Research Proposal

Anchoring in Context: Decoy Effects and Dynamic Reference Distortion in Evaluation

Jorge D. Ballestero

Behavioral Economics

Columbia University

May 9, 2025

Abstract

How do evaluators make decisions when faced with structured but sequential information? This paper investigates how dynamic anchoring and trait-level salience shape candidate evaluations in high-stakes screening environments. I propose two behavioral models: a dynamic anchoring model, which captures how evaluators update reference points over time, and a trait salience model, which models how attention weights shift across observable attributes. While prior work focused on narrative or interview-based evaluations to highlight subjective biases in high-stakes evaluation, this study uses purely quantitative resumes to isolate cognitive distortions in numeric data.

Through an online experiment, participants evaluate sequences of fictional resumes under time pressure. Experimental treatments manipulate resume ordering, the presence of decoy candidates, and access to an ideal reference profile. The results will test whether biases persist between traits, notably though evolving reference points, and whether introducing a stable benchmark can mitigate distortions in perception. This project contributes to literatures on bounded rationality, decision-making under uncertainty, and behavioral design in evaluation contexts such as hiring, admissions, and peer review.

Contents

1. Introduction

Last year, Columbia University received just under 60,000 applications for the Class of 2029. Admissions officers filtered through dozens of variables and essays to shortlist—and ultimately accept—a much smaller portion (about 4.27%). In cases like these, where judgments of quality depend on subjective interpretations of both qualitative and quantitative data, it is essential to understand the cognitive mechanisms that influence—and bias—these evaluations.

A common feature of academic and professional selection processes is the sequential screening of CVs. Recruiters, admissions officers, and other evaluators judge candidate quality based on a set of observable traits. Yet when faced with hundreds of applications, evaluators often do not spend equal time on each file, and may satisfice—choosing a "good enough" candidate rather than exhaustively identifying the best available one. The relevance of sequential decision-making in such contexts is underscored in Radbruch and Schiprowski 2024, which examines empirical interview data and identify a recurring concern: "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" (p. 1226). Instead of independently assessing each candidate against a fixed rubric, evaluators may fall victim to anchoring, implicitly comparing new candidates to recent ones.

Research in cognitive psychology and behavioral economics suggests that the frequency of heuristic-based judgments increases when decisions are made quickly and under cognitive load. In short: the less time spent on a decision, the more likely it is that evaluators exhibit bounded rationality. Given that recruiters spend, on average, seven seconds per resume (Ladders Inc. 2020), the anchoring effects observed by Radbruch and Schiprowski may be even more pronounced in resume screening settings.

Ultimately, this raises a central question: How does local context shape perceived quality in sequential evaluations? Through which cognitive channels—dynamic anchoring and trait-level salience—do these distortions emerge in high-stakes sequential evaluation? I specifically focus on resume evaluations, a domain where trait-level information is highly structured and cognitively legible, yet still susceptible to context-driven distortion.

This paper sets out to propose a generalizable model for how evaluators construct reference points, prioritize traits, and adopt heuristics when judging individuals under information constraints. Specifically, I argue that evaluators do not rely on static benchmarks ("the perfect candidate"), but rather build dynamic, local, and malleable anchors based on the distribution and order of resumes they evaluate.

For this, I develop two complementary mechanisms. The first, a dynamic anchoring model, captures how evaluators update global reference points over time and distort utility relative to these evolving benchmarks. The second, a trait salience model, describes how evaluators allocate attention across candidate features, over-weighting those that are locally volatile or surprising relative to prior expectations. Separating these models enables empirical testing of both contrast effects and trait-level biases, and allows for cleaner experimental design.

To test this framework, I propose an online experiment in which participants evaluate sets of resumes under manipulated conditions. Treatments will include: (1) the presence or absence of a strategically dominated "decoy" resume, (2) randomized orderings to test for sequential anchoring, (3) the presence of a standardized ideal resume for reference at the start of the search, among others to be determined. Primary outcome measures will include candidate ratings, binary hiring decisions, and self-reported decision strategies.

Ultimately, the contribution of this paper is twofold: first, to expand decoy theory into a dynamic, high-stakes social judgment framework; and second, to show that "biases" may also be strategically and purposely employed—not only as a result of internal heuristics but also by the structure of the evaluative environment itself. This framework has broader relevance to admissions, peer review, and other real-world decisions where value is relational, sequential, and shaped by context.

The models generate several testable predictions. First,Iexpect that candidate evaluations will be systematically distorted by the quality of previously seen resumes—producing a negative autocorrelation in sequential ratings (anchoring effect). Second, I predict that traits which deviate from expectations or vary locally will carry disproportionate weight in shaping perceived quality (trait salience effect). Most critically, I hypothesize that introducing a standardized "ideal" candidate at the beginning of the evaluation sequence will attenuate both distortions. Specifically, I predict that access to a stable reference will reduce volatility in global and trait-specific reference points, anchoring evaluators to the quality of the ideal candidate over time. Moreover, by providing a visible benchmark, this intervention may flatten salience weights—causing evaluators to treat each trait more evenly, converging toward a uniform weighting as cognitive contrast is suppressed.

