

Interview Sequences and the Formation of Subjective Assessments

Jonas Radbruch

Humboldt Universität zu Berlin, Berlin, Germany

and

Amelie Schiprowski

University of Bonn, Bonn, Germany

First version received August 2022; Editorial decision January 2024; Accepted April 2024 (Eds.)

Interviewing is a decisive stage of most processes that match candidates to firms and organizations. This article 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.

Key words: Interviews, Hiring, Contrast effect, Memory

JEL codes: D91, J20, M51

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.* [Wexley et al., 1972](#); [Pepitone and DiNubile, 1976](#); [Kenrick and Gutierrez, 1980](#)) and a few real-world applications, such as speed dating ([Bhargava and Fisman, 2014](#)), housing choices ([Simonsohn and Loewenstein, 2006](#)), and financial markets ([Hartzmark and 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 article is to provide large-scale field evidence on the quantitative importance and behavioural 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 analyse 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 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 2-day workshops, where evaluators conduct 12 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

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.

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 article. When analysing the hiring data, we stick to the same pre-registered specifications unless we need to adapt them to the slightly different institutional setup.

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 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 behavioural 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 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 favouring 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 behavioural mechanisms can also lead to a negative autocorrelation in decisions. We assess the potential relevance of sequential learning

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 [Supplementary Appendix A](#).

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 behavioural 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 article 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 labour 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 [Supplementary Appendix A](#) for a detailed overview). Initial evidence of contrast effects comes from laboratory experiments (*e.g.* [Wexley *et al.*, 1972](#); [Pepitone and DiNubile, 1976](#); [Kenrick and Gutierrez, 1980](#)). Existing field studies have used data on rental choices ([Bordalo *et al.*, 2019](#); [Simonsohn, 2006](#); [Simonsohn and Loewenstein, 2006](#)), a speed dating field experiment ([Bhargava and Fisman, 2014](#)) and financial market prices ([Hartzmark and 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 positive autocorrelation in physician decisions, and [Bhuller and Sigstad \(2023\)](#) show that judges change their sentencing behaviour in response to recent reversals of their decisions. In this study, we provide evidence that sequential contrast effects produce significant distortions in labour 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

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 $\approx 9,000$ interviews, our two datasets include $\approx 3,000$ evaluators from two distinct processes, conducting a total of $\approx 37,000$ 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.* \(2020\)](#).

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

the decision environment, the role of memory and the potential for policy interventions by firms and organizations.

More broadly, this article relates to field evidence on reference-dependent decision-making (for an overview, see [Donoghue and Sprenger, 2018](#)), and backward-looking, adaptive reference points in particular (*e.g.* [Thakral and 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 and Kahana, 2024](#)).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.

2.1.1. 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 signalling value and access to a network of high-ability peers and alumni. [Supplementary Appendix B.1](#) provides further information on the program.

The admission process is organized through 2-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 1 or 2 years. About half of the evaluators have undergone a 2-day interviewer training program. A workshop organizer from the study grant foundation is constantly present to lead and moderate the workshop.

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](#)).

2.1.2. 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 [Supplementary Appendix Figure B.1](#)).

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.

2.1.3. 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

2.1.4. 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 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.

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

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 behaviour 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.

2.2. *Setting 2: hiring process*

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

2.2.1. 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.

2.2.2. 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 \times year \times location cells, conditional on candidate gender.

2.2.3. 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.

2.2.4. 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

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.

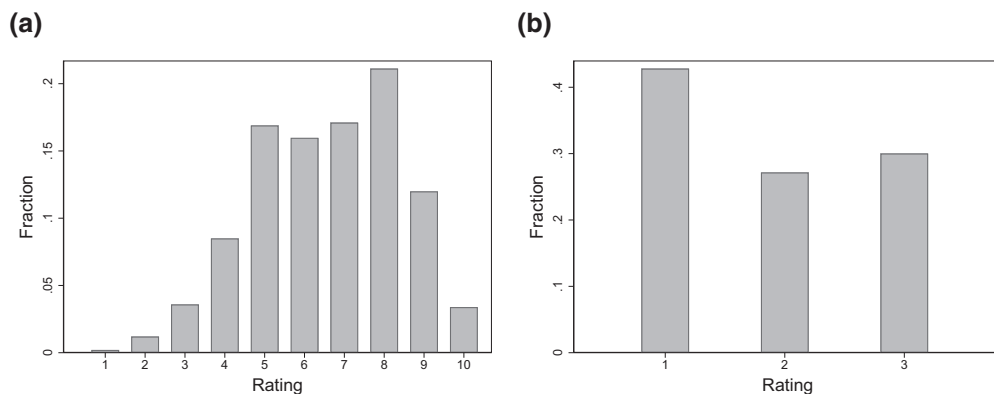


FIGURE 1

Distribution of interview ratings (a) admission process and (b) hiring process

Notes: (a) Shows the distribution of interview ratings in the study grant program ($N = 29,466$). A rating of ≥ 8 points expresses a yes vote. (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

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 versus permanent). Observed evaluator characteristics include gender, managerial responsibility, and interview experience.

