THIS MANUSCRIPT IS CURRENTLY UNDER REVIEW

Supporting Information

First Impressions Emerge Through Dynamic Trajectories of Social Inference

Public repository of data and analysis code: https://osf.io/uytzb

Table of Contents

Text Preprocessing Pipeline..…….……..…….……...…….……. 3

Trait Projection…….…….……..…….……...…….……. 4

Text Classification…….…….……..…….……...…….……. 5

Valence Scoring…….…….……..…….……...…….……. 6

Data Exclusion…….…….……..…….……...…….……. 7

Segment Boundary Selection in Study 4..…….……...…….…...….. 9

Study 1 Supplementary Analyses……..……...…….…..……...…….……...… 10

Interrater Reliability…..……..…….…….. 10

Correspondence Analyses Within Endpoint Judgments……...…… 10

Correspondence Analyses Between Moment-to-Moment and Endpoint Judgments…….. 11

Permutation Testing of Correspondence Analyses……...…… 12

Study 2 Supplementary Analyses……..……...…….…..……...…….…….. 14

Semantic Coherence Analyses…….…….……..…... 14

RSA……....……..….. 15

Study 3 Supplementary Analyses……..……...…….…..……...…….……..... 17

Semantic Coherence Analyses …….……..…… 17

RSA..……..……..…….….. 18

Mediation Analysis..……..……. 19

Study 4 Supplementary Analyses……..……...…….…..……...…….…….. 24

Semantic Coherence Analyses …….……...…24

RSA..……..……..…….….. 25

Segment-Level Analysis…..……..……. 27

Mediation Analysis..……..……. 28

Supplementary Tables …….……..…..……...31

Supplementary Figures …….……..…..……44

References ……..…….……..…..…… 47

Text Preprocessing Pipeline

Raw text entries were spell-corrected using a three-stage procedure designed to handle the domain-specific language that generic spell-checkers often mishandle. In the first stage, entries were checked against a curated pre-mapping dictionary that mapped common misspellings, concatenated compounds, and orthographic variants to canonical forms (e.g., "selfcentered" was mapped to "self-centered"). This dictionary also included a small set of identity mappings for words at risk of being inappropriately "corrected" toward a different, meaning-altering term (e.g., "indifferent" was mapped to itself to prevent correction toward "different"); these bypassed spell correction entirely. In the second stage, entries not resolved by the pre-mapping dictionary were checked against a domain whitelist of terms relevant to social perception and professional evaluation (e.g., "gaslighting," "performative") that generic spell-checkers often incorrectly split into unrelated word fragments; whitelisted entries bypassed word segmentation and proceeded directly to typo correction. In the third stage, genuinely concatenated inputs were segmented into component words, and each resulting token was corrected via dictionary lookup with a maximum edit distance of two characters.

Text Classification (SADCAT)

For a descriptive understanding of the range of text entries participants used, and to permit supplemental analyses corroborating effects using only socially relevant entries, each descriptor token was classified via the SADCAT dictionary used in prior social perception research (1). This dictionary assigns non-mutually-exclusive category flags across a set of validated social-perception subdictionaries, which for interpretability were grouped into five broader categories reported in the main text: *Personality Traits* (sociability, morality, ability, assertiveness, warmth, competence), *Mental States/Emotions*, *Appearance/Physical* (beauty, appearance, health), *Social Categories* (social groups, occupation, geography, and demographic subdictionaries), and *Other* (status, beliefs, deviance, and residual content). A token was coded as impression-relevant if it matched any category besides *Other*. Preprocessing artifacts, such as single-word fragments produced by spell-correction (e.g., "self" surviving from "self-centered"), were excluded prior to classification. Uncategorized words ranged between 6.9% and 11.4% across Studies 1–4. The predominant classification was Personality Traits, followed by Mental States/Emotions, with last two non-other classifications varying in ranking between studies (see Table S1).

Trait Projection

To situate free-response descriptors within the same seven-dimensional trait space as the rating scales, each descriptor embedding was projected onto seven semantic axes, one per trait dimension. Seed words for each axis were drawn from established psychometric lexicons: Warmth and Competence seeds were sourced from the SADCAT dictionary (1), and Big Five seeds from 100 unipolar markers (2). Each axis was constructed using a geometrically balanced protocol: for each pole of each trait, the full candidate seed list was filtered to the 10 seeds with highest cosine similarity to a canonical anchor word (e.g., "warm" and "cold" for Warmth; "competent" and "incompetent" for Competence), yielding exactly 10 seeds per pole for each axis (see Table S4). The trait axis was then defined as the normalized difference between the mean embeddings of the two poles, producing symmetrical semantic vectors in the 300-dimensional space. The embeddings of each participant's free-response descriptors, per video, were averaged and projected onto each axis via dot product, yielding a continuous trait-congruence score per dimension. These projection scores served as the semantic analog of the slider ratings in temporal correspondence analyses.

Valence Scoring

Each descriptor was assigned a continuous valence score using the NRC Valence, Arousal, and Dominance Lexicon (3). The NRC VAD lexicon was selected because it provides continuous rather than integer-valued scores, offers broad vocabulary coverage for social-perceptual terms, and distinguishes valence from arousal and dominance. Valence scores serve two roles across the analyses. First, in Study 1, the average descriptor valence was included as a covariate for both the convergence and trajectory alignment analyses to test whether key findings persist beyond shared affective information. That is, we sought to confirm that the semantic content captured in endpoint judgments carry information above and beyond the positivity or negativity of their impression of the target. Second, in Studies 2–4, valence trajectories were also constructed using the same linear interpolation and dynamic time warping procedure applied to semantic impression trajectories, with absolute difference in valence scores serving as the local cost function in place of cosine distance, reflecting the one-dimensional nature of the valence measure. This structurally equivalent comparison allows us to assess whether the temporal organization of successive inferences (i.e., impression trajectories) influence eventual impressions and decisions due to the unfolding of semantic content while controlling for any shared affective dynamics.

Data Exclusions

Exclusions were applied hierarchically at the descriptor, trial, and participant levels. Descriptor-level preprocessing is detailed in the Text Preprocessing Pipeline above. For Studies 2–4, following preprocessing and removal of descriptors unavailable in the FastText embedding model, participant-video trials containing fewer than three unique descriptor-entry timepoints were excluded because they provided insufficient observations for trajectory construction. Trajectory and endpoint-rating data were then synchronized such that each retained trial contained both forms of data. Finally, participants with fewer than three remaining valid video trials were excluded from analysis.

Study 1. No participants were excluded from the real-time judgment sample, yielding a final sample of 202 participants. In the endpoint-judgment task, 204 participants entered the exclusion pipeline. No participants were excluded for completing fewer than five video trials. Two participants were excluded because they produced an average of fewer than two descriptors per video, yielding a final sample of 202 participants. Following FastText filtering (14 words, 0.23%, excluded), no participant-video observations were lost entirely to OOV removal.

Study 2. The working sample comprised 215 participants with both trajectory and endpoint-rating data. Following descriptor-level preprocessing, one participant did not proceed to trajectory construction. One additional participant was manually excluded because of a data-collection error during the study, leaving 213 participants. Removal of descriptors absent from the FastText vocabulary eliminated 10 participant-video trajectories entirely (no trajectories were lost at this step in Studies 3 or 4). A further 451 participant-video trajectories retaining fewer than three unique entry timepoints were excluded; 28 participants had no valid trajectories remaining at this stage, leaving 185 participants represented in the trajectory data. After

trajectory and endpoint-rating data were synchronized, 1,369 matched participant-video trials remained. Thirty-five additional participants were excluded because they retained fewer than three valid trials. The final analytic sample comprised 150 participants and 1,326 participant-video trials (65 of 215 participants excluded, 30.2%).

Study 3. Of 391 participants who generated trajectory data, 13 discontinued the study before completing their first trial and were therefore not included in the 378-participant working sample. Following descriptor preprocessing and FastText filtering, 469 participant-video trajectories retaining fewer than three unique entry timepoints were excluded; 22 participants had no valid trajectories remaining at this stage, leaving 356 participants represented in the trajectory data. After trajectory and endpoint-rating data were synchronized, 2,753 matched participant-video trials remained. Fifty additional participants were excluded because they retained fewer than three valid trials. The final analytic sample comprised 306 participants and 2,682 participant-video trials (72 of 378 participants excluded, 19.0%).

Study 4. The working sample comprised 231 participants with both trajectory and endpoint-rating data. Following descriptor preprocessing and FastText filtering, 222 participant-video trajectories retaining fewer than three unique entry timepoints were excluded; 5 participants had no valid trajectories remaining at this stage, leaving 226 participants represented in the trajectory data. After trajectory and endpoint-rating data were synchronized, 1,626 matched participant-video trials remained. Eleven additional participants were excluded because they retained fewer than three valid trials. The final analytic sample comprised 215 participants and 1,611 participant-video trials (16 of 231 participants excluded, 6.9%).

Segment Boundary Selection in Study 4

For each video, sentence boundaries were identified from a WhisperX transcription of the audio, which provided sentence-level timestamps. Most videos began and ended with brief introductory and concluding remarks (e.g., “Hi, my name is…”). The sentences comprising these remarks were excluded from the segmentation procedure and instead fixed in place, such that every viewing order retained the same introductory and concluding segments in their original positions. Only sentence-end timestamps falling within the remaining body of the video were treated as candidate cut points for segmentation and reordering.

We then constructed a kernel density estimate of Study 3 participants' pause locations across the body of the video, using a fixed Gaussian bandwidth (4 s) along the video timeline; each candidate sentence boundary was then scored by the local pause density within a ±2 s window, z-scored within video. Three boundaries were selected to (a) minimize the summed pause-density score at the chosen cut points while (b) penalizing segment-length imbalances among the four resulting body segments (regularized deviations from the target length of 25% of the body duration), subject to minimum and maximum segment-length constraints (10–45 s).