2. Literature Review

The decoy effect was originally formalized by Huber, Payne, and Puto 1982, who demonstrated that the presence of an asymmetrically dominated alternative could reliably shift preferences between two competing options. The core mechanism behind this effect is often attributed to violations of the Independence of Irrelevant Alternatives (IIA), as the inclusion of a seemingly irrelevant third option changes the decision weight assigned to others.

This paper builds on three intersecting literatures: sequential evaluation and contrast effects, trait salience and attention allocation, and cognitive reference points in judgment under uncertainty.

The foundation for sequential distortion comes from research on anchoring and contrast effects. Radbruch and Schiprowski 2024 show that interviewers systematically rate candidates lower when they follow strong predecessors, suggesting that evaluators anchor their expectations on recent experiences. Similar findings have been observed in domains such as figure skating (Damisch, Mussweiler, and Plessner 2006) and academic grading (Gonzalez and Wu 2012), highlighting the robustness of recency-based comparison effects. Radbruch and Schiprowski discover that it is not uncommon for evaluators to assess and compare the current candidate to the previous one, instead of a generalizable benchmark, stable throughout all evaluation periods. The type of benchmark they—recruiters and evaluators—employ instead, perhaps due to the difficulty of processing sequential information in a way that—theoretically should—give each instance equal weight, is one formed by active recall, prioritizing more recent candidates.

Trait-level salience draws on work in attribute substitution (Kahneman and Frederick 2002) and decision-by-salience (Bordalo, Gennaioli, and Shleifer 2013). These models show that evaluators overweight features that deviate from expectations or stand out in local context. In complementing Radbruch and Shiprowski's findings with literature on trait-level salience, it is possible to determine the underlying components of the active-recall benchmark used by evaluators. This current study formalizes these ideas using an updating salience-weighted aggregation function, based on both expectation violation and recent contrast. The formal modeling of reference-dependent preferences draws from Kőszegi and Rabin 2006, who propose that individuals experience utility relative to expectations, not outcomes alone. Their framework introduces dynamic reference points shaped by rational expectations, producing psychological distortions such as loss aversion and sensitivity to deviations from prior beliefs. The current paper builds on this idea by modeling evaluators' utility as a function of perceived candidate quality relative to a moving benchmark—one shaped by recent evaluations.

Finally, the idea of stabilizing evaluations and contrast effects using external benchmarks connects to early findings by Wexley et al. 1972. Using videotaped interviews, they manipulated the order in which applicants of varying quality were presented and found that ratings of an average applicant varied dramatically depending on whether they were preceded by two strong or two weak candidates. In the most extreme case, contrast effects explained up to 80% of the variance in the evaluation of average applicants. This suggests that evaluators' judgments are not only relative, but highly sensitive to sequential framing. Nevertheless, they showed that the presence of an"evaluation guide" reduced inter-rater variance in structured interviews.

This paper extends that idea, modeling ideal candidates as anchors that may reduce reference-point volatility and attenuate salience-driven distortions. Rather than offering a static rubric to facilitate evaluation, the ideal resume in this study operates as a cognitive benchmark in an environment characterized by bounded attention. In doing so, this paper reinterprets Wexley et al.'s evaluation guide as a psychological stabilizer—one that may not only improve consistency across raters, but also consistency within raters over time. This study predicts that the presence of the ideal resume as a constant benchmark should reduce reference-point variations, as evaluators will continuously have access to a low-effort and highly comparable object. While the authors' evaluation guide was a rubric—a guide to faciliate subjective analysis—this paper's ideal resume serves as a constant point of comparison that seeks to limit reference-point updating. Morever, unlike their study, which focused on inter-rater reliability, the present design tests intra-rater volatility and distortion in the evaluation of structured, numeric information.

Together, these literatures motivate a novel empirical design that decomposes multiple cognitive mechanisms underlying judgment bias in structured, high-stakes environments.

3. Model and Design

3.1. Research Question and Scope

This paper proposes a model that integrates local contrast, prior expectations, and complexity to explain judgment distortions in resume evaluations. It focuses on three interrelated questions: (1) How do evaluators dynamically update reference points over the course of sequential evaluation? (2) Do evaluators disproportionately weight traits that are locally volatile or surprising? (3) Under what conditions do evaluators adopt strategic heuristics, and how are these decisions moderated by task complexity and stakes?