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 (Figure 1a), 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 favour of admission. In the hiring process (Figure 1b), about 30% of interviews result in a rating of 3 points, corresponding to a recommendation to hire the candidate.

[Supplementary Appendix C](#) provides additional summary statistics. [Supplementary Appendix 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). [Supplementary Appendix 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 analyse 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 independently by different evaluators based on another interview or a group discussion.¹³ The rationale for using TPA as a quality measure is 2-fold. 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 [Supplementary 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 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 [Supplementary 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

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. [Supplementary Appendix 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 article in four leading journals in economics.

15. The sets of candidates seen by two evaluators never overlap in the hiring process and almost never in the admission process (see [Supplementary 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				
<i>Guryan et al. (2009)</i>				
Leave-one-out mean	0.002* (0.001)	−0.001 (0.001)	−0.012 (0.027)	−0.005 (0.028)
R^2 (within)	0.998	0.998	0.710	0.707
<i>Jochmans (2023)</i>				
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				
<i>Guryan et al. (2009)</i>				
Lag ($t - 1$)	0.000 (0.006)	0.002 (0.006)	0.024 (0.017)	0.013 (0.023)
R^2 (within)	0.009	0.024	0.002	0.000
<i>Jochmans (2023)</i>				
Test statistic	0.915	0.960	1.426	0.967
p -value	0.360	0.337	0.154	0.333
N	26,970	26,970	5,165	5,165

Notes: 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 and 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 ($N = 312/N = 63$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 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 powered 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, [Supplementary 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 analyse 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

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 $t + k$:

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

The outcome variable $Y_{i,t}$ is the standardized rating made by the evaluator i of the candidate interviewed in period t . $\text{TPA}_{i,t+k}$, $k \in \{-11, \dots, -1, 1, \dots, 11\}$, 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 (Supplementary Appendix 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 , $k \in \{-11, \dots, -1, 1, \dots, 11\}$, 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

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 article, including the main hypothesis and the econometric specifications. Prior to pre-registration, we had access to data for the 2012/13 academic year. These “pilot” data are no longer contained in the estimation sample. When analysing the hiring data, we stick to the same pre-registered specifications, unless we need to adapt them due to the slightly different institutional setup.

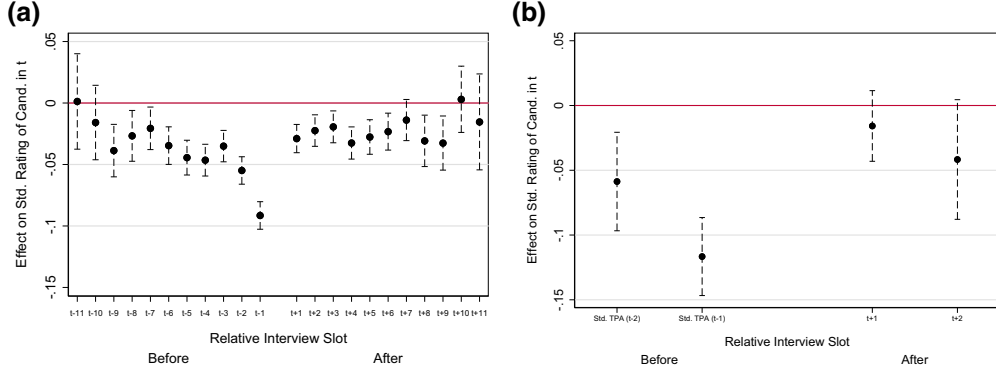


FIGURE 2

Effect of candidate quality in $t+k$ on Std. Rating of candidate in t (a) Admission process and (b) Hiring process

Notes: 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 ($N = 312/N = 63$). [Supplementary Appendix Tables D.1 and D.2](#) report the corresponding coefficients and p -values

effects with candidate pool (*i.e.* year \times location \times position type) fixed effects and cluster standard errors at that level ($N = 63$). The vector $X_{i,t}$ includes candidate and evaluator characteristics ([Supplementary Appendix 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). [Supplementary 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 [Supplementary Appendix 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. [Supplementary Appendix Figure D.2\(a\) and \(b\)](#) shows that the previous candidate's influence is not an artefact of sampling, as it persists when we estimate the influence of other candidates in a single regression, using a homogeneous subsample of interview slots. [Supplementary 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-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 [Supplementary 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 [Supplementary Appendix Table D.3 Panel A](#)). As shown in [Supplementary 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: [Supplementary Appendix D](#) includes several placebo and robustness checks for both data sets. [Supplementary 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 distribution of placebo coefficients (see [Supplementary Appendix D](#) for technical details). [Supplementary 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.