To assess the sensitivity of the selected boundaries to the specific sample of Study 3 pause data, we performed a participant-level bootstrap (1,000 resamples per video) and recorded, for each selected boundary, the proportion of bootstrap resamples that reselected the identical candidate sentence break. The segment boundaries for the eight selected videos were reselected in at least 50% of bootstrap resamples (stability range = .59–1.00; M = .85).

Study 1 Supplementary Analyses

Interrater Reliability

The interrater reliability of the moment-to-moment trait ratings in the real-time judgment task were assessed using Spearman-Brown split-half reliabilities across the seven trait dimensions. Specifically, for each of 1,000 iterations, participants' data (second-by-second time series) were randomly divided into two halves within each video and trait. Each half's moment-to-moment ratings were averaged to produce a mean trajectory, and the two mean trajectories were correlated using Pearson r. The resulting correlations were Fisher-z transformed, averaged across iterations, and then averaged across videos within each trait. Spearman-Brown correction was applied to the trait-level aggregates to yield the estimated reliability of the full sample's mean trajectory. Table S2 reports split-half reliabilities for each trait dimension.

Interrater reliability for the endpoint-judgment task relied on trait ratings following videos viewed in full duration. Interrater reliability was assessed using ICC(1, k), estimated from a one-way random effects model with a random intercept for video. The value for k was set to the harmonic mean number of raters per video (k = 10.3). A rater-level random effect was not included because the truncation design ensured that each participant viewed at most one video at full length, contributing a single observation and leaving rater-level variance inestimable. Table S3 reports ICC(1, k) reliabilities for each trait dimension.

Correspondence Analyses Within Endpoint Judgments

The main text reports significant convergence between truncated-viewing impressions and full-video impressions followed a significant logarithmic pattern for both trait ratings (Pearson r with full-video mean profile; $b = 0.08$, $SE = 0.01$, $t(1,738) = 6.65$, $p < .001$, ΔAIC =

9.6) and free-response descriptor similarity (cosine similarity between embeddings; b = 0.02, *SE* < 0.01, $t(1,737) = 9.32$, $p < .001$, ΔAIC = -16.3).

Using only socially-relevant descriptors (excluding artifact words and non-social vocabulary), the convergence effect remained significant, $b = 0.017$, SE = 0.002, $t(1,724) = 7.56$, $p < .001$. When the average descriptor valence score was included as a covariate in the model, the convergence effect again remained significant, $b = 0.017$, SE = 0.002, $t(1,739) = 8.82$, $p < .001$. Thus, the correspondence between final descriptors following truncated viewing and final descriptors entered after viewing the target video in full duration remained significant after controlling for the correspondence in the valence of those descriptors, demonstrating that convergence toward the full-video impression is not reducible to shared affective content.

Correspondence Analyses Between Moment-to-Moment and Endpoint Judgments

The main text reports that aggregated moment-to-moment continuous trait ratings were significantly predicted by the consolidated endpoint judgments of independent perceivers whose viewing was truncated at matched timepoints. This temporal correspondence was demonstrated using two complementary indices of endpoint impressions as predictors: trait ratings following truncated videos (b = 5.46, SE = 0.21, t(1,103) = 25.70, p < .001) and free-response descriptor projections onto trait dimensions (b = 63.26, SE = 5.16, t(966) = 12.26, p < .001). Both models regressed the aggregated continuous tracking score (at the video × trait × timepoint level) on the respective endpoint impression measure, with random intercepts for video and trait.

These correspondences survived controlling for participants' early impressions from the first 20% of viewing, ruling out a pure primacy account. A primacy account would attribute these correspondences to shared sensitivity to early-encounter information rather than genuine moment-to-moment integration of information across the entire encounter. If both impressions

are primarily driven by early-encounter information integration, temporal correspondence between endpoint judgments and continuously tracked trait ratings might emerge without any real-time tracking of information later in the encounter. To rule this out, we included as a covariate the continuously tracked ratings from the first 20% of the video length. This variable captures the early-encounter impression signal, offering us a test of whether temporal correspondence between continuously tracked ratings and endpoint judgments persists beyond this early impression alone. With the primacy covariate included, the effect of truncation ratings indeed remained significant ($b = 3.54$, $SE = 0.20$, $t(1,117) = 18.05$, $p < .001$), as did the effect of descriptor projections ($b = 47.70$, $SE = 3.89$, $t(925) = 12.25$, $p < .001$). The correspondence also remained significant when applying a more conservative filtering approach. Restricting projections to socially-relevant descriptors, the correspondence remained significant (b = 61.47, SE = 5.02, $t(977) = 12.26$, p < .001).

Including the average descriptor valence as a covariate, the correspondence again remained significant ($b = 64.07$, $SE = 5.02$, $t(915) = 11.33$, $p < .001$). The temporal alignment is therefore not attributable solely to shared affective content.

Permutation Testing of Correspondence Analyses

To test whether the observed correspondence between endpoint judgments and continuously tracked trait ratings reflected genuine alignment, we conducted nonparametric permutation testing. For each of 1,000 permutations, endpoint judgments were shuffled within each video across all trait × cutoff timepoint combinations (leaving the continuously tracked ratings intact), breaking their temporal alignment with continuously tracked ratings. The observed *t*-statistic for the fixed effect of endpoint judgments was then compared against the

resulting null distribution. Empirical *p*-values reflected the proportion of permuted *t*-statistics equal to or greater than the observed value.

Study 2 Supplementary Analyses

Semantic Coherence Analyses

As we report in the main text, participants' sequences of free-response descriptors were significantly more semantically coherent to prior entries than chance. Additionally, this effect was characterized by a logarithmically increasing coherence over the course of the encounter. When semantic coherence scores were recomputed using only socially-relevant descriptors, the semantic coherence effect remained significantly above chance ($b = 1.07$, $SE = 0.06$, $t(30.0) = 17.67$, $p < .001$). To confirm that coherence reflects genuine semantic continuity rather than an artifact of verbosity, we included the mean number of words per entry as a covariate (mean-centered); if more words per entry were driving coherence by producing more stable averaged embeddings, controlling for this should attenuate the effect. Overall coherence remained well-above chance after this control ($b = 1.21$, $SE = 0.06$, $t(29.3) = 18.98$, $p < .001$).

To confirm that coherence reflects semantic content above and beyond shared affective information, we included the shift in valence over time as a covariate — defined as the difference between the current entry's average valence and the average valence of all prior timepoints, capturing how much the valence score of each new entry coheres to the running affective context. Overall semantic coherence remained significantly above chance after controlling for shifts in valence over time ($b = 1.29$, $SE = 0.07$, $t(25.4) = 18.82$, $p < .001$), confirming that descriptors are more semantically tethered to prior inferences than chance over and above any tendency for affective information to persist or cohere across successive inferences.

The temporal growth effect remained significant when coherence scores were restricted to socially-relevant descriptors, confirming that growth is not driven by incidental non-social vocabulary ($b = 0.10$, $SE = 0.02$, $t(6,484) = 4.08$, $p < .001$). It also remained significant when the

mean number of words per entry was included as a covariate, confirming that the temporal pattern is not reducible to participants simply typing more words as the encounter progressed (b = 0.07, SE = 0.02, t(7,198) = 3.27, p = .001).

To test whether growth in coherence over time emerges above and beyond changes in growth in coherence of valence, we included valence shift alongside log time as a covariate to the model. Here, the temporal growth effect became non-significant (b = -0.03, SE = 0.02, t(6,766) = -1.29, p = .197), while shifts in valence was a significant predictor (b = 0.04, SE = 0.02, t(6,646) = 2.21, p = .027). This indicates that the logarithmic increase in coherence over time is strongly accounted for by systematic changes in affective information across the encounter. However, see Study 3 for a replication of the growth effect that does survive controlling for growing coherence in valence over time.

RSA

To test whether the effect of Impression Trajectory DM was reducible to aggregate semantic content from early in the video (i.e., reducible to a mere primacy effect), we replaced the Content-Only DM computed from the full trajectory with a Primacy Content-Only DM. This control was computed as the cosine distance between the mean embeddings of descriptors entered only during the first 20% of video duration. The effect of the Impression Trajectory DM remained significant after controlling for this primacy-limited content control (b = 0.481, SE = 0.022, t(66,800) = 21.87, p < .001), indicating that the predictive contribution of the trajectory's temporal and semantic organization cannot be reduced only to the influence of early impression content. When DMs were recomputed using only socially-relevant descriptors, the effect of the Impression Trajectory DM remained significant (b = 0.145, SE = 0.028, t(70,670) = 5.20, p < .001).

Two operationalizations of valence were tested here. When the pairwise absolute difference in average descriptor valence scores was added as a covariate to the model, the effect of Impression Trajectory DM remained significant ($b = 0.139$, $SE = 0.028$, $t(59{,}910) = 5.01$, $p < .001$). A more stringent control was also implemented by computing dynamic time warping (DTW) on 1-dimensional valence trajectories using the same 100-bin interpolation and 20% Sakoe-Chiba band as the semantic trajectories, with absolute difference in valence as the local cost function. This approach removes not only shared overall valence levels but shared temporal dynamics of valence. Under this control, the effect of Impression Trajectory DM was attenuated but remained significant ($b = 0.081$, $SE = 0.028$, $t(60{,}240) = 2.88$, $p = .004$), confirming that the semantic content and temporal organization of construals carry predictive information beyond their affective information in how they explain final impressions.

We additionally examined the sensitivity of the Impression Trajectory DM effect to the selected Sakoe–Chiba band width. The band width governs how much temporal flexibility DTW permits when aligning two trajectories; wider bands allow correspondingly more temporal warping. We systematically varied the band from 5% to 50% of the 100-bin time grid in 5% increments, yielding ten parameterizations spanning a narrow local-alignment constraint to near-unconstrained warping. Across all ten band widths, the Impression Trajectory DM effect was stable and highly significant (range of $b = 0.124$–0.128, all ps $< .001$; Table S10), indicating that the effect does not depend on a hyper-particular band selection. The 20% band used throughout the main analyses is consistent with standard practice in DTW applications to behavioral and time-series data.

