Spontaneous mental restructuring in the Connections game

Solim LeGris (solim.legris@nyu.edu)[1], **Marie Komischke**[1] **& Todd M. Gureckis**[1,2]
[1]Department of Psychology, New York University
[2]Center for Data Science, New York University

Abstract

Insight is often theorized to arise from mental restructuring—abandoning an initial conceptual framework to adopt a new one. However, because of its covert nature, directly observing this process has proven difficult. We developed a modified version of the NYT Connections game, where participants organized sixteen sequentially-presented words into four categories. By manipulating word order, we induced "semantic lures"—compelling but incorrect groupings—and varied whether participants received feedback after incorrect submissions. The order manipulation reliably modulated lure formation, but did not modulate insight reports on its own. A dissociation emerged at the level of individual trajectories: among participants who formed and then abandoned the lure, those who did so without any feedback reported insight credibly more often than those who restructured following an external error signal. Our proposed paradigm provides directly observable evidence that the "aha" experience tracks the trajectory of restructuring rather than simply reaching a solution.

Keywords: insight; problem solving; games; reasoning

Introduction

Insight relies on the ability to restructure one's conceptual framework for thinking about a problem (Knoblich et al., 1999; Ohlsson, 1984; Ohlsson, 1992; Öllinger et al., 2014). Although several proposals exist (Friston et al., 2017; Löwe et al., 2024), an important question remains as to what computational details might explain how people restructure their representations on nontrivial tasks. We posit that progress on this question depends in part on having empirical paradigms that can constrain models at the level of representational change itself, rather than at the level of solution outcomes alone. Yet, testing the representational restructuring claim empirically has proven to be particularly difficult, since most of the thinking relevant for providing evidence of the phenomenon occurs covertly. Although prior work has sought to explore this phenomenon through process-tracing experiments (Berckley & Hattie, 2023; Jones, 2003; LeGris et al., 2024; Stephen & Dixon, 2008), much of the literature has used outcome-only paradigms (e.g., riddles, compound remote associates, verbal problems, or arithmetic problems) prohibiting the measurement of process-level signatures of restructuring. Prior paradigms that induce representational change have made important progress, but to the best of our knowledge have presupposed it in the construction of their stimuli, correlating proxies of it post-hoc with insight (e.g., word vector distance; Chesebrough et al., 2023). In this work, we propose a paradigm that directly traces representational change, and we attempt to manipulate it experimentally.

We instantiate this approach using a task that has grown in popularity (over 3.3 billion plays in 2024; Peters, 2025) in part because of the routine feeling of insight it generates: the New York Times (NYT) *Connections* game. In this puzzle game, players must organize sixteen words into four disjoint groups, each sharing a common theme. Prior to the rise of large language models (LLMs) post-trained to incentivize "reasoning" (Guo et al., 2025), the task was found to be difficult for LLMs, precisely because it requires more than simply identifying low-level features that leverage semantic similarity (Todd et al., 2024). Notably, many puzzles in this task induce a tension between the surface similarity of the words evoking possible partitions, and global constraints that make those partitions unlikely. More concretely, it is often the case that semantically coherent groupings conflict with the constraints of the game: either too many words fall under a possible category, or not enough (see Figure 1 for an example of the former). This feature, paired with its ability to elicit "aha" moments, makes *Connections* valuable for studying insight in people: puzzle solvers may initially consider semantic groupings that have high conceptual cohesiveness despite being globally incorrect, thereby requiring restructuring to succeed at the task.

The present research

In the present study, we designed a variation of the *Connections* game in which we directly observe sequences of guesses and group formation, measure timing and "change-of-thinking", and manipulate the order of presentation of words. The proposed paradigm is unlike many other insight problem-solving experiments: instead of presupposing restructuring based on task structure and observed outcome, it affords direct observation of representational change. Inspired by other work (Clapper & Bower, 1994; Elio & Anderson, 1984), we use stimuli sequences with predetermined orderings. The sequences are designed using vector word embeddings, either to facilitate correct groupings early, or conversely to encourage incorrect groupings that will later need to be revised. Throughout, we treat insight self-reports not as the object of study per se, but as a correlated readout of restructuring, in line with mechanistic accounts of the "aha" experience (Öllinger et al., 2014). Importantly, this readout is not a mere artifact: insight phenomenology has been shown to correlate

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

with genuine understanding and discovery (Barot et al., 2024; Laukkonen, 2024).