3.2. Theoretical Framework

3.2.1 Dynamic Anchoring Model

(1) Global Anchoring Model

In order to capture shifts in local reference points, this model includes evolving anchors, capturing changes in expectations across candidates. Let R_t denote the evolving reference point at time t for the overall perceived quality q_t of a candidate. Let R_t evolve recursively as:

$$R_t = \alpha \cdot q_{t-1} + (1 - \alpha) \cdot R_{t-1} \tag{1}$$

With $\alpha \in [0, 1]$ representing the recency bias parameter in anchor updating, and q_{t-1} representing the perceived overall quality of the previous candidate.

The utility of the candidate is subsequently evaluated—and distorted relative to its anchor—as:

$$U_t = q_t + \psi(q_t - R_t) \tag{2}$$

We may further parametrize this distortion:

$$U_t = q_t + \lambda \cdot \psi(q_t - R_t) \tag{3}$$

Here, $\psi(\cdot)$ is a distortion function capturing the shape of evaluative bias—for instance, asymmetry in how evaluators respond to candidates above versus below their reference point. The scalar $\lambda > 0$ represents the magnitude or sensitivity of this distortion. A higher λ implies that evaluators place greater psychological weight on deviations from the reference point, amplifying perceived gains or losses. This separation allows us to empirically distinguish whether treatments reduce bias intensity (via λ) or alter the form of bias (via ψ).

This model attempts to capture evolutions in quality perceptions influenced by local reference points. More importantly, however, it sets out to formalize "surprise" with respect to expectations and its relationship to utility. Indeed, if our subjects were perfectly rational Econs, one could write $U_t = q_t$ and call it a day. However, this simplified equation fails to incorporate the weight of psychological anchors in the perception of quality. If a subject is expecting a 50% on an exam, his actual grade of 75% might feel to him like an 80%. However, if he was expecting a 90%, then that same grade might feel like a punch in the gut—or a 60%.

(2) Trait-Level Anchoring Extension

We hypothesize that a similar anchoring process may occur at the level of individual traits. That is, evaluators form dynamic expectations not only for overall quality but also for each specific trait. These expectations may distort how traits are perceived before aggregation.

Let q_{tk} denote the value of trait k for candidate t. I will define a trait-specific reference point as:

$$R_{tk} = \alpha_k \cdot q_{(t-1)k} + (1 - \alpha_k) \cdot R_{(t-1)k} \tag{4}$$

Evaluators may implicitly compare the current trait value to this evolving anchor. The distorted trait value becomes:

$$\tilde{q}_{tk} = q_{tk} + \lambda_k \cdot \psi(q_k - R_{tk}) \tag{5}$$

For parsimony, I assume a common distortion function $\psi(\cdot)$ across all traits, reflecting a shared cognitive response to deviations from reference points—such as loss aversion or asymmetric sensitivity. While a trait-specific distortion magnitude λ_k could allow different attributes (e.g., GPA vs. extracurriculars) to vary in susceptibility to bias, I maintain a single λ in the core model for tractability and defer λ_k to potential model expansions.

Note that equations (4) and (5) are identical to (1) and (3), respectively. The recency bias parameter α_k is now specific to the trait in question. This allows to test whether categorical variables (i.e., High/Med/Low)[1] are more susceptible to heuristics and biases than strictly numerical ones (e.g., GPA). These adjusted trait

[1]While these variables are framed categorically, 'Very Low' corresponds to 1/5, 'Low' to 2/5, and so on

values can then be ultimately fed into the trait weighting and salience model (3.2.2) to compute overall perceived quality:

$$q_t = \sum_{k=1}^{K} w_k^t \cdot \tilde{q}_{tk} \tag{6}$$

This formulation allows for trait-specific contrast effects. For example, if the GPA of recent candidates was particularly strong, a merely "good" GPA may be undervalued—even if the rest of the resume is unchanged. This maintains the "surprise effect" from the global anchoring model, while maintaining our hypothesis that specific traits are perceived differently depending on the evaluator's reference points. This allows for more testable implications and a greater understanding of the underlying mechanisms that contribute to perceived quality.

3.2.2 Trait-Level Salience

We assume that each resume consists of K observable traits (e.g., GPA, Standardized Evaluations). We further assume that traits are not evaluated equally—evaluators assign salience weights to each trait k based on two psychologically grounded components:

(1) Expectation Violation

$$\Delta_k^t = |q_{tk} - \mu_k| \tag{7}$$

This term captures how much a trait deviates from an evaluators prior expectations[2] about such trait μ_k.

(2) Local Contrast (Spillover)

$$Var_k^{(t-1,t-2)} = |q_{(t-1)k} - q_{(t-2)k}| \tag{8}$$

This term captures how volatile a trait has been across recent candidates, increasing attention to unstable dimensions.