Study 3 Supplementary Analyses

Semantic Coherence Analyses

As we report in the main text, participants' sequences of free-response descriptors were significantly more semantically coherent to prior entries than chance, and this effect was characterized by logarithmically increasing coherence over the course of the encounter. We conducted the same robustness checks as in Study 2 on both the main finding of semantic coherence and its logarithmic growth over time.

When semantic coherence scores were recomputed using only socially-relevant descriptors, the effect remained significantly above chance ($b = 1.01$, $SE = 0.04$, $t(138.2) = 28.50$, $p < .001$). Overall coherence similarly remained well above chance when the mean number of words per entry was included as a covariate ($b = 1.12$, $SE = 0.03$, $t(167.6) = 34.13$, $p < .001$), confirming that coherence does not reflect an artifact of verbosity producing more stable averaged embeddings.

Including shifts in valence as a covariate, overall semantic coherence remained significantly above chance ($b = 1.15$, $SE = 0.03$, $t(224.0) = 33.71$, $p < .001$), suggesting that successive inferences are more semantically tethered to prior inferences than chance over and above any tendency for affective information to persist over time.

The temporal growth effect remained significant when coherence scores were restricted to socially-relevant descriptors ($b = 0.48$, $SE = 0.03$, $t(9,228.7) = 18.36$, $p < .001$), and when the mean number of words per entry was included as a covariate ($b = 0.30$, $SE = 0.02$, $t(9,798.2) = 12.53$, $p < .001$), confirming that growth is not driven by increasing verbosity across the encounter.

In Study 3, the growth effect remained significant after including shifts in valence as a covariate ($b = 0.47$, $SE = 0.02$, $t(9,229.7) = 19.02$, $p < .001$), indicating that the logarithmic growth in semantic coherence reflects increasing semantic coherence over time that is not reducible to changes in affective information.

RSA

Predicting Ratings Dissimilarity

Applying the primacy control established in Study 2, the effect of Impression Trajectory DM remained significant ($b = 0.929$, $SE = 0.013$, $t(183,000) = 73.68$, $p < .001$), indicating that the predictive contribution of temporal trajectory organization cannot be reduced to the disproportionate influence of early impression content. The primacy Content-only control itself was a significant negative predictor ($b = -0.051$, $SE = 0.007$, $t(188,100) = -7.73$, $p < .001$).

The same vocabulary and valence robustness checks applied to Study 2 were conducted for Study 3. When Impression Trajectory and Content-Only DMs were recomputed using only socially-relevant descriptors, the effect of the Impression Trajectory DM remained significant ($b = 0.292$, $SE = 0.017$, $t(193,800) = 17.52$, $p < .001$).

Applying the mean valence covariate, the effect of the Impression Trajectory DM remained significant ($b = 0.345$, $SE = 0.016$, $t(189,900) = 21.08$, $p < .001$). Finally, under the more stringent DTW-based valence trajectory control, the effect was attenuated but remained significant ($b = 0.219$, $SE = 0.017$, $t(188,400) = 13.24$, $p < .001$), confirming that impression trajectory similarity predicts rating similarity through mechanisms beyond shared affective tone.

We once again tested the sensitivity of the effect of Impression Trajectory DM to the selected Sakoe–Chiba band width by systematically varying the band from 5% to 50% of the 100-bin time grid in 5% increments. Across all ten band widths, the effect of Impression Trajectory DM

was stable and highly significant (range of b = 0.392–0.398, all ps < .001; Table S11), indicating that the effect does not depend on a hyper-particular band selection.

Predicting Decision Disagreement

The same robustness framework was applied to the decision disagreement RSA. Applying the primacy control, the effect of Impression Trajectory DM remained significant (b = 4.542, SE = 0.071, z = 64.03, p < .001), indicating the predictive contribution of the trajectory cannot be reduced to early impression content. When Impression Trajectory and Content-Only DMs were recomputed using only socially-relevant descriptors, the effect remained significant (b = 1.362, SE = 0.095, z = 14.40, p < .001).

Adding the pairwise absolute difference in mean valence as a covariate, the effect of Impression Trajectory DM remained significant (b = 2.017, SE = 0.099, z = 20.36, p < .001). Under the more stringent DTW-based valence trajectory control, the effect was attenuated but remained significant (b = 0.981, SE = 0.099, z = 9.93, p < .001), with valence trajectory dissimilarity carrying substantial predictive weight (z = 109.80). The attenuation pattern was parallel to that observed for the ratings RSA, indicating that shared valence arcs explain a meaningful portion of why similar trajectories lead to similar decisions, while the surviving residual effect of the Impression Trajectory DM confirms that semantic trajectory organization contributes to decision disagreement through mechanisms beyond shared affective dynamics.

Mediation Analysis

The main text reports the mediation results without content-only dissimilarity as a covariate, isolating the total trajectory-to-ratings-to-decision pathway. Trajectory dissimilarity significantly predicted ratings dissimilarity (path a: b = 0.891, SE = 0.012, t = 77.43, p < .001), and ratings dissimilarity in turn significantly predicted decision disagreement in the outcome

model controlling for trajectory dissimilarity (path b: b = 1.364, SE = 0.014, z = 95.74, p < .001). The indirect effect was significant (indirect effect = 0.023, 95% CI [0.021, 0.024], p < .001), with approximately 25% of the total effect operating through trait-rating dissimilarity (proportion mediated = 0.25, 95% CI [0.24, 0.26]). The direct effect of trajectory dissimilarity on decision disagreement remained significant after accounting for the mediator (path c': b = 3.653, SE = 0.070, z = 52.35, p < .001; log-odds scale), indicating partial rather than full mediation. The total effect was b = 4.574, SE = 0.068, z = 67.72, p < .001.

To isolate the unique temporal trajectory contribution beyond aggregate semantic content, we added content-only dissimilarity as a covariate. Trajectory dissimilarity continued to predict ratings dissimilarity (path a: b = 0.394, SE = 0.017, t = 23.60, p < .001), and ratings dissimilarity predicted decision disagreement (path b: b = 1.340, SE = 0.014, z = 93.65, p < .001). The indirect effect remained significant (indirect effect = 0.010, 95% CI [0.009, 0.011], p < .001), with a proportion mediated of 0.23 (95% CI [0.20, 0.26]). The direct effect remained significant (path c': b = 1.783, SE = 0.100, z = 17.80, p < .001). The attenuation of the path a coefficient (from b = 0.891 to b = 0.394) reflects the substantial correlation between trajectory and content-only dissimilarity (r = 0.741); with content-only dissimilarity partialled out, the remaining trajectory variance that predicts ratings dissimilarity is smaller but still highly significant. The stability of the proportion mediated across both specifications — 25% without content-only dissimilarity, 23% with — indicates that the relative balance between the indirect and direct pathways is not an artifact of collinearity between trajectory dissimilarity and content-only dissimilarity. Content-only dissimilarity was itself a significant predictor in both the mediator model (b = 0.740, SE = 0.018, t = 40.85, p < .001) and the outcome model (b = 2.795, SE = 0.109, z = 25.74, p < .001).

Using only socially-relevant descriptors (with content-only dissimilarity as a covariate), trajectory dissimilarity predicted ratings dissimilarity (path a: b = 0.292, SE = 0.017, t = 17.53, p < .001), and ratings dissimilarity predicted decision disagreement (path b: b = 1.350, SE = 0.014, z = 94.49, p < .001). The indirect effect was significant (indirect effect = 0.007, 95% CI [0.006, 0.008], p < .001), with a proportion mediated of 0.27 (95% CI [0.23, 0.32]). The direct effect remained significant (path c': b = 1.050, SE = 0.101, z = 10.42, p < .001).

We also applied the two valence controls introduced in Study 2 to the mediation model, adding the corresponding valence covariate to both the mediator and outcome equations alongside content-only dissimilarity.

Controlling for aggregate valence dissimilarity alongside content-only dissimilarity, trajectory dissimilarity continued to predict ratings dissimilarity (path a: b = 0.345, SE = 0.016, t = 21.08, p < .001), and ratings dissimilarity continued to predict decision disagreement (path b: b = 1.069, SE = 0.015, z = 72.44, p < .001). The indirect effect remained significant (indirect effect = 0.007, 95% CI [0.006, 0.008], p < .001), with a proportion mediated of 0.17 (95% CI [0.15, 0.19]). The direct effect remained significant (path c': b = 1.789, SE = 0.103, z = 17.44, p < .001). Mean valence dissimilarity was itself a strong predictor in both the mediator model (b = 0.295, SE = 0.003, t = 94.39, p < .001) and the outcome model (b = 2.272, SE = 0.021, z = 107.09, p < .001). Notably, content-only dissimilarity became non-significant in the outcome model once valence was included (b = 0.029, SE = 0.114, z = 0.26, p = .796), suggesting that content-only dissimilarity's contribution to decision disagreement was substantially carried by aggregate affective differences rather than semantic content per se.

Under the more stringent DTW-based valence trajectory control, trajectory dissimilarity continued to predict ratings dissimilarity (path a: b = 0.219, SE = 0.017, t = 13.24, p < .001), and

ratings dissimilarity continued to predict decision disagreement (path b: b = 1.128, SE = 0.015, z = 77.02, p < .001). The indirect effect remained significant (indirect effect = 0.005, 95% CI [0.004, 0.005], p < .001), with a proportion mediated of 0.24 (95% CI [0.19, 0.30]). The direct effect remained significant (path c': b = 0.808, SE = 0.102, z = 7.89, p < .001). The valence trajectory covariate was significant in both the mediator (b = 0.274, SE = 0.003, t = 83.57, p < .001) and outcome models (b = 2.074, SE = 0.022, z = 95.71, p < .001). Because valence trajectories require at least two valid valence timepoints per participant-video, this specification was fit on a slightly reduced sample (N = 193,626 observations from 45,659 participant pairs). The attenuation of path a from 0.394 under content-only dissimilarity control alone to 0.219 here reflects the substantial structural overlap between the semantic-embedding and one-dimensional valence trajectories, since both are DTW distances computed over the same temporal bins. That the indirect effect survives even this aggressive partialling is the strongest evidence against a purely affective interpretation of the trajectory measure.