We hypothesized that word order would modulate what representations participants would form, and that in doing so we could influence the likelihood that representational change would be required to solve a given puzzle. To do so, we developed two main conditions alongside two control conditions. The two main conditions consisted of one that aimed to facilitate solution finding (i.e., *anti-lure*), and one that aimed to inhibit it (i.e., *luring*). We predicted higher rates of restructuring, and accordingly more frequent insight reports, in the *luring* condition than in the *anti-lure* condition. Although we do not find a direct modulation of insight rates between these two conditions, we find that the manipulation credibly modulates the tendency of participants to form the lure category. More importantly, analyses of word-grouping trajectories validate the experimental paradigm. We find credible differences in insight reports emerging directly from whether participants adopt the lure category and later abandon it, and this signature persists when solving outcome is held constant. Among participants who solved the puzzle after breaking out of the lure, those who did so spontaneously reported insight credibly more often than those who restructured following an external error signal. This dissociation would otherwise be invisible in paradigms that cannot trace the solution process. Taken together, our findings provide direct behavioral observation of the link between restructuring and insight at the level of individual trajectories and suggest that intrinsically guided self-correction, not merely reaching the solution, underlies the "aha" experience.

Methods

In the NYT *Connections* game, the puzzles vary in overall difficulty, and the individual categories within each puzzle also differ in how difficult they are to identify. Easier categories involve identifying intuitive semantic categories like "colors" or "quantity words", while more difficult ones require identifying higher-level concepts like shared morphological structure. Importantly, this task requires constraint satisfaction since it imposes a global partition constraint: exactly four categories with four words each must be submitted.

Instead of presenting all the words at once, participants see one word at a time in an experimentally determined order (Figure 1). This allows us to influence incremental hypothesis formation and track restructuring directly. Akin to a structured think-aloud protocol, at each word, we ask participants to provide natural language hypotheses and confidence ratings capturing their belief about the underlying categories, as well as report how much each word changed their thinking (1–5). Finally, we track every word placement and reorganization event throughout the experiment to obtain direct behavioral signatures of restructuring. Bonus payments were used to incentivize efficient solving and early labeling.

Figure 1: **Example puzzle sequences.** Each panel compares the order of the first 8 words for one of the puzzles we used in our experiment. The colors indicate the true underlying category to be inferred for that puzzle. **(A)** In the *luring* condition, participants were shown the three most similar words that are *not* in the same true underlying category as the first three words. Here the words "COUPLE", "HANDFUL" and "FEW" suggest "quantities", which turns out to be incorrect. **(B)** In the anti-lure condition, participants are shown consecutive pairs of words from the same true underlying category to encourage forming the right categories from the start. The same three lure words as in the luring condition are presented consecutively at the end of the sequence (not shown here).

Experimental design

We employed a 2×4 between-subjects design crossing feedback (present, absent) with sequence condition (*luring*, *anti-lure*, *facilitated*, *dispersed*), with participants nested within 16 puzzles. Each participant was randomly assigned to one puzzle, yielding 128 unique cells.

Sequence manipulation. We manipulated word order to control the likelihood that participants would form a misleading initial grouping, which we call a "lure". We hypothesized that this manipulation would induce differences in the need for reorganization and consequently insight. In the *luring* condition, a group of three semantically similar words is first presented. Importantly, each word belongs to a different true latent category. Conversely, in the *anti-lure* condition, semantically similar pairs of words from the ground-truth categories are consecutively presented from easiest to hardest before the lure words, inoculating the participant against the lure. Finally, in the *facilitated* condition, each correct group of four words is shown consecutively, ordered in increasing group difficulty, while the *dispersed* condition minimizes consecutive semantic similarity between words. Our primary comparison of interest is *luring* vs. *anti-lure*; *facilitated* and *dispersed* serve as easy and neutral controls.

Feedback manipulation. Participants either received feedback after each submission, or submitted once without any opportunity for revision. When receiving feedback, participants were told how many boxes out of four were correct, and they were given unlimited retries, with a minimum of eight retries before being given the opportunity to give up. This condition tests how an explicit error signal might modulate restructuring and its relation to insight.

