancy. The hiring process is highly competitive. It has high stakes for both the company, whose success builds on the human capital of its employees, and the candidates, who are applying to high-earning jobs with starting wages in the top 10% of the overall German wage distribution. An employment spell at the company is often a stepping stone to top management positions at other firms. Candidates for internships are university students, and candidates for permanent positions are mostly recent graduates. Prior to the interview stage, candidates have been pre-selected by the HR department based on their written application. Evaluators are consultants at the company, who have all gone through professional interviewer training and conduct interviews on a regular basis throughout the year.

Interview Process The process is organized through interview days at different locations, with a varying number of candidates and evaluators. The median interview day in our data includes eight candidates and eight evaluators. Typically, candidates have three independent one-to-one interviews, and evaluators interview three candidates per interview day. The assignment of candidates to interview days and evaluators as well as the allocation of time slots is exogenously determined by the HR department. The pool of candidates that can be assigned to an evaluator at a given time slot is defined by the location of the interview, the application

time, and the type of position (internship versus permanent). Furthermore, the HR department takes into account the gender of the candidates as it tries to ensure that each female candidate is interviewed by one female evaluator. Therefore, we consider the assignment process to be quasi-random within position × year × location cells, conditional on candidate gender.

Assessment and Hiring Decision The company's assessment process is highly standardized. Evaluators give sub-ratings on several dimensions of cognitive and non-cognitive ability. The cognitive dimensions have a focus on mathematical and analytical skills, while the non-cognitive dimensions are related to leadership and teamwork skills. Evaluators summarize their assessments in an overall rating on a three-point scale. A rating of three points expresses the recommendation to hire a candidate.

Evaluators enter their assessments in the applicant tracking system after every interview or after their last interview, without any explicit encouragement to re-adjust ratings after the last interview. There is no discussion of candidates during the interviewing phase. After all interviews have been conducted, hiring decisions are made at a final committee meeting. There are no fixed cut-off rules regarding the translation of ratings into hiring decisions. Moreover, committees do not face a quota at the level of the interview day, since the company hires consultants on a rolling basis.

Data Source The data cover all interviews for internships and permanent positions from January 2017 to April 2022.[12] They contain 8,423 interviews conducted by 357 distinct evaluators with 3,308 candidates on 461 interview days. We observe the assessment outcome of each interview, as well as the final hiring outcome of each candidate. The data allow reconstructing the order (but not the time stamp) of the interviews. Moreover, they report candidates' gender, study field, high-school GPA and aspired type of position (internship vs. permanent). Observed evaluator characteristics include gender, managerial responsibility, and interview experience.

[12] We drop 48 observations due to missing information on assessments and 718 observations due to missing information on the ordering of candidates within a sequence. 654 observations are excluded because the evaluator conducted only one interview on the given interview day.

3 Data

In this section, we provide descriptive statistics on both data sources, explain our baseline measure of candidate quality, and perform randomization checks.

3.1 Descriptive Statistics

Figure 1 plots the sample distribution of interview ratings in the two processes. In the admission process (Panel a), ratings range from 1 to 10, and the average rating is 6.6, with a standard deviation of 1.8. About 37% of the interviews result in a rating of 8 points or more, implying a vote in favor of admission. In the hiring process (Panel b), about 30% of interviews result in a rating of 3 points, corresponding to a recommendation to hire the candidate.

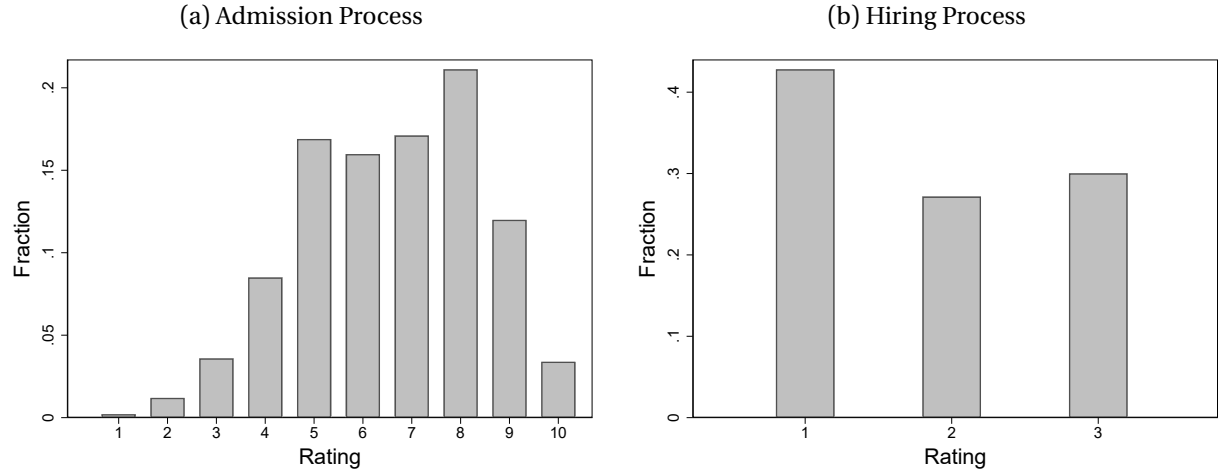
Appendix C provides additional summary statistics on interview outcomes. Figures C.1 and C.2 document substantial heterogeneity in the share of positive assessments per interview sequence and in the share of accepted candidates per workshop or interview day. The average workshop has an admission rate of 0.25 (SD: 0.07), while the average interview day has a job offer rate of 0.29 (SD: 0.17). Tables C.1 and C.2 report summary statistics on the characteristics of candidates and evaluators in the two processes.

3.2 Measurement of Candidate Quality Through Third-Party Assessments

Our aim is to analyze how a candidate's assessment changes when the quality of another candidate in the same interview sequence increases. In the context that we study, "quality" describes how well a candidate meets the respective admission or hiring criteria. True candidate quality is unobserved by design, otherwise conducting interviews would be unnecessary. Therefore, any quality measure must be thought of as an approximation.

Our preferred approximation is based on the third-party assessment (TPA) of a candidate's quality. We specify TPA as the average of the candidate's other two ratings, which were made

Figure 1: Distribution of Interview Ratings



Note: Panel (a) shows the distribution of interview ratings in the study grant program (N=29,466). A rating of ≥ 8 points expresses a yes vote. Panel (b) shows the distribution of interview ratings in the hiring process (N=8,423). A rating of 3 points expresses a recommendation to hire the candidate.

independently by different evaluators based on another interview or a group discussion.[13] The rationale for using TPA as a quality measure is twofold. First, all evaluators use the same criteria of candidate quality. This results in a strong correlation between ratings, despite the fact that evaluators differ in their leniency and see the same candidate in different contexts. The correlation between ratings and TPA is about 0.36 in the admission process and 0.25 in the hiring process (see Appendix Table C.3).[14] Second, while all evaluators measure the selection criteria with noise, their individual noise terms are independent of one another. Crucially, when two evaluators assess the same candidate, they are influenced by different sets of

[13] An alternative approach to measure candidate quality is based on predetermined characteristics, such as GPA. However, GPA is a weak predictor of assessments for two main reasons: first, candidates are pre-selected on having a strong GPA, which strongly limits the amount of variation in GPA in the sample; and second, selection criteria place equal weight on cognitive and social skills, which further reduces the relevance of GPA. Table C.3 illustrates that there is a positive but weak correlation of ratings with GPA in both processes. TPA exhibits an up to ten times stronger correlation and explains significantly more variation in the data. Nevertheless, we complement the main results with robustness checks where quality is predicted based on predetermined candidate characteristics (including GPA).

[14] For comparison, Card et al. (2019) document a correlation of about 0.25 between two referee reports of the same paper in four leading journals in economics.

other candidates, and different previous candidates in particular.[15] Moreover, both processes preclude any discussion of candidates before the final committee meeting (see Section 2 for details).[16] In Appendix Tables C.4 and C.5, we empirically assess a direct implication of the independence assumption. The idea is that we expect an evaluator's characteristics to correlate with her rating of a candidate. For instance, female evaluators give higher average ratings in both processes. Conversely, evaluator characteristics should not correlate with the candidate's TPA, i.e., with the other two evaluators' average assessment of the same candidate. In line with this intuition, the tables show that a candidate's rating — but not her TPA — correlates with the characteristics of the evaluator who made the rating. Additional evidence of the independence assumption will be provided with the randomization checks (Section 3.3), showing that the TPA measures of candidates within the same interview sequence are uncorrelated.

3.3 Randomization Checks

Our analysis relies on the assumption that candidates are as good as randomly assigned to and ordered within interview sequences, conditional on gender and randomization units (i.e., admission workshops or candidate pools).

Table 2 reports results from two randomization checks for each of the two assumptions. In Panel A, we test for a relationship between an individual's quality and the leave-one-out mean quality of the other candidates assigned to the same evaluator, using TPA measures as well as predictions based on observed characteristics. Similar to studies in the peer effects literature, it is necessary to correct for a bias arising from a mechanical negative correlation of candidate quality within randomization units. Intuitively, a candidate cannot be assigned to herself, implying that her quality will be negatively correlated with the quality of her potential 'peers' in the presence of fixed effects for the unit of randomization. A first approach to correct for this

[15] The sets of candidates seen by two evaluators never overlap in the hiring process and almost never in the admission process (see Appendix B.2). In both processes, two evaluators never see the same pair of candidates in the same order.

[16] One incidence where an evaluator changes her rating following the arguments of another evaluator is the discussion of marginal candidates in the final committee meeting of the study grant program (see Section 2). We will show that the results are robust to excluding marginal candidates.

Table 2: Assessment of Quasi-Random Assignment & Ordering

	Admission Process		Hiring Process	
	(1) Std. TPA	(2) Std. Predicted Rating	(3) Std. TPA	(4) Std. Predicted Rating
Panel A: Quasi-Random Assignment				
Guryan et al. (2009)				
Leave-one-out mean	0.002*	-0.001	-0.012	-0.005
	(0.001)	(0.001)	(0.027)	(0.028)
R^2 (within)	0.998	0.998	0.710	0.707
Jochmans (2023)				
test statistic	0.695	-0.048	0.710	1.037
p-value	0.487	0.962	0.478	0.300
Panel B: Quasi-Random Ordering				
Guryan et al. (2009)				
Lag (t-1)	0.000	0.002	0.024	0.013
	(0.006)	(0.006)	(0.017)	(0.023)
R^2 (within)	0.009	0.024	0.002	0.000
Jochmans (2023)				
test statistic	0.915	0.960	1.426	0.967
p-value	0.360	0.337	0.154	0.333
N	26970	26970	5165	5165

Note: TPA = third-party assessment of candidate quality (see section 3.2 for details). Panel A presents tests for a relationship between an individual's quality and the leave-one-out mean quality of the other candidates assigned to the same interview sequence. The test proposed by Guryan et al. (2009) controls for the leave-one-out mean quality at the workshop or candidate pool level. This test has limited power in the admission process (Columns 1 & 2) due to limited variation in the size of workshops. Therefore, we additionally provide test statistics and p-values from an alternative bias-corrected test for random peer assignment developed by Jochmans (2023), which does not require variation in the size of randomization units. In Panel B, we test for a relationship between the quality of the current and the previous candidate, conditional on the leave-one-out mean quality at the sequence level. All regressions control for gender and workshop/candidate pool fixed effects. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). $^{*}\ p < 0.10,^{**}\ p < 0.05,^{***}\ p < 0.01$.

exclusion bias was proposed by Guryan et al. (2009), who suggest controlling for the quality of the other candidates in the randomization unit (leave-one-out mean). This test performs well when there is sufficient variation in the size of randomization units. As revealed by the high R^2-values in Columns 1 and 2, this condition fails to hold in the admission process. In the hiring process, where candidate pools exhibit more variation in size, the test is better pow-

ered and shows no indication of candidate sorting by quality. The table additionally reports test statistics and p-values from an alternative bias-corrected test by Jochmans (2023), which does not require variation in the size of randomization units. In both processes, the test results do not reject the hypothesis of quasi-random assignment. As further evidence of random assignment, Appendix Tables C.6 and C.7 show that candidate characteristics are unrelated to the characteristics of assigned evaluators.

In Panel B, we assess the quasi-random ordering of candidates within sequences by testing for a relationship between the current and the previous candidate's measured quality. We now control for exclusion bias using the sequence-level leave-one-out mean quality, as candidates in the same sequence define the pool of potential previous candidates. None of the estimates suggests that candidates are systematically ordered with respect to their quality. Test statistics based on Jochmans (2023) are equally in line with the hypothesis of quasi-random ordering. Section 4 provides placebo checks that further support the assumption of quasi-random ordering.

4 Empirical Analysis

In this section, we provide empirical evidence on the interdependence of candidate assessments within interview sequences. In Section 4.1, we analyze how a candidate's assessment changes if another candidate's measured quality increases, depending on the relative position of her interview. In section 4.2, we estimate the autocorrelation in admission votes and hiring recommendations. Section 4.3 quantifies the effect on final admission and hiring decisions.[17]

[17] The analyses in this section are pre-registered for the study grant data. 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 data for the 2012/13 academic year. This "pilot" data is no longer contained in the estimation sample. When analyzing the hiring data, we stick to the same pre-registered specifications, unless we need to adapt them due to the slightly different institutional setup.

4.1 Influence of the Interview Sequence

4.1.1 Econometric Specification

In the following, we first describe the (pre-registered) main specification, which we apply to the admissions data. We then outline how we adjust the specification to the hiring data.

Main Specification (Admission Process) We use the following regression model to estimate how the assessment of a candidate interviewed in period t is affected by the measured quality of the candidate interviewed in another period $t+k$:

$$Y_{i,t} = \beta_k \, \text{TPA}_{i,t+k} + \gamma \, \text{TPA}_{i,t} + \pi_k \, \overline{\text{TPA}}_{i,-\{t,t+k\}} + X'_{i,t} \, \sigma + \eta_w + \epsilon_{i,t} \tag{1}$$

The outcome variable $Y_{i,t}$ is the standardized rating made by the evaluator i of the candidate interviewed in period t. $\text{TPA}_{i,t+k}, k \in \{-11, ..., -1, 1, ..., 11\}$, is the standardized third-party assessment of the candidate interviewed by evaluator i at time $t+k$ (see Section 3.2 for details). The coefficient of interest, β_k, measures the influence of $\text{TPA}_{i,t+k}$ on the rating of the candidate interviewed in t.

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

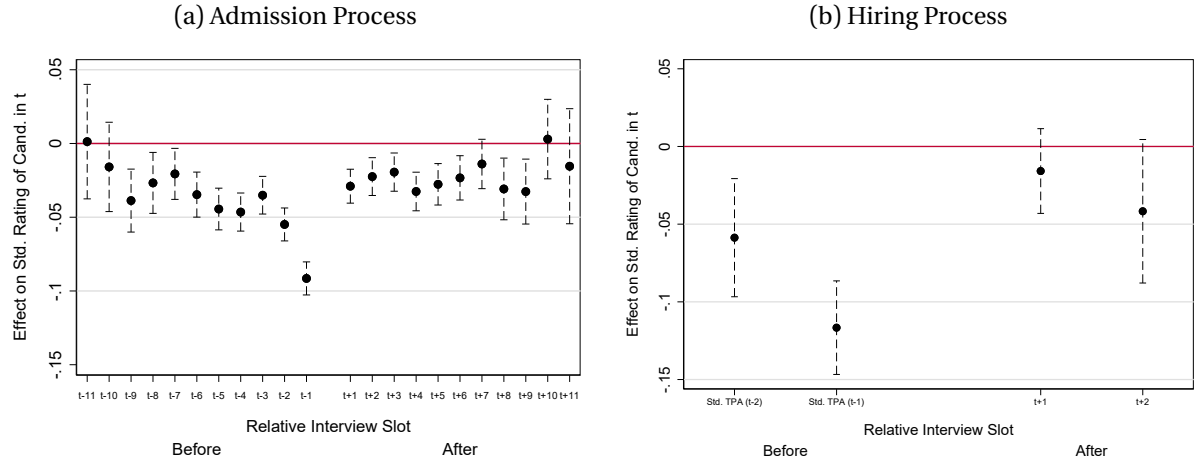
For each value of $k, k \in \{-11, ..., -1, 1, ..., 11\}$, we perform a separate estimation of equation 1, including all candidates for whom period $t+k$ exists. This allows us to use all available data for each value of k, but means that estimates for different values of k are partially based on different interview slots. As robustness checks, we additionally estimate single regressions with a subset of leads and lags.

Adjustments to Hiring Process We estimate the same specification for the hiring process, with the following setup-specific adjustments: first, k only takes values from -2 to +2, as the typical interview sequence includes three interviews. Second, due to these shorter sequences, we do not control for the leave-two-out mean $\overline{\text{TPA}}_{i,-\{t,t+k\}}$. Third, we replace the workshop fixed effects with candidate pool (i.e., year×location×position type) fixed effects and cluster standard errors at that level (N=63). The vector $X_{i,t}$ includes candidate and evaluator characteristics (Table C.2), order indicators and quarter fixed effects. As above, we first estimate a separate regression for each value of k using all available data. In a robustness check, we estimate the influence of the previous two candidates in a single regression, based on the sample of all third interviews.

4.1.2 Results

Admission Process Figure 2 (a) plots the estimates of β_k from equation 1. Appendix Table D.1 reports the corresponding coefficients and p-values (including Bonferroni adjustments). We make three main observations. First, the rating of a candidate decreases in the measured quality of the other candidates seen by the same evaluator. Second, both candidates interviewed before t ($k < 0$) and candidates interviewed afterwards ($k > 0$) have an influence, suggesting that evaluators adjust their ratings after having seen everyone. Third, the influence of the previous candidate strikingly stands out, being about three times stronger than that of the average other candidate in the sequence. As shown in Table D.3 (Panel A), a one standard deviation increase in the previous candidate's quality measure is about 25% as influential as a one standard deviation decrease in a candidate's own quality measure. Moreover, the effect compares to the influence of a one standard deviation change in the other candidates' average TPA, i.e., the sequence leave-two-out mean TPA. Appendix Figures D.2 (a) and (b) show that the previous candidate's influence is not an artifact of sampling, as it persists when we estimate the influence of other candidates in a single regression, using a homogeneous subsample of interview slots. Appendix Figure D.3 provides evidence that the overall negative influence of the other candidates can be captured by controlling for the average quality of the sequence (leave-

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



Note: Estimates are based on equation 1. The coefficients measure how the standardized TPA of the candidate interviewed in $t+k$ affects the standardized overall rating of the candidate in t. TPA = third-party assessment of candidate quality (see section 3.2 for details). Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). Appendix Tables D.1 and D.2 report the corresponding coefficients and p-values.

one-out mean TPA). Taken together, the results document two separate effects: an influence of the other candidates' average quality and an additional influence of recently observed quality.

Hiring Process Figure 2 (b) and the corresponding Appendix Table D.2 provide evidence that the influence of the previous candidate also stands out in the hiring process, where the evaluators are trained to conduct structured interviews and do so on a regular basis. We observe a strong relationship between the previous candidate's TPA and the current candidate's rating, which exceeds the influence of the other candidates in the sequence. The influence of the previous candidate's TPA is about half as strong as the influence of a candidate's own TPA (see Panel A of Table D.3). As shown in Appendix Figure D.2 (c), this result is robust to estimating the influence of the previous two candidates in a single regression.

Placebo and Robustness Checks Appendix D includes several placebo and robustness checks for both data sets. Appendix Tables D.1 and D.2 and Figure D.1 report results from a bootstrap procedure where we reshuffle the order of interviews in each sequence and estimate a dis-

tribution of placebo coefficients (see Appendix D for technical details). Appendix Figure D.4 shows the results of using TPA_t as an outcome, documenting the absence of a conditional correlation between TPA_t and TPA_{t+k} throughout the interview sequence.

Appendix Table D.3 reports the effects of previous, own and leave-two-out mean quality (estimated with and without control variables), and their robustness to changes in the sampling and estimation procedure. In particular, the results are robust to the exclusion of marginal candidates in the admission data and the exclusion of interview sequences with only two candidates in the hiring data. Moreover, regressions with interviewer and candidate fixed effects yield very similar estimates. Table D.4 documents the results' robustness to using different measures of candidate quality, including a prediction based on observable characteristics. It shows that the estimated relative importance of own versus previous quality is robust across quality measures, ranging from 0.18 to 0.28 in the admission process and from 0.42 to 0.53 in the hiring process. The same holds true when using an instrumental variable strategy, where one quality measure serves as an instrument for the other (Table D.5).

4.2 Autocorrelation in Evaluator Decisions

This section complements the causal evidence on the influence of the previous candidate with an estimate of the autocorrelation in binary admission votes and hiring recommendations. The appeal of the autocorrelation is that it directly reflects the evaluator's own perception of candidates, as opposed to the assessment of a third party. A potential drawback is that the autocorrelation may also contain the current candidate's influence on the previous candidate, due to the possibility of ex-post corrections. However, the previous analysis revealed that only the previous — and not the next — candidate has an influence that extends beyond contributing to the average quality of the interview sequence.