Overall, the indirect effect was statistically significant, and the direct effect remained significant in every case, confirming partial mediation throughout. Path b was stable across specifications that did not include a valence covariate (b ≈ 1.34 to 1.36) but attenuated (1.07 and 1.13) once valence was partialled out, indicating that a portion of the ratings-to-decision association reflects shared affective content. Proportion mediated was stable within the Content-only and descriptor-filtering robustness checks (23–27%) but shifted downward under mean-valence control (17%), reflecting the absorption of indirect-path variance by valence rather than instability of the underlying mediation structure. Most importantly, the indirect effect survived both valence controls — including the DTW-based valence trajectory control, which partials out a measure structurally analogous to the primary trajectory predictor — providing

strong evidence that the mediation pathway from semantic trajectories through final trait impressions to recruitment decisions is not reducible to affective dynamics. Table S18 reports the complete results across all robustness specifications.

Study 4 Supplementary Analyses

Semantic Coherence Analyses

As we report in the main text, participants' sequences of free-response descriptors were significantly more semantically coherent to prior entries than chance, and this effect was characterized by logarithmically increasing coherence over the course of the encounter. We conducted the same robustness checks as in Studies 2 and 3 on both the main finding of semantic coherence and its logarithmic growth over time.

When semantic coherence scores were recomputed using only socially-relevant descriptors, the effect remained significantly above chance ($b = 0.98$, $SE = 0.04$, $t(35.2) = 23.13$, $p < .001$). Overall coherence similarly remained well above chance when the mean number of words per entry was included as a covariate ($b = 1.14$, $SE = 0.05$, $t(22.6) = 25.40$, $p < .001$), confirming that coherence does not reflect an artifact of verbosity producing more stable averaged embeddings.

Including shifts in valence as a covariate, overall semantic coherence remained significantly above chance ($b = 1.16$, $SE = 0.05$, $t(23.3) = 23.98$, $p < .001$), suggesting that successive inferences are more semantically tethered to prior inferences than chance over and above any tendency for affective information to persist over time.

The temporal growth effect remained significant when coherence scores were restricted to socially-relevant descriptors ($b = 0.28$, $SE = 0.03$, $t(5{,}638.2) = 8.85$, $p < .001$), and when the mean number of words per entry was included as a covariate ($b = 0.25$, $SE = 0.03$, $t(6{,}180.9) = 8.91$, $p < .001$), confirming that growth is not driven by increasing verbosity across the encounter.

As in Study 3, the growth effect also remained significant after including shifts in valence as a covariate ($b = 0.29$, $SE = 0.03$, $t(5{,}892.7) = 10.02$, $p < .001$), indicating that the logarithmic growth in semantic coherence reflects increasing semantic coherence over time that is not reducible to changes in affective information.

Coherence After Reordering

Because Study 4 experimentally reordered the video segments, we additionally tested whether the coherence effect and its growth depended on whether participants viewed the original (canonical) sequence or one of the nine derangements. Neither the overall level of coherence ($b = 0.17$, $SE = 0.12$, $t(6{,}079.4) = 1.40$, $p = .161$) nor its logarithmic growth rate ($b = -0.13$, $SE = 0.10$, $t(6{,}064.0) = -1.33$, $p = .185$) differed significantly between canonical and reordered presentations, while the growth effect itself remained significant within the model ($b = 0.42$, $SE = 0.09$, $t(6{,}076.8) = 4.50$, $p < .001$). Coherence therefore accrued at a comparable rate under both presentation formats.

RSA

Predicting Ratings Dissimilarity

Applying the primacy control established in Studies 2 and 3, the effect of Impression Trajectory DM remained significant ($b = 0.847$, $SE = 0.014$, $t(147{,}700) = 60.25$, $p < .001$), indicating that the predictive contribution of temporal trajectory organization cannot be reduced to the disproportionate influence of early impression content. The primacy Content-only control itself was a significant negative predictor ($b = -0.036$, $SE = 0.007$, $t(155{,}000) = -4.94$, $p < .001$). As in Study 3, the primacy-limited centroid was substantially correlated with both the full Content-Only DM ($r = .545$) and the Impression Trajectory DM ($r = .400$).

The same vocabulary and valence robustness checks applied to Studies 2 and 3 were conducted for Study 4. When Impression Trajectory and Content-Only DMs were both recomputed using only socially-relevant descriptors, the effect of the Impression Trajectory DM remained significant ($b = 0.357$, $SE = 0.018$, $t(155{,}200) = 19.57$, $p < .001$), as did the corresponding Content-Only DM ($b = 0.557$, $SE = 0.019$, $t(156{,}900) = 30.09$, $p < .001$).

Applying the mean valence covariate, the effect of the Impression Trajectory DM remained significant ($b = 0.351$, $SE = 0.018$, $t(153{,}000) = 19.44$, $p < .001$). Finally, under the more stringent DTW-based valence trajectory control, the effect was attenuated but remained significant ($b = 0.251$, $SE = 0.018$, $t(152{,}700) = 13.74$, $p < .001$), confirming that impression trajectory similarity predicts rating similarity through mechanisms beyond shared affective tone. The valence trajectory covariate was itself a strong predictor ($b = 0.268$, $SE = 0.004$, $t(156{,}400) = 73.08$, $p < .001$).

We again varied the Sakoe–Chiba band from 5% to 50% of the 100-bin time grid in 5% increments. The effect of Impression Trajectory DM was stable and highly significant across all ten band widths (range of $b = 0.415$–0.420, all ps $< .001$; Table S12).

Predicting Decision Disagreement

The same robustness framework was applied to the decision disagreement RSA. Applying the primacy control, the effect of Impression Trajectory DM remained significant ($b = 5.139$, $SE = 0.082$, $z = 63.06$, $p < .001$), indicating the predictive contribution of the trajectory cannot be reduced to early impression content. When Impression Trajectory and Content-Only DMs were recomputed using only socially-relevant descriptors, the effect remained significant ($b = 2.681$, $SE = 0.107$, $z = 25.02$, $p < .001$).

Adding the pairwise absolute difference in mean valence as a covariate, the effect of Impression Trajectory DM remained significant ($b = 2.680$, $SE = 0.113$, $z = 23.81$, $p < .001$). Under the more stringent DTW-based valence trajectory control, the effect was attenuated but remained significant ($b = 1.879$, $SE = 0.112$, $z = 16.80$, $p < .001$), with valence trajectory dissimilarity carrying substantial predictive weight ($z = 95.13$). As in Studies 2 and 3, the attenuation pattern was parallel to that observed for the ratings RSA, indicating that shared valence arcs explain a meaningful portion of why similar trajectories lead to similar decisions, while the surviving residual effect of the Impression Trajectory DM confirms that semantic trajectory organization contributes to decision disagreement through mechanisms beyond shared affective dynamics.

Notably, under the mean valence control the Content-Only DM was reduced ($b = 0.309$, $SE = 0.123$, $z = 2.51$, $p = .012$), paralleling the pattern reported for Study 3, in which content-only dissimilarity's contribution to decision disagreement was substantially carried by aggregate affective differences rather than semantic content per se.

Segment-Level Analyses

The main text reports that descriptors generated for an identical video segment were more similar when the pair of participants had encountered the same ordered sequence of preceding segments. We conducted the same robustness checks on this segment-level effect as on the trajectory-level RSA.

The effect of preceding segments survived restricting descriptors to socially-relevant tokens ($b = 0.023$, $SE = 0.001$, $t(177{,}100) = 26.90$, $p < .001$) and survived controlling for pairwise valence dissimilarity within the segment ($b = 0.020$, $SE = 0.001$, $t(187{,}800) = 26.41$, $p <$

.001), with the valence covariate itself a strong predictor ($b = 0.105$, $SE = 0.001$, $t(187{,}700) = 140.54$, $p < .001$).

We additionally examined whether the preceding-segment effect could be explained by differences in how *much* participants wrote about each segment, rather than differences in what they inferred. Using the absolute difference in descriptor counts per segment as a sparsity-robust dependent variable — a measure that remains defined even when a segment elicited no usable descriptors — the preceding-segment effect emerged in the *opposite* direction from the content effect: pairs sharing the same preceding segments differed slightly *more* in the number of descriptors produced ($b = -0.073$, $SE = 0.009$, $t(193{,}400) = -7.77$, $p < .001$; estimated marginal means = 1.38 for pairs sharing the same preceding segments and 1.31 for pairs with different preceding segments). Because greater content similarity among pairs sharing the same preceding segments co-occurred with *greater* dissimilarity in descriptor quantity, the segment-level content effect cannot be attributed to shared verbosity.

Mediation Analysis

The main text reports the indirect pathways from the order manipulation through trajectory structure to endpoint impressions and decisions. Mediation was estimated via quasi-Bayesian Monte Carlo simulation (5,000 draws), sampling the *a* path (X → M) and *b* path (M → Y | X) independently from their respective approximate posteriors, forming the indirect effect as the product of draws, and computing percentile confidence intervals and two-tailed Monte Carlo *p*-values from the resulting distribution. Because crossed video × dyad random effects preclude a standard structural equation model, all mediator and outcome models were fitted as mixed-effects models with crossed random intercepts for video and dyad; continuous outcomes used linear mixed-effects models and binary outcomes used logistic mixed-effects

models. For the decision outcomes, the *a* path is on the raw mediator scale while the *b* path is on the logit scale, so the indirect effect is a logit-scale approximation and should be interpreted in terms of direction and whether the interval excludes zero rather than as a precise effect size.

Trajectory-Level Mediator

The main text reports the *a* path and the indirect effects for this model. The remaining paths are as follows. With trajectory dissimilarity as the mediator, trajectory dissimilarity strongly predicted final-impression dissimilarity (path *b*: $b = 0.828$, $SE = 0.013$, $p < .001$), while the direct effect of viewing-order pair type was non-significant once the mediator was included (path c': $b = 0.002$, $SE = 0.003$, $p = .560$), indicating that the order effect on endpoint impressions was fully carried by trajectory structure.