Stimuli

To generate high-quality "lure" categories, we analyzed a dataset of ~1,000 NYT puzzles. We aimed to identify three-word combinations that were semantically cohesive despite belonging to different latent categories. We quantified this

cohesiveness using average pairwise cosine similarity derived from the Google News 300d word2vec model (Mikolov et al., 2013), ultimately selecting the top forty puzzles with the strongest lures. We then generated word sequences for each condition and puzzle in this set, and used an LLM-based in-silico pilot to retain the sixteen puzzles that produced the largest restructuring effects.

Sequence generation. For each puzzle, we generated four orderings of the words. To generate the *dispersed* sequences, we devised a greedy search algorithm with random restarts satisfying two constraints: no consecutive words should be of the same true latent category, and the sum of consecutive pairwise word vector similarities should be minimal. At each restart, the algorithm picks a word at random, then picks randomly from the top three words satisfying the similarity constraint locally. For each puzzle, we ran the algorithm with 10M restarts, and picked the best sequence. To generate the *luring* sequences, we optimized the triads to maximize consecutive similarity, then we applied the same algorithm as for the *dispersed* sequences to the remaining words in the sequence. *Anti-lure* sequences proceed in three blocks: the two most prototypical non-lure exemplars from each category open the trial, followed by a third exemplar per category, with the lure words and any remaining slots presented at the end. Finally, *facilitated* sequences present each category as one contiguous block, ordered from most to least cohesive under word vector pairwise similarity, with the most prototypical exemplar leading each block.

LLM simulations. To simulate which sequences were most likely to require restructuring due to the lure, we conducted pilot simulations of our task using a closed-source LLM. Recent work validates the use of LLMs as "synthetic participants", demonstrating that they can replicate human cognitive behavioral patterns in text-based tasks to some extent (Binz & Schulz, 2023). For each puzzle, we prompted Gemini 3 Flash with each word sequentially, and asked the language model to perform the same actions as people in our experiment: place the current word in one of four boxes, move words between boxes if necessary, and emit a natural language hypothesis and confidence rating for each box. We set the reasoning token budget to low, and the temperature parameter to the default value of 1.0. For each puzzle and sequence condition, we sampled ten runs.

Selection criteria. We computed a step count statistic using only successful attempts, restricting the count to moves between boxes. We selected the top sixteen puzzles from the pilot simulations according to three criteria: (1) mean accuracy $\geq 20\%$ in both *luring* and *anti-lure* conditions; (2) positive step difference (*luring* > *anti-lure*); and (3) high combined rank on step difference (*luring* > *anti-lure*) and Cohen's d. The selected puzzles showed a mean step difference of 5.84 (SD = 2.75) and mean $d = 1.72$ (SD = 0.91, range 0.58–3.68), indicating potential for restructuring effects.

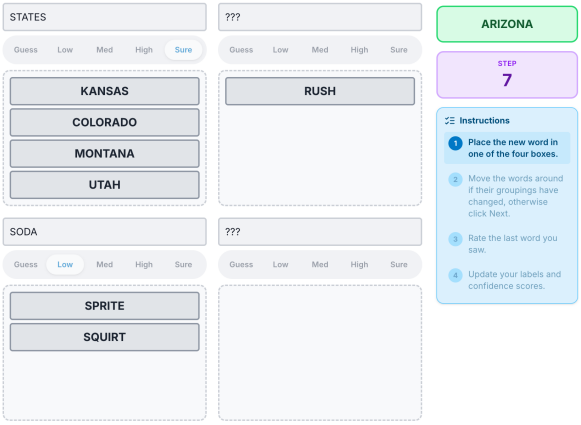

Figure 2: **Experiment web interface.** The puzzle-solving interface is displayed as seen during the "word accumulation" phase of the main task. The main area displays a 2×2 tiling of boxes where participants organize words into categories. Each box contains a text field for entering category labels (e.g., "STATES", "SODA"), and radio buttons for confidence ratings (Guess, Low, Med, High, Sure). The words to be placed (here "ARIZONA") appear in the top-right corner one at a time, with the step counter displayed below. The right panel provides step-by-step instructions, with the current action highlighted. This example shows a participant mid-puzzle who has formed a tentative category: "STATES" (high confidence) containing U.S. state names. The current word raises a potential issue with the hypothesized category, since it already contains 4 words.

Data collection

The experiment was implemented using the Smile framework (https://smile.gureckislab.org/) and deployed on Prolific. All code, stimuli, and de-identified participant data are available on GitHub (https://github.com/Le-Gris/connections-insight).

