Goal-Oriented Navigation in Carousel-Based Recommender Systems

Behnam Rahdari[1], Peter Brusilovsky[2]

[1]Stanford University, USA
[2]University of Pittsburgh, USA

Abstract

Carousel-based interfaces are widely used in recommender systems to present multiple recommendation lists simultaneously. In goal-oriented search, users must evaluate candidate items against specific criteria across multiple rows. We examine the role of carousel titles and interactive in-row sorting in a controlled recipe retrieval experiment (47 participants, 141 task rounds) across three interface conditions: baseline, hidden titles, and in-row sorting. The results show that in-row sorting significantly reduced task completion time (an estimated 31.6% proportional reduction in mixed-effects modeling) with descriptive trends suggesting more uniform completion times across cognitive styles, although sorting also increased selection mistakes relative to the baseline. Furthermore, hiding carousel titles significantly increased incorrect item selections, demonstrating the role of carousel titles in guiding user search. These findings provide empirical insights into how interface design and user control influence search performance in multi-list recommender systems.

Keywords

Carousel-based Interfaces, Interactive User Interfaces, Navigability, Goal-Oriented Search, Look-up Search

1. Introduction

Recommender systems increasingly present recommendations using carousel-based interfaces rather than a single ranked list [1, 2]. In these systems, horizontally scrollable carousels display multiple recommendation lists simultaneously [3]. Users scan vertically across category topics and scroll horizontally to inspect items within each row. While carousel interfaces are common in practice, their design has developed largely outside formal evaluation frameworks [4]. While most recommender systems research focuses on algorithmic accuracy or overall engagement metrics [5], the interface layout directly influences how users explore recommendations and how much effort they spend searching [6]. Direct empirical evidence remains limited on how specific carousel features affect goal-oriented search performance.

Navigation is especially impacted by the structured retrieval tasks. In goal-oriented search, users must find items that satisfy specific criteria [7]. Carousel-based interfaces provide multiple category entry points, but they can also scatter candidate items across multiple rows. This requires users to coordinate vertical scanning with horizontal scrolling, which can increase effort and cognitive demand. User goals consequently shape interaction behavior. While casual browsing allows for open exploration, goal-oriented search requires active filtering. For example, a user looking for a specific recipe must evaluate multiple constraints, such as ingredients and preparation time. In a carousel interface, tracking these criteria across separate rows increases the cognitive demands of the task.

Interface design choices directly influence this search behavior [8]. Carousel titles provide thematic context, helping users decide which rows to explore without inspecting every item. Hiding these labels forces users to infer category topics from individual items. In contrast, interactive controls allow users to adjust recommendations to their immediate needs [9, 10]. Sorting items within a row transfers organizational effort from the user to the system [11, 12].

IntRS'26: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 28, 2026, Minneapolis.
behnamr@stanford.edu (B. Rahdari); peterb@pitt.edu (P. Brusilovsky)
0000-0001-6514-912X (B. Rahdari); 0000-0002-1902-1464 (P. Brusilovsky)

This study investigates user navigation in carousel-based recommender interfaces. We evaluated two interface interventions in a controlled recipe-search experiment: hiding category labels and adding in-row sorting controls. Forty-seven participants completed 141 search tasks across three interface conditions: a standard baseline, a hidden-label condition where labels appeared only when clicked, and a sortable condition allowing in-row reordering. To evaluate navigation effort, we measured task completion time, user clicks, and incorrect item choices. We also collected cognitive style measures and usability ratings. In-row sorting reduced task duration by an estimated 31.6% in mixed-effects modeling (a 24.16-second descriptive difference compared to the baseline). Descriptive patterns also suggest that in-row sorting helped stabilize search durations across users with differing cognitive styles. In contrast, hiding labels significantly increased incorrect item choices, highlighting the role of category labels in supporting efficient navigation. These findings provide empirical guidance for designing carousel-based recommender interfaces.

2. Related Work

This section reviews research on navigability in information systems, carousel-based recommender interfaces, and cognitive factors influencing multi-list interaction.

2.1. Navigability and Interface-Aware Information Systems

Navigability describes the ease with which users can locate relevant items within an information system [6]. While earlier research focused on linear lists [7, 13], modern platforms increasingly organize content into two-dimensional layouts. Recent surveys emphasize that recommender systems need to be interface-aware, accounting for the spatial organization and interaction constraints of multi-list presentations rather than assuming isolated ranked lists [14, 15]. Empirical studies in information seeking show that navigational cues directly influence user search paths and efficiency [16, 17]. We build on this work by examining how multi-list carousel layouts and in-row sorting controls affect search performance.

2.2. Carousel-Based Recommender Interfaces

Carousel-based interfaces present recommendations across multiple horizontal rows, where each row represents a distinct category or algorithm output [3]. Prior work has focused primarily on algorithmic item selection and offline optimization for multi-list layouts. For example, carousel item selection has been modeled using multi-armed bandits [5, 18], combinatorial optimization under capacity constraints [19], and whole-page reranking models [20, 21]. To evaluate these layouts offline, researchers have proposed multi-row click models [2, 22] and coverage-based metrics [23, 24, 25].

In addition to algorithmic optimization, user studies and log analyses show that two-dimensional layouts introduce distinct vertical and horizontal position biases [26, 27]. However, evaluating interactive carousel interfaces in user studies remains challenging due to the need for specialized software infrastructure. Platforms like CARE [4] have begun addressing this gap by supporting controlled experimental environments and interaction logging for carousel layouts.