The same pattern held for recruitment decisions: trajectory dissimilarity predicted decision disagreement (path *b*: $b = 4.936$, $SE = 0.075$, $p < .001$; log-odds), with a non-significant direct effect (path c': $b = -0.022$, $SE = 0.018$, $p = .218$).

Segment-Level Mediators

Parallel mediation analyses using segment-level content as the mediator showed the same pattern. Using each pair's mean segment-content dissimilarity as the mediator, order pair type predicted segment-content dissimilarity (path *a*: $b = 0.009$, $SE = 0.001$, $p < .001$), which predicted final-impression dissimilarity (path *b*: $b = 0.342$, $SE = 0.009$, $p < .001$), yielding a significant indirect effect (indirect = 0.0032, 95% CI [0.0025, 0.0040], $p < .001$) with a non-significant direct effect (path c': $b = -0.002$, $SE = 0.003$, $p = .558$). The corresponding decision model also showed a significant indirect effect (path *a*: $b = 0.009$, $SE = 0.001$, $p < .001$; path *b*: $b = 2.249$, $SE = 0.053$, $p < .001$; indirect = 0.0210, 95% CI [0.0163, 0.0257], $p < .001$; path c': $b = -0.038$, $SE = 0.020$, $p = .063$).

Finally, we tested the segment-level pathway directly, using whether a pair shared the same ordered sequence of preceding segments as the predictor and segment-content dissimilarity as the mediator. Differing preceding segments predicted greater segment-content dissimilarity (path *a*: $b = 0.022$, $SE = 0.001$, $p < .001$), which predicted greater final-impression dissimilarity (path *b*: $b = 0.248$, $SE = 0.006$, $p < .001$), producing a significant indirect effect (indirect = 0.0054, 95% CI [0.0049, 0.0058], $p < .001$) with a non-significant direct effect (path c': $b = -0.003$, $SE = 0.002$, $p = .119$). The decision model likewise showed a significant indirect effect (path *a*: $b = 0.022$, $SE = 0.001$, $p < .001$; path *b*: $b = 1.945$, $SE = 0.042$, $p < .001$; indirect = 0.0419, 95% CI [0.0385, 0.0454], $p < .001$). In this model alone the direct effect also reached significance, in the negative direction (path c': $b = -0.056$, $SE = 0.015$, $p < .001$). Across all six specifications, the indirect effect was significant and in the expected direction, and the direct effect was non-significant in five of six.

Supplementary Tables

Table S1

Word Classification Distribution by Study

Category	Study 1	Study 2	Study 3	Study 4
Personality Traits	3,477 (58.1%)	8,801 (44.1%)	18,978 (72.8%)	11,045 (71.8%)
Mental States	2,169 (36.2%)	7,955 (39.9%)	5,033 (19.3%)	2,491 (16.2%)
Appearance	688 (11.5%)	2,288 (11.5%)	1,235 (4.7%)	867 (5.6%)
Social Categories	78 (1.3%)	1,540 (7.7%)	1,635 (6.3%)	1,112 (7.2%)
Other	586 (9.8%)	1,826 (9.2%)	1,647 (6.3%)	1,150 (7.5%)

Note. Categories are non-mutually exclusive.

Table S2

Split-Half Reliabilities for Continuous Trait Assessments (Study 1)

Trait Dimension	Mean *r*	*Spearman-Brown*	N_{videos}
Agreeableness	0.470	0.640	10
Competence	0.471	0.640	10
Conscientiousness	0.353	0.521	10
Extraversion	0.452	0.623	10
Neuroticism	0.795	0.886	10
Openness	0.430	0.602	10
Warmth	0.866	0.928	10

Table S3

Interrater Reliability (ICC) for Endpoint Judgments (Study 1)

Trait Dimension	ICC(average)	σ^2 (Video)	σ^2 (Resid)	k (harmonic)
Openness	0.593	0.880	6.247	10.3
Conscientiousness	0.698	1.337	5.996	10.3
Extroversion	0.835	2.774	5.662	10.3
Agreeableness	0.562	0.743	5.997	10.3
Neuroticism	0.596	1.028	7.189	10.3
Warmth	0.557	0.770	6.344	10.3
Competence	0.782	1.414	4.085	10.3

Note. ICC(average) represents the reliability of the mean rating across k raters.

Table S4

Seed Words for Trait Dimension Projection

Trait Dimension	Positive Pole Seeds	Negative Pole Seeds
Agreeableness	agreeable, helpful, sympathetic, considerate, generous, kind, polite, cooperative, unselfish, trusting	disagreeable, rude, unsympathetic, harsh, cruel, uncooperative, distrustful, selfish, cold, demanding
Competence	competent, skilled, knowledgeable, proficient, capable, confident, talented, adept, intelligent, competence	incompetent, inept, incompetency, incompetence, unintelligent, ineffective, negligent, inattentive, inexperienced, ineffectual
Conscientiousness	conscientious, diligent, careful, thorough, efficient, systematic, disciplined, practical, neat, organized	disorganized, undisciplined, messy, inefficient, sloppy, inconsistent, lazy, impractical, careless, negligent
Extroversion	extroverted, sociable, talkative, assertive, energetic, enthusiastic, outgoing, communicative, active, bold	introverted, shy, timid, quiet, bashful, inhibited, reserved, silent, withdrawn
Neuroticism	neurotic, temperamental, irritable, fretful, moody, emotional, anxious, insecure, jealous, envious	calm, tranquil, peaceful, placid, relaxed, steady, unemotional, stable, secure
Openness	open, deep, creative, original, complex, imaginative, inventive, philosophical, intellectual, artistic	conventional, traditional, simple, conservative, unimaginative, unintellectual, uncreative, shallow, unoriginal
Warmth	warm, warmth, gentle, warmhearted, hospitable, friendly, pleasant, cordial, affectionate, congenial	cold, chilly, frigid, frosty, icy, dry, dark, coldhearted, harsh, wintry

Table S5

Study 2 Linear Mixed-Effects RSA: Full Model

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.63	0.03	23.68	10.66	< .001
Trajectory dissimilarity	0.13	0.03	4.61	60490	< .001
Content-only dissimilarity	0.63	0.03	18.88	60170	< .001

Note. Random effects: dyad variance = 0.0135 (SD = 0.1160); video variance = 0.0065 (SD = 0.0806); Residual variance = 0.1795 (SD = 0.4236). N = 84,946 (11,148 dyads, 10 videos). DV = 1 − Pearson r (trait profile dissimilarity). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S6

Study 2 Linear Mixed-Effects RSA: Primacy Control

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.60	0.03	22.57	10.68	< .001
Trajectory dissimilarity	0.48	0.02	21.87	66800	< .001
Primacy Content-only dissimilarity	0.01	0.01	0.90	76770	.366

Note. Random effects: dyad variance = 0.0149 (SD = 0.1219); video variance = 0.0065 (SD = 0.0808); Residual variance = 0.1794 (SD = 0.4236). N = 84,300 (11,148 dyads, 10 videos). DV = 1 − Pearson r (trait profile dissimilarity). Primacy Content-only dissimilarity computed from the first 20% of each trajectory. Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S7

Study 2 Linear Mixed-Effects RSA: Socially-Relevant Descriptors Only

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.64	0.03	24.29	10.56	< .001
Trajectory dissimilarity	0.15	0.03	5.20	70670	< .001
Content-only dissimilarity	0.51	0.03	15.97	67370	< .001

Note. Random effects: dyad variance = 0.0140 (SD = 0.1181); video variance = 0.0064 (SD = 0.0801); Residual variance = 0.1794 (SD = 0.4235). N = 84,946 (11,148 dyads, 10 videos). DV = 1 − Pearson r (trait profile dissimilarity). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S8

Study 2 Linear Mixed-Effects RSA: Mean Valence Covariate

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.60	0.03	21.60	10.51	< .001
Trajectory dissimilarity	0.14	0.03	5.01	59910	< .001
Content-only dissimilarity	0.35	0.03	10.37	60530	< .001
Valence dissimilarity (mean)	0.21	0.01	36.32	84670	< .001

Note. Random effects: dyad variance = 0.0127 (SD = 0.1127); video variance = 0.0071 (SD = 0.0842); Residual variance = 0.1771 (SD = 0.4208). N = 84,946 (11,148 dyads, 10 videos). DV = 1 − Pearson r (trait profile dissimilarity). Valence dissimilarity = absolute difference in mean NRC VAD valence. Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S9

Study 2 Linear Mixed-Effects RSA: Valence Trajectory DTW Covariate

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.61	0.03	22.00	10.53	< .001
Trajectory dissimilarity	0.08	0.03	2.88	60240	.004
Content-only dissimilarity	0.45	0.03	13.08	61500	< .001
Valence dissimilarity (DTW)	0.17	0.01	26.21	84790	< .001

Note. Random effects: dyad variance = 0.0131 (SD = 0.1146); video variance = 0.0071 (SD = 0.0840); Residual variance = 0.1781 (SD = 0.4220). N = 84,818 (11,148 dyads, 10 videos). DV = 1 − Pearson r (trait profile dissimilarity). Valence dissimilarity = DTW distance between 1-D valence trajectories (abs-diff cost, 20% Sakoe-Chiba band). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S10

Study 2 Grid Search Robustness Check for DTW Sakoe-Chiba Band Width

Band Width	*b*	*SE*	*t*	*p*
5%	0.132	0.028	4.71	<.001
10%	0.131	0.028	4.68	<.001
15%	0.130	0.028	4.63	<.001
20%	0.129	0.028	4.61	<.001
25%	0.129	0.028	4.60	<.001
30%	0.129	0.028	4.60	<.001
35%	0.129	0.028	4.60	<.001
40%	0.129	0.028	4.60	<.001
45%	0.130	0.028	4.61	<.001
50%	0.130	0.028	4.62	<.001

Note. Sensitivity of the trajectory dissimilarity effect to Sakoe–Chiba band width. Regression coefficients (*b*) and standard errors (*SE*) for trajectory dissimilarity predicting impression dissimilarity (1−Pearson *r*) across ten band widths (5–50% of the 100-bin time grid), using linear interpolation throughout. All models include Content-only dissimilarity as a covariate and crossed random intercepts for video and dyad. The shaded row corresponds to the 20% band used in the main analysis.