Participants. We recruited 248 participants online (49.8% male, 49% female, 1.2% other). Participants varied in age from 20.4 to 72.2 years (M=40.99, SD=11.93), and in word puzzle experience (36.9% Sometimes, 30.5% Rarely, 16.5% Often, 9.3% Every day, 6.8% Never). We chose to only recruit participants located within the United States, since some puzzles require American-centric knowledge. Participants were compensated $7.50 for an expected 30 minutes of their time, plus a potential performance bonus of up to $3.50 calculated using a logarithmic decay function based on the number of steps and the number of attempts to success, where applicable. Participants were randomly assigned to the feedback (N=131) or no-feedback (N=117) condition, one of four sequence conditions ($N_{\text{luring}} = 60$, $N_{\text{anti-lure}} = 49$, $N_{\text{dispersed}} = 69$, $N_{\text{facilitated}} = 70$), and one of sixteen puzzles. To proceed to the main task, each participant was asked to complete a comprehensive tutorial, after which a short quiz was administered to ensure understanding. If a participant failed any of the eight questions on the quiz, they were automatically taken back to the tutorial until they succeeded.

Procedure. After completing the quiz, participants were taken to the main task interface (see Figure 2). At the start of the game, participants were shown a 2×2 tiling of four empty boxes. Each box has an editable text field at the top and radio buttons for rating confidence. Words are shown one at a time to the participant in the top right corner, beside the top-right box tile. Below the word tile, a step counter indicates how many steps have been taken so far, alongside a summary of step-by-step instructions. The task can be understood as having two phases: "word accumulation" followed by a "free deliberation" period.

In the first phase of the task, the participant sees the sixteen words sequentially. The first step for each word involves placing it in one of four boxes. Second, the participant can optionally move any of the words on the board between boxes if their belief about the correct partition changes. At the next step, the participant is asked to rate how much the word changed their thinking (1–5). Finally, the participant is asked to update the label for each box that currently contains words, or skip to the next box if their hypothesis has not changed. Additionally, participants are asked to give a confidence rating for their current label for each box (1–5). Hypothesis generation and confidence ratings are encouraged through visual cues. Throughout the first phase, the instruction for the current step in the summary box is highlighted to remind the participant what they should do next (see Figure 2).

In the second phase, the participant is given unlimited time to freely move words between boxes, update the boxes' labels, and change confidence ratings. The step counter continues to increase as words are moved around. The participant can submit (or resubmit in the feedback condition) once satisfied with their solution.

We recorded several behavioral measures and collected insight self-reports during debriefing.

- **Step count:** Total placement and movement actions.
- **Score:** Number of correctly grouped categories (0–4) in each submission.
- **Change-of-thinking:** Per-word rating (1–5) of how much each word changed the participant's thinking.
- **Confidence:** Per-box rating (1–5) of certainty in the current category label and partition.
- **Hypotheses:** Per-box verbal hypothesis.
- **Insight:** Self-report of whether the participant experienced an "aha moment", collected after the task.

In this work, we restrict our analyses to insight self-reports, word placement dynamics, score and change-of-thinking.

Results

All models were fit using Bayesian mixed-effects models via the `bambi` package (Capretto et al., 2022) with weakly informative priors. We report posterior means and 95% Highest Density Intervals (HDI); effects are considered credible when the HDI excludes zero. Puzzle is treated as a random effect in all our analyses, and all models reported converged. Sample size was determined a priori to obtain approximately two participants per cell of the $2 \times 4 \times 16$ design.

Lure engagement groups. Participants who never grouped two or more lure words together were classified as "no-lure"; those who formed a partial (2/3) or full (3/3) lure category and subsequently abandoned it were classified as "broke out"; and those whose final solution contained the lure words were classified as "stuck." None of the participants in the *FB-signal* group were found to be in the "no-lure" group, leaving the no-lure × FB-signal cell structurally empty in subsequent trajectory × feedback analyses. Across the full sample, 135/248 (54.4%) participants formed and subsequently abandoned a lure category. The rate varied by feedback group: 42/117 (35.9%) in *No-FB*, 30/39 (76.9%) in *FB-silent*, and 63/92 (68.5%) in *FB-signal*.