2.3. Cognitive Factors and User Control

Task complexity and interface cues shape how users navigate structured systems. Removing titles or category headers deprives users of contextual cues, increasing the effort needed to locate target items [8, 11]. Visual search studies demonstrate that clear structural cues allow users to scan and filter content more efficiently [28].

Providing interactive controls gives users the ability to adjust recommendations according to immediate task requirements [9, 12, 29]. In the domain of food and recipe recommendations, multi-list interfaces have been shown to support diverse user decision-making styles and health goals [30, 31]. Interactive

mechanisms such as in-row sorting allow users to dynamically reorganize candidate items [10]. We extend this literature by empirically evaluating how carousel titles and in-row sorting interact with user cognitive styles during goal-oriented recipe retrieval.

3. Method

This section describes the experimental methodology used to evaluate goal-oriented navigation in carousel-based recommender interfaces. We first detail the search system implementation, including dataset curation, dense retrieval pipelines, and interface conditions. We then present the user study evaluation, describing the experimental design, task structure, participant recruitment, and telemetry data filtering. The experimental system was built upon CARE [4], an open-source evaluation framework for carousel-based recommender systems. We extended CARE to support recipe retrieval from an external database and semantic clustering of recipes.

3.0.1. Dataset and Retrieval Pipeline

The recipe corpus was derived from the Food.com dataset [30]. We filtered the raw dataset to remove incomplete entries, recipes with missing preparation steps, and invalid metadata. The resulting experimental corpus comprises 23,000 recipes with structured ingredient lists, preparation times, and user review counts![32].

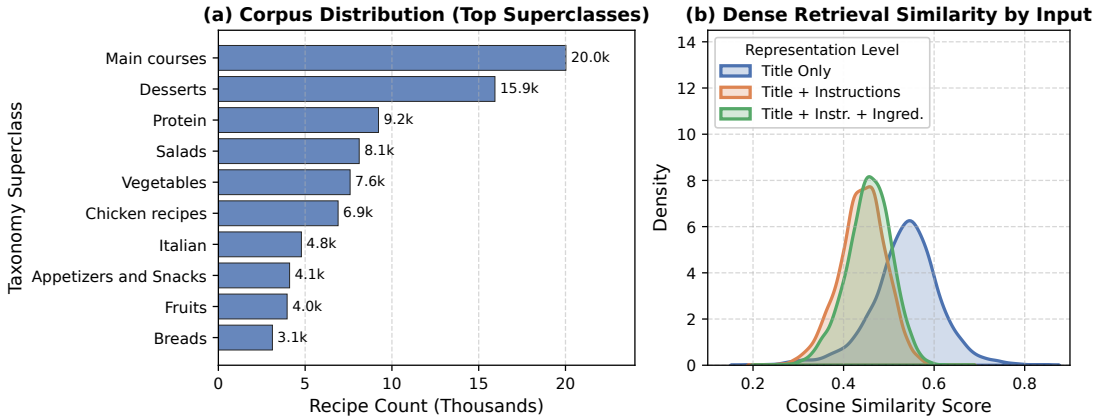

Figure 1: Recipe characteristics: (a) distribution of recipe counts across the top taxonomy superclasses, and (b) probability density of semantic similarity scores across inputs.

The retrieval pipeline organizes candidate recipes into thematical carousels. Figure 1 shows the distribution of recipe superclasses across the corpus. Category taxonomies were extracted from the hierarchical menu structure of Food.com. Each recipe title, ingredient set, and preparation instruction text was converted into a 768-dimensional dense vector using the SentenceTransformer model `all-mpnet-base-v2`. Category labels were embedded using the same encoder model.

Candidate items are retrieved dynamically using dense vector similarity search. As shown in Figure 2, precomputed recipe embeddings are indexed using FAISS. For a query q with embedding $\mathbf{e}_q = f_\theta(q)$, the system retrieves the top candidate set $\mathcal{R}_q \subset \mathcal{R}$ from recipe corpus $\mathcal{R}$ via cosine similarity.

Given category taxonomy $\mathcal{C} = \{c_1, \ldots, c_M\}$ with category embeddings $\mathbf{e}_{c_k} = f_\theta(c_k)$, candidate recipes are assigned to carousel rows $\mathcal{R}_{c_k}$ based on nearest category similarity:

$$\mathcal{R}_{c_k} = \left\{ r \in \mathcal{R}_q \mid \operatorname{argmax}_{c_m \in \mathcal{C}} \left(\frac{\mathbf{e}_r \cdot \mathbf{e}_{c_m}}{\|\mathbf{e}_r\| \|\mathbf{e}_{c_m}\|} \right) = c_k \right\} \tag{1}$$