Table S11

Study 3 Grid Search Robustness Check for DTW Sakoe-Chiba Band Width

Band Width	*b*	*SE*	*t*	*p*
5%	0.398	0.017	23.78	<.001
10%	0.396	0.017	23.71	<.001
15%	0.395	0.017	23.65	<.001
20%	0.394	0.017	23.60	<.001
25%	0.394	0.017	23.55	<.001
30%	0.393	0.017	23.52	<.001
35%	0.393	0.017	23.50	<.001
40%	0.393	0.017	23.49	<.001
45%	0.392	0.017	23.47	<.001
50%	0.392	0.017	23.45	<.001

Note. Sensitivity of the trajectory dissimilarity effect to Sakoe–Chiba band width. Regression coefficients (*b*) and standard errors (*SE*) for trajectory dissimilarity predicting impression dissimilarity (1−Pearson *r*) across ten band widths (5–50% of the 100-bin time grid), using linear interpolation throughout. All models include Content-only dissimilarity as a covariate and crossed random intercepts for video and dyad. The shaded row corresponds to the 20% band used in the main analysis.

Table S12

Study 4 Grid Search Robustness Check for DTW Sakoe-Chiba Band Width

Band Width	*b*	*SE*	*t*	*p*
5%	0.420	0.018	22.79	<.001
10%	0.418	0.018	22.68	<.001
15%	0.417	0.018	22.61	<.001
20%	0.416	0.018	22.55	<.001
25%	0.415	0.018	22.52	<.001
30%	0.415	0.018	22.50	<.001
35%	0.415	0.018	22.50	<.001
40%	0.415	0.018	22.50	<.001
45%	0.415	0.018	22.50	<.001
50%	0.415	0.018	22.50	<.001

Note. Sensitivity of the trajectory dissimilarity effect to Sakoe–Chiba band width. Regression coefficients (b) and standard errors (SE) for trajectory dissimilarity predicting impression dissimilarity (1−Pearson r) across ten band widths (5–50% of the 100-bin time grid), using linear interpolation throughout. All models include Content-only dissimilarity as a covariate and crossed random intercepts for video and dyad. The shaded row corresponds to the 20% band used in the main analysis.

Table S13

Study 3 Linear Mixed-Effects Ratings RSA: Full Model

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.42	0.02	25.44	19.87	< .001
Trajectory dissimilarity	0.39	0.02	23.60	190129.91	< .001
Content-only dissimilarity	0.74	0.02	40.85	192584.00	< .001

Note. Random effects: dyad variance = 0.0257 (SD = 0.1602); video variance = 0.0046 (SD = 0.0675); Residual variance = 0.1161 (SD = 0.3408). N = 194,225 (45,665 dyads, 18 videos). DV = 1 − Pearson r (trait profile dissimilarity). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S14

Study 3 Linear Mixed-Effects Ratings RSA: Primacy Control

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.40	0.02	22.35	19.61	< .001
Trajectory dissimilarity	0.93	0.01	73.68	183031.53	< .001
Primacy Content-only dissimilarity	-0.05	0.01	-7.73	188102.39	< .001

Note. Random effects: dyad variance = 0.0260 (SD = 0.1613); video variance = 0.0053 (SD = 0.0725); Residual variance = 0.1166 (SD = 0.3415). N = 188,357 (45,456 dyads, 18 videos). DV = 1 − Pearson r (trait profile dissimilarity). Primacy Content-only dissimilarity computed from the first 20% of each trajectory. Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S15

Study 3 Linear Mixed-Effects Ratings RSA: Socially-Relevant Descriptors Only

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.46	0.02	27.77	19.73	< .001
Trajectory dissimilarity	0.29	0.02	17.53	193774.14	< .001
Content-only dissimilarity	0.73	0.02	42.63	193497.99	< .001

Note. Random effects: dyad variance = 0.0255 (SD = 0.1597); video variance = 0.0046 (SD = 0.0675); Residual variance = 0.1166 (SD = 0.3414). N = 194,225 (45,665 dyads, 18 videos). DV = 1 − Pearson r (trait profile dissimilarity). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S16

Study 3 Linear Mixed-Effects Ratings RSA: Mean Valence Covariate

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.43	0.01	31.46	21.26	< .001
Trajectory dissimilarity	0.34	0.02	21.08	189866.12	< .001
Content-only dissimilarity	0.33	0.02	18.29	192194.68	< .001
Valence dissimilarity (mean)	0.29	0.00	94.39	193579.95	< .001

Note. Random effects: dyad variance = 0.0244 (SD = 0.1563); video variance = 0.0030 (SD = 0.0546); Residual variance = 0.1111 (SD = 0.3333). N = 194,225 (45,665 dyads, 18 videos). DV = 1 − Pearson r (trait profile dissimilarity). Valence dissimilarity = absolute difference in mean NRC VAD valence. Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S17

Study 3 Linear Mixed-Effects Ratings RSA: Valence Trajectory DTW Covariate

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.44	0.01	31.58	21.17	< .001
Trajectory dissimilarity	0.22	0.02	13.24	188415.17	< .001
Content-only dissimilarity	0.49	0.02	26.78	191753.46	< .001
Valence dissimilarity (DTW)	0.27	0.00	83.57	193191.47	< .001

Note. Random effects: dyad variance = 0.0244 (SD = 0.1562); video variance = 0.0031 (SD = 0.0555); Residual variance = 0.1123 (SD = 0.3351). N = 193,626 (45,659 dyads, 18 videos). DV = 1 − Pearson r (trait profile dissimilarity). Valence dissimilarity = DTW distance between 1-D valence trajectories (abs-diff cost, 20% Sakoe-Chiba band). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S18

Study 3 Controls and Robustness Analyses for Mediation RSA

Specification	Path a (SE)	Path b (SE)	Path c’ (SE)	ACME [95% CI]	Prop. Mediated [95% CI]
MAIN	0.891 (0.011)	1.364 (0.014)	3.653 (0.070)	0.023 [0.021, 0.024]	0.249 [0.240, 0.259]
CONTENT	0.394 (0.017)	1.340 (0.014)	1.783 (0.100)	0.010 [0.009, 0.011]	0.229 [0.205, 0.256]
IMP-REL	0.292 (0.017)	1.350 (0.014)	1.050 (0.101)	0.007 [0.006, 0.008]	0.274 [0.232, 0.322]
VAL-MEAN	0.345 (0.016)	1.069 (0.015)	1.789 (0.103)	0.007 [0.006, 0.008]	0.171 [0.151, 0.193]
VAL-TRAJ	0.219 (0.017)	1.127 (0.015)	0.807 (0.102)	0.005 [0.004, 0.005]	0.236 [0.189, 0.297]

Note. Path a (trajectory → ratings dissimilarity) coefficients are on the identity scale (linear mediator model). Paths b and c' are on the log-odds scale (logistic outcome model). ACME and proportion mediated are on the probability scale, representing the change in P(decision disagreement) for a one-SD increase in trajectory dissimilarity. All specifications include crossed random intercepts for video and dyad. CONTENT adds the cosine distance between participants' average semantic content as a covariate in both models; all subsequent specifications also include this CONTENT covariate. IMP-REL restricts trajectory and CONTENT computation to socially-relevant descriptors (personality, emotion, appearance, and social category terms). VAL-MEAN adds the absolute difference in mean NRC-VAD valence score between the two participants in each dyad as a scalar covariate. VAL-TRAJ adds the DTW distance between participants' one-dimensional NRC-VAD valence trajectories as a covariate, computed using the same 100-bin interpolation and 20% Sakoe-Chiba band as the semantic trajectories but with absolute valence difference as the local cost function. The shaded row (MAIN) corresponds to the primary specification reported in the main text. All ACMEs and proportions mediated: $p < .001$.

Table S19

Impression Dimensions Predicted by Recruitment Decision (Study 3)

Dimension	*b*	*SE*	*t*	*df*	*p*	p_{FDR}
Openness	1.962	0.073	26.85	2547	< .001	< .001
Conscientiousness	1.739	0.074	23.62	2342	< .001	< .001
Extroversion	1.939	0.100	19.33	2611	< .001	< .001
Agreeableness	1.621	0.072	22.47	2431	< .001	< .001
Neuroticism [R]	0.883	0.087	10.20	1076	< .001	< .001
Warmth	1.934	0.080	24.19	2558	< .001	< .001
Competence	2.246	0.068	33.10	2521	< .001	< .001
Confidence	2.754	0.086	31.94	2610	< .001	< .001
Leadership	3.178	0.086	36.77	2603	< .001	< .001
Ambitiousness	2.239	0.070	31.94	2531	< .001	< .001
Trustworthiness	2.235	0.069	32.33	2234	< .001	< .001

Note. Each row reports the fixed-effect coefficient for Recruitment Decision (Recruit = 1, Reject = 0) from a separate linear mixed-effects model: Dimension ~ Decision + (1 | Participant) + (1 | Video). Betas are unstandardized; SE = standard error; df = Satterthwaite degrees of freedom; p = two-tailed p-value; p_{FDR} = p-value after Benjamini-Hochberg false discovery rate correction across all 11 tests. Neuroticism is reverse-coded prior to modeling, indicated by "[R]" next to the label.