4.2.1 Econometric Specification

We estimate the autocorrelation using the following specification:

Table 3: Autocorrelation in Evaluator Decisions

	Admission Process			Hiring Process	
	(1) Yes (t)	(2) Yes (t)	(3) Rank(t)	(4) Yes (t)	(5) Yes (t)
Yes (t-1)	-0.056***	-0.057***	-0.406***	-0.127***	-0.131***
	(0.006)	(0.006)	(0.042)	(0.013)	(0.013)
Controls	No	Yes	Yes	No	Yes
Outcome Mean	0.37	0.37	6.43	0.31	0.31
N	26970	26970	26970	5165	5165

Note: Estimates are based on equation 2. In the admission process, "Yes" describes a vote in favor of admitting the candidate. In the hiring process, "Yes" describes a recommendation to hire the candidate. All regressions include workshop (Columns 1-3) or candidate pool (Columns 4-5) fixed effects, as well as the evaluator's leave-one-out mean decision. Controls include candidate characteristics, evaluator characteristics, and interview order. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). $^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$.

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

$Y_{i,t}$ denotes evaluator i's binary decision (admission vote or hiring recommendation) on the candidate in t. $Y_{i,t-1}$ denotes evaluator i's decision on the candidate in $t-1$. To control for evaluator leniency and the average strength of the other candidates, we include the evaluator's leave-one-out mean decision and rating, excluding the candidate in t ($\overline{Y}_{i,-t}$). In the admission process, $\overline{Y}_{i,-t}$ is computed at the level of the evaluator's interview sequence. In the hiring process, where the sequence includes at most three candidates, $\overline{Y}_{i,-t}$ is computed over all interviews conducted by the evaluator in the same year. As before, the specification controls for evaluator and candidate characteristics ($X_{i,t}$) and includes workshop/candidate pool fixed effects.

4.2.2 Results

Table 3 reports the estimates of the autocorrelation in evaluator decisions for both datasets.

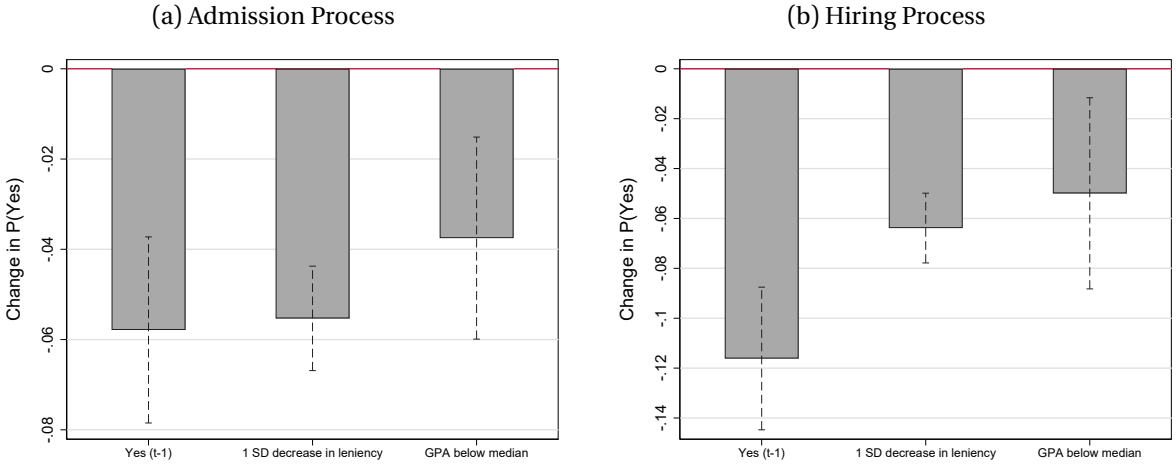
Admission Process Columns 1 (without controls) and 2 (with controls) show that the probability of receiving a yes vote decreases by about 6 percentage points (15% relative to the mean) if the previous candidate receives a yes vote. As reported in Column 3, candidates who follow a candidate with a yes vote move down by about 0.4 ranks on average in the evaluator's distribution of ratings given to the candidates in the sequence.

Hiring Process Turning to the hiring process, Columns 5 and 6 show that the evaluator's decisions exhibit a negative autocorrelation of about 12.5 percentage points (40% relative to the mean). On average, this estimate strongly exceeds the estimated autocorrelation in the admission process. Additional analyses in Section 5 (Figure 4) will show that this difference can be explained by the different lengths of interview sequences.

Comparison to Other Determinants Figure 3 illustrates a comparison of the autocorrelation to the influence of candidate GPA and evaluator leniency, measured as the share of yes votes given to candidates in prior interview sequences. In the admission (hiring) process, the absolute size of the autocorrelation roughly corresponds to the influence of a one (two) standard deviation change in evaluator leniency. In both settings, the autocorrelation is about 30% larger than the coefficient on a median split of candidate GPA.

Robustness and Additional Analyses Appendix E contains several robustness checks and additional results. Table E.1 documents that the estimated autocorrelation is robust to the inclusion of candidate fixed effects. Coefficients become more negative after the introduction of evaluator fixed effects — in line with a downward bias in autoregressive models estimated on finite panels (Nickell, 1981). Figure E.1 shows that the size of the autocorrelation strongly weakens beyond t-1. Finally, Figure E.2 reports the results from a back-of-the-envelope calculation regarding the share of evaluator decisions that are reversed due to the autocorrelation.

Figure 3: Influence of Previous Decision, Evaluator Leniency and Candidate GPA



Note: Regressions only include evaluators who have conducted at least five interviews in the past. Leniency describes the share of yes votes given to candidates in past interview sequences. Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

4.3 Impact on Admission and Hiring Outcomes

Having identified a strong influence of the previous candidate on the single interview assessment, we now estimate the impact on final admission or hiring decisions. In both settings, every candidate receives three independent assessments, two of which can be influenced by a previous candidate.[18] Columns 1 and 3 of Table 4 report how the average measured quality (TPA) of the two preceding candidates affects the final admission or hiring probability. We find that a one standard deviation increase in the average TPA of the two preceding candidates reduces the probability of admission by about 2.8 percentage points and the hiring probability by about 3.2 percentage points. In both processes, the effect roughly corresponds to a 10% change relative to the outcome mean.

Columns 2 and 4 report how the number of yes votes given to the two previous candidates affect the final outcomes. Estimates show that an additional yes vote given to one of the previous candidates reduces the admission probability by about 4.3 percentage points and the

[18] In the admission process, every candidate receives two interview assessments and an additional assessment based on a group discussion (see Section 2 for details). In the hiring process, every candidate has three interviews, two of which are preceded by another candidate (every candidate is once first in the sequence).

Table 4: Joint Impact of Previous Candidates on Final Admission and Hiring Outcome

	Admission Probability		Hiring Probability	
	(1)	(2)	(3)	(4)
Average TPA of Previous Candidates (Std.)	-0.028***		-0.037***	
	(0.004)		(0.010)	
No of Previous Candidates w/ Yes		-0.043***		-0.069***
		(0.006)		(0.018)
Outcome Mean	0.25	0.25	0.29	0.29
N	12237	12237	1925	1925

Note: The level of observation is the candidate. TPA = third-party assessment of candidate quality (see Section 3.2 for details). In both processes, every candidate receives three independent assessments, two of which can be influenced by a previous candidate. Therefore, the average TPA is based on two previous candidates, and the number of previous candidates with a yes vote ranges from 0 to 2. All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects. Controls include candidate characteristics. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

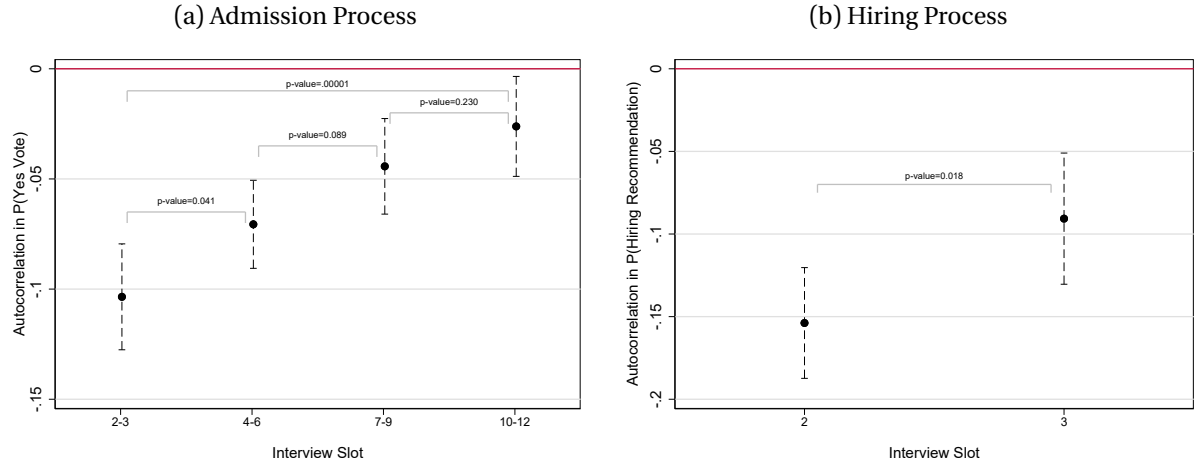
hiring probability by about 6.3 percentage points (≈ 20% relative to the mean). Overall, these estimates document that the influence of the previous candidate on individual interview assessments leads to quantitatively meaningful changes in final decisions with high stakes for both candidates and organizations.

5 The Role of Prior Experiences and Similarity

The results presented so far have demonstrated that the quality of the previous candidate has a large average effect on interview outcomes. From the perspective of firms and organizations, it is important to understand the conditions under which this influence is more or less pronounced. In this section, we investigate the role of the evaluators' prior experiences and of similarity between interviews. Beyond offering insights for organizational design, these analyses will also inform the discussion of the behavioral mechanism in Section 6.

Experience Within the Interview Sequence Over the course of the interview sequence, evaluators experience an increasing number of candidates. In Figure 4, we analyze how the influence of the previous candidate evolves over the sequence. In both settings, we find strong

Figure 4: Experience Within the Interview Sequence



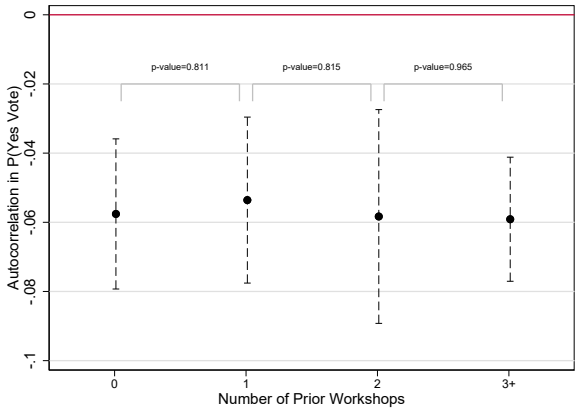
Note: The figure shows estimates of the autocorrelation based on equation 2, interacting the prior candidate's yes vote/hiring recommendation with the slot of the current interview. Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

evidence that the previous candidate's influence decreases while evaluators collect more interview experiences. In the admission process (Panel a), the autocorrelation weakens from about 10 percentage points in slots 2-3 to about 3 percentage points in slots 10-12. In the hiring process (Panel b), where sequences only include three candidates, it amounts to about 15 percentage points in the second slot and decreases to 9 percentage points in the third slot. This heterogeneity also reconciles differences in the average autocorrelation between the two processes (see Table 3). Notably, the average autocorrelation in the hiring process is roughly equivalent to the autocorrelation in the first three admission interviews of a given sequence.

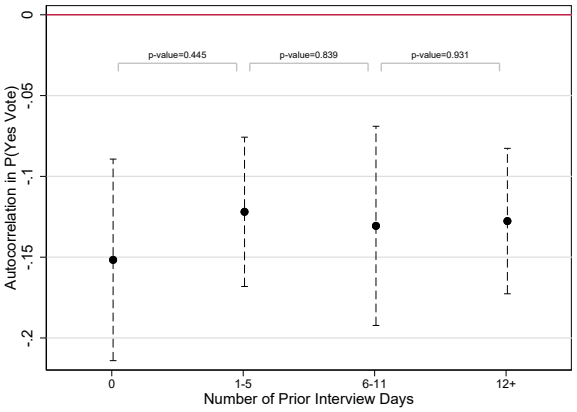
Experience Prior to the Interview Sequence Given the large role of within-sequence experience, a natural question is whether background experience acquired in prior sequences also mitigates the previous candidate's influence. Figure 5 illustrates that this is not the case. In both processes, the autocorrelation does not vary with the number of interview days or workshops that an evaluator has experienced. Two additional findings support the notion that past experiences do not matter for evaluations in the current sequence. First, Appendix Table F.1 shows that the average quality of candidates seen during a workshop in the previous academic year

Figure 5: Experience Prior to the Interview Sequence

(a) Admission Process



(b) Hiring Process

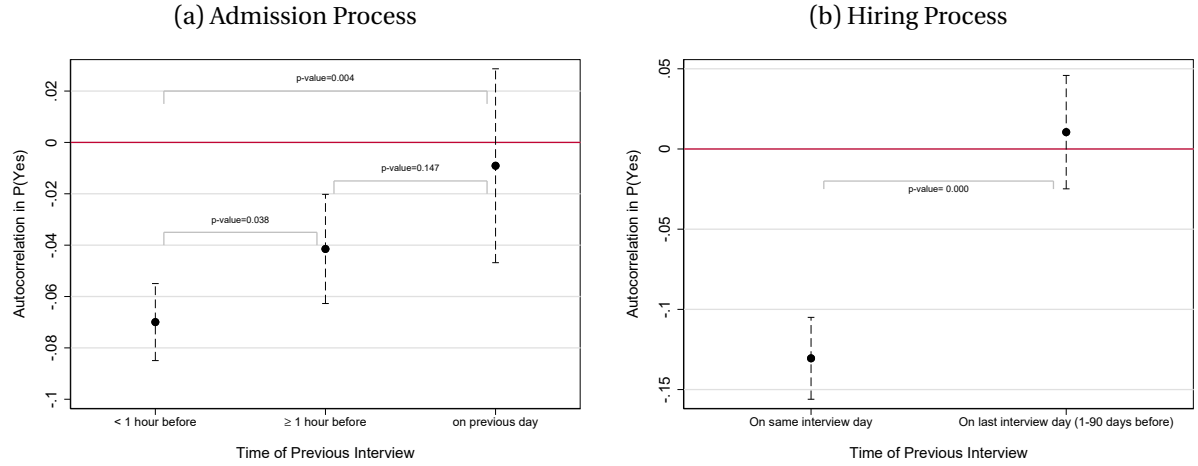


Note: The figure shows estimates of the autocorrelation based on equation 2, interacting the prior candidate's yes vote/hiring recommendation with the evaluator's number of past workshops/interview days. Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

(admission process) or during the last 365 days (hiring process) does not affect current ratings. Second, Appendix Table F.2 reports that the autocorrelation does not decrease with additional interviewer training, age, or managerial responsibility. This suggests that more background knowledge about (expected) candidate quality and the selection criteria do not mitigate the previous candidate's influence.

Time Distance Between Interviews We now study how the autocorrelation varies with the time distance between t and t-1. The results in Figure 6 (a) suggest that longer breaks weaken the autocorrelation in admission votes. The autocorrelation roughly decreases by half when there is an hour or more between two interviews, and approaches zero after a day change. In the hiring process, we do not observe the time gap between interviews on the same day. However, we can assess whether the first interview on a given interview day is influenced by the last interview on the previous interview day (within a range of 90 days). As shown in Panel (b), this is not the case. The data thus offer consistent evidence that only recent interview experiences matter and that the influence of prior interview experiences decreases with elapsed time.

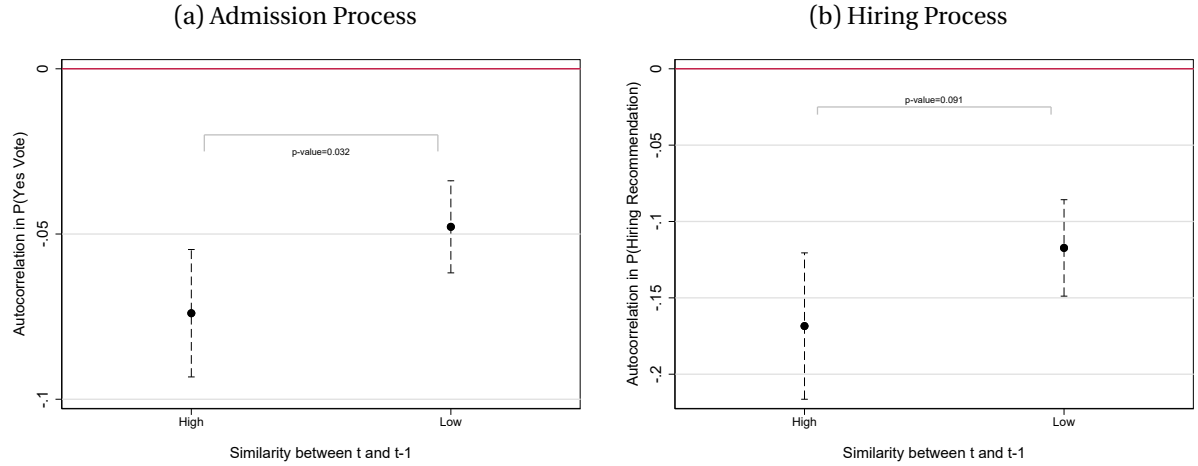
Figure 6: Time Between Interviews



Note: Panel (a) plots estimates of the autocorrelation in yes votes based on equation 2, interacting the prior candidate's yes vote with the time gap between the end of the interview in t-1 and the start of the interview in t. Panel (b) shows the autocorrelation in hiring recommendations for same-day interviews and the correlation between the recommendation given to the first candidate on a given interview day and the recommendation given to the last candidate on the last interview day. Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

Similarity Between Candidates The previous analyses focused on the role of time for the influence of previous interview experiences. We now assess whether the observable similarity of subsequent candidates matters. More specifically, we analyze how the autocorrelation differs depending on the similarity of two subsequent candidates in terms of their observed characteristics. In the study grant data, we construct a simple index, which is defined as the number of characteristics shared between the current and previous candidate (including gender, migration status, first-generation status, and study field). We interact a median split of this index with the vote of the previous candidate. Panel (a) of Figure 7 shows the result, revealing that the autocorrelation is significantly stronger when two subsequent candidates share more characteristics. In the hiring data (Panel b), we only observe gender and study field as relevant candidate characteristics. The results suggest that similarity along these dimensions also strengthens the influence of the previous candidate.

Figure 7: Observable Similarity of Candidates



Note: The figure shows estimates of the autocorrelation based on equation 2, interacting the prior candidate's yes vote/hiring recommendation with a median split of a similarity index, defined as the number of observable characteristics that the candidate in t and the candidate in t-1 have in common (gender, migration status, first-generation status and study field in Panel a; gender and study field in Panel b). Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

6 Behavioral Mechanism

The empirical results have documented two distinct effects: first, the individual assessment decreases in the average quality of the other candidates in the sequence; and second, the previous candidate's quality has a strong additional negative influence. There are several straightforward ways to explain the influence of the other candidates' average quality, such as learning about an uncertain evaluation threshold or an implicit target on the number of yes votes. The fact that both previous and subsequent candidates have an influence (Figure 2) suggests that this effect occurs after all candidates have been interviewed.

In this section, we discuss mechanisms that can explain the strong additional influence of the previous candidate and its heterogeneity. We first consider a contrast effect model where candidates are evaluated against a benchmark formed through the associative recall of prior interviews. We provide evidence that such a framework can explain the reduced-form findings and yields a good quantitative fit with the data. We then consider sequential learning and a gambler's fallacy as alternative explanations.

6.1 Contrast Effect with Associative Recall

Evaluators exhibit a contrast effect if they evaluate candidates relative to a quality norm or benchmark. The notion of contrast effects is well known in the economics and psychology literature (see Appendix A for an overview). However, it is conceptually less clear why contrasting focuses on recent and similar experiences. A straightforward explanation is offered by the concept of associative recall, which is a guiding principle in psychological research on memory (e.g., Kahana, 2012; Kahana et al., 2022) and has been incorporated into models of economic decision-making by Bordalo et al. (2020) and Wachter and Kahana (2023).[19] Under associative recall, evaluators retrieve prior interview experiences from memory based on their relative recency and similarity to the current interview situation.

In the following, we first describe a simple framework of contrast effects with associative recall, based on Bordalo et al. (2020). We then discuss its relation to our previous findings and provide additional reduced-form results on distinctive features of the framework. Finally, we summarize the results from a structural estimation evaluating the framework's quantitative fit with the data.