Both terms consist of variations in observable traits which induce perceived attention—salience weight—in the recruiter's evaluation of the candidates. For example, if the recruiter is expecting the average GPA of a candidate pool to be around 3.00, he will be surprised to examine an applicant with a GPA of 1.50, likely increasing the attention GPA will receive for the upcoming candidates. Moreover, if the subsequent candidate boasted a GPA of 4.00, Grade-Point-Average is likely now at the forefront of the recruiter's mind.

[2] Trait-level expectations μ_k are assumed fixed across the task. This reflects the idea that evaluators maintain stable priors (e.g., "a 3.5 GPA is strong") even if their reference point for the current applicant pool shifts over time.

Trait Weighting Function In order to determine the weight a trait in a given resume q_{tk} receives, I suggest a—once again—dynamically evolving weighting function:

$$w_k^t = \alpha_k \cdot \phi(\beta_1 \cdot \Delta_k^t + \beta_2 \cdot Var_t^{(t-1,t-2)}) + (1 - \alpha_k)w_k^{t-1} \tag{9}$$

With β_1 and β_2 being weights on the different salience components, ultimately derived through regression, and α_k corresponding to the recency weight on trait k. Moreover, ϕ is a convex distortion function, in order to amplify the deviance from expectations—and recent candidates.

Normalization We will normalize weights to sum to 1 as follows:

$$\tilde{w}_k^t = \frac{w_k^t}{\sum_{j=1}^K w_j^t} \tag{10}$$

Trait weights are normalized such that:

$$\sum_{k=1}^K w_k^t = 1 \quad \forall t$$

This ensures that the weighted sum of distorted trait values yields a well-scaled perceived quality score q_t.

Perceived Overall Quality Total perceived quality q_t^* is therefore the weighted sum of raw trait values:

$$q_t^* = \sum_{k=1}^K \tilde{w}_k^t \cdot q_{tk} \tag{11}$$

3.3. Experimental Design

This study implements an incentivized between-subjects online experiment designed to test how local context and reference frames shape evaluators' perceived quality of candidates in sequential decision environments. Participants evaluate fabricated, structured resumes under time pressure, with treatments designed to isolate the influence of dynamic anchoring, trait-level salience, and stabilizing interventions.

Each resume will be composed of standardized, easily comparable traits: GPA (0.00-4.00), SAT (0-1600), Years of Experience, Extracurricular Score (Very High/High/Med/Low/Very Low), Quantitative Assessments (1-5). These attributes are presented as plain text values in a consistent order. In maintaining the intelligibility of traits in a categorical fashion, evaluators will be able to determine—by virtue of their prior beliefs and reference points—each candidate's "fit" (i.e., their quality). This will be done in an attempt to reduce the subjective evaluations present in qualitative assessments, in order to highlight the effect of anchoring bias on cardinal information.

Each resume is displayed for exactly 7 seconds to simulate fast-paced, low-effort evaluation conditions similar to real-world CV screening. Primary outcome measures will be considered as a secondary treatment: both participants will be instructed to provide a score after seeing each resume. Half of the participants will also be prompted to respond to a binary hiring decision: to hire the candidate they just evaluated and end their sequence, or keep searching. This group—hereafter dubbed "satisficing group"—will be told that they will only be able to immediately hire the candidate they just saw. If they want to hire a previously evaluated candidate, they will have to wait until the end of the sequence. Both groups will be able to manually select their preferred candidate during a brief window at the end of the sequence.

Random participants from both groups will equally be able to "immediately reject" candidates. That is, removing them from consideration at the end of the sequence. The purpose of this feature is to test for participants' use of heuristics—do they mentally "discard" candidates below a certain threshold? Are resumes that immediately follow discarded candidates perceived better?

At random intervals, participants may be prompted with questions to measure their cognitive uncertainty. The questions will be phrased as follows: "You decided to shortlist/not shortlist the previous candidate. How sure are you of this decision—from 1 to 10?" It is also relevant to measure the self-reported weights of the different traits that the CV is comprised of. To do so, the participants will be prompted—again, at random intervals—to either (1) fill out a series of multiple choice questions about each trait or (2) manually input their estimated weights for each trait (Appendix D). In order to avoid unnecessarily influencing the salience of each trait, I would implement a pilot study to obtain a general understanding of these values without affecting the main experiment.

At the end of the resume sequence, participants will be asked questions about their experience. This is intended to provide information about deliberate use of heuristics ("I mentally rejected any resume with a GPA below 3.5"), uncertainty about their final decision: "Do you believe you hired the best applicant?" and more data on self-reported salience weights—the questions asked at random intervals during the pilot study will be asked at this stage in the main experiment.