[Supplementary 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. [Supplementary Appendix 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 ([Supplementary Appendix 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:

$$Y_{i,t} = \delta Y_{i,t-1} + \theta \bar{Y}_{i,-t} + X'_{i,t} \mu + \omega_w + \zeta_{i,t} \quad (2)$$

$Y_{i,t}$ 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 ($\bar{Y}_{i,-t}$). In the admission process, $\bar{Y}_{i,-t}$ is computed at the level of the evaluator's interview sequence. In the hiring process, where the sequence includes at most three candidates, $\bar{Y}_{i,-t}$ 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.

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*** (0.006)	-0.057*** (0.006)	-0.406*** (0.042)	-0.127*** (0.013)	-0.131*** (0.013)
Controls	No	Yes	Yes	No	Yes
Outcome mean	0.37	0.37	6.43	0.31	0.31
N	26,970	26,970	26,970	5,165	5,165

Notes: Estimates are based on equation (2). In the admission process, “Yes” describes a vote in favour 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 ($N = 312/N = 63$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

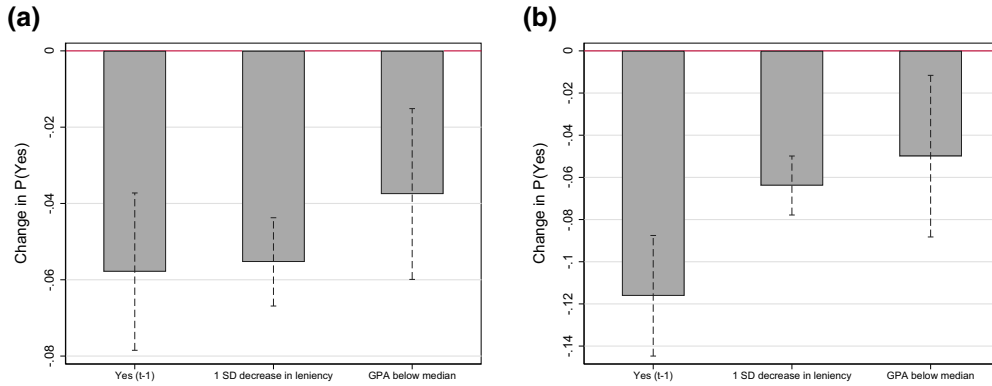


FIGURE 3

Influence of previous decision, evaluator leniency and candidate GPA (a) Admission process and (b) Hiring process

Notes: 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 ($N = 312/N = 63$)

4.2.2. 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

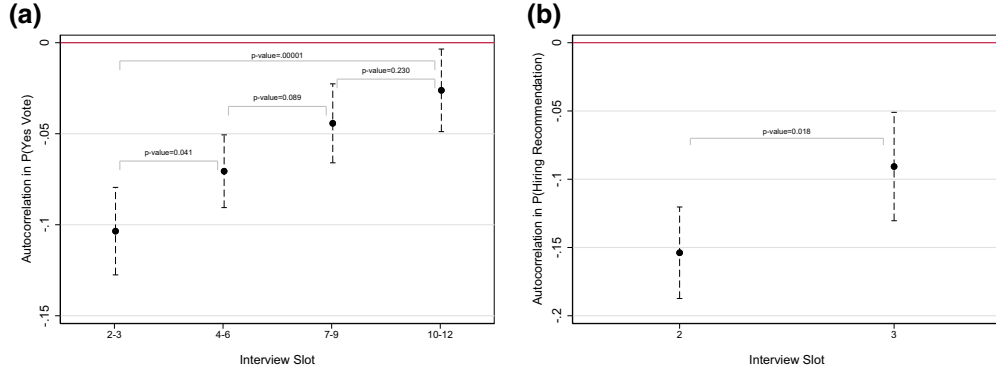


FIGURE 4

Experience within the interview sequence (a) Admission process and (b) Hiring process

Notes: 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 ($N = 312/N = 63$)

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: [Supplementary Appendix E](#) contains several robustness checks and additional results. [Supplementary Appendix 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). [Supplementary Appendix Figure E.1](#) shows that the size of the autocorrelation strongly weakens beyond $t - 1$. Finally, [Supplementary Appendix 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.

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 hiring probability by about 6.3 percentage points ($\approx 20\%$ relative to the mean). Overall, these estimates document that the influence of the previous candidate on individual interview assessments leads

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*** (0.004)		−0.037*** (0.010)	
No. of previous candidates w/ Yes		−0.043*** (0.006)		−0.069*** (0.018)
Outcome mean	0.25	0.25	0.29	0.29
<i>N</i>	12,237	12,237	1,925	1,925

Notes: 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 ($N = 312/N = 63$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 behavioural mechanism in Section 6.