6.1.1 Framework

We consider an evaluator who assesses a candidate interviewed at time t. The interview results in the following valuation of the candidate:[20]

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

The valuation V_t depends on the candidate's own quality as perceived by the evaluator ($\tilde{q}_t$) and its difference to a quality norm (q_t^n).[21] The extent to which this difference affects the val-

[19] Appendix A includes a more detailed overview on psychological memory research.

[20] For sake of simplicity, we focus on the instantaneous valuation of the candidate formed at the time of the interview t, thereby abstracting from any ex-post adjustments that can occur after seeing all candidates.

[21] We abstract from anchoring towards the norm, as present in Bordalo et al. (2020). Anchoring can lead to assimilation effects in the case of small quality differences. In the context of candidate selection, evaluators aim to differentiate candidates, making the incidence of assimilation effects unlikely. Nevertheless, we formally discuss an extension with anchoring in Appendix G.1 and provide a quantitative assessment in Appendix Section H.5.4.

uation is determined by the salience $\sigma(\tilde{q}_t, q_t^n)$, which increases in the size of the difference.[22] Evaluators form the quality norm q_t^n by recalling candidates seen in previous interviews. Recall is associative, meaning that a prior interview experience receives a higher weight if its context is more similar to the current one. The norm is thus a similarity-weighted average of previously observed candidate quality:

$$q_t^n = \sum_{l=1}^{t-1} \tilde{q}_{t-l} \times \omega_{t-l}, \quad \text{where } \omega_{t-l} = \frac{S(c_t, c_{t-l})}{\sum_{l=1}^{t-1} S(c_t, c_{t-l})}$$

In this expression, the function $S(c_t, c_{t-l})$ captures the contextual similarity between the current interview and the interview that took place in period $t-l$. Similarity $S(c_{t-l})$ decreases in the distance between interview contexts c_t and c_{t-l}, where context includes both the time of the interview and additional features such as the characteristics of candidates.[23] Importantly, similarity matters in relative terms: when the similarity of one interview increases, this reduces the extent to which another interview is retrieved from memory. In other words, the recall of one interview interferes with the recall of another.[24]

In summary, the framework predicts the occurrence of contrast effects through the interplay of associative recall, which determines the quality norm, and the attention to quality differences. The notion of a sequential contrast effect — i.e., contrasting with respect to the previous candidate — is naturally incorporated: due to their high contextual similarity, more recent interviews receive a strong weight in the quality norm.

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

[23] Bordalo et al. (2020) argue that "critically contextual stimuli, such as location and time, act as cues that trigger recall of similar past experiences" (p. 1401). The overview of Kahana et al. (2022) summarizes the finding that time and other contextual features determine recall as the laws of recency and similarity (see Appendix A for details). Note that the choice of referring to recency as a form of contextual similarity has the main purpose of treating the different determinants of recall within a single framework.

[24] The notion that forgetting over time results from competition between memories due to interference is a central theme in memory research (see, e.g., the overview by Kahana et al., 2022). Examples of experimental evidence on interference include Pantelis et al. (2008) and da Costa Pinto and Baddeley (1991).

6.1.2 Reduced-Form Evidence

It is straightforward to interpret the results from Sections 4 and 5 in light of the presented framework. Differences in relative timing determine the recall of prior candidates, which can explain why the previous candidate matters most, why the influence decreases when interviews are separated by longer breaks, and why experiences from past sequences do not play a role. Moreover, the relative weight of the previous candidate decreases when evaluators expand their memory database over the sequence, explaining the smaller influence in later slots.[25] Finally, additional dimensions of similarity augment the recall of the previous candidate, which implies that the previous candidate has a stronger influence when sharing observable characteristics with the current candidate.

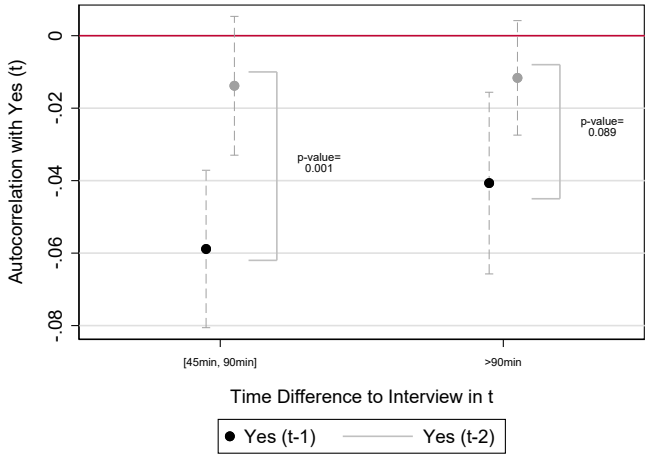
Additional Results: Interference A distinctive feature of models with associative recall is the notion of interference, whereby one memory disrupts the retrieval of other related memories, as described above. This notion has direct conjectures regarding the role of relative versus absolute recency and similarity, which we can take to the data.

Associative recall suggests that time differences between interviews matter in relative terms. The previous candidate has a strong influence because she is recalled without the interference of another interview in between. To assess this conjecture, we exploit the fact that the study grant data offer variation in both the absolute and the relative time difference between interviews. Thus, we can compare the influence of previous candidates whose interviews have on average the same absolute time distance to a given interview in t but a different relative distance (t-1 vs. t-2). More specifically, the idea is to compare (i) the influence of a candidate in t-1 who was interviewed τ minutes ago with (ii) the influence of a candidate in t-2 who was also interviewed τ minutes ago. The only difference between (i) and (ii) is whether another interview occurred during period τ.[26] Figure 8 provides strong evidence that the previous candidate is

[25] The intuition is that every interviewed candidate receives some positive weight, which mechanically reduces the weight of the previous candidate. Moreover, increasing the size of the memory database makes it more likely that other prior candidates interfere with the recall of the previous candidate.

[26] Note that the two effects need to be estimated using different sets of interviews in t, as it is not possible that both cases apply to the same interview.

Figure 8: The Role of Relative vs. Absolute Time Differences Between Interviews

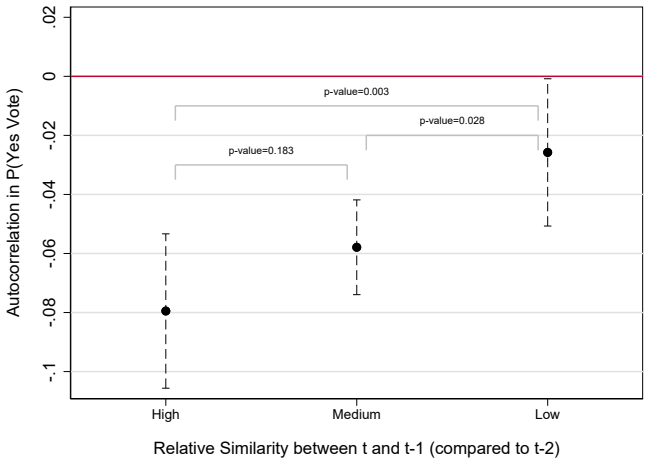


Note: The figure shows estimates of the autocorrelation based on equation 2. The black (gray) dots show the autocorrelation between the vote given to the candidate interviewed in t and the candidate interviewed in t-1 (t-2), depending on the time between the end of the interview in t-1 (t-2) and the start of the interview in t. Note that the autocorrelation with t-2 and t-1 are estimated on two different subsets of interviews. The cut-off at 45 minutes is chosen as the minimum time distance between t-2 and t. Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

influential due to her relative — rather than absolute — similarity in time. The autocorrelation between the vote in t and the vote in t-1 is significantly stronger than with the vote in t-2, although both t-1 and t-2 have the same absolute time difference τ relative to t (over an interval of 45 – 90 minutes). Moreover, the autocorrelation with an interview in t-1 that took place >90 minutes ago exceeds the autocorrelation with an interview in t-2 that took place ≤90 minutes ago. In other words, the relative recency is consistently more important than the absolute one, in line with the idea that the interview in t-1 interferes with the recall of the interview in t-2.

In Figure 9, we assess the role of interference for the role of observable similarity between candidates. Section 5 showed that the autocorrelation increases when two subsequent candidates share more characteristics. Associative recall predicts that the similarity of characteristics also matters in relative terms. Again, this is related to the notion of interference, where the recall of one experience (e.g., t-1) decreases when another experience (e.g., t-2) becomes more similar. In the study grant data, we can analyze how the influence of the previous candidate depends on the relative similarity of the candidates in t and t-1, compared to the similarity of

Figure 9: The Role of Relative Similarity Between Candidates



Note: The figure shows estimates of the autocorrelation based on equation 2, interacting the prior candidate's yes vote with the relative similarity of the candidate in t-1. High/medium/low relative similarity = the candidate interviewed in t-1 is more/equally/less similar to the candidate in t than the candidate interviewed in t-2. Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

t and t-2.[27] Specifically, we compare three cases: the candidate in t-1 is more similar, equally similar, and less similar to the candidate in t than the candidate in t-2. Figure 9 shows that as the relative similarity of t-1 decreases, the strength of the autocorrelation reduces from about 8 to about 2 percentage points, in line with the idea that similarity matters in relative terms.[28] Appendix Figure G.2 further supports this conjecture by showing that the relevant variation in the similarity index is not only driven by the (absolute) similarity between t and t-1 but also by the similarity between t and t-2. Panel (a) shows that the autocorrelation between t and t-1 is significantly weaker when the candidate in t-2 is more similar to the candidate in t. In Panel (b), we further split the middle group from Figure 9 into cases where both t-1 and t-2 are very similar to t, and cases where both are not. The pattern shows that the autocorrelation

[27] The same analysis would be severely underpowered in the hiring data. Given that the interviewers see at most three candidates, the analysis can only be conducted using observations from the third slot. However, there are only 516 individuals who are in the third slot of a sequence and follow a candidate with a positive hiring recommendation. Further dividing this group by relative similarity would result in unreasonably small cells.

[28] In Appendix Figure G.1, we perform the same exercise considering every characteristic separately. The overall pattern is consistent, although the single characteristics produce a less powerful variation than the joint index. A discussion of symmetric similarity by gender is provided in the working paper version (Radbruch & Schiprowski, 2020).

Table 5: Previous Candidate's Influence Within and Between Sub-Scores (Hiring Process)

	Cognitive Score	Non-Cognitive Score
	(1)	(2)
TPA, Cognitive (t-1)	-0.096***	-0.035**
	(0.015)	(0.014)
TPA, Non-Cognitive (t-1)	-0.035**	-0.065***
	(0.017)	(0.017)
p-value (coeff equality)	0.025	0.243
Outcome Mean	1.87	2.03
N	5155	5155

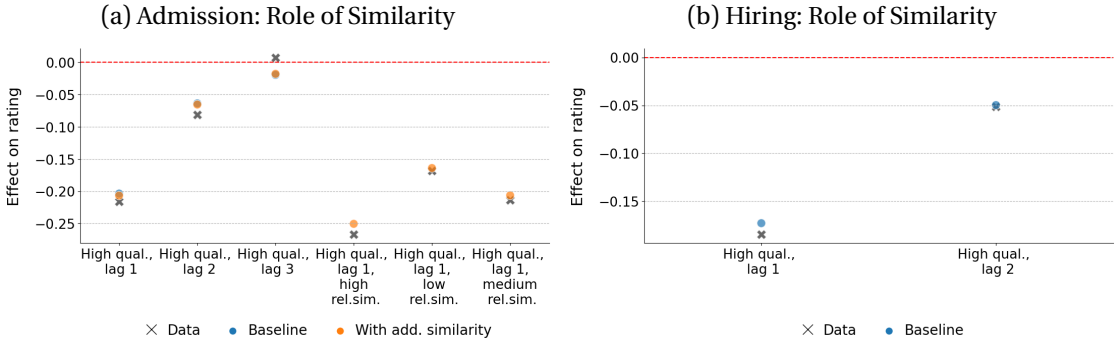
Note: TPA= "third-party-assessment". All regressions include candidate pool fixed effects and control for the candidate's own TPA measures. Additional controls include candidate characteristics, evaluator characteristics, and interview order. Standard errors are clustered at the candidate pool level (N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

remains unchanged when the absolute similarity of t-1 increases, but the relative similarity remains constant.

Additional Results: Contrasting Table 5 provides further evidence in favor of a contrast effect as the explanation. Research in psychology has argued that contrast effects occur through specific attributes of a given choice (see, e.g., Higgins et al., 1977; Simonsohn & Gino, 2013). Applied to our context, the quality of the previous candidate should matter more within rather than between attributes. We can test this conjecture in the hiring data, which report a candidate's cognitive and non-cognitive sub-scores. In line with the notion that contrast effects occur within quality attributes, Table 5 shows that the previous candidate's cognitive skills have a significantly stronger influence on the cognitive score than the previous candidate's non-cognitive skills, and vice versa.

Finally, Appendix Table G.1 demonstrates that the influence of the previous candidate is driven by large quality differences between t and $t-1$. This observation is in line with the framework presented above, where larger quality differences are more salient to the evaluator.

Figure 10: Empirical Moments and Model Fit: Influence of Previous Candidates



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

6.1.3 Structural Estimation and Quantitative Fit

To further strengthen the link between theory and empirics, we structurally estimate the framework using the method of simulated moments. While the reduced-form results have shown that the framework yields empirically relevant conjectures, the structural estimation also assesses its quantitative plausibility.

We present details on the model's parameterization, estimation and identification in Appendix H. Figure 10 presents the fit of the key simulated moments with their empirical counterparts. These moments describe how a candidate's rating reacts to the measured quality of the preceding candidates, depending on their time slots and, for the admission process, their relative observable similarity. We observe that the model estimates closely match the empirical influence of the previous candidates. Appendix Figure H.1 shows that this also holds true for other targeted moments.

Overall, a simple parameterization of the model provides a good quantitative fit of the previous candidates' influence. Moreover, we obtain very similar estimates of the key recall parameters across the two settings (see Appendix Table H.1), suggesting that the recall process might work very similarly across contexts. Benchmark models without associative recall re-

sult in a substantially worse fit with the data (see Appendix Figure H.2). Additional results and robustness checks, as well as a more detailed discussion, are provided in Appendix H.4.

6.2 Sequential Learning with Bayesian Updating

An alternative behavioral mechanism is sequential (Bayesian) learning about an admission or hiring threshold that depends on the average quality of previous candidates. In such a model, interviews with high-quality candidates increase the evaluator's belief about the threshold and thereby reduce the next candidate's rating. This behavior would need to occur despite the presence of well-defined selection criteria and the possibility to adjust ratings ex post to the average quality of all candidates in the sequence.

A standard model of Bayesian learning cannot explain the previous candidate's strong influence. In particular, such a model predicts the ordering of prior candidates to be irrelevant, which is not in line with the results presented in Section 4 (Figures 2 and D.2).[29] Nevertheless, one could posit a version of Bayesian learning where recent candidates receive a higher weight; for example, due to time-limited memory with exponential decay. Such a model would also need separate benchmarks for different candidate subgroups to account for the role of observable similarity.[30] However, it is unclear how the model would incorporate the fact that evaluators recall all candidates (or their average quality) at the end of the sequence to realize ex-post adjustments.

For a more general assessment of Bayesian learning, we investigate the role of evaluator experience and signal precision for the previous candidate's influence.[31] More experienced evaluators should hold better priors about the quality threshold and learn less from recent experiences. Against this conjecture, the results show that experience, age, interviewer training or managerial responsibility do not mitigate the influence of the previous candidate (see Fig-

[29] The structural estimation further supports this argument, showing that a framework with perfect recall of all prior candidates does not provide a good fit with the empirical moments (see Appendix Figure H.2).

[30] An alternative explanation for the role of the previous candidate's similarity could be a preference for diversity in combination with limited recall of prior candidates. However, this would not explain why similarity matters in relative terms (see Figure 9).

[31] These empirical tests are inspired by Bhuller and Sigstad (2023), who investigate Bayesian learning as an explanation for judges' reactions to appeals.

ure 5 & Table F.2). Moreover, evaluators should place greater weight on more precise signals in a Bayesian learning process. As a proxy of signal precision, we measure the other two evaluators' (dis)agreement about the previous candidate's quality. The idea is that a signal about the previous candidate's quality is more precise if the other two evaluators agree in their assessment of that candidate. Appendix Table I.1 shows that the influence of the previous candidate does not vary with the measured precision of the signal in either of the two processes.

In summary, the results speak against a standard model of Bayesian learning as an explanation of the previous candidate's influence. While it is more difficult to rule out extensions with time-limited memory, we note that they are not consistent with all patterns in the data.

6.3 Gambler's Fallacy

The gambler's fallacy describes the mistaken belief that a 'good draw' should follow a 'bad draw' and vice versa (e.g., Rabin, 2002; Rabin & Vayanos, 2010).[32] Under the gambler's fallacy, evaluators hold downward (upward) biased priors about the next candidate's quality after having seen a strong (weak) candidate. If these biased priors have a strong influence on the posterior belief about a candidate —for example, due to high noise in the interview signal—, they could explain the autorcorrelation observed in the data.[33]

However, the findings presented so far are only partially in line with the predictions of a gambler's fallacy. In particular, a gambler's fallacy where evaluators expect overall quality reversals does not explain why the previous candidate's influence is stronger within rather than between dimensions of candidate quality (Table 5), nor why it is reinforced by observable similarity (Figure 7). To make a gambler's fallacy consistent with these findings, one would need to assume, for example, that evaluators form their priors within each dimension of quality and candidate sub-group separately. However, even such a specific version would not explain the

[32] An overview on studies of the gambler's fallacy is provided by Oskarsson et al. (2009). Much of the laboratory evidence is based on tasks where subjects are asked to produce or recognize random sequences (e.g. Rapoport & Budescu, 1992, 1997; Bar-Hillel & Wagenaar, 1991). An example of early field evidence is Clotfelter and Cook (1993).

[33] A related mechanism is a backward-looking form of narrow bracketing, similar to Simonsohn and Gino (2013), where evaluators target a number of positive assessments. Similar arguments that speak against a gamblers fallacy also make it unlikely that narrow bracketing can explain the previous candidate's influence.

role of relative similarity (Figure 9).

Two additional empirical results speak against a gambler's fallacy. First, Appendix Table I.2 shows that the influence of the previous candidate's quality measure persists after controlling for the previous decision. This rules out a simple gambler's fallacy model (Rabin, 2002), where evaluators expect binary reversals, but not a more complicated version with beliefs about continuous quality (Rabin & Vayanos, 2010). Second, the gambler's fallacy predicts 'streaks' to matter in the sense that evaluators find three positive decisions in a row more unlikely than two. As a result, a positive decision in t-1 should have a stronger influence on the decision in t when the decision in t-2 was also positive. This is not the case in the contrast effect model, where the two previous candidates separately influence the quality benchmark. The results shown in Appendix Table I.3 do not support the relevance of streaks. In both settings, we find no evidence that two prior yes votes reduce the current decision more than a single one, nor that the effects of candidate quality in t-2 and t-1 reinforce each other.

7 Policy Responses

Irrespective of its behavioral mechanism, the influence of the previous candidate creates significant distortions in hiring and admission decisions. These distortions occur in professional processes, where evaluators have access to objective evaluation criteria and hold generic information about potential biases. In this section, we assess potential policy responses. First, we provide evidence that an information treatment carried out by the study grant program did not reduce the previous candidate's influence. Second, we explore an ordering algorithm that minimizes the observable similarity of subsequent candidates. Third, we simulate how the impact of interview-level contrast effects reduces when organizations collect more independent assessments per candidate. Finally, we discuss a procedure to flag assessments that are susceptible to a consequential influence of contrast effects.

Table 6: Effect of Information Treatment (Study Grant Program)

	Simple Diff		Diff-in-Diff	
	(1) Yes(t)	(2) Yes(t)	(3) Yes(t)	(4) Yes(t)
Yes (t-1)	-0.054***	-0.059***	-0.051***	-0.055***
	(0.018)	(0.019)	(0.010)	(0.009)
Yes (t-1) × 2022/23			0.008	0.006
			(0.021)	(0.021)
Yes (t-1) × Jan-Mar			-0.014	-0.009
			(0.013)	(0.012)
Yes (t-1) × Jan-Mar × 2022/23	-0.027	-0.017	-0.012	-0.009
	(0.026)	(0.026)	(0.030)	(0.029)
Controls	No	Yes	No	Yes
Outcome mean	0.39	0.39	0.37	0.37
N	6136	6136	33106	33106

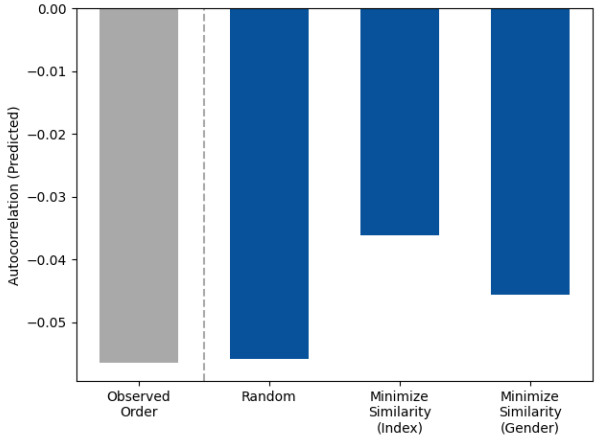
Note: Admission workshops take place from October to March. The information treatment was implemented in the second half of the academic year 2022/23 (January-March). In Columns 3-4, the academic years 2013/14 to 2016/17 serve as the control group."Yes" describes a vote in favor of admitting the candidate. All regressions include workshop fixed effects. Appendix Table J.1 shows results using the TPA measure. Standard errors are clustered at the workshop level (N=78 in Columns 1-2; N=390 in Columns 3-4). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

7.1 Information & Awareness

A popular approach to reduce biases in subjective assessments is the creation of awareness via training and information. To assess the impact of awareness on the incidence of contrast effects, we evaluate an information treatment that the study grant program implemented in the second half of the 2022/23 admission season (January-March).[34] Specifically, workshop organizers received updated guidelines for the pre-interview briefing of evaluators. The new guidelines included information about the concept of the contrast effect, a brief summary of our key findings and strategies to counteract contrast effects. This low-key implementation was chosen to respect time and human resource constraints within the organization. Appendix J.1 provides further details on the intervention. Importantly, no other changes in the admission process occurred simultaneously.