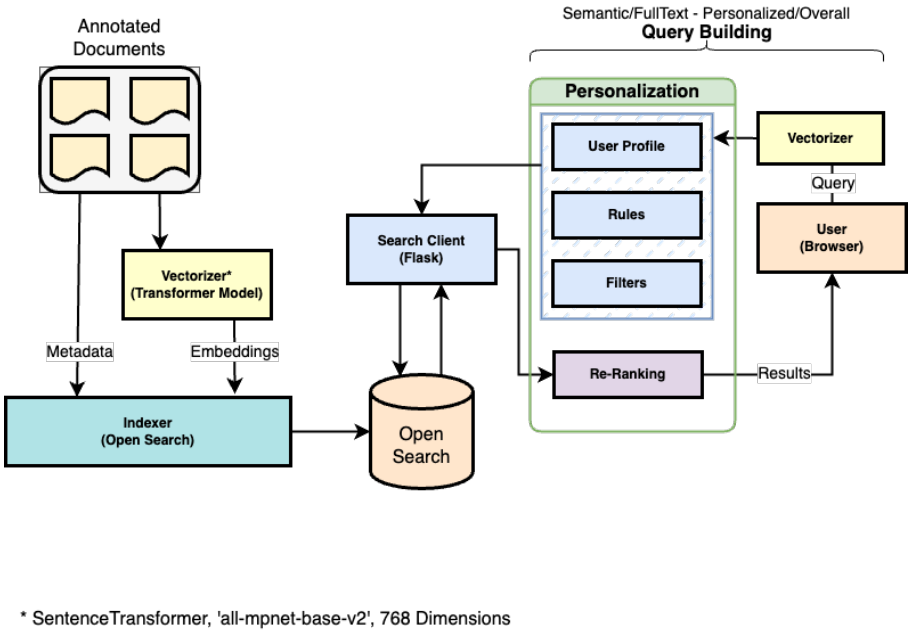

Figure 2: System architecture for dense retrieval, FAISS indexing, and carousel generation.

3.0.2. Interface Conditions

In the experimental platform, we implemented three interface variations to evaluate how they influence user during a goal-oriented search task.

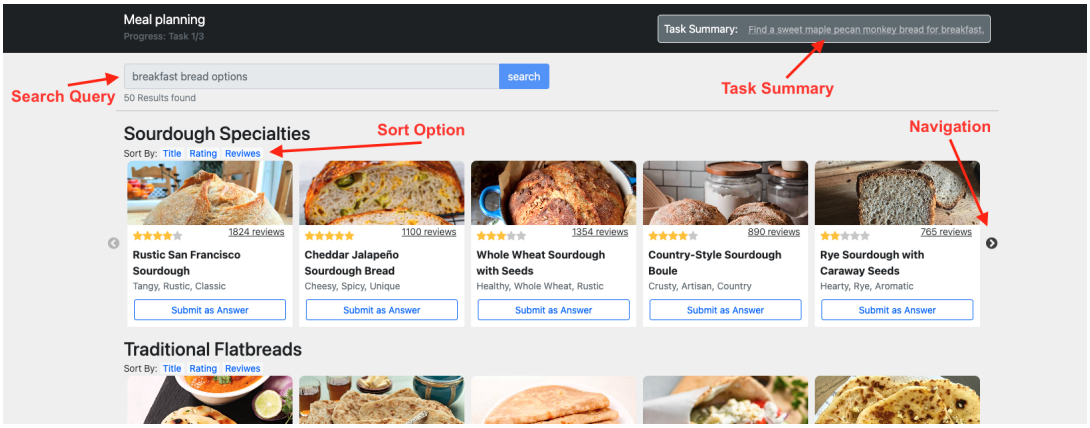

Figure 3: Experimental interface layout illustrating query input, task guidance box, horizontal recommendation carousels, and in-row sorting controls.

Figure 3 illustrates the main layout of the experimental interface. For each carousel row k, the horizontal sequence of items is defined by a ranking permutation $\pi_k : \mathcal{R}_{c_k} \to \{1, \ldots, |\mathcal{R}_{c_k}|\}$.

The baseline condition (*baseline*) presents a standard carousel interface where each horizontal row displays its category title. Items within each row appear in a fixed order determined by default algorithmic similarity $\pi_k(r) = \text{rank}_{r \in \mathcal{R}_{c_k}}(\text{sim}(\mathbf{e}_q, \mathbf{e}_r))$. The hidden title condition (*hidden*) conceals category titles by default. Users must click an explicit reveal button to display the title. This condition serves as an ablation to evaluate the navigational value of category titles while retaining the default item permutation $\pi_k(r)$. The in-row sort condition (*sortable*) combine interactive sorting controls within the header of each carousel row. Users can dynamically reorder items in an individual row by attribute $A \in \{\text{Rating}, \text{Review Count}, \text{Title}\}$, updating the row permutation to $\pi_k^{(A)}(r) = \text{rank}_{r \in \mathcal{R}_{c_k}}(A(r))$. The sorting interaction operates only on the selected carousel row. All conditions standardize horizontal navigation to directional arrow buttons, disabling drag-and-drop gestures to maintain uniformity.

3.1. User Study

We conducted an online user study to evaluate user navigation efficiency across the three interface conditions during structured retrieval tasks.

3.1.1. Design and Tasks

The study employed a mixed-design approach. Interface condition was evaluated as a between-subjects factor across three groups (*baseline*, *hidden*, *sortable*). Task difficulty and round sequence were evaluated across three consecutive retrieval rounds, while task domain was evaluated as a counterbalanced within-subjects factor.