3.3.1 Experimental Conditions

Participants are randomly assigned to one of four groups:

- **Control (C)**: The resumes are shown in random order, with no reference candidate and no manipulations.
- **Ideal Anchor (IA)**: Before viewing any resumes, participants are shown a standardized "ideal candidate" resume with excellent scores in all traits. The resume remains visible throughout the task. This condition tests whether a stable external reference reduces reference point volatility and normalizes

trait salience weights.

- **Sequential Contrast (SC)**: Resume order is manipulated to create high–low–high contrast sequences, testing whether perceived quality is distorted by the quality of immediately prior candidates (anchoring via recency).
- **Decoy + Target (DT)**: A target resume is preceded by a strategically dominated resume (decoy) that is strictly worse on multiple dimensions. This setup tests for decoy effects—whether a weak but similar resume boosts the evaluation of its successor.

Each sequence contains 20 resumes, selected and ordered according to condition. In SC and DT treatments, candidate order is fixed to activate the relevant bias. In Control and IA, sequences are randomized.

In order to obtain more reliable data, participants will be incentivized to make the right choices by being offered a bonus in the following way:

- An evaluated resume will be selected at random. If your rating of that candidate is close to a predefined standard, you receive a bonus.[3]

3.3.2 Outcome Measures

Primary dependent variables include:

- Continuous perceived quality score (1–10)
- Binary hiring decision (shortlist/not shortlist, hired/not hired)
- Trait weights (self-reported and then estimated via regression on candidate attributes)
- Reference point variations (estimated from evaluation patterns over time)

Post-task survey questions will also be beneficial to understand the self-awareness of participants in both their use of heuristics and biased decision-making.

3.3.3 Hypotheses

Each treatment maps onto a testable hypothesis:

[3] The predefined standard is simply the normalized average of the candidate's five traits. For instance, GPA: 3.70, SAT: 1560, Years of Experience: 3, Extracurricular Score: Low (2/5), Quantitative Assessment: 4/5, would correspond to a total score of: 0.74/1, or 7.4/10. Participants will be asked to give the candidates a score from 1-10. If their scoring of the randomly selected candidate is within 0.5 (or 0.05 out of 1) of the predefined score, they will receive a bonus.

- **H1 (Dynamic Anchoring)**: Candidate evaluations will exhibit negative autocorrelation with prior candidate quality.

- **H2 (Trait Salience)**: Traits that deviate from prior expectations or local distributions will be overweighted in evaluation.

- **H3 (Ideal Anchor Effect)**: Access to a stable ideal candidate will reduce reference point variation (sensitivity) and flatten trait salience weights toward uniformity.

- **H4 (Decoy Effect)**: The presence of a strictly dominated decoy will increase perceived quality of a target candidate.

3.4. Data Collection and Empirical Strategy

Participants will be recruited via an online platform such as Prolific or MTurk, targeting English-speaking adults with basic familiarity with resume-style evaluations. Each participant will be randomly assigned to one of the four experimental treatments (Control, Ideal Anchor, Sequential Contrast, Decoy + Target) and to one of two evaluation formats (scoring and satisficing or just scoring). A total of 200 participants (50 per treatment $\times$ 2 evaluation types) will be targeted to ensure sufficient power for within—and between—treatment comparisons.

To empirically test the hypotheses, the following strategies will be used:

- **Anchoring (H1)**: Estimate negative autocorrelation in quality ratings using the regression:

$$q_{it} = \beta_0 + \beta_1 q_{i(t-1)} + \varepsilon_{it}$$

  where q_{it} is the perceived quality score for resume t by participant i. This is a simplified version of the regression used in Radbruch and Schiprowski 2024. In a similar vein, however, a significantly negative β_1 would support dynamic anchoring. Standard errors are clustered at the participant level to account for within-subject correlation across resume evaluations.

- **Trait Salience (H2)**: Estimate a weighted linear regression of ratings on traits:

$$q_{it} = \sum_k w_k q_{kit} + \varepsilon_{it}$$

  and compare inferred weights w_k across treatments and time periods to detect shifts in trait importance due to local contrast or expectation violation. While trait-level salience weights w_k^t are endogenously modeled as functions of expectation violations and local contrast, I aim to infer them using reduced-form regressions of participant ratings on observable trait values.

- **Ideal Anchor Effect (H3)**: Compare the variance in reference points and trait weights between the Ideal Anchor and Control groups. Hypothesis: the Ideal Anchor group will show reduced volatility and flatter weight distributions.
- **Decoy Effect (H4)**: Use a difference-in-differences framework comparing the evaluation of the target resume with and without the presence of a dominated decoy in adjacent positions.