[34] The evaluation of the intervention was pre-registered at https://osf.io/n6ru3.

Figure 11: Simulation Results: Reordering of Candidates (Admission Process)



Note: The figure shows the simulated autocorrelation in yes votes under different ordering schemes. The gray bar shows the estimated autocorrelation. The blue bars show the simulated autocorrelations based on (i) a random reordering of candidates, (ii) a reordering that minimizes the relative observable similarity, and (iii) a reordering that minimizes the relative similarity by gender.

We evaluate the intervention using additional data for the academic year 2022/23. Table 6 reports the results, based on the autocorrelation (Equation 2). Appendix Table J.1 additionally shows results based on the TPA measure. We estimate the effect of the intervention with both a simple before-after comparison (Columns 1-2) and a difference-in-differences specification, where previous academic years from our main dataset serve as the control group (Columns 3-4). The results suggest that the intervention did not significantly alter the size of the autocorrelation. More specifically, the estimates and their standard errors rule out that the autocorrelation reduced by 50% or more, indicating that light information treatments are insufficient to significantly counteract contrast effects.

7.2 Reordering Candidates

A second possible intervention targets the sequencing of interviews. The results have shown that the previous candidate has a stronger influence when the (relative) similarity to the current candidate is high (see Figures 7 and 9). Based on this result, we explore the potential to minimize the average autocorrelation by reducing the relative similarity between subsequent

candidates. Due to the short sequences in the hiring process, we only perform this analysis for the admission process.

To reorder candidates within interview sequences, we use a greedy algorithm that starts with a random candidate and iteratively adds the candidate with the lowest relative similarity to the previously added candidate.[35] We calculate the resulting average autocorrelation based on the shares of subsequent candidates with a high, medium or low relative similarity, and the estimated autocorrelation for these three groups (based on Figure 9). Figure 11 illustrates the results of this procedure. The gray bar shows the autocorrelation in yes votes, as observed in the data. A random reordering — which we run as an implementation check — leaves the autocorrelation unchanged. In turn, minimizing the relative similarity of subsequent candidates within sequences reduces the average autocorrelation by about 40%. To inform settings where fewer candidate characteristics are observed, we also simulate a reordering based solely on gender (using the estimates from Appendix Figure G.1 a). This leads to a reduction by about 20%. Overall, these results offer a simple proof-of-concept that reordering candidates — especially when based on a comprehensive set of characteristics — can potentially reduce contrasting against the previous candidate.

7.3 Increasing the Number of Independent Interviews

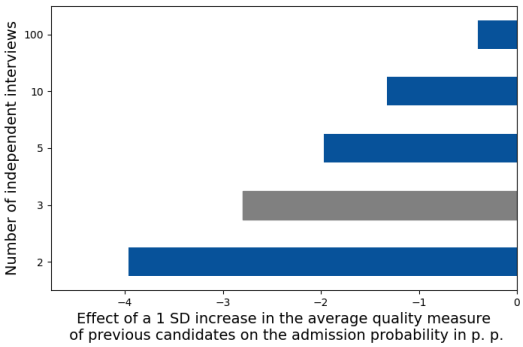
An alternative approach to mitigate distortions in final decisions takes the evaluator-level effect as given and increases the number of independent interviews per candidate. The intuition is that independent biases in individual assessments cancel out in the aggregate. More specifically, the individual-specific average quality of previous candidates converges to the population average as the number of independent interviews increases. We conduct a simulation exercise to understand how quickly this process mitigates the impact on final decisions. Details on the simulation procedure are provided in Appendix J.2.

Figure 12 illustrates the simulation results, which quantify the impact of a one standard deviation change in the average quality of an individual's previous candidates on the admis-

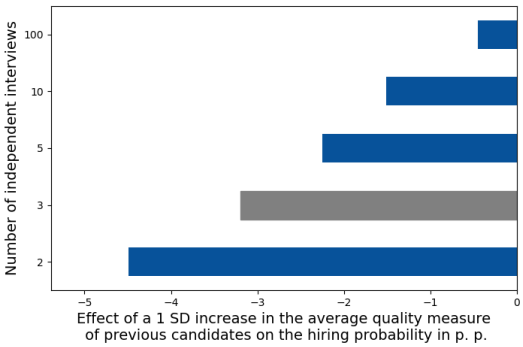
[35] Note that this is a heuristic approach that serves as a proof-of-concept regarding the feasibility of reducing the autocorrelation with a simple reordering algorithm.

Figure 12: Simulation Results: Increasing the Number of Independent Assessments

(a) Admission Process



(b) Hiring Process



Note: The gray bar shows the estimated influence of the previous candidates' average quality on the admission or hiring probability. The blue bars illustrate the simulated impact under a varying number of interviews. Note that we simulate N-1 assessments as being influenced by a previous candidate, as is the case in the two processes. Details on the simulation procedure are provided in Appendix J.2.

sion or hiring probability. As expected, the impact decreases as the number of interviews increases, although the rate of decrease is rather slow. To reduce the impact by half relative to our benchmark of three independent assessments, both organizations would have to conduct about ten interviews per candidate. This illustrates that the collection of multiple assessments can help reduce the impact of individual-level errors, although complete elimination may not be realistic due to the costs of additional assessments.

7.4 Flagging Interview Assessments

Finally, organizations can introduce straightforward flagging procedures into their assessment systems, to identify hiring or admission decisions which might have been altered by contrast effects. Such a procedure could alert organizations to the need for collecting additional assessments on specific candidates, or prompt deeper committee discussions about them. In its simplest form, a flagging procedure would highlight assessments which were made after seeing a particularly strong or weak candidate, and which are pivotal to the committee's final

decision. The cut-offs for flagging candidates need to trade off the likelihood of making Type I and Type II errors with the costs of spending more time and effort on specific candidates.

8 Conclusion

Using data on interviews from two high-stakes selection processes, this paper shows that candidate assessments are negatively influenced by the quality of the previous candidate in the interview sequence. This influence is sizable compared to other determinants, such as the candidate's own quality or the average quality of the other candidates in the same sequence. It is particularly pronounced at the beginning of the interview sequence and when subsequent candidates are observably similar. Additional reduced-form and structural results support a contrast effect model where the benchmark for current evaluations is formed through the associative recall of prior candidates.

As the strong influence of the previous candidate creates significant distortions in admission and hiring decisions, we explore potential policy responses for firms and organizations. We find that a light information treatment was not effective in mitigating the influence. Simulations suggest that the reordering of candidates based on their similarity could reduce the average influence. Furthermore, collecting multiple independent assessments per candidate reduces the impact of individual contrast effects on final decisions, albeit at a slow rate. As the collection of independent assessments usually involves high costs, organizations would benefit from concentrating such efforts on decisions with a high risk of reversal due to contrast effects. We propose a simple flagging procedure to identify such decisions.

Beyond these interventions, organizations can complement subjective interview assessments with an increasing number of alternative tools, such as job-testing technologies or selection algorithms. Previous research suggests that these can improve match quality (Hoffman et al., 2018), and promote diversity when designed accordingly (Bergman et al., 2020). Determining how to optimally combine objective and subjective information about candidates seems an important avenue for future research.

References

Afrouzi, H., Kwon, S., Landier, A., Ma, Y., & Thesmar, D. (2023). Overreaction in expectations: Evidence and theory. *Quarterly Journal of Economics, 138*(3), 1713–1764.

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

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

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics, 12*(4), 428–454.

Bergman, P., Li, D., & Raymond, L. (2020). Hiring as exploration. *NBER Working Paper No. 27736*.

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

Bhuller, M., Dahl, G. B., Løken, K. V., & Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy, 128*(4), 1269–1324.

Bhuller, M., & Sigstad, H. (2023). Feedback and learning: The causal effects of reversals on judicial decision-making. *mimeo*.

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

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

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

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*(1), 51–62.

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

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

Clotfelter, C. T., & Cook, P. J. (1993). Notes: The "gambler's fallacy" in lottery play. *Management Science, 39*(12), 1521–1525.

Costa Pinto, A. da, & Baddeley, A. D. (1991). Where did you park your car? analysis of a naturalistic long-term recency effect. *European Journal of Cognitive Psychology, 3*(3), 297–313.

DellaVigna, S., Heining, J., Schmieder, J. F., & Trenkle, S. (2022). Evidence on job search models from a survey of unemployed workers in Germany. *Quarterly Journal of Economics, 137*(2), 1181–1232.

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, Vol. 1). North-Holland.

Ebbinghaus, H. (1913). Retention and obliviscence as a function of the time. In H. Ebbinghaus, H. A. Ruger, & C. E. Bussenius (Eds.), *Memory: A contribution to experimental psychology* (pp. 62–80). Teachers College Press.

Enke, B., Schwerter, F., & Zimmermann, F. (2020). Associative memory and belief formation. *Discussion Paper Series – CRC TR 224 No. 148*.

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

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

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology, 66*(3), 325–331.

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.

Higgins, T. E., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology, 13*, 141–154.

Hoffman, M., Kahn, L., & Li, D. (2018). Discretion in hiring. *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.

Jin, L., Tang, R., Ye, H., Yi, J., & Zhong, S. (2023). Path dependency in physician decisions. *Review of Economic Studies*, fortcoming.

Jochmans, K. (2023). Testing random assignment to peer groups. *Journal of Applied Econometrics, 38*(3), 321–333.

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

Kahana, M. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*, 107–138.

Kahana, M., Diamond, N. B., & Aka, A. (2022). Laws of human memory. *PsyArXiv*.

Kenrick, D. T., & Gutierres, S. E. (1980). Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology*, *38*(1), 131–140.

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: Divisive normalization in parietal cortex. *Journal of Neuroscience*, *31*(29), 10627–10639.

Malmendier, U., & Wachter, J. A. (2021). Memory of past experiences and economic decisions. *mimeo*.

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.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? judging sequences of binary events. *Psychological bulletin*, *135*(2), 262.

Pantelis, P. C., Van Vugt, M. K., Sekuler, R., Wilson, H. R., & Kahana, M. J. (2008). Why are some people's names easier to learn than others? the effects of face similarity on memory for face-name associations. *Memory & cognition*, *36*, 1182–1195.

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American*, *219*(6), 84–93.

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. *Quarterly Journal of Economics*, (Vol. 117, No. 3), 775–816.

Rabin, M., & Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, *77*(2), 730–778.

Radbruch, J., & Schiprowski, A. (2020). Interview sequences and the formation of subjective assessments. *ECONtribute Discussion Papers Series No. 45*.

Rangel, A., & Clithero, J. A. (2014). Chapter 8 - the computation of stimulus values in simple choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (Second Edition, pp. 125–148). Academic Press.

Rapoport, A., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*(3), 352.

Rapoport, A., & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review*, *104*(3), 603.

Sherman, S. J., Ahlm, K., Berman, L., & Lynn, S. (1978). Contrast effects and their relationship to subsequent behavior. *Journal of Experimental Social Psychology*, *14*(4), 340–350.

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.

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

Thakral, N., & Tô, L. T. (2021). Daily labor supply and adaptive reference points. *American Economic Review*, *111*(8), 2417–2443.

Wachter, J. A., & Kahana, M. J. (2023). A retrieved-context theory of financial decisions. *Quarterly Journal of Economics*.

Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology*, *56*(1), 45.

Appendix – for Online Publication

A Additional Material: Related Literature

Relation to Psychological Literature on Contrast Effects and Memory

In Section 6, we use a contrast effect model with associative recall to explain the negative influence of the previous candidate's quality. This mechanism relates to two strands of literature in psychology: research on contrast effects as an established bias in the perception and evaluation of alternatives and research on memory and the recall of experiences.

Contrast Effects Contrast effects refer to the phenomenon in which a stimulus is perceived or evaluated relative to previous or surrounding stimuli. The influence of contrast effects has been documented in several areas of visual and sensory perception, with the Ebbinghaus or Titchener circles illusion being a prominent example. In the context of social judgments, laboratory studies have provided evidence of contrast effects in the evaluation of performance (Wexley et al., 1972), physical attractiveness (Kenrick & Gutierres, 1980), crimes (Pepitone & DiNubile, 1976; Parducci, 1968), and the relevance of social issues (Sherman et al., 1978). More recent literature — mostly in economics — has documented contrast effects in field data. We provide an overview of these studies in Table A.1.

The question of relative perception has also been studied in neuroscience, where contrast effects have been formalized using divisive normalization models (e.g., Carandini & Heeger, 2012). According to these models, contrast effects originate from the neural processing mechanisms of the brain. They emerge because a neuron's response to a given stimulus is influenced by the background activity of surrounding neurons in the reference population. For example, a neuron reacts more strongly to a bright light spot when neighboring neurons respond to darker spots at the same time. Through this normalization process, the brain interprets stimulus values based on their context. Divisive normalization models have been supported by laboratory studies that show that neuronal activity adapts to the context in which a choice option is evaluated (e.g., Rangel & Clithero, 2014; Louie et al., 2011).

Memory and Associative Recall To understand why contrasting arises specifically against more recent and similar candidates, we rely on the concept of associative recall. This concept has been extensively studied in the psychological literature on memory (see,e.g., Kahana, 2012) and has more recently been applied to economic decision-making by Bordalo et al. (2020) and Wachter and Kahana (2023). In the following, we offer a brief overview of key

memory effects that are relevant to understanding the nature of contrast effects in interviewing. In particular, we summarize key findings regarding the recall of information, which are described in more detail in Kahana (2012), Kahana (2020) and Kahana et al. (2022).[1]

An important premise in memory research is that the recall of experiences is context-dependent. Specifically, an experience is more likely to be recalled if the current context is more similar to the context of that experience. In the framework of Bordalo et al. (2020), "similarity" encompasses various contextual variables such as time, location, and other attributes. The relevant underlying findings of psychology are summarized by Kahana et al. (2022) as the laws of recency and similarity. In the following, we summarize the two concepts and their usage in the psychology literature.

The law of recency states that more recent experiences are recalled more strongly. The first observations that recall diminishes over time dates back to the seminal work by Ebbinghaus (1913). Contemporary research in psychology identifies two primary mechanisms driving recency effects (Kahana et al., 2022). The first mechanism is related to interference, where the recall of an experience is blocked by the incidence of subsequent, similar experiences.[2] Thus, forgetting is not merely driven by the passage of time, but rather by the incidence of competing experiences. Supporting this idea, da Costa Pinto and Baddeley (1991) show that the recall of a car's parking location remains consistent over time unless other parking experiences intervene. The second mechanism relates to changes in context which occur as time progresses. As the context changes, a given experience becomes increasingly dissimilar to the present, reducing the recall of that experience. This notion is closely related to the law of similarity, which states that the recall of an experience or item increases when it is more similar to the current situation (Kahana et al., 2022). Similarity is defined by the context, including time, space, semantics, or other attributes. Several experimental findings support the notion that individuals are more likely to recall experiences or items that they observed in a similar context (e.g., Godden & Baddeley, 1975). At the same time, other (less similar) items are recalled less, likely due to interference (e.g., Pantelis et al., 2008).

[1] An alternative overview of memory research and its application to economic decision-making can be found in Malmendier and Wachter (2021).

[2] More generally, interference can also arise from earlier memories (also called proactive interference), where earlier events block the recall of more recent events.

Table A.1: Field Evidence on Contrast Effects

	Setting	Indiv.-Level Data?†	N	Main Effect	Role of Experience	Role of Similarity between t and t-1	Effect Within vs. Across Choice Attributes
Simonsohn (2006)	Housing choices	Yes	1,067 movers	Movers from cities with higher average commuting time choose longer commutes in their destination city.	• Evidence of adjustment to commuting times in destination city	not analyzed	not analyzed
Loewenstein, Simonsohn (2006); Bordalo et al. (2019)	Housing choices	Yes	646 movers/ 2,773 movers	Movers from more expensive cities pay higher rents in their destination city.	• Evidence of adjustment to prices in destination city	not analyzed	not analyzed
Chen, Moskowitz, Shue (2016)	asylum decisions, loan decisions, baseball umpires	Yes	672 decision makers > 1 Mio. decisions	Negative autocorrelation in binary decision outcomes (attributed to a gambler's fallacy)	• Experience mitigates autocorrelation among judge and loan officer decisions.	• Effect stronger when asylum applicants are of the same origin.	not analyzed
Bhargava, Fisman (2013)	Speed dating experiment	Yes	474 speed daters 7,684 decisions	Dating outcome in t decreases in attractiveness in t-1 (only for male daters).	• Effect weakens over dating sequence • Effect close to zero for weekly speed daters; no difference between daters with bi-weekly/monthly/less experience	not analyzed	not analyzed
Hartzmark, Shue (2018)	Financial markets	No	75,897 returns	Stock market returns neg. related to earnings surprises announced by large firms on previous day.	not analyzed	• Stronger effect for same-industry announcements	not analyzed
This study	Admission & hiring interviews	Yes	2,853 evaluators 37,899 decisions	Assessment of candidate in t decreases in quality of candidate in t-1.	• Effect weakens over interview sequence (Fig. 4) • No heterogeneity by prior experience (Fig. 5) or interviewer training status (Table F.2)	• Effect decreases with break between t and t-1 (Fig. 6) • Effect increases in (relative) similarity in candidate characteristics between t and t-1 (Fig. 7 & 9)	• Stronger effect within than across dimensions of candidate quality (Table 5)

Note: †: Individual-level data = data at the level of the decision maker.

B Additional Material: Institutional Setting

B.1 Background Information on Study Grant Program

Candidates at the admission workshops apply for a large merit-based study grant program for German university students. The grant provides a lump sum payment of 300 euros per month. Depending on their parental income, recipients additionally receive a stipend that covers up to their entire living costs. Additional financial support is offered when spending a semester abroad. Moreover, the program offers a large, cost-free course program that includes language classes abroad, summer schools, and career workshops. Finally, its benefits include various networking opportunities and a high signaling value.

B.2 Illustration of Admission Workshop Schedule

Figure B.1: Illustration of Admission Workshop 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

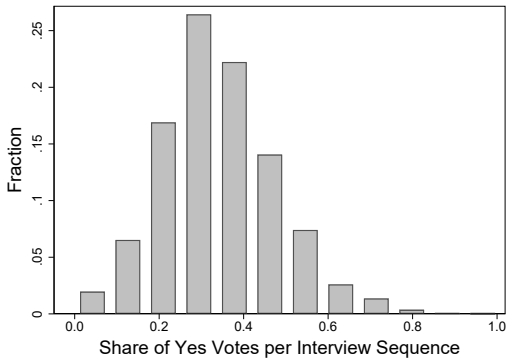
Note: The timetable illustrates the assignment of candidates to evaluators and time slots. Candidates are identified by an ID between 1 and 48. Evaluators are identified by an ID between A and H at the respective time slot. When a candidate ID appears in a slot denoted "Group", this means that the candidate presents in front of her group and moderates the subsequent discussion. If not noted otherwise, there is a 5 minutes break after every interview or group discussion. If more or fewer than 48 candidates attend the workshop, the schedule gets slightly adjusted. The empirical analysis relies on the actual schedule with the actual number of participants.

C Additional Material: Data

Overview Figures C.1 and C.2 show distributions of yes vote shares per interview sequence and of admission/hiring rates per workshop/day. Tables C.1 and C.2 provide summary statistics on candidate and evaluator characteristics. Table C.3 documents the correlation between ratings and GPA vs. TPA as measures of candidate quality. Tables C.4 and C.5 provide evidence that an evaluator's characteristics only influence her own rating of a given candidate, without any spillover on the other two evaluators' ratings (TPA). Tables C.6 and C.7 show that there is no indication of systematic sorting of candidates to evaluators.