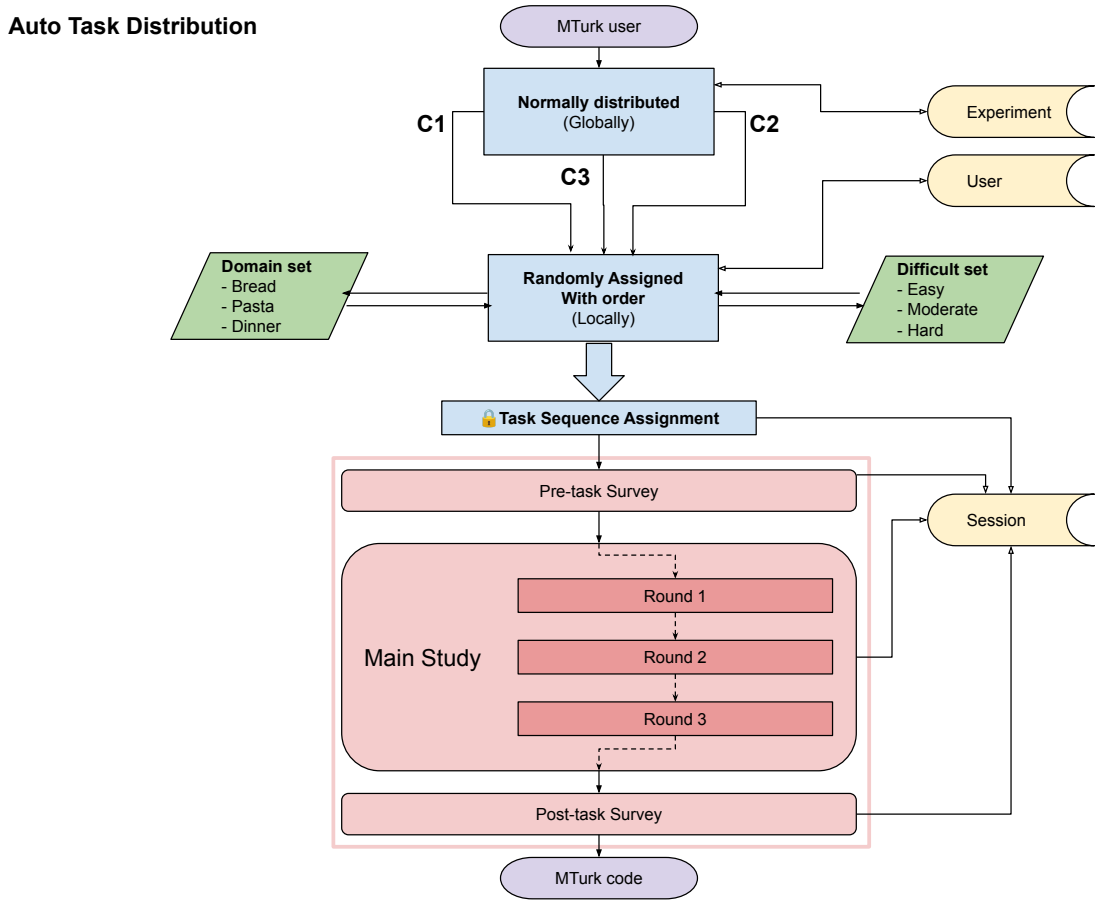

Figure 4: Experimental workflow diagram outlining onboarding, pre-task survey, randomized task rounds, and post-task evaluation.

Figure 4 shows the overall user study procedure. Each participant completed an interactive onboarding tutorial corresponding to their assigned interface condition. Participants then completed a pre-task questionnaire and three consecutive retrieval tasks.

Task requirements combined three complexity tiers with three domain contexts (bread, pasta, and dinner recipes). Task complexity was assigned in a fixed ascending progression across rounds: Round 1 presented an easy task (locating an item matching a single criterion), Round 2 presented a moderate task (satisfying two criteria, such as locating the most-reviewed recipe in a specific category), and Round 3 presented a hard task (satisfying three joint criteria, such as identifying the highest-rated five-star item in a target category). Task domains were counterbalanced and randomized across rounds to minimize domain order effects. Figure 5 displays an example task prompt.

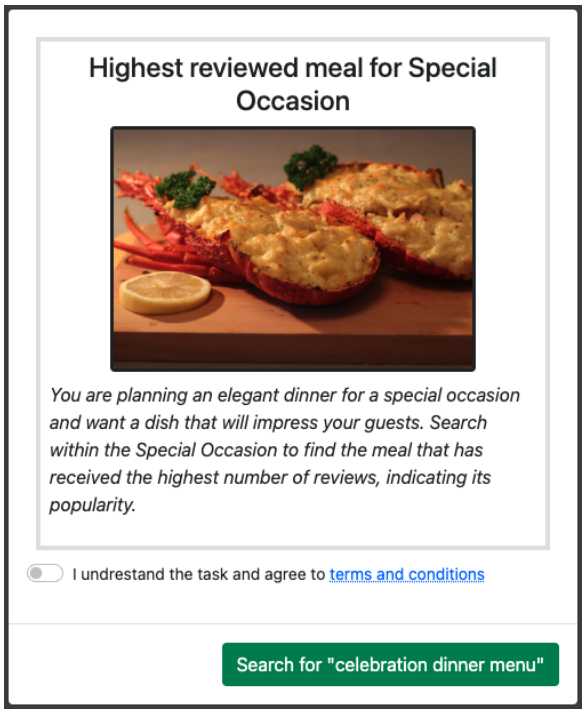

Figure 5: Example task prompt depicting a moderate complexity search task within the dinner domain.

3.1.2. Participants

During pre-study planning, sample size requirements were estimated using G*Power for a mixed-design repeated-measures ANOVA across three between-subjects interface groups and three repeated task rounds. Assuming a medium effect size ($f = 0.25$), $\alpha = 0.05$, and a statistical power of 0.90, the power calculation indicated a minimum requirement of 45 participants. For statistical inference, we subsequently employed linear mixed-effects models and generalized linear models to account for participant clustering and non-normal error distributions.