Asymmetric Distortion (ψ): To test for the asymmetric distortion implied by $\psi(\cdot)$, I estimate the following regression:

$$q_{it} = \beta_0 + \beta_1 \cdot (q_{it} - R_{it}) \cdot 1(q_{it} \geq R_{it}) + \beta_2 \cdot (q_{it} - R_{it}) \cdot 1(q_{it} < R_{it}) + \varepsilon_{it} \tag{12}$$

Here, q_{it} is the perceived quality score of resume t evaluated by participant i, and R_{it} is the participant's internal reference point at time t, proxied by a moving average or exponentially weighted average of previous scores. The indicator functions $1(q_{it} \geq R_{it})$ and $1(q_{it} < R_{it})$ separate positive and negative deviations from the anchor.

A significant difference that takes the form of $|\beta_2| > |\beta_1|$ would support the model's prediction of asymmetric distortion, in which losses below the reference point loom larger than equivalent gains above it—corresponding to a distortion function with $\theta > 1$.

Robustness checks will include:

- Splitting the sample by evaluation mode to test whether scoring-only participants exhibit different biases than participants with the ability to satisfice
- Conditioning on cognitive certainty and self-reported strategies

3.5. Limitations

First, the artificial nature of the task may limit external validity. While resume screening may take only a few seconds, real-world hiring decisions are rarely made under such strict temporal constraints, and evaluators often have access to richer, more qualitative data (e.g., essays, recommendation letters, interviews). While this design choice is deliberate—to isolate distortions in numeric and structured traits—it may understate the role of qualitative judgment or overstate the impact of contrast effects in more deliberative settings.

Second, participants in online experiments may not approach the task with the same level of engagement or accountability as real recruiters or admissions officers. Although incentivization is used to mitigate this concern, it remains possible that participants treat the task more as a game than a serious evaluative process.

Finally, while the model includes both global and trait-specific reference points, and allows for endogenous salience weighting, it does not account for interaction effects between traits (e.g., GPA and SAT may be interpreted jointly) or unobserved biases. These may play a more prominent role in actual decision-making environments.

Despite these limitations, the experimental design enables precise identification of sequential and trait-level distortions, providing a valuable framework for understanding bounded rationality in structured evaluation environments.

4. Potential Contributions

4.1. Contributions to Literature

This paper contributes to the behavioral economics and judgment and decision-making literatures in several ways.

First, it introduces a dual-mechanism framework for distortion in sequential evaluations: a dynamic anchoring model capturing reference point updating over time, and a trait salience model capturing shifts in evaluative attention. These mechanisms are typically studied in isolation; this paper brings them together in a unified experimental design.

Second, it extends theories of salience and contrast effects to settings involving structured, quantitative data. Unlike previous work focusing on interviews or narrative evaluations, this study shows that even numeric attributes—such as GPA or standardized test scores—are subject to distortion when viewed in sequence or framed against recent alternatives.

Third, the paper tests a relatively novel intervention: presenting an "ideal candidate" to stabilize reference points and reduce volatility in trait weighting.[4] This mechanism, grounded in earlier work on structured evaluation templates, is adapted in this study to experimentally evaluate high-stakes screening settings.

Finally, the study speaks to real-world applications in hiring, admissions, and institutional decision-making. By identifying cognitive distortions in early-stage evaluations and testing plausible mitigations, the paper offers both theoretical insight and practical guidance for improving judgment under uncertainty.

4.2. Contributions to Policymaking

As a result of the experiment, I expect that including an ideal candidate as a stable reference will reduce volatility in anchors and allow for more stable decision-making (H3). As mentioned above, I also predict that

[4] While Wexley et al. made an evaluation guide available in a similar way, I do not know if a fictitious "ideal candidate", i.e., 'what you—the recruiter—should be looking for' has been extensively attempted in the literature.

even categorical and numerical data will be perceptually distorted and contribute to potential preference reversals and biased judgments (H1, H2, H4). As a result, I believe that it is generally impossible to completely remove biases from subjective evaluations, especially in low-effort, high-stakes contexts such as the ones examined in this study. However, I propose two methods to reduce the impact of these biases as much as possible.

In their study, Radbruch and Schiprowski 2024 discovered that "information treatments are insufficient to significantly counteract contrast effects". If the negative effects of biased decision-making cannot be substantially resolved in cases where decision-makers are aware of such biases, I suggest limiting the number of situations in which biases may cloud judgment. In hiring and screening contexts, evaluators must comprehensively go through hundreds of applications, each with repetitive data and—at times—difficult-to-compare information (essays, cover letters). While some instances call for subjective interpretations, there is no need to let bounded rationality affect other cases—in which subjective analysis can only fog evaluations. For instance, an interviewer might notice relevant elements that would not be picked up by a conversation transcript: body language, perceived confidence, charisma, facial expressions; which highlights the essential nature of personal interpretation. However, a 3.5 GPA is a 3.5 GPA, regardless of the fact that it may precede a 3.7, or succeed a 3.9. In limiting subjective judgments to situations that warrant it, the effect of biases on general evaluations of applicants can be strongly reduced.