Figure C.1: Distribution of Yes Vote Shares & Admission Rates (Admission Process)

(a) Share of Yes Votes per Interview Sequence



(b) Share of Admissions per Workshop

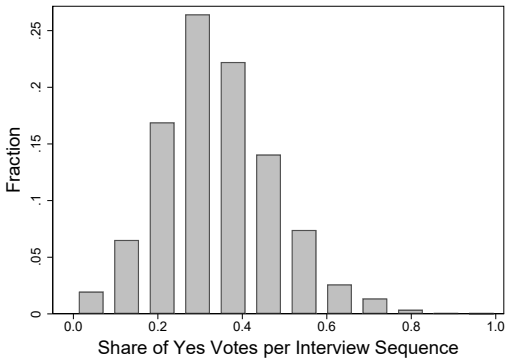
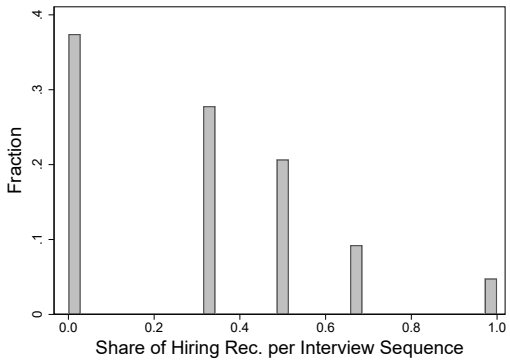


Figure C.2: Distribution of Hiring Recommendation Shares & Offer Rates (Hiring Process)

(a) Share of Hiring Rec. per Interview Sequence



(b) Share of Offers per Interview Day

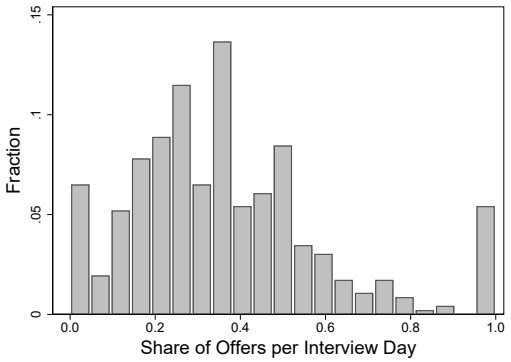


Table C.1: Summary Statistics on Evaluator & Candidate Characteristics (Admission Process)

	Evaluators		
	N	Mean	SD
Female	2496	0.48	0.50
Age	2496	42.02	11.57
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	0.27	0.44
Experience: 1	2496	0.21	0.41
Experience: 2	2496	0.15	0.36
Experience: 3+	2496	0.37	0.48
No of interviews	2496	11.81	0.71
	Candidates		
	N	Mean	SD
Female	14733	0.55	0.50
Age	14733	19.58	1.18
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 C.2: Summary Statistics on Evaluator & Candidate Characteristics (Hiring Process)

	Evaluators		
	N	Mean	SD
Female	357	0.35	0.48
Management Responsibility	357	0.69	0.46
Years of Evaluator Experience	357	1.40	1.71
Interview in Prev. 30 Days	357	0.23	0.42
	Candidates		
	N	Mean	SD
Female	3313	0.42	0.49
Internship Application	3313	0.35	0.48
Field: STEM	3313	0.25	0.43
Field: Business	3313	0.58	0.49
Field Missing	3313	0.09	0.28
Field: Soc. Sciences, Humanities	3313	0.08	0.27
High GPA	3313	0.37	0.48

Table C.3: Correlation of Interview Ratings with Candidate GPA and TPA

	Std. Rating, Admission Process			Std. Rating, Hiring Process		
	(1)	(2)	(3)	(4)	(5)	(6)
GPA (Std.)	0.066***		0.034***	0.074***		0.053***
	(0.008)		(0.006)	(0.021)		(0.019)
TPA (Std.)		0.366***	0.363***		0.250***	0.247***
		(0.006)	(0.006)		(0.017)	(0.017)
R-Squared	0.004	0.130	0.131	0.004	0.060	0.062
N	26970	26970	26970	5165	5165	5165

Note: TPA = third-party assessment of candidate quality (see Section 3.2 for details). All regressions include workshop (Columns 1-3) or candidate pool (Columns 4-6) fixed effects. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table C.4: Influence of Evaluator Characteristics on Rating and TPA (Admission Process)

	Rating (Std.)	TPA (Std.)
	(1)	(2)
Female	0.044***	0.005
	(0.014)	(0.013)
Age	0.004***	0.001
	(0.001)	(0.001)
Field: Social Sciences	0.019	0.010
	(0.024)	(0.021)
Field: STEM	0.033**	0.019
	(0.016)	(0.015)
Field: Medicine	0.017	-0.023
	(0.026)	(0.028)
Field: Others	0.042	0.037
	(0.061)	(0.068)
Experience > 1 Workshop	-0.129***	-0.018
	(0.015)	(0.014)
p-value (joint significance)	0.00	0.56
N	26970	26970

Note: Humanities is the omitted study field. Regressions control for candidate characteristics. Standard errors are clustered at the workshop level (N=312). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table C.5: Influence of Evaluator Characteristics on Rating and TPA (Hiring Process)

	Overall Rating		Avg. Sub-Rating	
	(1) Std. Rating	(2) Std. TPA	(3) Std. Rating	(4) Std. TPA
Female	0.147***	-0.034	0.170***	-0.038
	(0.039)	(0.025)	(0.034)	(0.024)
Management Responsibility	0.020	-0.007	0.038	-0.021
	(0.044)	(0.040)	(0.042)	(0.041)
Above Median Experience	-0.096***	0.020	-0.129***	0.025
	(0.034)	(0.025)	(0.035)	(0.023)
Interview in Prev. 30 Days	-0.060*	0.016	-0.072*	0.010
	(0.034)	(0.032)	(0.037)	(0.028)
p-value (joint significance)	0.00	0.51	0.00	0.35
N	5165	5165	5165	5165

Note: Regressions control for candidate characteristics. Standard errors are clustered at the candidate pool level (N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table C.6: Relationship between Evaluator and Candidate Characteristics (Admission Process)

	Evaluator Characteristic			
	(1) Female	(2) Age	(3) Experience	(4) Field: STEM
Female	0.005	-0.107	-0.020	-0.003
	(0.004)	(0.085)	(0.023)	(0.004)
Age	0.005*	0.023	0.014	-0.002
	(0.002)	(0.056)	(0.015)	(0.002)
GPA > Median	0.000	0.061	0.020	-0.001
	(0.006)	(0.118)	(0.035)	(0.006)
Migration Background	0.002	0.115	0.039	-0.007
	(0.007)	(0.168)	(0.042)	(0.007)
1st Generation Student	-0.008	-0.003	0.049	0.007
	(0.005)	(0.130)	(0.038)	(0.005)
Field: STEM	0.007	0.041	0.014	-0.009*
	(0.006)	(0.125)	(0.034)	(0.005)
p-value (joint significance)	0.29	0.79	0.38	0.49
Outcome Mean	0.48	42.00	2.59	0.36
N	29466	29466	29466	29466

Note: Quasi-random assignment to evaluators is conditional on gender. Regressions include workshop fixed effects. Standard errors are clustered at the workshop level (N=312). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table C.7: Relationship between Evaluator and Candidate Characteristics (Hiring Process)

	Evaluator Characteristic		
	(1) Female	(2) Manager	(3) Experience
Female	0.150***	-0.015*	0.031
	(0.012)	(0.009)	(0.047)
Field: STEM	-0.005	0.011	0.040
	(0.015)	(0.016)	(0.066)
Field: Business	-0.002	-0.013	-0.032
	(0.014)	(0.016)	(0.071)
High GPA	0.003	-0.009	-0.037
	(0.009)	(0.014)	(0.044)
p-value joint significance (excl. gender)	0.98	0.18	0.43
Outcome Mean	0.31	0.69	2.00
N	8437	8437	8437

Note: Quasi-random assignment to evaluators is conditional on gender. Regressions include candidate pool fixed effects. Standard errors are clustered at the candidate pool level (N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

D Additional Material: Influence of the Interview Sequence

Overview Table D.1 (admission process) reports the estimates illustrated in Figures 2 (a) and Figure D.3. Table D.2 (hiring process) reports the estimates illustrated in Figure 2 (b). Both tables include a p-value from a placebo test assessing the null hypothesis that a given candidate's influence exceeds the influence of the average, based on a bootstrap procedure of random re-shuffling (see details below). Figure D.1 shows the distributions of the estimated placebo coefficients for the candidate in $t-1$. Figure D.2 shows estimates of the other candidates' influence, based on a homogeneous sample containing a subset of interviews. Figure D.3 is analogous to Figure 2 (a), but additionally controls for the average influence of all other candidates in the sequence using their leave-one-out mean TPA. Figure D.4 provides an additional placebo check, documenting the absence of an autocorrelation in TPA throughout the sequence.

Tables D.3 and D.4 report the coefficients for $k = -1$ and show their robustness to alternative specifications and the use of alternative quality measures. The results in Table D.5 are based on an instrumental variable approach to the measurement of candidate quality, where one quality measurement serves as an instrument for the other.

Details on Placebo Check (Bootstrap Procedure) Tables D.1 and D.2 report p-values from a placebo check based on a bootstrap procedure which repeatedly reshuffles the order of interviews in each sequence. In the placebo dataset, all information is kept; the only difference is the order of candidates within the sequence. Using the placebo dataset, we estimate how the measured quality (TPA) of the candidate in period $t-k$ affects the rating of the candidate in t based on equation 1, but using the placebo lags and leads (and the accordingly adjusted leave-two-out mean TPA). We repeat this procedure 2000 times and compute the resulting distribution of t-values.

This distribution yields the test statistic under the null hypothesis that the quality of the candidate in t+k does not affect the rating of the current candidate more than the average other candidate. This means that the adjusted p-values measure if the estimated influence of a candidate exceeds that of the average other candidate in the sequence. In Panel (a), we expect the average candidate to have an influence of -0.036, which is the weighted average of the estimated coefficients. In Panel (b), where regressions control for the leave-one-out mean TPA of the interview sequence, we expect the average candidate to have an influence of -0.002.

This bootstrap procedure maintains all possible violations of assumptions in the data. We simply compute the adjusted p-value as the percentage of t-values in the placebo distribution that are in absolute terms larger than the empirical counterpart. Figure D.1 shows the distributions of the estimated placebo coefficients for the candidate in $t-1$. The dashed line indicates the actual coefficient estimate.

Table D.1: Coefficients and p-Values Corresponding to Figures 2 (a) and D.3

	β_{-11}	β_{-10}	β_{-9}	β_{-8}	β_{-7}	β_{-6}	β_{-5}	β_{-4}	β_{-3}	β_{-2}	β_{-1}	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
Panel A: Leave-two-out control																						
β_k	0.001	-0.016	-0.039***	-0.027**	-0.021**	-0.035***	-0.044***	-0.046***	-0.035***	-0.055***	-0.091***	-0.029***	-0.022***	-0.019***	-0.033***	-0.028***	-0.023***	-0.014	-0.031***	-0.033***	0.003	-0.015
	(0.020)	(0.015)	(0.011)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.009)	(0.011)	(0.011)	(0.014)	(0.020)
p-value (Bonf. adj.)	1.00	1.00	0.01	0.25	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.08	0.00	0.00	0.05	1.00	0.09	0.08	1.00	1.00
p-value (Placebo)	0.99	0.90	0.26	0.81	0.95	0.43	0.06	0.02	0.49	0.00	0.00	0.84	0.99	0.99	0.54	0.79	0.90	0.99	0.71	0.50	0.99	0.78
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	2142	4523	7003	9498	11994	14490	16986	19482	21978	24474	26970	26970	24474	21978	19482	16986	14490	11994	9498	7003	4523	2142
Panel B: Leave-one-out control																						
β_k	0.034	0.019	-0.004	0.007	0.014	-0.002	-0.012	-0.013*	0.000	-0.021***	-0.062***	0.006	0.012*	0.016**	0.002	0.008	0.014*	0.023***	0.002	-0.003	0.033**	0.017
	(0.021)	(0.015)	(0.011)	(0.011)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.009)	(0.011)	(0.012)	(0.014)	(0.021)
p-value (Bonf. adj.)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.00	1.00	1.00	0.45	1.00	1.00	1.00	0.19	1.00	1.00	0.44	1.00
p-value (Placebo)	0.11	0.19	0.71	0.52	0.12	0.80	0.11	0.06	1.00	0.00	0.00	0.27	0.07	0.02	0.75	0.27	0.08	0.01	0.88	0.81	0.02	0.43
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	2142	4523	7003	9498	11994	14490	16986	19482	21978	24474	26970	26970	24474	21978	19482	16986	14490	11994	9498	7003	4523	2142

Note: This table reports the coefficients and p-values corresponding to Figures 2 (a) and D.3. All regressions include workshop fixed effects and control variables as specified in equation (1). Standard errors are clustered at the workshop level (N=312). The table reports two additional sets of p-values.The first set relies on a Bonferroni adjustment to account for multiple hypothesis testing. The second set relies on a placebo exercise that randomly re-shuffles the order of interviews (see introduction to Appendix D for details). $^{*}\ p<0.10,^{**}\ p<0.05,^{***}\ p<0.01$.

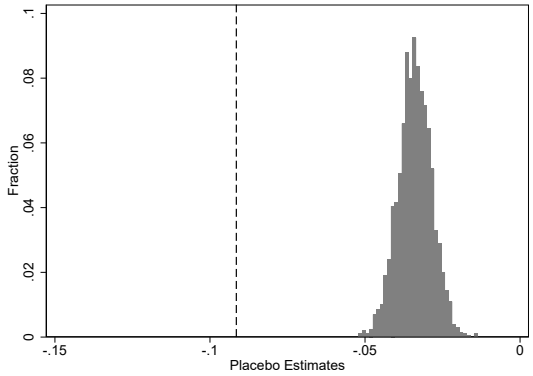
Table D.2: Coefficients and p-Values Corresponding to Figure 2 (b)

	β_{-2}	β_{-1}	β_1	β_2
β_k	-0.059***	-0.117***	-0.016	-0.042*
	(0.019)	(0.015)	(0.014)	(0.023)
p-value (Bonf. adj.)	0.01	0.00	1.00	0.30
p-value (Placebo)	0.40	0.00	1.00	0.86
Controls	Yes	Yes	Yes	Yes
N	1893	5165	5165	1893

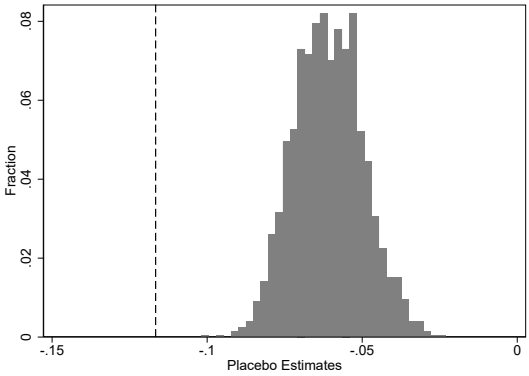
Note: This table reports the coefficients and p-values corresponding to Figures 2 (b). All regressions include candidate pool fixed effects and control variables as specified in equation (1). Standard errors are clustered at the candidate pool level (N=63). The table reports two additional sets of p-values.The first set relies on a Bonferroni adjustment to account for multiple hypothesis testing. The second set relies on a placebo exercise that randomly re-shuffles the order of interviews (see introduction to Appendix D for details). $^{*}\ p<0.10,^{**}\ p<0.05,^{***}\ p<0.01$.

Figure D.1: Distribution of Placebo Values for $t-1$

(a) Admission Process



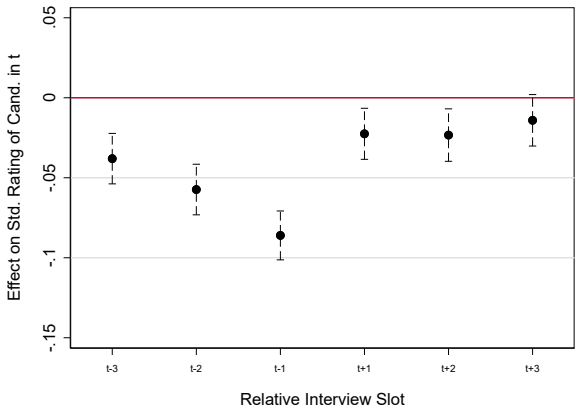
(b) Hiring Process



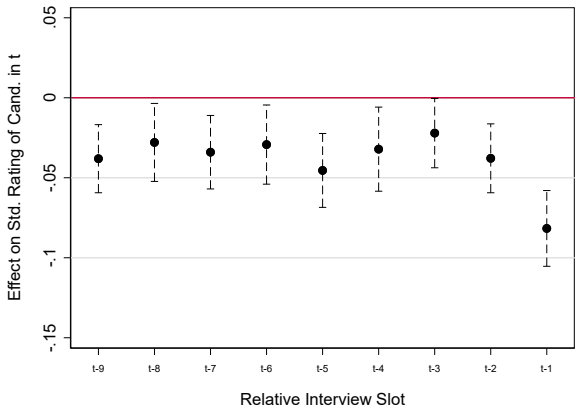
Note: The dashed lines indicate the estimated influence of the previous candidate in the data. To obtain placebo estimates, we randomly reshuffle the order of interviews for each interviewer and estimate equation 1 for k=t-1 using the placebo ordering. The histogram illustrates the distribution of placebo estimates resulting from 2,000 repetitions. The average placebo estimate corresponds to the expected influence of a randomly drawn candidate in the sequence.

Figure D.2: Effect of Candidate Quality in $t+k$ on Std. Rating of Candidate in t: Results from Single Regressions

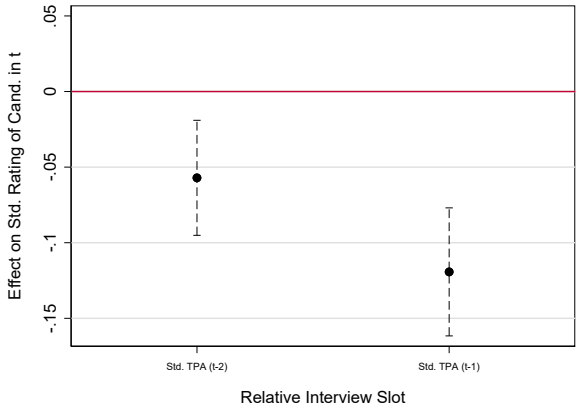
(a) Admission: Interviews with 3 lags and 3 leads



(b) Admission: Interviews with at least 9 leads

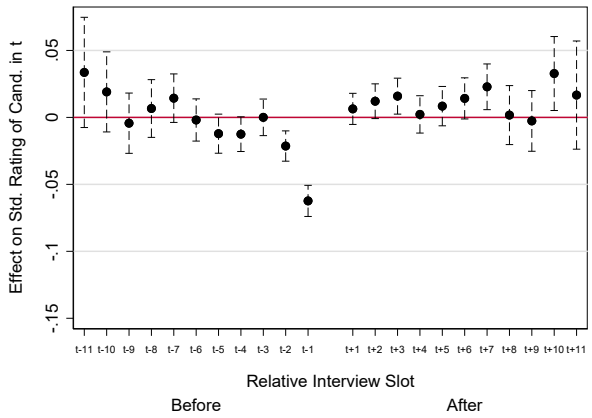


(c) Hiring: Interviews with 2 lags



Note: Estimates result from a single regression on the subset of slots that have three lags and three leads in Panel (a)), 9 lags in Panel (b), and two lags in Panel (c). 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 3.2 for details). Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level (Panel (a)), N=312) or candidate pool level (Panel (b), N=63).

Figure D.3: Influence of Candidate Quality in $t+k$ on Std. Rating of Candidate in t, Conditional on Leave-One-Out Mean Quality

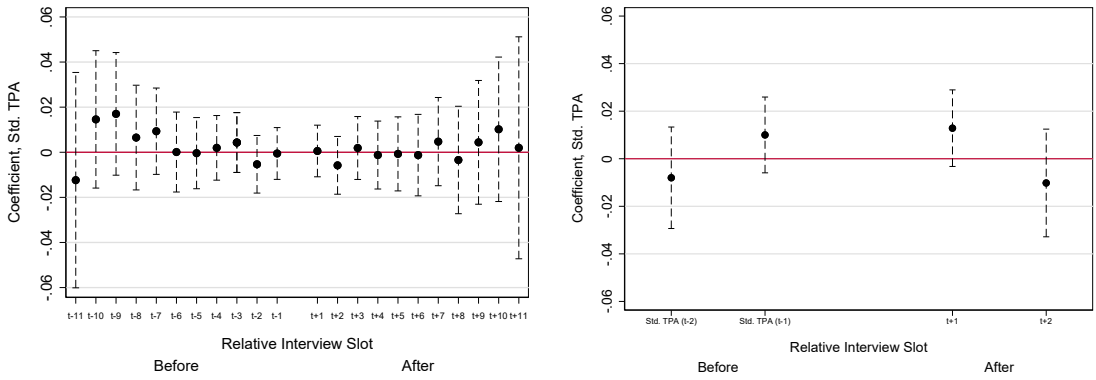


Note: This figure shows the estimated coefficients β_k from equation 1, resulting from separate regressions for each value of $k = \{-11,...,-1,1,...,11\}$, but conditioning on the leave-one-out mean. Therefore, it estimates the additional effect of the candidate interviewed in $t+k$, beyond her contribution to the average quality of the sequence (leave-one-out mean, excluding the candidate in t). 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 3.2 for details). Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level. Appendix Table D.1 reports the corresponding coefficients and p-values.