Sequence manipulation modulates lure formation. We fit a mixed-effects model: `lure_formed ~ sequence_condition + (1 | puzzle)`, Bernoulli, $N = 248$ across $J = 16$ puzzles, with *luring* as the reference level. We find lure formation to be 15.0 percentage points more likely under *luring* condition than *anti-lure* (HDI $[0.034, 0.260]$, $P(> 0) = 0.998$). All credible separation came from cross-cluster comparisons between lure-promoting (*luring*, *dispersed*) and lure-suppressing (*facilitated*, *anti-lure*) conditions; within-cluster contrasts were not credible. The design effectively yielded a 2-level lure-promoting / lure-suppressing contrast rather than four cleanly separated levels. We also find substantial puzzle-level heterogeneity in how likely one is to form a lure category, $\sigma = 1.01$ (HDI $[0.18, 1.88]$). This result is comparable in magnitude to the *luring-anti-lure* effect itself, motivating the inclusion of puzzle as a random effect across all subsequent models. Having determined that the manipulation modulates lure formation, we next ask whether it produces corresponding effects on self-reported insight.

Sequence condition has no marginal effect on insight. We fit a mixed-effects model: `insight ~ sequence_condition + (1 | puzzle)`, Bernoulli, $N = 236$ across $J = 16$ puzzles, with *luring* as the reference level. We find no credible main effect of sequence condition on insight. Participants in the *luring* condition were found to be 6.4 percentage points more likely than *anti-lure* participants to report an insight, but this difference was not credible under our model (HDI $[-0.102, 0.244]$, $P(> 0) = 0.77$). This was the largest difference, and all other pairwise contrasts had HDIs that include zero. Puzzle-level heterogeneity in insight reports was smaller than for lure formation, $\sigma = 0.54$ (HDI [0.01, 0.97]) versus $\sigma = 1.01$, suggesting that insight is more participant-driven than puzzle-driven. This result suggests that producing more lure formation does not, on its own, produce more insight reports. To investigate this further, we examined whether insight effects emerge conditional on lure trajectory and error signal exposure.