5.1. *Experience within the interview sequence*

Over the course of the interview sequence, evaluators experience an increasing number of candidates. In Figure 4, we analyse how the influence of the previous candidate evolves over the sequence. In both settings, we find strong evidence that the previous candidate's influence decreases while evaluators collect more interview experiences. In the admission process (Figure 4a), 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 (Figure 4b), 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.

5.2. *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

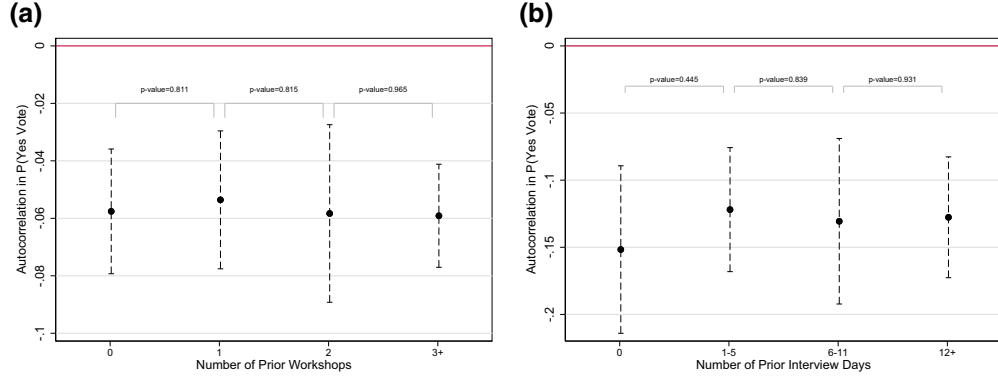


FIGURE 5

Experience prior to the interview sequence (a) Admission process and (b) Hiring process

Notes: 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 ($N = 312/N = 63$)

current sequence. First, [Supplementary Appendix Table F.1](#) shows that the average quality of candidates seen during a workshop in the previous academic year (admission process) or during the last 365 days (hiring process) does not affect current ratings. Second, [Supplementary 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.

5.3. 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 Figure 6(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.

5.4. 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 analyse 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. Figure 7(a) shows the result, revealing that the autocorrelation is significantly stronger when two subsequent candidates share more characteristics. In the hiring data (Figure 7(b)), we only observe gender and

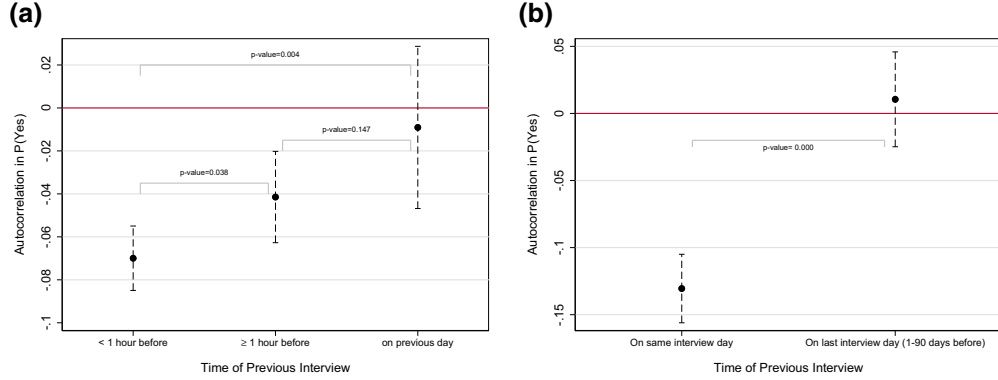


FIGURE 6

Time between interviews (a) Admission process and (b) Hiring process

Notes: 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 ($N = 312/N = 63$)

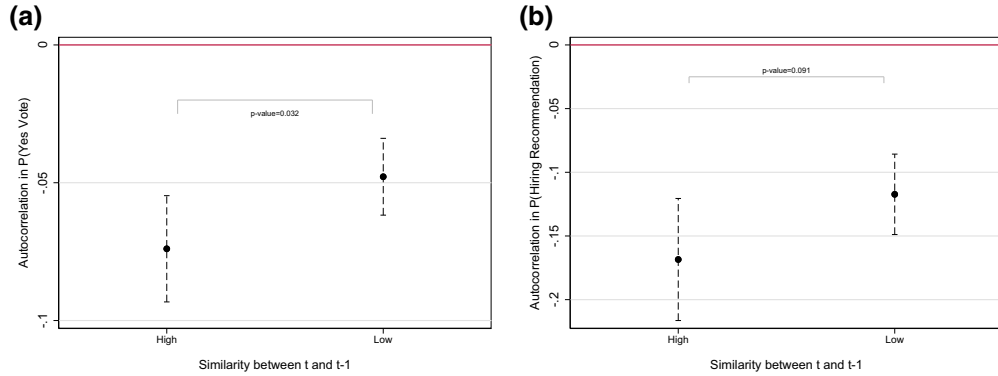


FIGURE 7

Observable similarity of candidates (a) Admission process and (b) Hiring process

Notes: 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 (a); gender and study field in (b)). Dashed lines show 95% confidence intervals. Standard errors are clustered at the workshop/candidate pool level ($N = 312/N = 63$)

study field as relevant candidate characteristics. The results suggest that similarity along these dimensions also strengthens the influence of the previous candidate.