Figure D.4: Placebo Check: Relationship between TPA_t and TPA_{t+k}

(a) Admission Process

(b) Hiring Process



Note: The coefficients measure how the standardized TPA of the candidate interviewed in period $t+k$ is related to the standardized TPA of the candidate interviewed in period t. TPA = third-party assessment of candidate quality (see Section 3.2 for details). The underlying regression is analogous to the main regression described by equation 1, where the outcome variable is replaced by TPA_t. All regressions include workshop fixed effects (Panel (a))) or candidate pool fixed effects (Panel (b)). Controls include candidate characteristics, evaluator characteristics and interview order. Moreover, we control for exclusion bias using the leave-one-out mean TPA at the interview sequence level in Panel (a) and at the candidate pool level in Panel (b). Coefficients on lags and leads are by construction close to symmetric, as they capture the autocorrelation in TPA. Small differences are due to the control variables. Dashed lines show 95% confidence intervals, with standard errors clustered at the workshop level (Panel (a)), N=312) or candidate pool level (Panel (b), N=63).

Table D.3: Influence of Previous Candidate Quality: Robustness to Sample & Specification

	Admission Process Std. Rating (t)		Hiring Process Std. Rating (t)	
	(1)	(2)	(3)	(4)
Panel A: Baseline				
Std. TPA (t-1)	-0.093***	-0.091***	-0.113***	-0.117***
	(0.006)	(0.006)	(0.015)	(0.015)
Std. TPA (t)	0.360***	0.349***	0.251***	0.234***
	(0.006)	(0.006)	(0.017)	(0.018)
Std. leave-two-out mean TPA	-0.102***	-0.102***		
	(0.008)	(0.008)		
Panel B: Without marginal candidates				
Std. TPA (t-1)	-0.094***	-0.094***		
	(0.007)	(0.007)		
Std. TPA (t)	0.346***	0.333***		
	(0.008)	(0.008)		
Std. leave-two-out mean TPA	-0.101***	-0.102***		
	(0.010)	(0.009)		
Panel C: Only interview sequences with 3 candidates				
Std. TPA (t-1)			-0.116***	-0.116***
			(0.019)	(0.018)
Std. TPA (t)			0.266***	0.251***
			(0.018)	(0.019)
Panel D: Estimation with interviewer FE				
Std. TPA (t-1)	-0.062***	-0.061***	-0.107***	-0.113***
	(0.006)	(0.006)	(0.012)	(0.012)
Std. TPA (t)	0.389***	0.376***	0.258***	0.239***
	(0.006)	(0.006)	(0.014)	(0.014)
Panel E: Estimation with candidate FE				
Std. TPA (t-1)	-0.084***	-0.082***	-0.128***	-0.121***
	(0.007)	(0.007)	(0.024)	(0.023)
Std. leave-two-out mean TPA	-0.069***	-0.070***		
	(0.009)	(0.009)		
Controls	No	Yes	No	Yes
N	26970	26970	5165	5165

Note: TPA = third-party assessment of candidate quality (see Section 3.2 for details). In Columns 1 and 2, the leave-two-out mean is computed at the level of the evaluator's interview sequence. All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects and an indicator of gender. 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 changed during the final committee meeting. In Panel D, the leave-two-out mean TPA is omitted due to collinearity with interviewer fixed effects. In Panel E, the candidate's own TPA is omitted due to collinearity with candidate fixed effects. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table D.4: Influence of Previous Candidate Quality: Robustness to Quality Measure

	Admission Process Std. Rating (t)		Hiring Process Std. Rating (t)	
	(1)	(2)	(3)	(4)
Panel A: Baseline (Quality measure = average TPA)				
Std. TPA (t-1)	-0.093***	-0.091***	-0.113***	-0.117***
	(0.006)	(0.006)	(0.015)	(0.015)
Std. TPA (t)	0.360***	0.349***	0.251***	0.234***
	(0.006)	(0.006)	(0.017)	(0.018)
Std. leave-two-out mean TPA	-0.102***	-0.102***		
	(0.008)	(0.008)		
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$			0.450	0.499
Panel B: Quality measure = TPA from group discussion rating only				
Std. TPA (t-1)	-0.057***	-0.056***		
	(0.006)	(0.006)		
Std. TPA (t)	0.212***	0.201***		
	(0.007)	(0.007)		
Std. leave-two-out mean TPA	-0.070***	-0.069***		
	(0.008)	(0.008)		
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$	0.268	0.278		
Panel C: Quality measure = TPA from other interview rating only				
Std. TPA (t-1)	-0.084***	-0.083***		
	(0.006)	(0.006)		
Std. TPA (t)	0.344***	0.330***		
	(0.007)	(0.007)		
Std. leave-two-out mean TPA	-0.081***	-0.082***		
	(0.007)	(0.007)		
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$	0.244	0.253		
Panel D: Quality measure = average sub-score				
Std. TPA (t-1)			-0.120***	-0.123***
			(0.014)	(0.014)
Std. TPA (t)			0.286***	0.268***
			(0.014)	(0.015)
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$			0.421	0.457
Panel E: Quality measure = predicted rating based on candidate characteristics				
Std. Pred. Rating (t-1)	-0.041***	-0.040***	-0.047***	-0.041***
	(0.007)	(0.007)	(0.013)	(0.014)
Std. Pred. Rating (t)	0.221***	0.219***	0.101***	0.102***
	(0.007)	(0.007)	(0.016)	(0.015)
Std. leave-two-out mean Pred. Rating	-0.053***	-0.059***		
	(0.010)	(0.010)		
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$	0.185	0.183	0.463	0.407
Panel F: Quality measure = Average residualized TPA				
Std. Residualized TPA (t-1)	-0.077***	-0.077***	-0.105***	-0.121***
	(0.006)	(0.006)	(0.016)	(0.016)
Std. Residualized TPA (t)	0.267***	0.256***	0.251***	0.223***
	(0.007)	(0.006)	(0.019)	(0.020)
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$	0.290	0.300	0.419	0.542
Controls	No	Yes	No	Yes
N	26970	26970	5165	5165

Note: TPA = third-party assessment of candidate quality (see Section 3.2 for details). The leave-two-out mean is computed at the level of the evaluator's interview sequence. All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects and a gender indicator. Controls include candidate characteristics, evaluator characteristics and interview order. In Panel E, we predict ratings by regressing the rating on characteristics of the candidates, while leaving out the workshop/interview day itself. In Panel F, the main TPA measure is residualized from the influence of the previous candidate's quality, using a sample split approach, where the split is done at the half of the sample period. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). $^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$.

Table D.5: Influence of Previous Candidate Quality: IV Estimation

	Admission Process Std. Rating (t)		Hiring Process Std. Rating (t)	
	(1)	(2)	(3)	(4)
Panel A: Instrument one TPA measure with other TPA measure				
Std. TPA (t-1)	-0.295***	-0.286***	-0.544***	-0.563***
	(0.034)	(0.034)	(0.095)	(0.101)
Std. TPA (t)	1.010***	1.013***	0.940***	0.937***
	(0.014)	(0.015)	(0.062)	(0.070)
Std. leave-two-out mean TPA	-0.283***	-0.282***		
	(0.035)	(0.034)		
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$	0.292	0.283	0.578	0.601
F-stat (weak ID)	215.22	223.78	65.18	51.59
Panel B: Instrument average TPA measure with predicted rating				
Std. TPA (t-1)	-0.178***	-0.173***	-0.368***	-0.353***
	(0.028)	(0.028)	(0.125)	(0.119)
Std. TPA (t)	0.867***	0.877***	0.851***	0.850***
	(0.022)	(0.022)	(0.113)	(0.104)
Std. leave-two-out mean TPA	-0.173***	-0.194***		
	(0.032)	(0.032)		
$\frac{\text{coeff(t-1)}}{\text{coeff(t)}}$	0.205	0.197	0.433	0.415
F-stat (weak ID)	215.78	214.84	26.22	20.83
Controls	No	Yes	No	Yes
N	26970	26970	5165	5165

Note: TPA = third-party assessment of candidate quality (see Section 3.2 for details). Panel A reports results from using one TPA measure to instrument the other. In the admission process, we use the TPA from the group discussion to instrument the TPA from the other interview. In the hiring process, we use the rating from the candidate's first interview as an IV for the rating from her second or third interview. Panel B reports results from using the predicted quality measure (based on predetermined characteristics) to instrument the baseline TPA measure. All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects and a gender indicator. Controls include candidate characteristics, evaluator characteristics and interview order. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E Additional Material: Autocorrelation

Overview Table E.1 shows estimates of the autocorrelation from specifications with candidate or evaluator fixed effects. Figure E.1 shows the size of the autocorrelation beyond t-1. Figure E.2 documents the autocorrelation and share of reversed decisions for different segments of the quality distribution.

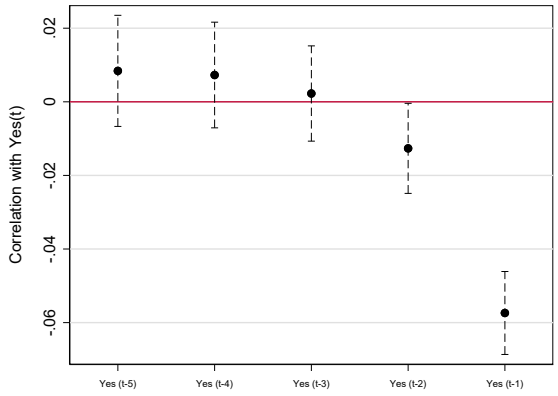
Table E.1: Robustness of Autocorrelation to Candidate and Interviewer FE

	Admission Process		Hiring Process	
	(1) Yes (t)	(2) Yes (t)	(3) Yes (t)	(4) Yes (t)
Yes (t-1)	-0.054***	-0.139***	-0.105***	-0.114***
	(0.008)	(0.006)	(0.019)	(0.013)
Candidate FE	Yes	No	Yes	No
Interviewer FE	No	Yes	No	Yes
Outcome Mean	0.37	0.37	0.31	0.31
N	26970	26970	5165	5165

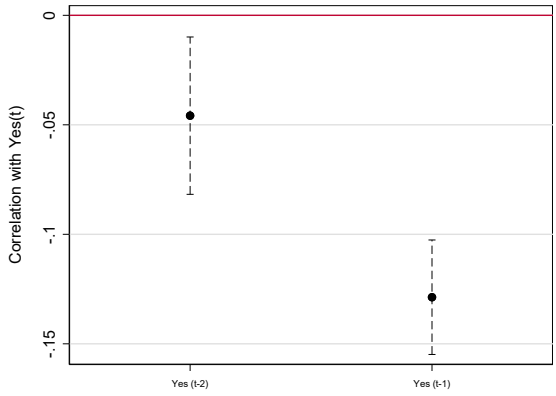
Note: Estimates are based on equation 2, replacing candidate pool fixed effects with candidate fixed effects (Columns 1 & 3) or evaluator fixed effects (Columns 2 & 4). All regression include the evaluator's leave-one-out mean assessment, as well as controls for candidate characteristics, evaluator characteristics and interview order. The increase in the estimated size of the autocorrelation in Columns 2& 4 is expected, given the downward bias in autoregressive models estimated on finite panels (Nickell, 1981). Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). $^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$.

Figure E.1: Autocorrelation Beyond t-1

(a) Admission Process



(b) Hiring Process



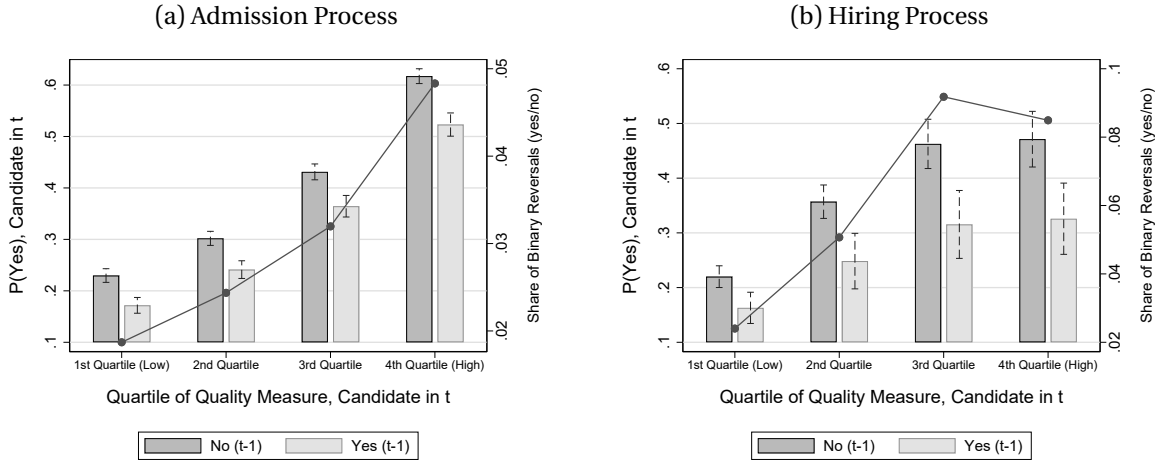
Note: Each coefficient results from a separate regression of equation 2, where the evaluator's vote/recommendation on the candidate interviewed in t is related to her vote/recommendation on the candidate interviewed in a given previous period. Dashed lines show The dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63).

Binary Decision Reversals by Candidate Quality Based on the estimated autocorrelaton, we perform a back-of-the-envelope calculation that captures the reversal of evaluator decisions from a yes to a no vote and vice versa for different segments of candidate quality.[1]

Our computation of reversals follows Chen et al. (2016). Consider that the probability of receiving a yes vote or hiring recommendation can be described by the linear probability model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$. 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, we distinguish two types of reversals: if the previous candidate received a 'no', the negative autocorrelation increases the current candidate's probability of a 'yes' by $\beta_0 - P(Y = 1)$. If the previous candidate received a 'yes', the probability to receive a 'yes' changes by $P(Y = 1) - (\beta_0 + \beta_1)$. The expected number of reversals is the weighted instance of the two cases: $(\beta_0 - P(Y = 1))P(Y_{t-1} = 0) + (P(Y = 1) - (\beta_0 + \beta_1))P(Y_{t-1} = 1)$.

Figure E.2 illustrates the autocorrelation and decision reversals by quartiles of measured candidate quality. In the admission process (Panel (a))), the share of binary reversals is about 2% for candidates in the lowest quartile and increases to 5% in the highest quartile. A similar overall pattern can be observed in the hiring process (Panel (b)). However, stronger autocorrelation in this process also induces more reversals (up to 10% for candidates of above median quality). In general, the quantification exercise reveals that the autocorrelation is strong enough to shift a significant fraction of candidates across the threshold for a given binary decision.

Figure E.2: Autocorrelation and Decision Reversals by Candidate Quality



Note: Candidate quality is measured through an independent third-party assessment (see Section 3.2 for details). The dashed lines show 95% confidence intervals. The black dots display the back-of-the envelope reversal rate, i.e., the share of binary evaluator decisions which switch due to the autocorrelation.

[1]Note that this reversal rate captures only the yes-no margin of the previous candidate's influence, and thereby only provides a lower bound of the overall impact along the full rating scale.

F Additional Material: Role of Experience and Similarity

Table F.1: Influence of Average Candidate Quality in Past Year

	Yes (t)	
	(1) Admission Process	(2) Hiring Process
Standardized values of TPA	0.148***	0.088***
	(0.006)	(0.008)
Std. TPA (t-1)	-0.038***	-0.044***
	(0.007)	(0.007)
Leave-two-out Mean TPA (std.)	-0.044***	
	(0.007)	
Average TPA in Prev. Year (Std.)	0.002	0.008
	(0.008)	(0.007)
Controls	Yes	Yes
Outcome Mean	0.35	0.30
N	5549	4583

Note: "Average TPA in Prev. Year"= average TPA of candidates seen in the previous academic year (admission process) or during the past 365 days (hiring process). Regressions only include evaluators who conducted interviews in the past year. All regression include the evaluator's leave-one-out mean assessment, as well as controls for candidate characteristics, evaluator characteristics and interview order. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table F.2: Heterogeneity: Evaluator Characteristics

	Admission Process			Hiring Process	
	(1) Yes (t)	(2) Yes (t)	(3) Yes (t)	(4) Yes (t)	(5) Yes (t)
Yes (t-1)	-0.053***	-0.058***	-0.066***	-0.141***	-0.097***
	(0.008)	(0.007)	(0.008)	(0.018)	(0.023)
Yes (t-1) x Female Evaluator	-0.009			0.038	
	(0.012)			(0.031)	
Yes (t-1) x Interviewer Training		0.003			
		(0.015)			
Yes (t-1) x Age above Median			0.018		
			(0.012)		
Yes (t-1) x Managerial Responsibility					-0.047
					(0.030)
Controls	Yes	Yes	Yes	Yes	Yes
Outcome Mean	0.37	0.37	0.37	0.31	0.31
N	26970	26970	26970	5165	5165

Note: "Training" equals one if the evaluator participated in a 2-days interviewer training offered by the study grant program. All regression include the evaluator's leave-one-out mean assessment, as well as controls for candidate characteristics, evaluator characteristics and interview order. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

G Additional Material: Contrast Effect with Associative Recall

G.1 Additional Material: Contrast Effects Framework

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

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

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

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

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

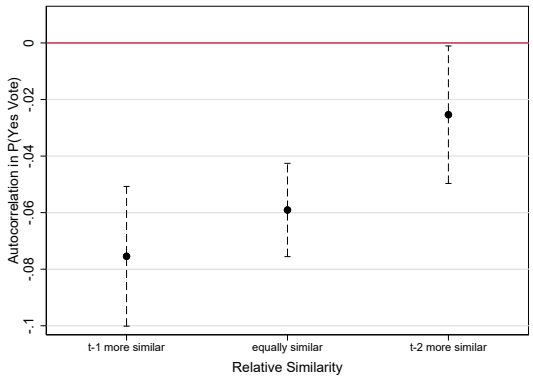
The relative importance of anchoring versus contrasting depends on the size of the quality difference $q_t - q_t^n$ and its implied salience. If the difference is small, it does not capture the evaluator's attention. Anchoring is thus relatively more important and can lead to the assimilation of two subsequent candidates. For larger differences, contrasting as described by the second and third part dominates. Note that the salience of small quality differences is likely to be setting-specific and depend on the evaluator's decision problem. Given that the goal of interviewing is to differentiate between candidates, it is likely that evaluators pay strong attention even to small quality differences in this setting.

In Appendix H.5.4, we provide estimates of a model variant with anchoring to the norm. The model overall provides a slightly lower fit with the data.

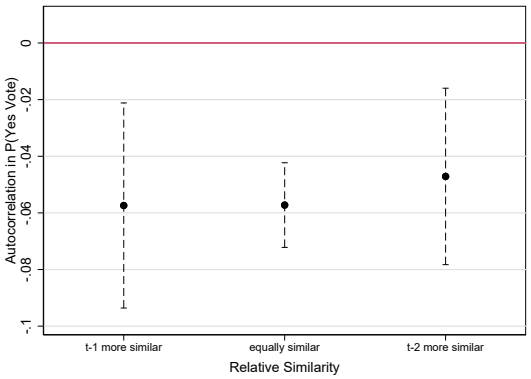
G.2 Additional Material: Reduced-form Evidence

Figure G.1: The Role of Relative Similarity in Characteristics (Admission Process)

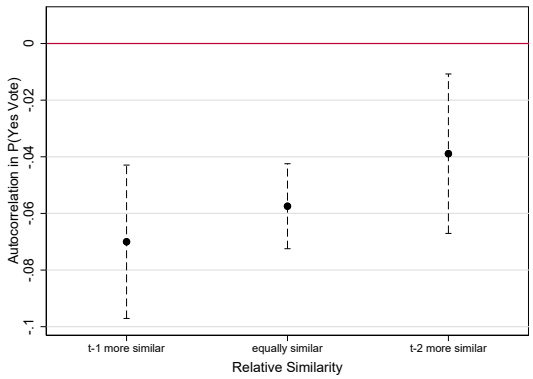
(a) Gender



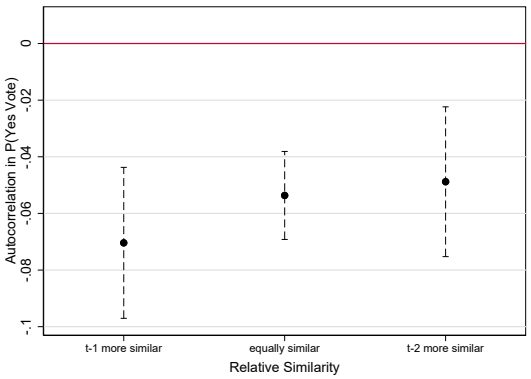
(b) Migration Background



(c) 1st Generation Student



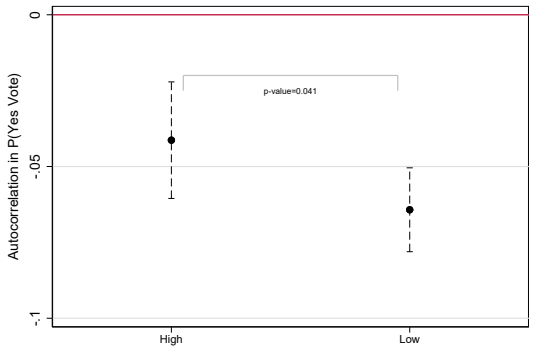
(d) Study Field



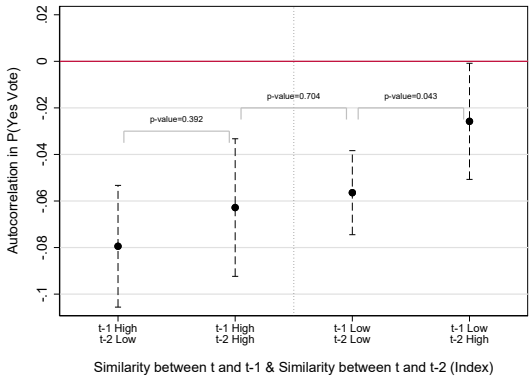
Note: The figures present estimates of the autocorrelation between t and t-1 based on equation 2, where the vote of the candidate in t-1 is interacted with her relative similarity to the candidate in t in a given observable characteristics. “t-1 more similar" = the candidate in t-1, but not the candidate 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. The dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop level (N=312).