Spontaneous restructuring tracks insight beyond solv-

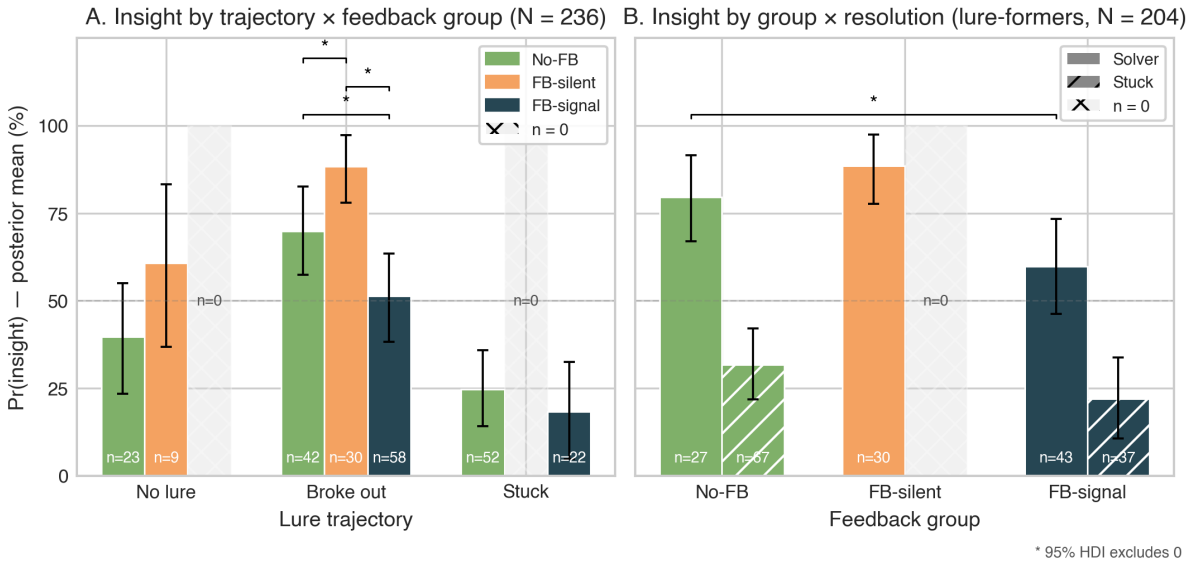

Figure 3: **Insight report rates by lure trajectory and feedback exposure.** **(A)** Posterior insight probabilities for the full sample (N=236), crossing lure trajectory (*No lure*, *Broke out*, *Stuck*) with feedback group: *No-FB* (no-feedback condition), *FB-silent* (feedback condition; solved on first attempt and never triggered the signal), *FB-signal* (feedback condition; received at least one error signal). Hatched grey cells correspond to empty cells: *FB-signal* × *No lure* is empty because all *FB-signal* formed the lure, and *FB-silent* × *Stuck* is empty by construction (*FB-silent* is defined by first-attempt success). **(B)** Same posterior probabilities restricted to lure-formers (N=204), crossing feedback group with final outcome (*Solver* vs. *Stuck*). *FB-silent* × *Stuck* is a structural zero. Bars show posterior means; error bars are 95% Highest Density Intervals; cell counts (n) are printed inside each bar. Asterisks mark contrasts whose 95% HDI excludes zero.

ing alone. To test whether engagement with the lure modulates the effect of feedback on insight, we crossed lure trajectory with feedback group and fit a mixed-effects model: `insight ~ lure_trajectory * feedback_group + (1 | puzzle)`, Bernoulli, $N = 236$ across $J = 16$ puzzles. Within the "broke out" trajectory, the two feedback groups who restructured without an external error signal (*FB-silent* at 0.88, HDI $[0.78, 0.97]$; *No-FB* at 0.70, HDI $[0.57, 0.83]$) reported insight more often than the group whose restructuring was prompted by a signal (*FB-signal* at 0.51, HDI $[0.38, 0.63]$). "Broke out" participants in *No-FB* reported insight 18 percentage points more often than those in *FB-signal* (HDI $[0.009, 0.361]$, $P(> 0) = 0.98$), and those in *FB-silent* reported insight 37 percentage points more often than those in *FB-signal* (HDI $[0.201, 0.532]$, $P(> 0) = 1.00$). No such effect emerged among *stuck* participants (*No-FB* vs. *FB-signal*: +0.064, HDI $[-0.115, 0.234]$, $P(> 0) = 0.78$), localizing the gap to participants who actually restructured (Figure 3A).

Because *FB-silent* is selected post-randomization on first-attempt success, the *FB-silent* vs. *FB-signal* contrast risks confounding signal exposure with puzzle difficulty: using each puzzle's *No-FB* completion rate as a proxy, we found that *FB-silent* participants are concentrated on easier puzzles ($\gamma = +2.45$, HDI $[+0.83, +4.31]$, $P(> 0) = 0.998$). To isolate the signal effect from this selection bias, we compared *No-FB* and *FB-signal* directly while holding outcome constant by restricting to solvers: `insight ~ feedback_group + (1 | puzzle)`, $N = 81$ across $J = 16$ puzzles. Solvers in *No-FB* reported insight at 0.78 (HDI $[0.66, 0.90]$) versus 0.60 (HDI $[0.45, 0.73]$) in *FB-signal*, an 18 percentage-point contrast (HDI $[-0.002, 0.352]$, $P(> 0) = 0.97$); restricting further to solvers who broke out of the lure ($N = 70$, $J = 14$) sharpened it to 23 percentage points (HDI $[+0.038, +0.392]$, $P(> 0) = 0.99$; Figure 3B). Although sequence condition alone does not modulate insight, how participants engage with the lure and whether they receive an error signal jointly do: it is self-driven restructuring, not solving *per se*, that tracks insight.

Exploratory analyses. We conducted several exploratory analyses to characterize the restructuring process in greater detail. After each word, participants rated their change-of-thinking (1–5 scale; $M = 2.55$, $SD = 1.45$). We fit a mixed-effects model: `change_of_thinking_rating ~ position_centered * sequence_condition + (1 | participant) + (1 | puzzle)`, Gaussian, $N = 3{,}698$ ratings from 248 participants across $J = 16$ puzzles. Change-of-thinking ratings credibly increased over word positions ($b = 0.06$, 95% HDI $[0.04, 0.08]$), suggesting that change-of-thinking is more likely to occur as the puzzle progresses and constraints become more obvious. Notably, participants in the *dispersed* condition showed a flatter change-of-thinking trajectory compared with the *luring* condition ($b = -0.03$, 95% HDI $[-0.06, -0.01]$). We also