6. BEHAVIOURAL 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 [Supplementary 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 \(2024\)](#).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 candidate²⁰ :

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

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 valuation 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

19. [Supplementary 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 [Supplementary Appendix G.1](#) and provide a quantitative assessment in [Supplementary Appendix H](#).

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 $\lim_{x/y \rightarrow \infty} \sigma(x/y, 1) = \sigma$.

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.

6.1.2. Qualitative fit with main results. 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.

6.1.3. 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 ($t - 1$

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 [Supplementary 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).

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.

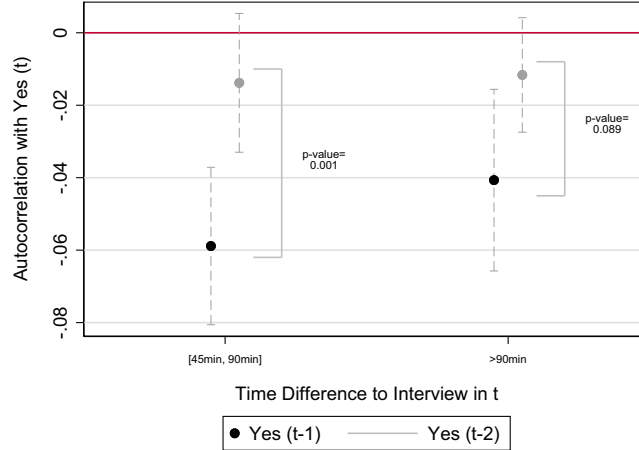


FIGURE 8

The role of relative versus absolute time differences between interviews

Notes: 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 ($N = 312/N = 63$)

versus $t-2$). 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 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 analyse 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 t and $t-2$.²⁷ Specifically, we compare three cases: the candidate in $t-1$ is more similar, equally

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.

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.

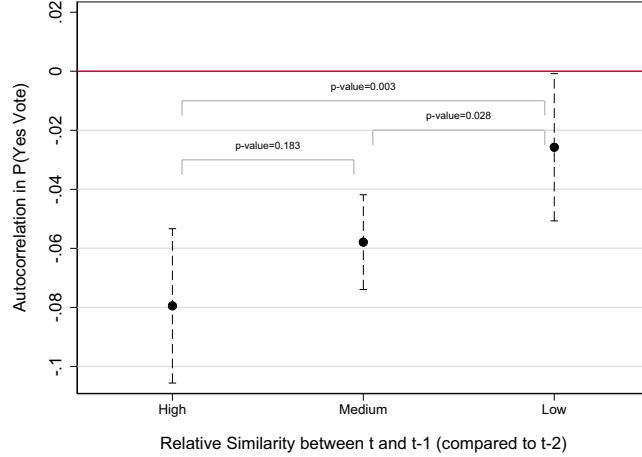


FIGURE 9

The role of relative similarity between candidates

Notes: 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 ($N = 312/N = 63$)

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

6.1.4. Additional results: contrasting. Table 5 provides further evidence in favour 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 and 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.

28. In [Supplementary 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 and Schiprowski, 2020](#)).

TABLE 5
Previous candidate's influence within and between sub-scores (hiring process)

	Cognitive score (1)	Non-cognitive score (2)
TPA, cognitive ($t - 1$)	-0.096*** (0.015)	-0.035** (0.014)
TPA, non-cognitive ($t - 1$)	-0.035** (0.017)	-0.065*** (0.017)
p -value (coeff equality)	0.025	0.243
Outcome mean	1.87	2.03
N	5,155	5,155

Notes: 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 ($N = 63$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

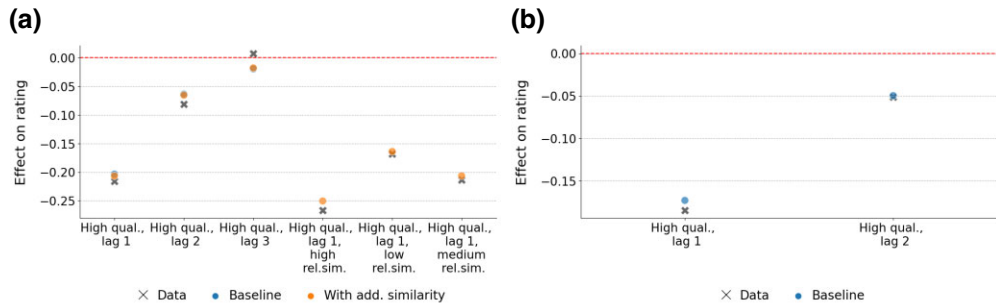


FIGURE 10

Empirical moments and model fit: influence of previous candidates (a) Admission: role of similarity and (b) Hiring: role of similarity