A total of 47 validated participants completed the study across the three between-subjects conditions: baseline ($n = 13$), hidden titles ($n = 17$), and in-row sort ($n = 17$). Participants were recruited through Amazon Mechanical Turk (MTurk). Eligibility was restricted to adult workers residing in the United States (age ≥ 18) with historical task approval rates above 95%. The study protocol received institutional review board (IRB) approval under an exempt determination for minimal-risk online user studies. All participants provided digital informed consent prior to the study and received financial compensation consistent with platform standards. Any incomplete submissions were excluded during quality screening.

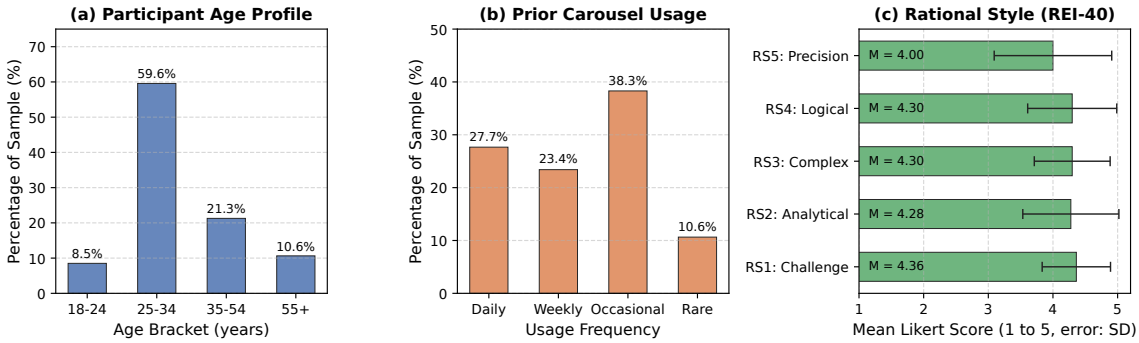

Figure 6: Participant profile across validated study participants ($N = 47$): (a) age bracket distribution, (b) prior interface and carousel usage frequency, and (c) baseline rational cognitive style scores (REI-40) with standard deviations.

Figure 6 summarizes participant demographics across completed task rounds. The pre-task survey assessed cognitive decision-making styles using the Rational-Experiential Inventory (REI-40) [33], measuring Rational Ability and Engagement items (RS1–RS5, such as enjoyment of intellectual challenge) and Experiential items (ES1–ES5, such as reliance on intuitive impressions). Post-task questionnaires gathered demographic information (DEM1–DEM5), digital interface confidence, and System Usability Scale (SUS1–SUS5) evaluations.

3.1.3. Data Collection and Filtering

The CARE platform logged client-side telemetry. Logged metrics included task duration, click counts, horizontal scroll events, and item selection mistakes. Selection mistakes were logged whenever an item submitted by a participant failed to meet the specified task criteria.

A total of 71 participants initially enrolled in the study. We excluded sessions that were abandoned before completing all three rounds. Task durations exceeding a 5-minute cap (300 seconds) were capped to minimize the influence of extreme idle outliers. The final validated dataset contains 141 complete task rounds from 47 unique participants ($n = 13$ baseline, $n = 17$ hidden titles, $n = 17$ in-row sort), with each participant contributing exactly three task rounds and full survey responses. In subsequent distributional boxplots (Figure 7), an additional visual filter (Duration ≤ 250 s, Clicks ≤ 100, leaving $N = 133$ rounds) is applied solely to improve plot readability by omitting extreme graphical outliers.

4. Results

This section presents the evaluation of goal-oriented search across the three carousel interface conditions. We report performance, interaction metrics, error rates, learning effects, and cognitive traits.

4.1. Performance Metrics

Task performance varied across the three interface conditions ($n = 13$ baseline, $n = 17$ hidden, $n = 17$ sortable). Table summary statistics evaluate duration, click volume, and scroll activity across all 141 task rounds.

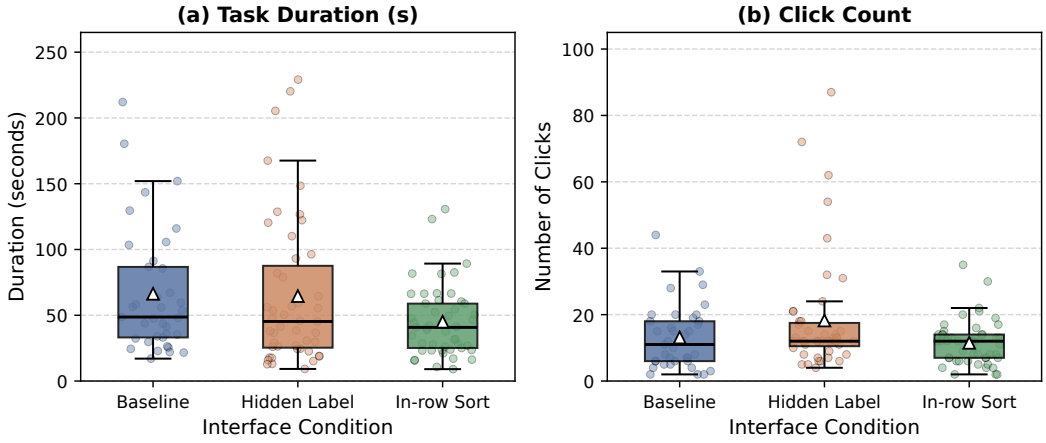

Figure 7: Distributional boxplots with overlaid stripplots for (a) Task Duration and (b) Click Count across interface conditions ($N = 133$, filtered to Duration ≤ 250 s and Clicks ≤ 100). Triangles indicate condition means. Horizontal lines within boxes indicate medians.