Figure G.2: Additional Evidence on the Role of Relative Similarity (Admission Process)

(a) Role of Similarity (t-2)



(b) Role of Similarity t-1 vs. t-2



Note: The figures present estimates of the autocorrelation between t and t-1 based on equation 2, where the vote of the candidate in t-1 is interacted with different categories measuring the similarity between the candidate in t and the candidates in t-1 and/or t-2. High/low similarity is defined by a median split of a similarity index, counting the number of observable characteristics which the candidate in t and the candidate in t-1/t-2 have in common (gender, migration status, first generation status and study field). The dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop level (N=312).

Table G.1: Previous Candidate's Influence and Size of the Quality Difference

	Yes (t)	
	(1) Admission Process	(2) Hiring Process
Std. TPA (t-1)	-0.012**	-0.016
	(0.006)	(0.013)
Std. TPA (t-1) x Large Difference	-0.018**	-0.051***
	(0.007)	(0.017)
Controls	Yes	Yes
Outcome Mean	0.37	0.31
N	26970	5165

Note: TPA = third-party assessment of candidate quality (see Section 3.2 for details). "Large Difference" equals one if the absolute TPA difference between the candidate in t and t-1 is equal to or larger than the median. All regressions include workshop (Column 1) or candidate pool (Column 2) fixed effects. Controls include candidate characteristics, evaluator characteristics and interview order. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

H Additional Material: Structural Estimation

H.1 Parameterization

Based on the framework in Section 6.1, we parameterize an evaluator's instantaneous valuation of a candidate interviewed in period t as:[1]

$$V_t = \alpha \times \tilde{q}_t + \sigma(\tilde{q}_t, q_t^n) \times (\tilde{q}_t - q_t^n) + u_t \tag{3}$$

The valuation depends on the candidate's own measured quality $\tilde{q}_t$ and on the difference between $\tilde{q}_t$ and the quality norm q_t^n, multiplied by its salience $\sigma(\tilde{q}_t, q_t^n)$. u_t denotes a normally distributed implementation error (see Section H.2 for details).

We parameterize the quality norm q_t^n as a weighted average of the qualities of the preceding candidates: $q_t^n(c_t) = \sum_{l=1}^{t-1} \tilde{q}_{t-l}\omega_{t-l}$. The weight ω_{t-l} of the candidate observed in the period $t-l$, for $l = 1, 2, \ldots, t-1$, is determined by her relative similarity to the current candidate: $\omega_{t-l} = \frac{S_{t-l}(c_t, c_{t-l})}{\sum_{m=1}^{t-1} S_{t-m}(c_t, c_{t-m})}$, where $S_{t-l}(c_t, c_{t-l})$ denotes the similarity between t and $t-l$. In the baseline specification, similarity is only defined by the time dimension. Similarity in time exponentially decreases in the lag l between two interviews, at a rate determined by the parameter δ_1: $S_{t-l}^{time}(l) = e^{-\delta_1(l-1)}$.

For the admission process, we also estimate an extended specification in which similarity additionally depends on observable candidate characteristics. These two dimensions of similarity are multiplicatively separable, such that similarity is defined by $S_{t-l}(c_t, c_{t-l}) = S_{t-l}^{time}(l) \times S_{t-l}^{char}((\mathbb{1}_{diff})_{t-l})$. For simplicity, the similarity in characteristics enters in binary terms: $S_{t-l}^{char}((\mathbb{1}_{diff})_{t-l}) = e^{-\delta_2(\mathbb{1}_{diff})_{t-l}}$. The indicator $(\mathbb{1}_{diff})_{t-l}$ equals one if the candidate in $t-l$ differs from the current candidate in terms of her observable characteristics.[2] If this is the case, the similarity in time is multiplied by the factor $e^{-\delta_2} < 1$ and, thus, absolute similarity is reduced.

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

The data do not report the evaluators' valuations, but their ratings on a discrete 1-10 scale

[1] In the admission process, the final valuation is arguably composed of the instantaneous valuation as modeled here and an additional term that captures the possibility of ex post adjustments, i.e., the influence of the average quality of the other candidates in the sequence. For the sake of simplicity, we abstract from this in the estimation and calculate the empirical moments conditional on the leave-one-out mean quality of the sequence.

[2] In line with the reduced-form analysis in Section 5, we construct an index counting the number of shared observable characteristics. $(\mathbb{1}_{diff})_{t-l}$ equals one if two candidates share not more than the median number (i.e., two) of characteristics.

in the admission process and a 1-3 scale in the hiring process. Therefore, we need to map the latent valuations to the observed ratings. To this end, we bin the ordered (simulated) valuations into groups corresponding to the share of candidates that receive a given rating in the observed distribution.[3]

In addition to the model presented, we estimate two benchmark models. First, we estimate a model with $\delta_1 = 0$ and $\delta_2 = 0$. This eradicates associative recall, such that all previous candidates receive the same weight in the norm. Second, we replace the quality norm with the expected quality (i.e., the sample average). This eradicates recall in general.

H.2 Estimation

We use a minimum-distance estimator to estimate the model described in Section H.1. Let $m(\xi)$ denote the vector of simulated moments as a function of the model parameters, and $\hat{m}$ the vector of moments observed in the data. The estimator chooses the parameter vector $\hat{\xi}$ that minimizes the distance $(m(\xi) - \hat{m})' W (m(\xi) - \hat{m})$. As a weighting matrix W, we use the diagonal of the inverse of the variance-covariance matrix.[4]

We estimate the variance of the parameters using

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

where $\hat{G} \equiv \nabla_\xi m(\xi)$, $\hat{\Sigma} \equiv Var(m(\xi))$,[5] J_m is the number of empirical observations used to calculate the moment and J_s is the corresponding number of simulated observations.

We calibrate the standard deviation of the error term to a value of 1.68 for the admission process and 0.68 for the hiring process, which corresponds to the respective standard deviation of the residual rating (conditional on own, previous and leave-one-out mean measured quality). We fix the draw of errors across all estimations.

In every simulation step, we simulate a population of 10,000 evaluators, each interviewing 12 candidates in the admission process, and 40,000 evaluators who interview 3 candidates in the hiring process. We solve the minimization problem using a Python implementation of the DFO-LS algorithm (Gabler, 2021). We impose the following box constraints on the parameters: $\alpha \geq 0$ (own quality), $\sigma > 0$, $\delta_1 > 0$ and $e^{-\delta_2} \in (0, 1]$. We calibrate θ to different values.

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

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

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

To increase our confidence in the identification and estimation of the structural parameters, we simulate a set of 3,000 evaluators with given model parameters, which corresponds roughly to our actual sample size. We start our estimation procedure at a perturbed initial value and check that the estimator is able to back out the original parameters that were used to simulate the data.

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

H.3 Identification

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

The variation in the influence of the preceding candidates identifies the importance of relative time for recall based on similarity, as described by δ_1. More precisely, the moments describe how the ratings respond to the quality of the candidates in the three preceding interview slots. We capture this through the coefficients of separate OLS regressions that link a current candidate's rating to a dummy indicating whether a given previous candidate was of high measured quality (i.e., in the highest quality quartile).

The role of similarity in candidate characteristics, as captured by δ_2, is identified through moments that describe the interaction between the influence of the previous candidate and her relative similarity in terms of observed characteristics. We only estimate δ_2 for the admission process, where we observe enough characteristics of the candidates and have the statistical power to measure their impact. A distinctive feature of the model is that similarity enters in relative terms. To capture this notion, we allow the influence of the previous candidate to depend on how similar the candidate in t-1 is compared to the candidate in t-2, who is still recent and provides a possible point of comparison in case the candidate in t-1 lacks similarity. We construct three cases (in analogy to Figure G.1): (i) the candidate in t-1 is more similar to the candidate in t (high relative similarity); (ii) the candidates in t-1 and t-2 are equally similar to the candidate in t (medium relative similarity); (iii) the candidate in t-2 is more similar to the candidate in t (low relative similarity). Again, we use the coefficients of an OLS regression that links the current candidate's rating to the previous candidate's quality and interacted with her relative similarity. The model parameter δ_2 determines how the influence of the previous candidate differs between these three cases.

To identify the parameters of the salience function, we use the coefficients of a regression

estimating how assessments react to the difference between the current candidate's measured quality and the measured quality of the previous candidate (controlling for its own quality). As discussed above, the parameters σ and θ determine the shape of this relationship. To simplify the identification of σ, we calibrate θ, thus fixing the range where the differences are not fully salient. In the main estimation, we use $\theta = 70$. We explore the robustness of this calibration in Section H.5.1.

Finally, we identify α through moments that describe how the rating of a candidate varies with her own quality measure. More specifically, we use the average rating per quartile of the current candidate's measured quality.

We restrict the estimation of moments to candidates for whom we observe a candidate 3 periods before in the admission process and 2 periods before in the hiring process. To account for the level of randomization and in line with the reduced-form analysis, all moments are computed conditional on workshop/candidate pool fixed effects and candidate characteristics. Moreover, moments that target δ_1, δ_2, and σ control for their own measured quality in both processes and leave-one-out mean measured quality in the admission process.

H.4 Results

Parameter Estimates & Model Fit Table H.1 presents the parameter estimates for both processes. Figure H.1 illustrates the model fit with the empirical moments.

Overall, the results show that a simple parameterization of the evaluators' recall process can provide quantitatively reasonable predictions regarding the influence of previous candidates. The baseline estimates of δ_1 are strikingly consistent across the two settings, ranging between 1.2 in the admission process (Column 1) and 1.5 in the hiring process (Column 3). They imply a strong decline in similarity with an increasing time lag, resulting in a high weight of the previous candidate in the norm.

When adding the second dimension of similarity for the admission process (Column 2), the interpretation of δ_1 changes. It now describes how similarity in time evolves for candidates with high similarity in characteristics. The estimate of $e^{-\delta_2}$, i.e., the factor by which similarity in time is multiplied if two candidates are observationally more different, amounts to about 0.2. Jointly, the estimates of δ_1 and δ_2 imply that the previous candidate's weight varies strongly with her relative similarity in characteristics: the previous candidate has a weight of about 93% if she is more similar to the current candidate than the candidate in t-2, but only about 48% if her relative similarity is low (i.e., the candidate in t-2 is more similar). Figure H.1 reveals that the estimated recall process closely matches the corresponding empirical moments almost perfectly.

Another element of the model is the relationship between the differences in quality and the valuation, as described by the salience function. As discussed above, the proposed functional form is characterized by the parameters σ and θ. We estimate σ and calibrate θ to the

value of 70 (see Appendix H.5.1 for robustness checks with respect to different calibrations of θ). Note that the estimates of σ (and α) are not directly interpretable and comparable across the two settings, due to different quality measures and error structures. Panel (c) and Panel (d) of Figure H.1 show that the structural estimates provide overall a good fit with the empirical evidence. For the admission process, we note that they slightly over-predict the effect of large quality differences at the left boundary. One possible explanation is the nature of the evaluation process: while the model predicts a strong impact of large quality differences, interviewers might be reluctant to implement this in their ratings.

Robustness of Estimates Additionally, we provide estimates from robustness checks. In Section H.5.1 we show estimates for alternative calibrations of θ. In Section H.5.2, we use the identity matrix to weight the moments.

Comparison to Benchmark Models In Section H.5.3 we provide estimates from two benchmark models. In the first benchmark, we eradicate associative recall by setting $\delta_1 = \delta_2 = 0$. We thus assume that the quality norm is the average quality of all prior candidates. This assumption results in a considerably worse overall fit with the empirical moments (SSE=325 (61) vs. 129 (15) in the admission (hiring) process). The second benchmark assumes that the norm is not formed through recall, but instead consists of the expected candidate quality. This model has no chance of predicting the influence of previous candidates (see also Figure H.2), which further reduces the fit (SSE=576/SSE=203). Taken together, the estimation of the two benchmark models shows that associative recall is key to explain the empirical pattern.

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

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

Table H.6 provides the resulting parameter estimates and model fit. The estimates of the similarity parameters are close to those in the baseline model. The estimate of σ is naturally higher, as it now is the only parameter that captures the influence of own quality.

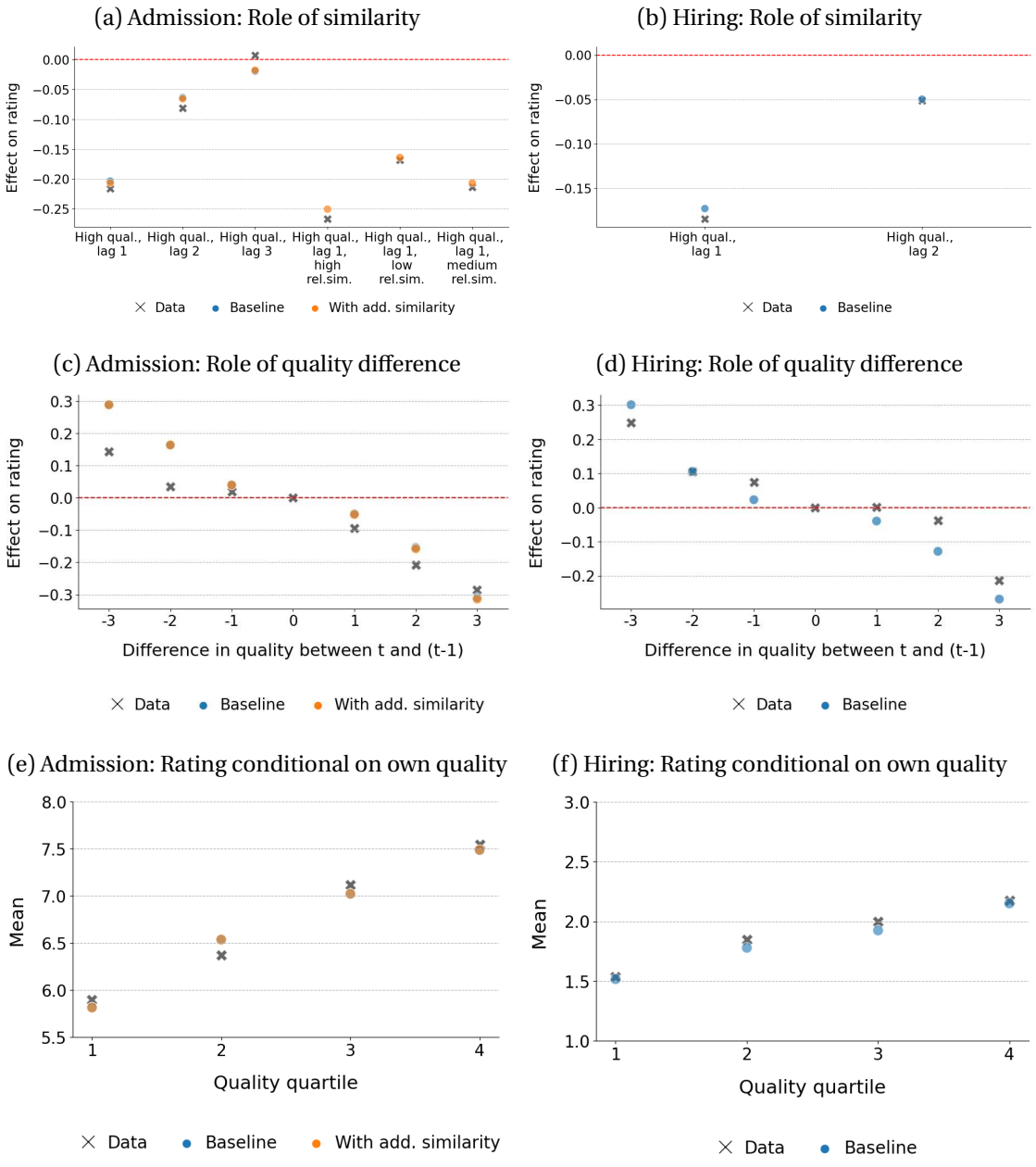
The model captures the relative importance of previous candidates equally well, but the role of quality differences is slightly worse. This is because the model predicts small assimilation effects that are not observable in the data.

Table H.1: Structural Estimates

	Admission		Hiring
	Baseline (1)	With add. similarity (2)	Baseline (3)
Recall Parameters			
δ_1	1.209	1.413	1.533
	(0.383)	(0.355)	(0.852)
$e^{-\delta_2}$		0.247	
		(0.313)	
Weights of Previous Candidates			
ω_{t-1}	0.721	0.724	0.822
ω_{t-2}	0.215	0.218	0.178
ω_{t-3}	0.064	0.057	
ω_{t-1} | high rel. sim		0.917	
ω_{t-1} | medium rel. sim		0.75	
ω_{t-1} | low rel. sim		0.473	
Valuation Parameters			
α	0.169	0.169	0.486
	(0.018)	(0.009)	(0.148)
σ	0.134	0.134	1.321
	(0.034)	(0.016)	(0.147)
$\theta^{\dagger}$	70.0	70.0	70.0
Weighted SSE	128.908	128.227	14.815
Number of moments	13	16	12

Note: The table shows estimates of the parameters in equation 3, with standard errors in brackets. Columns (1) and (3) report estimates from the baseline model, where similarity is only based on the time dimension. Column (2) reports estimates from an extension that includes similarity in terms of candidate characteristics. The second salience parameter θ is calibrated to 70. The weights describe the weight that a previously interviewed candidate receives in the quality norm. They are calculated for a candidate at position 4 (3) in the interview sequence for the admission (hiring) process. "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/equally/less similar to the candidate in t than the candidate in t-2. The estimates of σ and α are not comparable across the two processes, due to different quality measures and error structures. Estimation is based on the method of simulated moments (see Appendix H.2 for details). † = calibrated. SSE= Sum of Squared Errors.

Figure H.1: Empirical Moments and Model Fit: Main Specifications



Note: This figure illustrates the model fit for the estimates reported in Table H.1. In Panels (a) and (b), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. "rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/equally/less similar to the candidate in t than the candidate in t-2. In Panels (c) and (d), the moments describe the effect of a given quality difference between the candidates in t and t-1. In Panels (e) and (f), they describe the average rating conditional on the quartile of own measured quality..

H.5 Robustness of Estimates and Additional Specifications

H.5.1 Calibration of θ

Table H.2: Structural Estimates: Calibration of θ (Admission Process)

	Admission					
	Baseline			With add. similarity		
	(1)	(2)	(3)	(4)	(5)	(6)
Recall Parameters						
δ_1	1.225	1.256	1.209	1.374	1.466	1.413
	(0.318)	(0.258)	(0.383)	(0.417)	(0.254)	(0.355)
$e^{-\delta_2}$				0.247	0.216	0.247
				(0.081)	(0.110)	(0.313)
Weights of Previous Candidates						
ω_{t-1}	0.725	0.732	0.721	0.716	0.728	0.724
ω_{t-2}	0.213	0.209	0.215	0.223	0.216	0.218
ω_{t-3}	0.063	0.059	0.064	0.061	0.055	0.057
ω_{t-1} | high rel. sim				0.913	0.929	0.917
ω_{t-1} | medium rel. sim				0.741	0.759	0.75
ω_{t-1} | low rel. sim				0.463	0.456	0.473
Valuation Parameters						
α	0.169	0.174	0.169	0.168	0.17	0.169
	(0.014)	(0.009)	(0.018)	(0.010)	(0.013)	(0.009)
σ	0.131	0.138	0.134	0.135	0.141	0.134
	(0.020)	(0.015)	(0.034)	(0.021)	(0.023)	(0.016)
$\theta^{\dagger}$	100.0	30.0	70.0	100.0	30.0	70.0
Weighted SSE	125.836	138.814	128.908	125.711	137.176	128.227
Number of moments	13	13	13	16	16	16

Note: The table shows estimates of the parameters in equation 3, with standard errors in brackets. Columns (1), (3) and (5) report estimates from the baseline model where similarity is only based on the time dimension. Columns (2), (4) and (6) report estimates from an extension that includes similarity in terms of candidate characteristics. The weights are calculated for a candidate at position 4 in the interview sequence. "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/equally/less similar to the candidate in t than the candidate in t-2. Estimation is based on the method of simulated moments (see Appendix H.2 for details). † = calibrated. SSE= Sum of Squared Errors.