Table S20

Study 4 Video Segmentation and Boundary Stability

Video	Body window (s)	Boundaries (s)	Segment durations (s)	Reselection
Video 1	8.2–68.3	23.5 / 45.7 / 57.6	15.3 / 22.2 / 12.0 / 10.7	.59 / .90 / 1.00
Video 2	12.3–118.7	46.4 / 89.6 / 105.4	34.1 / 43.2 / 15.8 / 13.2	.73 / .72 / .72
Video 3	8.8–125.4	51.7 / 93.9 / 106.8	42.9 / 42.2 / 12.9 / 18.6	.67 / .70 / .90
Video 4	6.5–126.9	39.5 / 79.5 / 104.5	33.0 / 40.0 / 24.9 / 22.4	.90 / .78 / .79
Video 5	18.4–95.0	39.1 / 65.3 / 77.8	20.7 / 26.1 / 12.6 / 17.1	.82 / 1.00 / .99
Video 6	1.4–77.9	29.7 / 54.9 / 67.0	28.3 / 25.2 / 12.1 / 10.8	.96 / 1.00 / 1.00
Video 7	5.5–104.0	28.2 / 64.2 / 88.6	22.7 / 36.0 / 24.4 / 15.3	.73 / .85 / .94
Video 8	6.1–92.1	24.1 / 52.1 / 81.7	18.0 / 28.0 / 29.6 / 10.4	.99 / .84 / 1.00

Note. Videos are anonymized and labeled Video 1-8. Body window = onset and offset (in s) of the reorderable body of each video, after the fixed introductory and concluding sentences were excluded from segmentation. Boundaries are the three selected sentence-end cut points, in original order; segment durations are the four resulting body segments, also in original order. Reselection = proportion of 1,000 participant-level bootstrap resamples of the Study 3 pause data in which the identical candidate sentence break was reselected, reported for each boundary in the same order. Boundaries were selected to minimize summed pause density at the cut points while penalizing segment-length imbalance, subject to 10–45 s segment-length constraints.

Table S21

Study 4 Linear Mixed-Effects Ratings RSA: Full Model

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.39	0.03	12.48	7.39	< .001
Trajectory dissimilarity	0.42	0.02	22.55	152776.81	< .001
Content-only dissimilarity	0.62	0.02	31.33	156228.68	< .001

Note. Random effects: dyad variance = 0.0266 (SD = 0.1632); video variance = 0.0075 (SD = 0.0867); Residual variance = 0.1110 (SD = 0.3332). N = 158,062 (23,003 dyads, 8 videos). DV = 1 − Pearson r (trait profile dissimilarity). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S22

Study 4 Linear Mixed-Effects Ratings RSA: Primacy Control

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.37	0.03	11.46	7.39	< .001
Trajectory dissimilarity	0.85	0.01	60.24	147675.29	< .001
Primacy Content-only dissimilarity	-0.04	0.01	-4.94	154995.97	< .001

Note. Random effects: dyad variance = 0.0266 (SD = 0.1630); video variance = 0.0080 (SD = 0.0894); Residual variance = 0.1122 (SD = 0.3349). N = 155,488 (22,997 dyads, 8 videos). DV = 1 − Pearson r (trait profile dissimilarity). Primacy Content-only dissimilarity computed from the first 20% of each trajectory. Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S23

Study 4 Linear Mixed-Effects Ratings RSA: Socially-Relevant Descriptors Only

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.42	0.03	13.25	7.36	< .001
Trajectory dissimilarity	0.36	0.02	19.57	155198.07	< .001
Content-only dissimilarity	0.56	0.02	30.09	156894.48	< .001

Note. Random effects: dyad variance = 0.0266 (SD = 0.1632); video variance = 0.0077 (SD = 0.0877); Residual variance = 0.1113 (SD = 0.3336). N = 157,862 (23,003 dyads, 8 videos). DV = 1 − Pearson r (trait profile dissimilarity). Trajectory and Content-only dissimilarity recomputed using only socially-relevant descriptors (personality, emotion, appearance, and social category terms). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S24

Study 4 Linear Mixed-Effects Ratings RSA: Mean Valence Covariate

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.39	0.03	14.70	7.51	< .001
Trajectory dissimilarity	0.35	0.02	19.44	153000.31	< .001
Content-only dissimilarity	0.25	0.02	12.79	155796.60	< .001
Valence dissimilarity (mean)	0.28	0.00	82.66	156238.72	< .001

Note. Random effects: dyad variance = 0.0260 (SD = 0.1611); video variance = 0.0056 (SD = 0.0745); Residual variance = 0.1063 (SD = 0.3260). N = 158,062 (23,003 dyads, 8 videos). DV = 1 − Pearson r (trait profile dissimilarity). Valence dissimilarity = absolute difference in mean NRC VAD valence. Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S25

Study 4 Linear Mixed-Effects Ratings RSA: Valence Trajectory DTW Covariate

Predictor	*b*	*SE*	*t*	*df*	*p*
Intercept	0.40	0.03	14.70	7.51	< .001
Trajectory dissimilarity	0.25	0.02	13.74	152669.76	< .001
Content-only dissimilarity	0.37	0.02	18.75	156308.37	< .001
Valence dissimilarity (DTW)	0.27	0.00	73.08	156396.44	< .001

Note. Random effects: dyad variance = 0.0264 (SD = 0.1626); video variance = 0.0057 (SD = 0.0752); Residual variance = 0.1071 (SD = 0.3273). N = 158,062 (23,003 dyads, 8 videos). DV = 1 − Pearson r (trait profile dissimilarity). Valence dissimilarity = DTW distance between 1-D valence trajectories (abs-diff cost, 20% Sakoe-Chiba band). Predictors on raw (unstandardized) scale. Random intercepts for video and dyad.

Table S26
Study 4 Indirect Effects of Viewing Order on Endpoint Impressions and Decisions

Specification	Path a (SE)	Path b (SE)	Path c′ (SE)	Indirect [95% CI]	*p*
Trajectory → Impressions	0.001 (0.001)	0.828 (0.013)	0.002 (0.003)	0.001 [<0.001, 0.002]	.006
Trajectory → Decisions	0.001 (0.001)	4.936 (0.075)	-0.022 (0.018)	0.007 [0.001, 0.012]	.013
Segment content → Impressions	0.009 (0.001)	0.342 (0.009)	-0.002 (0.003)	0.003 [0.003, 0.004]	< .001
Segment content → Decisions	0.009 (0.001)	2.249 (0.053)	-0.038 (0.020)	0.021 [0.016, 0.026]	< .001

Note. The predictor in every row is viewing-order pair type (between- vs. within-order pairs); the specification column names the mediator and the outcome. Impressions = final trait-profile dissimilarity (1 − Pearson r); Decisions = binary recruitment-decision disagreement. Each row is a separate mediation model estimated by quasi-Bayesian Monte Carlo simulation (5,000 draws), sampling the a and b paths independently from their approximate posteriors and forming the indirect effect as the product of draws; the confidence interval is a percentile interval and p is a two-tailed Monte Carlo p-value. For Impressions rows all paths are on the identity scale; for Decisions rows path a is on the raw mediator scale while paths b and c′ are on the log-odds scale, so the indirect effect is a logit-scale approximation and should be read for direction and interval coverage rather than as a precise effect size. All mediator and outcome models include crossed random intercepts for video and dyad. All rows are estimated at the participant-pair level. N = 158,062 and 161,498 for the trajectory mediator and 126,053 and 129,041 for the segment-content mediator (impressions and decisions, respectively); Ns differ because mediator coverage varies with descriptor sparsity.

Table S27

Study 4 Indirect Effects of Preceding-Segment Sequence on Endpoint Impressions and Decisions

Specification	Path a (SE)	Path b (SE)	Path c′ (SE)	Indirect [95% CI]	*p*
Impressions	0.022 (0.001)	0.248 (0.006)	-0.003 (0.002)	0.005 [0.005, 0.006]	< .001
Decisions	0.022 (0.001)	1.945 (0.042)	-0.056 (0.015)	0.042 [0.039, 0.045]	< .001

Note. The predictor in every row is whether a pair viewed the same ordered sequence of preceding segments, and the mediator is segment-content dissimilarity; the specification column names the outcome. Impressions = final trait-profile dissimilarity (1 − Pearson r); Decisions = binary recruitment-decision disagreement. Each row is a separate mediation model estimated by quasi-Bayesian Monte Carlo simulation (5,000 draws), sampling the a and b paths independently from their approximate posteriors and forming the indirect effect as the product of draws; the confidence interval is a percentile interval and p is a two-tailed Monte Carlo p-value. For Impressions rows all paths are on the identity scale; for Decisions rows path a is on the raw mediator scale while paths b and c′ are on the log-odds scale, so the indirect effect is a logit-scale approximation and should be read for direction and interval coverage rather than as a precise effect size. All mediator and outcome models include crossed random intercepts for video and dyad. All rows are estimated at the pair × segment level, since the sequence of preceding segments is defined separately for each segment position, and are therefore not directly comparable in precision to Table S26. N = 193,101 (impressions) and 198,686 (decisions).

Table S28

Sample Demographics

Study	N	*Age*	*% Female*	*% White*	*% Black*	*% Asian*
Study 1	202	32.5 (5.4)	44.1	67.7	16.4	8.0
Study 2	150	39.6 (12.3)	41.6	72.8	16.0	8.8
Study 3	306	31.5 (5.3)	51.0	68.0	25.3	5.1
Study 4	215	30.8 (5.9)	57.2	58.6	30.7	7.9

Note. Gender and race were collected as non-exclusive (multi-select) fields; percentages are computed among participants who reported that field and may sum to more than 100% for participants who selected more than one category. Percentages are not computed over the full analytic N because demographic reporting was optional.