Participants completed search tasks fastest in the in-row sort condition. Mean task duration reached 54.15 seconds ($SD = 52.46$ s, $Mdn = 41.16$ s) in the in-row sort condition. Participants in the baseline condition averaged 78.31 seconds ($SD = 69.75$ s, $Mdn = 53.56$ s), while participants in the hidden condition averaged 79.94 seconds ($SD = 77.19$ s, $Mdn = 47.42$ s). In-row sorting controls reduced average completion time by 24.16 seconds (30.9%) relative to the baseline condition.

Interaction volume followed a similar pattern. Participants in the in-row sort condition registered the lowest average click count ($M = 12.47$, $SD = 8.79$, $Mdn = 12.00$). The baseline condition yielded higher descriptive click counts ($M = 18.21$, $SD = 29.68$, $Mdn = 12.00$), and the hidden condition produced the highest click volume ($M = 23.59$, $SD = 28.74$, $Mdn = 12.00$). In-row sorting also reduced click variance, reducing standard deviation to 8.79 compared to approximately 29 in the other two conditions. Scroll counts averaged 23.31 ($SD = 73.60$) for in-row sort, 34.51 ($SD = 66.56$) for baseline, and 33.35 ($SD = 56.29$) for hidden.

Figure 7 presents the distributions for task duration and click counts. The in-row sort condition exhibits inter-quartile ranges across both metrics, whereas the hidden condition displays extended upper tails in duration and clicks.

4.2. Error Rates and Navigation Effort

Selection mistakes were logged whenever a submitted recipe failed to satisfy the required task criteria. Baseline participants had the lowest mistake count ($M = 3.05$, $SD = 4.49$, $Mdn = 0.0$). Participants in the in-row sort condition had an intermediate mistake average ($M = 3.88$, $SD = 5.13$, $Mdn = 0.0$). Participants in the hidden condition accumulated the highest error count ($M = 5.55$, $SD = 5.24$, $Mdn = 4.0$). Because mistake counts exhibited substantial overdispersion (Variance $\gg$ Mean), we evaluated predictors of task selection mistakes using a Negative Binomial generalized linear model (GLM) with a log link. The hidden condition significantly increased selection mistakes relative to the baseline condition ($\beta = 0.691$, $SE = 0.246$, $z = 2.81$, $p = 0.005$, representing an estimated 99.5% increase in expected mistakes). The in-row sort condition also exhibited a statistically significant positive coefficient relative to baseline ($\beta = 0.510$, $SE = 0.248$, $z = 2.06$, $p = 0.039$, representing an estimated 66.6% increase in expected mistakes). Cognitive reflection provided a significant protective effect against task errors. Higher rational scores on RS1 were associated with substantial reductions in selection mistakes ($\beta = -0.889$, $SE = 0.187$, $z = -4.76$, $p < 0.001$). Task sequence rounds did not significantly alter error frequencies in the model (Round 2 : $p = 0.732$; Round 3 : $p = 0.748$).

Table 1
Linear Mixed-Effects Models (LMM) for task performance metrics with participant random intercepts ($N = 47$ participants, 141 task rounds). Reference categories: Baseline condition, Round 1. Standard errors are in parentheses. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Predictor	log(Duration)	Duration (s)	Click Count	Scroll Count
Fixed Effects				
Intercept	4.62*** (0.28)	105.54 (66.93)	6.26 (20.32)	48.20*** (12.60)
Interface Condition				
In-row Sort	-0.38** (0.13)	-23.83 (20.23)	-6.20 (6.11)	-8.60 (9.45)
Hidden Label	+0.04 (0.14)	+1.96 (20.23)	+4.92 (6.11)	+1.80 (9.62)
Task Sequence (Round)				
Round 2	-0.31*** (0.08)	-26.37** (9.90)	+2.23 (4.48)	-8.40** (3.10)
Round 3	-0.52*** (0.08)	-31.75** (9.90)	+1.26 (4.48)	-14.20*** (3.12)
Cognitive Styles				
Rational Item (RS1)	-0.12*** (0.03)	-1.86 (15.35)	+2.55 (4.64)	-1.95 (1.45)
Experiential Item (ES1)	+0.03 (0.03)	+2.40 (2.52)	+1.82** (0.65)	+1.40 (1.38)
Variance Components				
$\sigma^2_{\text{participant}}$	0.14	2189.29	112.90	145.20
$\sigma^2_{\text{residual}}$	0.22	2301.84	471.86	342.60
Marginal R^2 / Conditional R^2	0.29 / 0.54	0.26 / 0.41	0.31 / 0.48	0.14 / 0.38

4.3. Sequence and Learning Effects

Task progression across rounds introduced noticeable familiarity effects. Because task difficulty increased monotonically across rounds (Round 1 = Easy, Round 2 = Moderate, Round 3 = Hard), the round term in the model captures the combined effects of increasing task complexity and sequence familiarity.

Table 1 reports the linear mixed-effects model (LMM) parameter estimates with participant random intercepts, controlling for repeated observations per participant.