Notes: The figure documents the model fit for the estimates reported in [Supplementary Appendix Table H.1](#). In (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 [Supplementary Appendix Figure H.1](#)

Finally, [Supplementary 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.

6.1.5. 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 [Supplementary 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. [Supplementary 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 [Supplementary Appendix Table H.1](#)), suggesting that the recall process might work very similarly across contexts. Benchmark models without associative recall result in a substantially worse fit with the data (see [Supplementary Appendix H.5.3](#)). Additional results and robustness checks, as well as a more detailed discussion, are provided in [Supplementary Appendix H.5](#).

6.2. *Sequential learning with Bayesian updating*

An alternative behavioural 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 behaviour 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 (Figure 2 and [Supplementary Appendix Figure 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 Figure 5 & [Supplementary Appendix 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. [Supplementary 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.

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 [Supplementary Appendix H.5.3](#)).

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.

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 and 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 autocorrelation 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 role of relative similarity (Figure 9).

Two additional empirical results speak against a gambler's fallacy. First, [Supplementary 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 and 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 [Supplementary 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 behavioural 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.

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. Bar-Hillel and Wagenaar, 1991; Rapoport and Budescu, 1992, 1997). 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.

TABLE 6
Effect of information treatment (study grant program)

	Simple Diff		Diff-in-Diff	
	(1) Yes(<i>t</i>)	(2) Yes(<i>t</i>)	(3) Yes(<i>t</i>)	(4) Yes(<i>t</i>)
Yes (<i>t</i> − 1)	−0.054*** (0.018)	−0.059*** (0.019)	−0.051*** (0.010)	−0.055*** (0.009)
Yes (<i>t</i> − 1) × 2022/23			0.008 (0.021)	0.006 (0.021)
Yes (<i>t</i> − 1) × Jan–Mar			−0.014 (0.013)	−0.009 (0.012)
Yes (<i>t</i> − 1) × Jan–Mar × 2022/23	−0.027 (0.026)	−0.017 (0.026)	−0.012 (0.030)	−0.009 (0.029)
Controls	No	Yes	No	Yes
Outcome mean	0.39	0.39	0.37	0.37
<i>N</i>	6,136	6,136	33,106	33,106

Notes: 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 and 4, the academic years 2013/14 to 2016/17 serve as the control group. “Yes” describes a vote in favour of admitting the candidate. All regressions include workshop fixed effects. [Supplementary Appendix Table J.1](#) shows results using the TPA measure. Standard errors are clustered at the workshop level (*N* = 78 in Columns 1 and 2; *N* = 390 in Columns 3 and 4). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7.1. Information and 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. [Supplementary Appendix J.1](#) provides further details on the intervention. Importantly, no other changes in the admission process occurred simultaneously.

We evaluate the intervention using additional data for the academic year 2022/23. Table 6 reports the results, based on the autocorrelation (equation 2). [Supplementary 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 and 2) and a difference-in-differences specification, where previous academic years from our main dataset serve as the control group (Columns 3 and 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

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

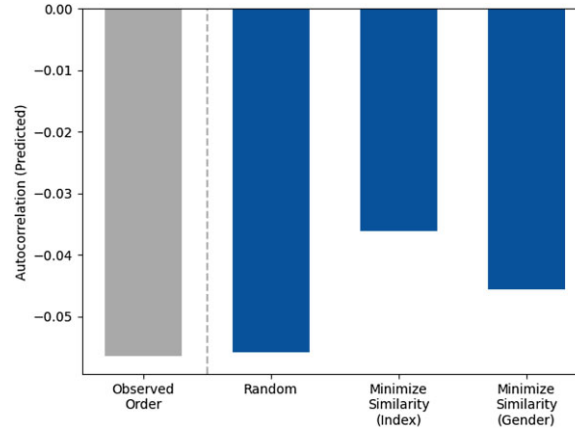


FIGURE 11

Simulation results: reordering of candidates (admission process)

Notes: The figure shows the simulated autocorrelation in yes votes under different ordering schemes. The light bar shows the estimated autocorrelation. The dark 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

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 [Supplementary 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

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.