In order to address the cases that cannot be replaced by cold, numerical comparison, I suggest a similar response to that of Radbruch and Schiprowski: increasing the number of independent subjective analyses. The authors argue that "independent biases in individual assessments cancel out in the aggregate". In conducting parallel judgments of the same candidate, a greater picture is formed—and one that is more reliable and close to reality. The main issue with this response is that increasing the number of independent assessments is costly in both money and time, and goes against trends of massive automatic screening and evaluation programs with artificial intelligence. Even so, I believe that some parts of the recruiter's decision-making cannot be delegated to numeric metrics; and that in those cases, increasing the number of independent judgments can ultimately lead to more informed—and better—decisions.

5. Conclusion

Evaluating human candidates is a complex task, shaped not only by the attributes of the candidates themselves but also by the cognitive and contextual frames through which they are viewed. This paper presents a formal and empirical investigation into two such frames: dynamic anchoring and trait salience.

Through an experiment simulating resume evaluation under time constraints, I test how evaluators update internal benchmarks over time, how they allocate weight across traits, and whether introducing a stable ideal candidate can mitigate distortions. The results of this study will clarify when and how contextual biases emerge, even in the presence of structured, quantitative information.

More broadly, this paper contributes to a growing literature that moves beyond “errors” in judgment to investigate how decision environments shape cognitive processing. It also opens new avenues for evaluating decision architecture in high-stakes social and institutional settings.

References

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2013). "Salience and consumer choice". In: *Journal of Political Economy* 121.5, pp. 803–843.

Damisch, Lysann, Thomas Mussweiler, and Henning Plessner (2006). "Judging the judges: A replication and extension of sequential effects in figure skating". In: *Journal of Sport & Exercise Psychology* 28.1, pp. 76–82.

Gonzalez, Richard and George Wu (2012). "Anchoring in grade predictions: Evidence from academic evaluations". In: *Journal of Behavioral Decision Making* 25.3, pp. 221–232.

Huber, Joel, John W. Payne, and Christopher Puto (1982). "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis". In: *Journal of Consumer Research* 9.1, pp. 90–98.

Kahneman, Daniel and Shane Frederick (2002). "Representativeness revisited: Attribute substitution in intuitive judgment". In: *Heuristics and Biases: The Psychology of Intuitive Judgment*. Ed. by Thomas Gilovich, Dale Griffin, and Daniel Kahneman. Cambridge University Press, pp. 49–81.

Kőszegi, Botond and Matthew Rabin (2006). "A model of reference-dependent preferences". In: *The Quarterly Journal of Economics* 121.4, pp. 1133–1165.

Ladders Inc. (2020). *Why do recruiters spend only 7.4 seconds on resumes?* https://www.theladders.com/career-advice/why-do-recruiters-spend-only-7-4-seconds-on-resumes. Accessed May 7, 2025.

Radbruch, Lea and Niklas Schiprowski (2024). "Contrast Effects in Sequential Interview Ratings". In: *Journal of Labor Economics* 42.1. Forthcoming, pp. 1221–1250.

Wexley, Kenneth N. et al. (1972). "Importance of contrast effects in employment interviews". In: *Journal of Applied Psychology* 56.1, pp. 45–48.

A. Summary of Core Functional Forms

For clarity and reference, I summarize the core recursive structures used in the model. These govern reference point formation, salience-based weighting, distortion, and perceived quality.

A.1. Reference Point Updating

Global and trait-specific reference points update as:

$$R_t = \alpha \cdot q_{t-1} + (1-\alpha) \cdot R_{t-1} \qquad R_{tk} = \alpha_k \cdot q_{t-1,k} + (1-\alpha_k) \cdot R_{t-1,k}$$

A.2. Salience-Based Trait Weighting

Weights evolve based on expectation violation and local trait volatility:

$$w_k^t = \alpha_k \cdot \phi(\beta_1 \cdot \Delta_k^t + \beta_2 \cdot \mathrm{Var}_k^{(t-1,t-2)}) + (1-\alpha_k) \cdot w_k^{t-1}$$

Where:

- $\Delta_k^t = |q_{tk} - \mu_k|$ is trait-level surprise,
- $\mathrm{Var}_k^{(t-1,t-2)}$ captures contrast in recent resumes,
- $\phi(\cdot)$ is a convex, non-linear transformation that amplifies deviations in salience signals. I propose $\phi(x) = x^\eta$ with $\eta > 1$ to allow salient traits—those that are either highly surprising with respect to priors or recently unstable—to receive disproportionately large attention weights. Higher values of η increase convexity and therefore large surprises dominate attention. For simplicity, however, keeping $\eta = 1$ (and therefore $\phi(x) = x$) allows for easier experimental validity without having to estimate η via nonlinear least squares post-experiment. Normalization is applied subsequently to ensure trait weights sum to one.