Task duration decreased across successive task rounds despite increasing task complexity (Round 2: $\beta = -26.37$, $SE = 9.90$, $p = 0.008$; Round 3: $\beta = -31.75$, $SE = 9.90$, $p = 0.001$), though the confounded experimental schedule prevents separating learning gains from difficulty progression. In the log-transformed duration model, the in-row sort condition maintained a statistically significant negative coefficient ($\beta = -0.38$, $SE = 0.13$, $p = 0.004$), representing an estimated 31.6% proportional reduction in task duration. In the raw duration model, the sortable coefficient was $\beta = -23.83$ ($SE = 20.23, p = 0.239$), reflecting high variance in raw duration times; the log-transformed model provides the appropriate specification for proportional completion time gains. For click count and scroll count, condition differences did not reach statistical significance in the mixed-effects models ($p > 0.05$).

4.4. Cognitive Style Interactions

Individual differences in problem-solving style interacted with interface conditions to influence task performance. We conducted exploratory comparisons across participant traits and survey items using Benjamini-Hochberg False Discovery Rate (FDR) adjustments.

Table 2
Exploratory comparative analysis across participant traits and survey items for task success with Benjamini-Hochberg False Discovery Rate (FDR) corrections ($N = 141$ task rounds). q-values represent adjusted significance thresholds ($q < 0.05$).

Variable / Trait	Survey Item	t-statistic	p-value	q-value (FDR)
Technology Usage Frequency	DEM3	-5.24	< 0.001	$< 0.001^*$
Enjoy Intellectual Challenge	RS1	-3.87	< 0.001	0.003^*
Education Level	DEM2	+3.04	0.004	0.030^*
System Usability: Easy to Use	SUS1	-2.57	0.014	0.072
System Usability: Well Integrated	SUS3	-2.50	0.016	0.072
Task Duration (s)	Duration	-2.34	0.021	0.082
Confidence in Digital Interfaces	DEM4	-2.31	0.025	0.082
Click Count	Clicks	-1.94	0.055	0.158
Analytical Problem Solving	RS4	-1.84	0.072	0.178
Prior Carousel Experience	DEM5	-1.79	0.077	0.178

Table 2 summarizes the FDR-corrected results for exploratory bivariate comparisons between zero-mistake rounds and error-prone rounds across all 141 task observations. Enjoyment of intellectual challenges (RS1) emerged as the primary cognitive style variable surviving correction ($q = 0.003$). Technology usage frequency (DEM3, $q < 0.001$) and education level (DEM2, $q = 0.030$) also remained statistically significant predictors. Note that because participant-level traits are evaluated across repeated rounds in this exploratory screening, these bivariate relationships should be interpreted alongside the participant-clustered models in Table 1.

Figure 8 illustrates the interaction between RS1 scores and interface condition on task duration. In the baseline and hidden conditions, participants with lower RS1 scores incurred higher task durations.

In contrast, the in-row sort condition visually exhibited more consistent task durations across all RS1 levels. While the interaction term in the mixed model did not achieve statistical significance (Sortable $\times$ RS1 : $p = 0.612$), these descriptive distributions suggest that in-row sorting controls may help offset individual differences in cognitive problem-solving styles during structured search tasks.

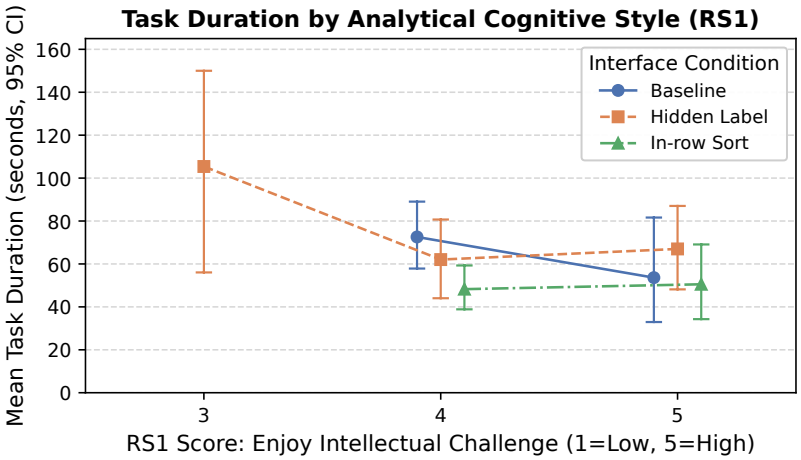

Figure 8: Mean task duration across discrete RS1 scores (Enjoy Intellectual Challenge) by interface condition with 95% confidence intervals. In-row sorting stabilizes navigation duration across differing cognitive profiles.

Subjective usability evaluations diverged from objective performance gains. System Usability Scale (SUS) composite ratings showed no statistically significant main effect across interface conditions ($F(2, 44) = 0.14, p = 0.871$). Median SUS scores were near 4.0 across all three experimental groups.

5. Discussion

This section discusses the empirical findings regarding carousel titles, user cognitive styles, and in-row sorting in multi-list recommender interfaces. We also examine the disconnect between behavioral efficiency and subjective usability, analyze algorithmic exposure trade-offs, and outline study limitations.

5.1. Carousel Titles and Navigation Friction

Concealing carousel titles created substantial navigation friction. According to Hick's Law [34], decision time increases when options lack clear hierarchical organization. Category titles provide critical visual signposts in two-dimensional layouts, allowing users to filter out irrelevant rows before inspecting individual items. When titles are concealed, users must examine individual recipe cards to infer row categories. This exploratory scanning significantly increased incorrect item selections ($p < 0.001$), corroborating prior observations on navigation effort in multi-list interfaces [6]. Designers of multi-list recommenders should include category titles to support efficient visual scanning.