Supplementary Figures

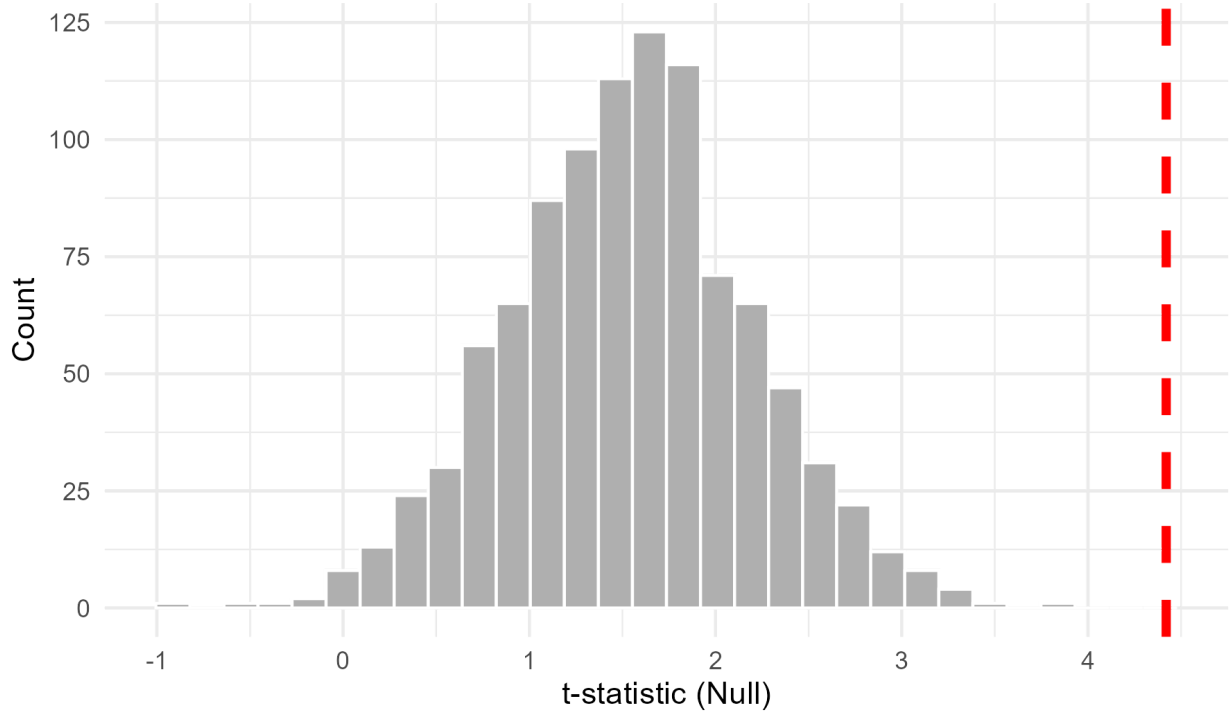

Figure S1. Study 2 distribution of trajectory dissimilarity regression coefficients (t-values) obtained under 1,000 permutation iterations in which each participant's word embeddings were randomly reassigned across their own observed timestamps before recomputing pairwise DTW cosine distances. This procedure preserves both the words produced and the moments at which they were entered, but destroys their temporal order. The red dashed line indicates the observed t-value from the main model ($t = 4.61$). No permuted coefficient exceeded the observed value, yielding $p < .001$.

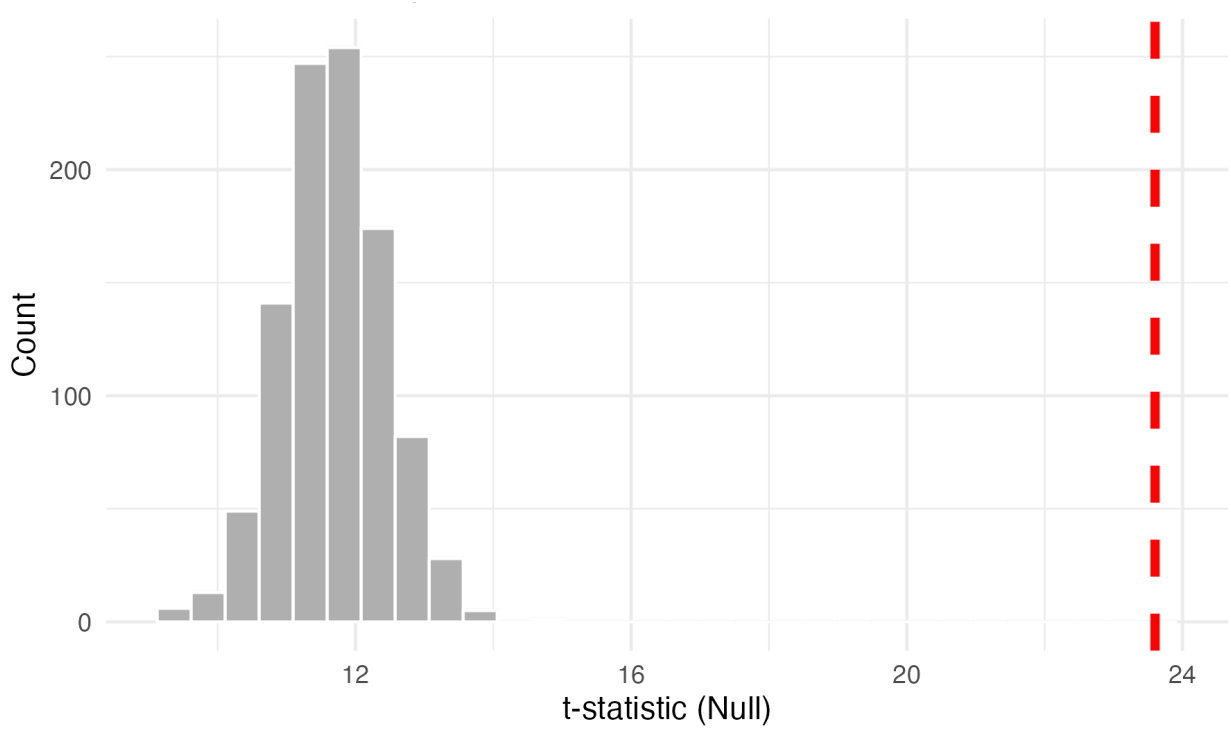

Figure S2. Study 3 distribution of t-statistics for the trajectory dissimilarity predictor obtained under 1,000 permutation iterations in which each participant's word embeddings were randomly reassigned across their own observed timestamps before recomputing pairwise DTW cosine distances. This procedure preserves both the words produced and the moments at which they were entered, but destroys their temporal order. The red dashed line indicates the observed t-value (t = 23.60). No permuted coefficient exceeded the observed value, yielding $p < .001$.

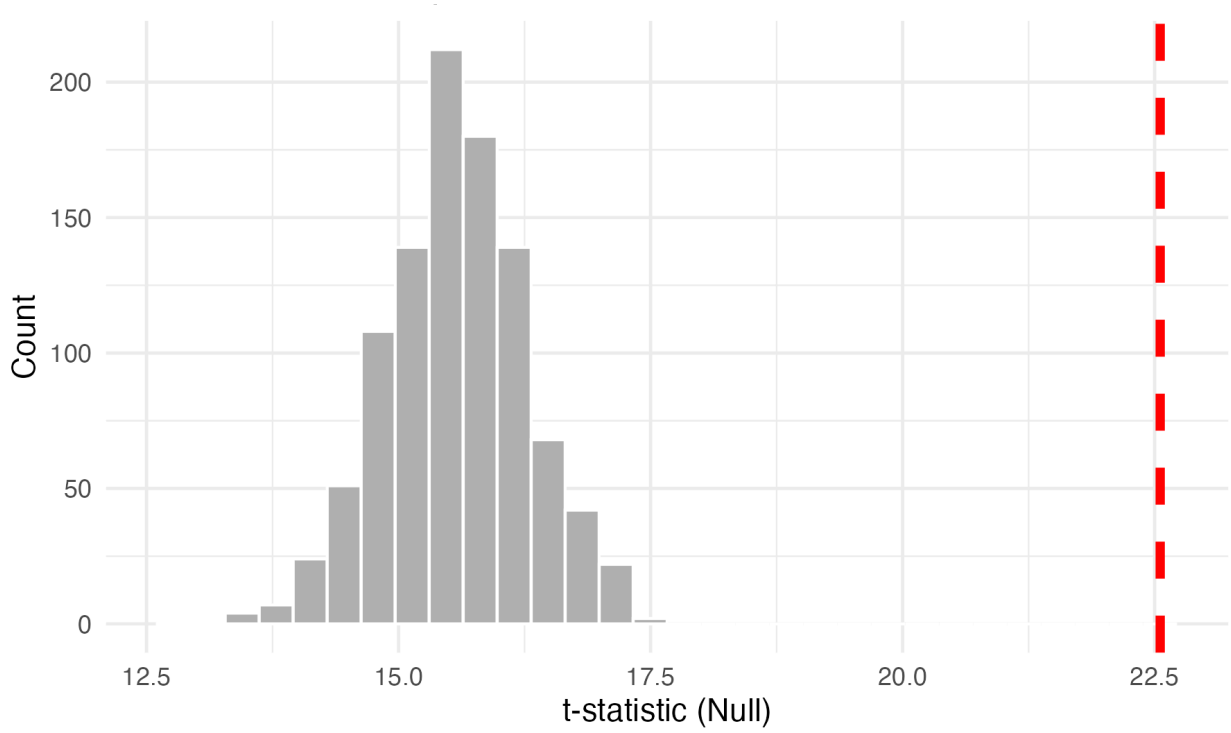

Figure S3. Study 4 distribution of t-statistics for the trajectory dissimilarity predictor obtained under 1,000 permutation iterations in which each participant's word embeddings were randomly reassigned across their own observed timestamps before recomputing pairwise DTW cosine distances. This procedure preserves both the words produced and the moments at which they were entered, but destroys their temporal order. The red dashed line indicates the observed t-value (t = 22.55). No permuted coefficient exceeded the observed value, yielding $p < .001$.

References

1. G. Nicolas, X. Bai, S. T. Fiske, A spontaneous stereotype content model: Taxonomy, properties, and prediction. *J. Pers. Soc. Psychol.* **123**, 1243–1263 (2022).

2. L. R. Goldberg, The development of markers for the Big-Five factor structure. *Psychol. Assess.* **4**, 26–42 (1992).

3. S. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych, Y. Miyao, Eds. (Association for Computational Linguistics, 2018), pp. 174–184.