A.3. Distortion and Perceived Utility

Trait-level distortion is modeled via:

$$\tilde{q}_{tk} = q_{tk} + \lambda \cdot \psi(q_{tk} - R_{tk}) \quad \text{with} \quad \psi(x) = \begin{cases} x & \text{if } x \geq 0 \\ \theta x & \text{if } x < 0 \end{cases}$$

Total perceived candidate quality is:

$$q_t^* = \sum_k w_k^t \cdot \tilde{q}_{tk}$$

B. Experimental Design Summary

B.1. Conditions and Evaluation Modes

Participants are randomly assigned to one of four treatment groups:

- **Control (C)** – 20 resumes in random order
- **Ideal Anchor (IA)** – An ideal candidate shown and visible throughout
- **Sequential Contrast (SC)** – Ordered to induce high-low-high patterns
- **Decoy + Target (DT)** – Strong candidate followed by a strategically dominated decoy

Each participant rates all 20 candidates on a 1–10 scale. Resume exposure time is fixed at **7 seconds per resume**.

B.2. Resume Format (What Participants See)

Each candidate profile includes five quantitative attributes:

- GPA (0.0–4.0)
- SAT Score (1250–1600)
- Quantitative Assessment Score (1–5)
- Work Experience (years)
- Extracurriculars Score (Very High/High/Medium/Low/Very Low)

B.3. Example Candidate Resume (All Treatments)

Each of the 20 candidates is shown in the following format, with numerical values randomly drawn from empirical distributions:

Candidate: C14
GPA: 3.62
SAT Score: 1430
Quantitative Assessment Score: 3
Experience: 2 years
Extracurriculars Score: High

Figure 1: Resume Format Shown to All Participants (7 Seconds Each)

B.4. Ideal Candidate (IA Condition)

In the IA condition, participants see the following candidate before evaluation begins:

Candidate: Ideal Candidate
GPA: 3.95
SAT Score: 1580
Quantitative Assessment Score: 5
Experience: 3 years
Extracurriculars Score: Very High

Figure 2: Ideal Candidate Resume (Visible Throughout IA Treatment)

B.5. Target vs. Decoy Comparison

Table 1: Target and Decoy Candidate Attributes (DT Treatment)

Trait	**Target**	**Decoy**
GPA	3.85	3.60
SAT Score	1540	1490
Quantitative Assessment	5	4
Experience (Years)	3	2
Extracurriculars Score	High	Med

The decoy is strictly dominated on all observable dimensions. This tests whether proximity to a worse alternative inflates perceived quality of the target candidate.

C. Incentives and Evaluation Instructions

Participants are informed that one candidate will be selected at random at the end of the experiment. If the participant's rating of that candidate is close to a predefined benchmark (the normalized average of the candidate's five traits), they receive a bonus. This encourages careful attention across all evaluations.

Each resume is rated independently. Some participants are asked to select the single strongest candidate at the end of the task.

D. Post-Evaluation Questions

Referenced in Section 3.3, participants respond to:

1. How confident were you in your previous candidate rating? (1-10 scale with a slider)
2. Which traits did you pay most attention to while evaluating resumes? How important was each trait in your scoring decision? (100 points to allocate between the five traits)
3. Did you compare candidates to a standard, to one another, or to the most recent applicant? (Participants answer these questions with: Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly Agree)
4. Did you use any mental benchmarks to facilitate the selection process? (i.e., discard all resumes with a GPA below 3.0)

These responses are used to estimate subjective trait priors μ_k, validate model predictions, and analyze behavioral spillover effects.

E. Potential Extensions and Robustness Checks

- Estimate heterogeneity in θ (loss aversion) across participants
- Estimate weight distortion with $\eta > 1$
- Include interaction between traits (e.g., SAT and GPA being correlated)

F. Instructions (Displayed at the Start of the Experiment)

You will be shown a sequence of 20 fictional candidate resumes. Each resume will include five traits:

- Grade Point Average (GPA)

- SAT Score
- Quantitative Assessment (1–5 scale)
- Years of Work Experience
- Extracurricular Score (Very High/High/Medium/Low/Very Low)

Each resume will be displayed for **7 seconds**, after which you will be asked to:

1. Provide a **quality rating** for the candidate on a 1–10 scale.
2. If you are certain you will not move forward with this candidate, you may **immediately reject** them, removing them from ulterior consideration.
3. (In some cases) Decide whether to **hire** the candidate immediately, or wait to see the rest.

At the end of the sequence, you will select one of the candidates you evaluated—and did not reject—to hire.

You may also be asked occasional questions about your confidence and which traits you considered most important.

Bonus: One resume will be selected at random at the end of the study. If your rating of that candidate is close to a pre-defined "expert" score, you will receive a monetary bonus.