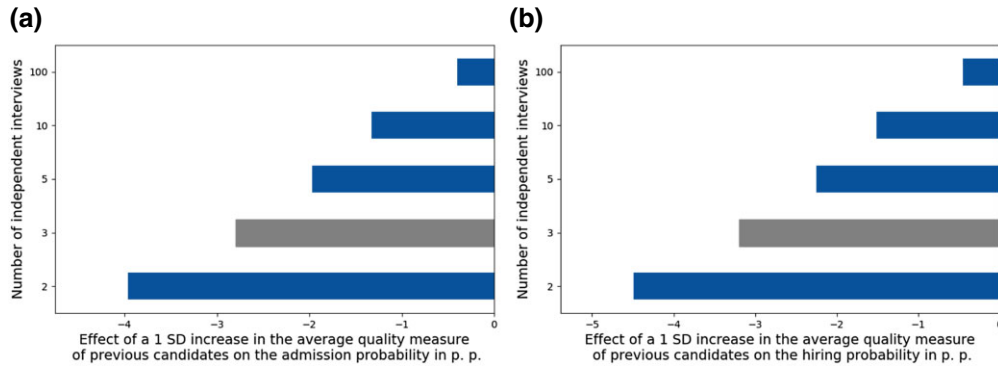


FIGURE 12

Simulation results: increasing the number of independent assessments (a) Admission process and (b) Hiring process

Notes: The light bar shows the estimated influence of the previous candidates' average quality on the admission or hiring probability. The dark 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 [Supplementary Appendix J.2](#)

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 [Supplementary 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 admission 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 article 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.

Acknowledgments. 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 number 280092119, Radbruch), and by the Open Access Publication Fund of the University of Bonn is gratefully acknowledged.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data availability statement

The data used in this article cannot be shared publicly, as they consist of proprietary data owned by a private company and a study grant program. All replication code, explanations of data construction, and empirical moments needed to replicate the structural analysis are available at the following DOI: <https://doi.org/10.5281/zenodo.10837168>.

REFERENCES

- AFROUZI, H., KWON, S., LANDIER, A., *et al.* (2023), “Overreaction in Expectations: Evidence and Theory”, *Quarterly Journal of Economics*, **138**, 1713–1764.
- AUTOR, D. H. and SCARBOROUGH, D. (2008), “Does Job Testing Harm Minority Workers? Evidence from Retail Establishments”, *Quarterly Journal of Economics*, **123**, 219–277.
- BAR-HILLEL, M. and WAGENAAR, W. A. (1991), “The Perception of Randomness”, *Advances in Applied Mathematics*, **12**, 428–454.
- BERGMAN, P., LI, D. and RAYMOND, L. (2020), “Hiring As Exploration” (Working Paper No. 27736, NBER).
- BHARGAVA, S. and FISMAN, R. (2014), “Contrast Effects in Sequential Decisions: Evidence from Speed Dating”, *Review of Economics and Statistics*, **96**, 444–457.
- BHULLER, M., DAHL, G. B., LØKEN, K. V., *et al.* (2020), “Incarceration, Recidivism, and Employment”, *Journal of Political Economy*, **128**, 1269–1324.
- BHULLER, M. and SIGSTAD, H. (2023), “Feedback and Learning: The Causal Effects of Reversals on Judicial Decision-Making” (mimeo).
- BORDALO, P., COFFMAN, K., GENNAIOLI, N., *et al.* (2021), “Memory and Representativeness”, *Psychological Review*, **128**, 71–85.
- BORDALO, P., GENNAIOLI, N. and SHLEIFER, A. (2019), “Memory and Reference Prices: An Application to Rental Choice”, *AEA Papers and Proceedings*, **109**, 572–76.