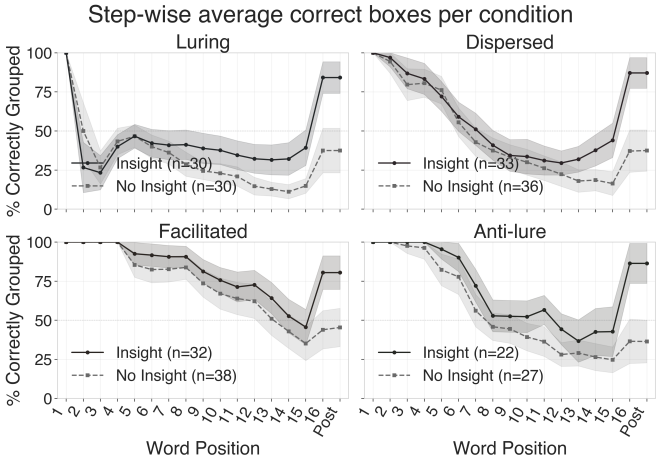

Figure 4: **Grouping correctness by step and insight.** Each panel shows mean % correctly grouped at each step for participants who reported insight vs. no insight, averaged across participants and puzzles (steps 1–16 and final "Post" submission). We observe characteristic trajectories indicating differences between conditions and insight groups in puzzle-solving dynamics.

present descriptive analyses of puzzle-solving dynamics by looking at the average number of correct boxes across participants at each step of the game. Figure 4 reveals distinct problem-solving signatures across conditions and a stark divergence by insight report. In the *luring* condition, we observe a characteristic U-shaped trajectory: grouping correctness initially declines as participants interact with the lure, followed by a sharp recovery towards the end—but only for those who report insight. In contrast, participants who report no insight exhibit a "flatline" trajectory, failing to recover from the lure's interference.

Discussion

We find that, in the no-feedback condition, a non-trivial fraction of participants who initially adopted the lure category subsequently restructured away from it. Although feedback was unavailable to revise incorrect groupings, the conjunction of accumulating semantic evidence inconsistent with an early hypothesis and an explicit partition constraint appears to have provided sufficient signal for restructuring.

More importantly, a striking pattern emerges once we condition on observed restructuring, as operationalized in this work. Among participants who broke out of the lure, those who did so without an error signal reported insight credibly more often than those whose restructuring was prompted by an external error signal—a contrast that survives holding final solving outcome constant. No analogous effect was observed among participants who remained stuck. The dissociation therefore implicates the dynamics of restructuring itself, not just whether a solution was found. Consistent with this localization, puzzle-level heterogeneity was substantially larger for lure formation than for insight reports, suggesting that the structural opportunity for restructuring is largely puzzle-driven, while whether that restructuring is experienced as insight is comparatively participant-driven. These findings align with longstanding characterizations of insight as marked by sudden, representational reorganization rather than incremental search (Chronicle et al., 2004; Gilhooly & Murphy, 2005; Kounios & Beeman, 2014).

A small fraction of participants reported insight without ever forming the lure explicitly. It remains unclear whether such insights are false alarms or resolutions over alternative or partial groupings. Participants on the no-lure trajectory may plausibly have encountered and resolved alternative ambiguities not captured by our analyses. We leave exploring this question to future work.

The present research has several further limitations. First, the sequence manipulation effectively collapsed into a two-level lure-promoting versus lure-suppressing contrast rather than the four cleanly separated conditions we had aimed at: *anti-lure* and *facilitated* did not differ credibly from one another, nor did *luring* and *dispersed*. The semantic salience of our lures appears to be strong enough that fine-grained ordering effects within each cluster are dominated by whether the lure words appear consecutively or are scattered throughout the sequence. Second, the *FB-silent* group is selected post-randomization on first-attempt success, introducing a potential confound. We addressed this with the outcome-constant *No-FB* versus *FB-signal* contrast, but contrasts involving *FB-silent* should be interpreted descriptively. Third, our paradigm trades ecological validity for process visibility: incremental presentation and per-step probes likely alter the dynamics of solving relative to the original game. Finally, much of the rich data we have collected from participants has yet to be analyzed. For instance, step-by-step natural language hypotheses are bound to be informative about the availability and difficulty of concepts for solving the task, and how this relates to restructuring.

As in our task, particularly in the no-feedback condition, many real-world problems have the flavor of unsupervised learning or inference over hidden latent structure. Difficult problems whose solutions were found through insight almost never involve an omniscient oracle who can guide the problem-solver or learner in their quest for solutions or answers. Particularly in science, constraints often appear as conflicts with new data that cannot be accounted for, a phenomenon that has been theorized to drive scientific revolutions, otherwise known as paradigm shifts (Kuhn, 1962). Thus, the scientist who notices discrepancies in how their model or theory accommodates new data must identify why or how violations arise, which can then lead to restructuring and insight. A formal account of the mechanisms underlying these abilities is still an open problem. By providing direct behavioral evidence of conditions under which insight arises, this project lays the groundwork for future modeling work to provide a computational account of restructuring in people and machines.