Table H.3: Structural Estimates: Calibration of θ (Hiring Process)

	Hiring		
	Baseline		
	(1)	(2)	(3)
Recall Parameters			
δ_1	1.424	1.416	1.533
	(0.700)	(1.115)	(0.852)
$e^{-\delta_2}$			
Weights of Previous Candidates			
ω_{t-1}	0.806	0.805	0.822
ω_{t-2}	0.194	0.195	0.178
ω_{t-1} | high rel. sim			
ω_{t-1} | medium rel. sim			
ω_{t-1} | low rel. sim			
Valuation Parameters			
α	0.467	0.478	0.486
	(0.189)	(0.226)	(0.148)
σ	1.293	1.628	1.321
	(0.185)	(0.268)	(0.147)
$\theta^{\dagger}$	100.0	30.0	70.0
Weighted SSE	15.072	15.987	14.815
Number of moments	12	12	12

Note: The table shows estimates of the parameters in equation 3, with standard errors in brackets. Columns (1) and (3) report estimates from the baseline model where similarity is only based on the time dimension. Column (2) reports estimates from an extension that includes similarity in terms of candidate characteristics. The weights are calculated for a candidate at position 3 in the interview sequence. "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). Estimation is based on the method of simulated moments (see Appendix H.2 for details). † = calibrated. SSE= Sum of Squared Errors.

H.5.2 Identity Matrix as Weighting Matrix

Table H.4: Structural Estimates: Identity Matrix

	Admission		Hiring
	Baseline	With add. similarity	Baseline
	(1)	(2)	(3)
Recall Parameters			
δ_1	2.069	1.2	1.026
	(0.533)	(0.585)	(0.733)
$e^{-\delta_2}$		0.263	
		(0.200)	
Weights of Previous Candidates			
ω_{t-1}	0.875	0.679	0.736
ω_{t-2}	0.111	0.243	0.264
ω_{t-3}	0.014	0.078	
ω_{t-1} | high rel. sim		0.887	
ω_{t-1} | medium rel. sim		0.699	
ω_{t-1} | low rel. sim		0.43	
Valuation Parameters			
α	0.199	0.177	0.577
	(0.012)	(0.014)	(0.196)
σ	0.084	0.125	1.225
	(0.016)	(0.021)	(0.297)
$\theta^{\dagger}$	70.0	70.0	70.0
Weighted SSE	0.076	0.083	0.023
Number of moments	13	16	12

Note: The table shows estimates of the parameters in equation 3, with standard errors in brackets. Columns (1) and (3) report estimates from the baseline model where similarity is only based on the time dimension. Column (2) reports estimates from an extension that includes similarity in terms of candidate characteristics. The second salience parameter θ is calibrated to 70. The weights are calculated for a candidate at position 4 (3) in the interview sequence for the admission(hiring) process. "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/equally/less similar to the candidate in t than the candidate in t-2. Estimation is based on the method of simulated moments (see Appendix H.2 for details). The weighting matrix is the identity matrix. † = calibrated. SSE= Sum of Squared Errors.

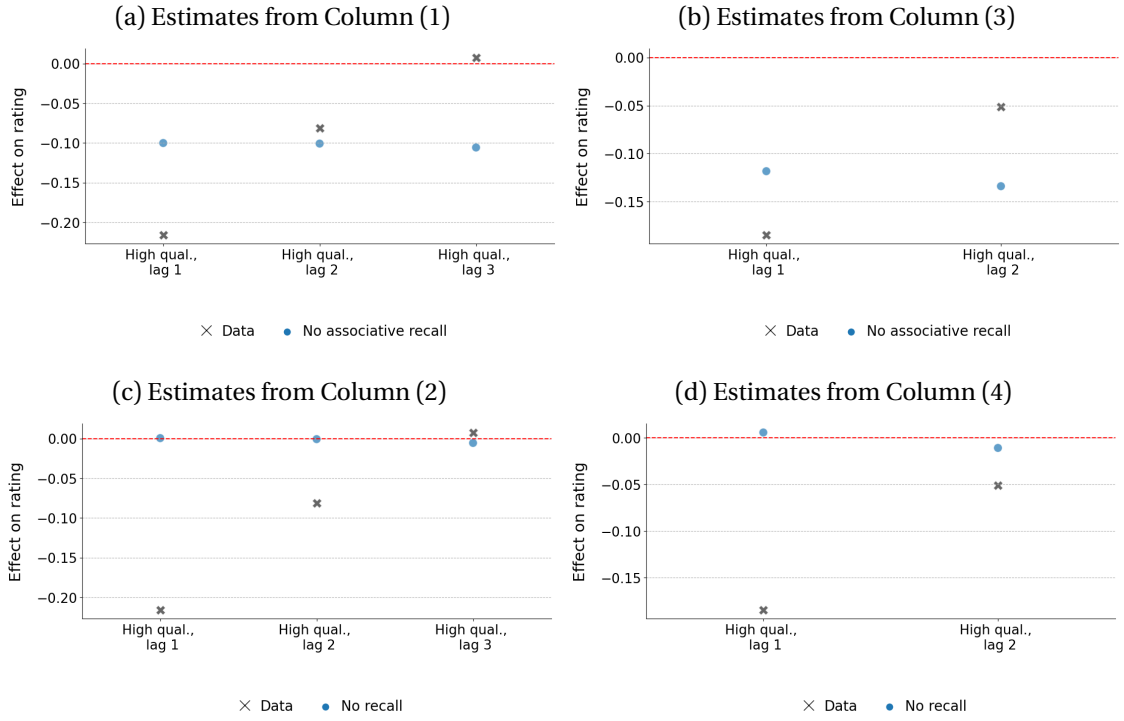
H.5.3 Benchmark Models

Table H.5: Structural Estimates: Benchmark Models

	Admission		Hiring	
	No associative recall	No recall	No associative recall	No recall
	(1)	(2)	(3)	(4)
Valuation Parameters				
α	0.108	0.241	0.403	1.04
	(0.014)	(0.016)	(0.088)	(0.186)
σ	0.275	0.0	1.516	0.179
	(0.032)	(0.025)	(0.127)	(0.420)
$\theta^{\dagger}$	70.0	70.0	70.0	70.0
Weighted SSE	325.379	575.8	61.432	203.284
Number of moments	13	13	12	12

Note: The table shows parameter estimates for two benchmark models. Columns (1) and (3) report estimates from a model where similarity plays no role for recall ($\delta_1 = 0$ and $\delta_2 = 0$). Columns (2) and (4) report estimates from a model where the norm is not formed through recall at all, but consists of the expected quality (sample average). Estimation is based on the method of simulated moments (see Appendix H.2 for details). The second salience parameter θ is calibrated to 70. † = calibrated. SSE= Sum of Squared Errors.

Figure H.2: Empirical Moments and Model Fit: Benchmark Models



Note: This figure illustrates the model fit for the estimates reported in Table H.5. The empirical moments describe the effect of following a high quality candidate, depending on similarity in time.

H.5.4 Model with Anchoring

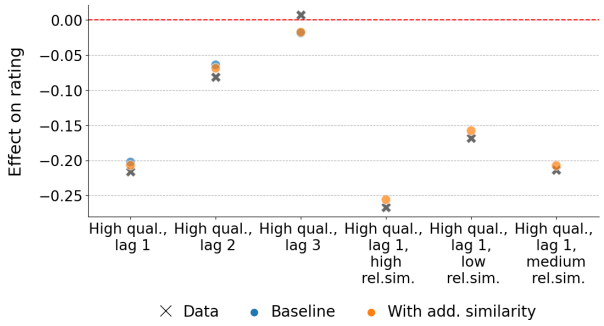
Table H.6: Structural Estimates: Model with Anchoring

	Admission		Hiring
	Baseline	With add. similarity	Baseline
	(1)	(2)	(3)
Recall Parameters			
δ_1	1.256	1.483	1.5
	(0.258)	(0.271)	(0.616)
$e^{-\delta_2}$		0.206	
		(0.077)	
Weights of Previous Candidates			
ω_{t-1}	0.732	0.729	0.818
ω_{t-2}	0.209	0.216	0.182
ω_{t-3}	0.059	0.055	
ω_{t-1} | high rel. sim		0.933	
ω_{t-1} | medium rel. sim		0.762	
ω_{t-1} | low rel. sim		0.449	
Valuation Parameters			
α	0.172	0.166	0.477
	(0.010)	(0.011)	(0.125)
σ	0.489	0.479	2.443
	(0.013)	(0.013)	(0.216)
$\theta^{\dagger}$	70.0	70.0	70.0
Weighted SSE	166.515	158.51	16.444
Number of moments	13	16	12

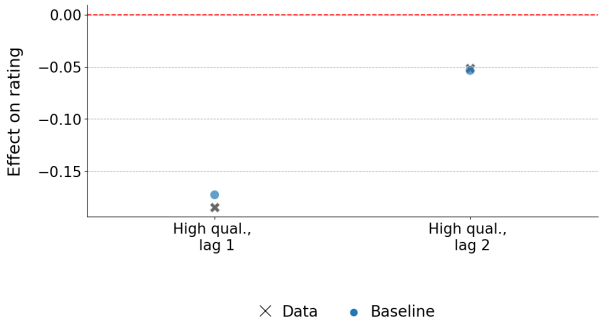
Note: The table shows parameter estimates of the parameters of a model variant with anchoring to the norm (see Appendix G.1 for a description). Columns (1) and (3) report estimates from the baseline model where similarity is only based on the time dimension. Column (2) reports estimates from an extension that includes similarity in terms of candidate characteristics. The second salience parameter θ is calibrated to 70. The weights are calculated for a candidate at position 4 (3) in the interview sequence for the admission(hiring) process. "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/equally/less similar to the candidate in t than the candidate in t-2. Estimation is based on the method of simulated moments (see Appendix H.2 for details). † = calibrated. SSE= Sum of Squared Errors.

Figure H.3: Empirical Moments and Model Fit: Model with Anchoring

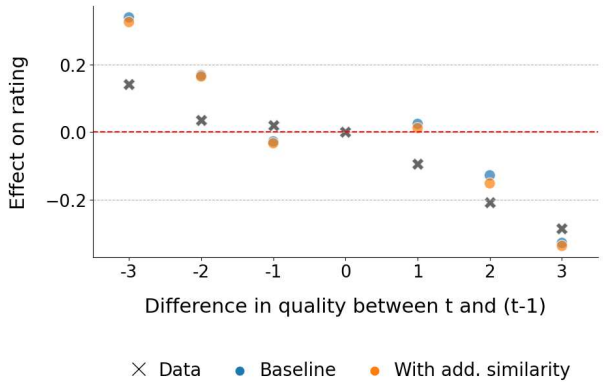
(a) Admission: Role of similarity



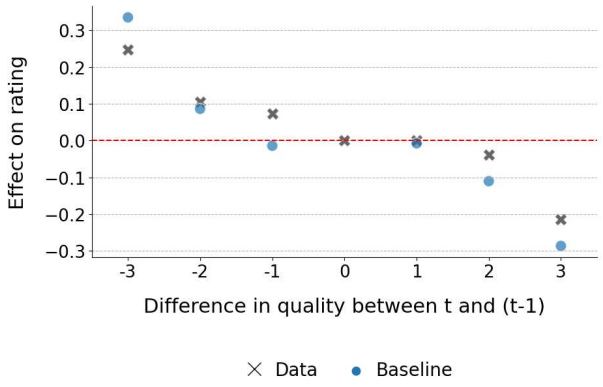
(b) Hiring: Role of similarity



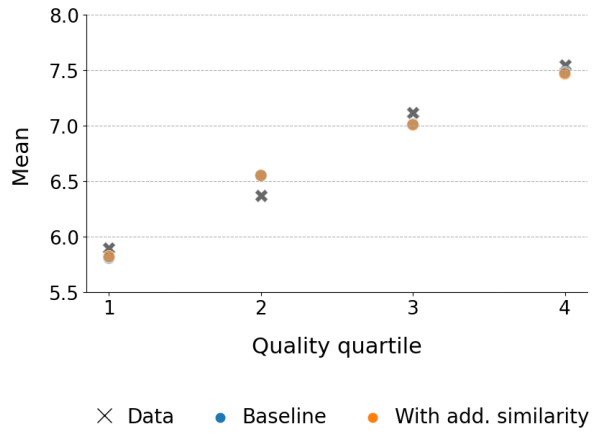
(c) Admission: Role of quality differences



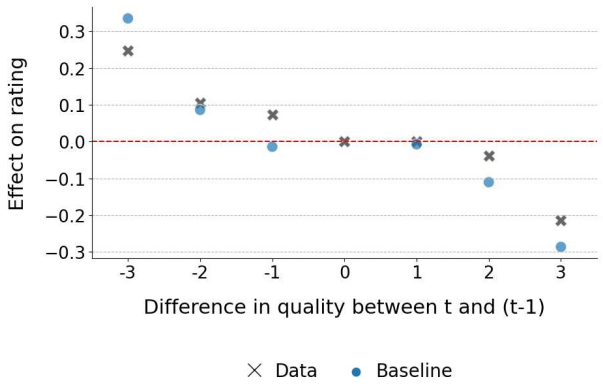
(d) Hiring: Role of quality differences



(e) Admission: Rating conditional on own quality



(f) Hiring: Rating conditional on own quality



Note: This figure documents the model fit for the estimates reported in Table H.6. In Panel (a)) and (b), the empirical moments describe the effect of following a high quality candidate, depending on similarity in time and observable characteristics. "Rel.sim." describes relative similarity in terms of observable characteristics (index including gender, study field, migration status, first generation status). "High/medium/low rel.sim." = the candidate in t-1 is more/equally/less similar to the candidate in t than the candidate in t-2. In Panels (c) and (d), the moments describe the effect of a given quality difference between the candidates in t and t-1. In Panels (e) and (f), they describe the average rating conditional on the quartile of own quality.

I Additional Material: Alternative Mechanisms

Table I.1: Test for the Role of Signal Precision

	Admission Process		Hiring Process	
	(1) Yes(t)	(2) Std. Rating (t)	(3) Yes(t)	(4) Std. Rating (t)
Yes (t-1)	-0.052***		-0.138***	
	(0.008)		(0.017)	
Yes (t-1) x Weak Signal (t-1)	0.001		0.028	
	(0.011)		(0.029)	
Std. TPA (t-1)		-0.087***		-0.121***
		(0.006)		(0.018)
Std. TPA (t-1) x Weak Signal (t-1)		-0.025		0.051
		(0.020)		(0.052)
Outcome Mean	0.37	0.00	0.31	0.01
N	26970	26970	5165	5165

Note: Weak signals are cases where the other two evaluators disagree in their assessment of the previous candidate's quality (i.e., one positive and one negative assessment). All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects, the evaluator's leave-one-out mean rating and share of yes votes, candidate characteristics (including TPA), evaluator characteristics and interview order dummies. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table I.2: Influence of Previous Quality Conditional on Previous Decision

	Admission Process		Hiring Process	
	(1) Yes(t)	(2) Yes(t)	(3) Yes(t)	(4) Yes(t)
Std. TPA (t-1)	-0.025***	-0.017***	-0.046***	-0.037***
	(0.003)	(0.003)	(0.007)	(0.007)
Yes (t-1)		-0.046***		-0.114***
		(0.006)		(0.013)
Outcome Mean	0.37	0.37	0.31	0.31
R-Squared	0.12	0.12	0.07	0.08
N	26970	26970	5165	5165

Note: All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects, the evaluator's leave-one-out mean rating and share of yes votes, candidate characteristics (including TPA), evaluator characteristics and interview order dummies. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table I.3: Test for Additional Influence of Streaks

	Admission Process		Hiring Process	
	(1) Yes(t)	(2) Std. Rating (t)	(3) Yes(t)	(4) Std. Rating (t)
Yes (t-1)	-0.065***		-0.108***	
	(0.008)		(0.023)	
Yes (t-2)	-0.029***		-0.076***	
	(0.008)		(0.023)	
Yes (t-1) x Yes (t-2)	0.018		0.035	
	(0.014)		(0.053)	
Std. TPA (t-1)		-0.086***		-0.120***
		(0.006)		(0.021)
Std. TPA (t-2)		-0.051***		-0.057***
		(0.006)		(0.019)
Std. TPA (t-1) × Std. TPA (t-2)		0.004		-0.009
		(0.006)		(0.024)
Outcome Mean	0.37	0.00	0.31	0.03
N	24474	24474	1893	1893

Note: 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. All regressions include workshop (Columns 1-2) or candidate pool (Columns 3-4) fixed effects, the evaluator's leave-one-out mean decision, candidate and evaluator characteristics and interview order dummies. Standard errors are clustered at the workshop/candidate pool level (N=312/N=63). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

J Additional Material: Policy Responses

J.1 Information & Awareness

Implementation Details

Email communication to workshop organizers:

Starting January 1, 2023, we have expanded the talking points for your briefing of evaluators. The briefing now includes information on potential evaluation errors due to sequential contrast effects. This text module was developed in collaboration with external researchers, whose team will investigate whether awareness among evaluators can reduce the influence of contrast effects. Therefore, we kindly ask you to follow the talking points as closely as possible during your briefings.

Text module in talking points for evaluator briefing:

(Sequential) Contrast Effect:

- Applicants receive higher ratings when directly following a weak candidate and receive lower ratings when following a strong candididate.
- Empirical research has shown that this effect creates significant impacts on admission decisions made in our process, based on data covering the period 2013/14–2016/17.
- For example, the data show that the likelihood of receiving a 'yes' vote depends negatively on the vote given to the previous applicant: the likelihood amounts to 39% among applicants who follow an applicant with a 'no' vote; compared to only 33% for among those who follow an applicant with a 'yes' vote.
- Advice on how to reduce contrast effects: Try to focus on the objective selection criteria when interviewing; change the order of the applicants when determining final ratings at the end; reflect on the circumstances of the interviews when you are deciding on your final ratings.

Table J.1: Effect of Information Treatment (Admission Process)

	Simple Diff		Diff-in-Diff	
	(1) Std. Rating (t)	(2) Std. Rating (t)	(3) Std. Rating (t)	(4) Std. Rating (t)
TPA (std.), t-1	-0.077***	-0.078***	-0.087***	-0.085***
	(0.020)	(0.020)	(0.009)	(0.008)
TPA (t-1) × Jan-Mar			-0.011	-0.011
			(0.011)	(0.011)
TPA (t-1) × 2022/23			0.009	0.008
			(0.022)	(0.021)
TPA (t-1) × Jan-Mar × 2022/23	-0.026	-0.024	-0.014	-0.014
	(0.025)	(0.025)	(0.027)	(0.027)
Controls	No	Yes	No	Yes
Outcome mean	0.01	0.01	0.00	0.00
N	6136	6136	33106	33106

Note: TPA = third-party assessment of candidate quality (see section 3.2 for details). Admission workshops take place from October to March. The information treatment was implemented in the second half of the academic year 2022/23 (January-March). In Columns 3-4, the academic years 2013/14-2016/17 serve as the control group. Appendix Table J.1 shows results using the TPA measure. All regressions include workshop fixed effects. Standard errors are clustered at the workshop level (N=78 in Columns 1-2; N=390 in Columns 3-4). * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

J.2 Increasing the Number of Independent Interviews

Details on Simulation To quantify how an increase in the number of independent interviews reduces the aggregate impact of contrast effects, we start by simulating the average quality of previous candidates. The simulations assume that every interview exposes candidates to a different preceding candidate. We simulate quality based on the empirical distribution of TPA in the data. Increasing the number of independent interviews per candidate mechanically reduces the variation in this average. Using the simulated average quality distributions, we compare how a one standard deviation change in the average quality of previous candidates affects the final decision across different scenarios. The computations are based on the estimates in Table 4 (Section 4.3), which link the average quality of two previous candidates to an individual's admission or hiring probability. Note that this final outcome is also influenced by a third independent assessment, which is not influenced by a previous candidate, either because it is based on the group discussion or because it the first interview in a sequence. Therefore, the simulations reported in Figure 12 express the impact of a 1 SD change in the average quality of the two previous candidates, in a process with one additional assessment.