- BORDALO, P., GENNAIOLI, N. and SHLEIFER, A. (2020), "Memory, Attention, and Choice", *Quarterly Journal of Economics*, **135**, 1399–1442.
- CARD, D., DELLAVIGNA, S., FUNK, P., *et al.* (2019), "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics*, **135**, 269–327.
- CHEN, D., MOSKOWITZ, T. J. and SHUE, K. (2016), "Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires", *Quarterly Journal of Economics*, **131**, 1181–1242.
- CLOTFELTER, C. T. and COOK, P. J. (1993), "Notes: The 'Gambler's Fallacy' in Lottery Play", *Management Science*, **39**, 1521–1525.
- DA COSTA PINTO, A. and BADDELEY, A. D. (1991), "Where did you Park Your Car? Analysis of a Naturalistic Long-Term Recency Effect", *European Journal of Cognitive Psychology*, **3**, 297–313.
- DELLAVIGNA, S., HEINING, J., SCHMIEDER, J. F., *et al.* (2022), "Evidence on Job Search Models from a Survey of Unemployed Workers in Germany", *Quarterly Journal of Economics*, **137**, 1181–1232.
- DONOGHUE, T. and SPRENGER, C. (2018), "Chapter 1 - Reference-Dependent Preferences", in Douglas Bernheim, B., Della Vigna, S. and Laibson, D. (eds) *Handbook of Behavioral Economics: Applications and Foundations I* (Vol. 1) (Amsterdam: North-Holland) 1–77.
- ENKE, B., SCHWERTER, F. and 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**, 545–579.
- GURYAN, J., KROFT, K. and NOTOWIDIGDO, M. J. (2009), "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments", *AEJ: Applied Economics*, **1**, 34–68.
- HARTZMARK, S. M. and SHUE, K. (2018), "A Tough act to Follow: Contrast Effects in Financial Markets", *Journal of Finance*, **73**, 1567–1613.
- HIGGINS, T. E., RHOLES, W. S. and JONES, C. R. (1977), "Category Accessibility and Impression Formation", *Journal of Experimental Social Psychology*, **13**, 141–154.
- HOFFMAN, M., KAHN, L. and LI, D. (2018), "Discretion in Hiring", *Quarterly Journal of Economics*, **133**, 765–800.
- HORTON, J. J. (2017), "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment", *Journal of Labor Economics*, **35**, 345–385.
- JIN, L., TANG, R., YE, H., *et al.* (2023), "Path Dependency in Physician Decisions", *The Review of Economic Studies*, **rdad096**. <https://doi.org/10.1093/restud/rdad096>.
- JOCHMANS, K. (2023), "Testing Random Assignment to Peer Groups", *Journal of Applied Econometrics*, **38**, 321–333.
- KAHANA, M. (2012), *Foundation of Human Memory* (Oxford: Oxford University Press).
- KAHANA, M., DIAMOND, N. B. and AKA, A. (2022), "Laws of Human Memory", PsyArXiv, <https://doi.org/10.31234/osf.io/aczu9>, preprint: not peer reviewed.
- KENRICK, D. T. and GUTIERRES, S. E. (1980), "Contrast Effects and Judgments of Physical Attractiveness: When Beauty Becomes a Social Problem", *Journal of Personality and Social Psychology*, **38**, 131–140.
- 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**, 1417–1426.
- OSKARSSON, A. T., VAN BOVEN, L., MCCLELLAND, G. H., *et al.* (2009), "What's Next? Judging Sequences of Binary Events", *Psychological Bulletin*, **135**, 262.
- PANTELOS, P. C., VAN VUGT, M. K., SEKULER, R., *et al.* (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.
- PEPITONE, A. and DINUBILE, M. (1976), "Contrast Effects in Judgments of Crime Severity and the Punishment of Criminal Violators", *Journal of Personality and Social Psychology*, **33**, 448–459.
- RABIN, M. (2002), "Inference by Believers in the Law of Small Numbers", *Quarterly Journal of Economics*, **117**, 775–816.
- RABIN, M. and VAYANOS, D. (2010), "The Gambler's and Hot-Hand Fallacies: Theory and Applications", *The Review of Economic Studies*, **77**, 730–778.
- RADBRUCH, J. and SCHIPROWSKI, A. (2020), "Interview Sequences and the Formation of Subjective Assessments" (ECONtribute Discussion Papers Series No. 45).
- RAPOPORT, A. and BUDESCU, D. V. (1992), "Generation of Random Series in Two-Person Strictly Competitive Games", *Journal of Experimental Psychology: General*, **121**, 352.
- RAPOPORT, A. and BUDESCU, D. V. (1997), "Randomization in Individual Choice Behavior", *Psychological Review*, **104**, 603.
- SIMONSOHN, U. (2006), "New Yorkers Commute More Everywhere: Contrast Effects in the Field", *The Review of Economics and Statistics*, **88**, 1–9.
- SIMONSOHN, U. and GINO, F. (2013), "Daily Horizons: Evidence of Narrow Bracketing in Judgment from 10 Years of MBA-Admission Interviews", *Psychological Science*, **24**, 219–224.
- SIMONSOHN, U. and LOEWENSTEIN, G. (2006), "Mistake #37: The Effect of Previously Encountered Prices on Current Housing Demand", *The Economic Journal*, **116**, 175–199.
- SINGH, M. (2021), "Heuristics in the Delivery Room", *Science*, **374**, 324–329.
- THAKRAL, N. and TÖ, L. T. (2021), "Daily Labor Supply and Adaptive Reference Points", *American Economic Review*, **111**, 2417–2443.

- WACHTER, J. A. and KAHANA, M. J. (2024), "A Retrieved-Context Theory of Financial Decisions", *Quarterly Journal of Economics*, **2**, 1095–1147.
- WEXLEY, K. N., YUKL, G. A., KOVACS, S. Z., *et al.* (1972), "Importance of Contrast Effects in Employment Interviews", *Journal of Applied Psychology*, **56**, 45.