Acknowledgments

We acknowledge the use of AI-assisted code editors for writing, editing, and debugging both experiment and analysis code.

References

Barot, C., Chevalier, L., Martin, L., & Izard, V. (2024). "now I get it!": Eureka experiences during the acquisition of mathematical concepts. *Open Mind*, *8*, 17–41. https://doi.org/10.1162/opmi_a_00116

Berckley, J., & Hattie, J. (2023). Making learning visible: Observable correlates of the aha! moment when moving from surface to deep thinking. *J. Creat. Behav.*, *57*(3), 439–449. https://doi.org/10.1002/jocb.589

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2022). Bambi: A simple interface for fitting bayesian linear models in python. *Journal of Statistical Software*, *103*(15), 1–29. https://doi.org/10.18637/jss.v103.i15

Chesebrough, C., Chrysikou, E. G., Holyoak, K. J., Zhang, F., & Kounios, J. (2023). Conceptual Change Induced by Analogical Reasoning Sparks Aha Moments. *Creativity Research Journal*, *35*(3), 499–521. https://doi.org/10.1080/10400419.2023.2188361

Chronicle, E. P., MacGregor, J. N., & Ormerod, T. C. (2004). What makes an insight problem? the roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *J. Exp. Psychol. Learn. Mem. Cogn.*, *30*(1), 14–27. https://doi.org/10.1037/0278-7393.30.1.14

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 443–460. https://doi.org/10.1037/0278-7393.20.2.443

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & Cognition*, *12*, 20–30. https://api.semanticscholar.org/CorpusID:34767954

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.*, *29*(10), 2633–2683. https://doi.org/10.1162/neco_a_00999

Gilhooly, K. J., & Murphy, P. (2005). Differentiating insight from non-insight problems. *Think. Reason.*, *11*(3), 279–302. https://doi.org/10.1080/13546780442000187

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., . . . Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*(8081), 633–638. https://doi.org/10.1038/s41586-025-09422-z

Jones, G. (2003). Testing two cognitive theories of insight. *J. Exp. Psychol. Learn. Mem. Cogn.*, *29*(5), 1017–1027. https://doi.org/10.1037/0278-7393.29.5.1017

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, memory, and cognition*, *25*(6), 1534.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annu. Rev. Psychol.*, *65*, 71–93. https://doi.org/10.1146/annurev-psych-010213-115154

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago.

Laukkonen, R. E. (2024). The adaptive function of insight. *The Emergence of Insight*, 183–198.

LeGris, S., Lake, B., & Gureckis, T. M. (2024). Predicting insight during physical reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Löwe, A. T., Touzo, L., Muhle-Karbe, P. S., Saxe, A. M., Summerfield, C., & Schuck, N. W. (2024). Abrupt and spontaneous strategy switches emerge in simple regularised neural networks. *PLoS Computational Biology*, *20*(10), e1012505.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ohlsson, S. (1984). Restructuring revisited. *Scand. J. Psychol.*, *25*(2), 117–129. https://doi.org/10.1111/j.1467-9450.1984.tb01005.x

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking*, *1*, 1–44.

Öllinger, M., Jones, G., & Knoblich, G. (2014). The dynamics of search, impasse, and representational change provide a coherent explanation of difficulty in the nine-dot problem. *Psychol. Res.*, *78*(2), 266–275. https://doi.org/10.1007/s00426-013-0494-8

Peters, J. (2025, June). The New York Times is starting to test its take on Scrabble. https://www.theverge.com/games/682063/the-new-york-times-nyt-games-crossplay-scrabble-pips

Stephen, D. G., & Dixon, J. A. (2008). The Self-Organization of insight: Entropy and power laws in problem solving. *The Journal of Problem Solving*, *2*(1), 6. https://doi.org/10.7771/1932-6246.1043

Todd, G., Merino, T., Earle, S., & Togelius, J. (2024). Missed connections: Lateral thinking puzzles for large language models. *2024 IEEE Conference on Games (CoG)*, 1–8. https://doi.org/10.1109/CoG60054.2024.10645557