5.2. Cognitive Style Interactions and User Control

In-row sorting controls helped narrow performance differences across problem-solving styles. Goal-oriented search in two-dimensional carousels requires evaluating candidate items against multiple constraints across several rows. In the baseline and hidden conditions, participants with lower rational style scores (RS1) exhibited higher task completion times. Providing dynamic in-row sorting allowed users to reorganize candidate items according to attributes like ratings or review counts. Visually, this interaction pattern appeared to moderate the completion time gap between heuristic and analytical searchers. Although formal interaction terms did not reach statistical significance in our sample ($N = 47$), giving users interactive controls to restructure recommendation outputs [9, 10] remains a promising avenue to support diverse problem-solving styles during structured search tasks.

However, empowering users with in-row sorting introduced a measurable speed-accuracy trade-off. While sorting accelerated task completion and narrowed duration gaps across cognitive styles, sortable participants exhibited a higher rate of selection mistakes than baseline participants ($\beta = 0.510, p = 0.039$, corresponding to an estimated 66.6% increase in expected mistakes). The ease of reordering

candidate rows appears to facilitate rapid, opportunistic decision-making, which can occasionally lead to submissions before secondary constraints are fully verified. Designers should consider lightweight confirmation cues or constraint highlighting when providing interactive sorting tools.

5.3. Disconnect Between Objective Metrics and Subjective Usability

The empirical findings reveal a disconnect between objective interaction efficiency and subjective usability evaluations. In-row sorting produced an average completion time reduction of 24.16 seconds (30.9%) and compressed click variance. However, post-task System Usability Scale (SUS) composite ratings remained indistinguishable across all three interface conditions. This divergence suggests that users do not readily perceive fine-grained efficiency improvements during short experimental tasks. Participants adapted their interaction strategies to overcome interface friction without registering explicit dissatisfaction. Usability evaluations of recommender interfaces should rely on objective telemetry alongside subjective questionnaires, as subjective ratings alone may fail to capture substantial differences in navigation effort.

5.4. Popularity Bias and Catalog Exposure Risks

Empowering users with sorting controls introduces systemic trade-offs for recommendation algorithms. Users in our study frequently sorted rows by average ratings or review volume to satisfy task constraints. This interaction pattern concentrates visual attention heavily on top-ranked head items. Recommender systems already suffer from popularity bias and feedback loops [5]. Giving users sorting controls can exacerbate exposure inequality by pushing niche items further out of view. As emphasized in recent interface-aware recommendation frameworks [14], interface designers must balance interactive user agency with catalog discovery goals. Future systems could combine user-driven sorting with diversity-preserving reranking algorithms to protect long-tail exposure.

5.5. Limitations and Future Work

Several limitations contextualize our findings. The study evaluated goal-oriented retrieval exclusively within the recipe domain. Tasks focused on structured constraint satisfaction rather than open-ended leisure browsing. Casual exploration in media streaming or e-commerce environments may prioritize content serendipity over retrieval speed.

Furthermore, task difficulty progressed in a fixed ascending sequence across rounds (Easy to Moderate to Hard). Although task domains were counterbalanced, this schedule completely confounds task difficulty with round sequence, preventing us from separating practice effects from task complexity in the repeated-measures models. We evaluated a single interaction pattern for concealing titles (click-to-reveal buttons). Alternative visual cues, such as hover-based tooltips, semi-transparent labels, or progressive disclosure mechanisms, might produce different interaction dynamics and error rates.

6. Conclusion

This study investigated how carousel titles and interactive in-row sorting influence goal-oriented search in carousel-based recommender interfaces. While recommender systems research predominantly treats the interface as a passive display for ranked lists, our empirical findings demonstrate that visual structure and interactive controls actively shape search strategies and user performance. Providing lightweight in-row sorting mechanisms reduced task duration by an estimated 31.6% in mixed-effects modeling, compressed interaction variance, and exhibited descriptive trends toward equalizing duration across different cognitive profiles. At the same time, this accelerated navigation introduced an empirical speed-accuracy trade-off, increasing selection mistakes relative to methodical baseline scanning (an estimated 66.6% increase in the Negative Binomial model). Conversely, concealing carousel titles nearly doubled selection errors (an estimated 99.5% increase), confirming that category labels serve as essential

visual signposts in two-dimensional recommendation grids. Finally, the disconnect between substantial objective efficiency gains and invariant subjective usability ratings highlights that interaction telemetry reveals operational friction that standard survey scales overlook.

7. Declaration on Generative AI

During the preparation of this work, ChatGPT was only used to identify and correct grammatical errors, typos, and other writing mistakes. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Y. Wang, D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, Q. Mei, Optimizing whole-page presentation for web search, ACM Transactions on the Web 12 (2018) 1–25.

[2] B. Rahdari, B. Kveton, P. Brusilovsky, The magic of carousels: Single vs. multi-list recommender systems, in: Proceedings of the 33rd ACM Conference on Hypertext and Social Media, HT '22, ACM, 2022, pp. 166–174.

[3] B. Loepp, Multi-list interfaces for recommender systems: Survey and future directions, Frontiers in Big Data 6 (2023) 1181287.

[4] B. Rahdari, P. Brusilovsky, CARE: An infrastructure for evaluation of carousel-based recommender interfaces, in: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24 Companion, ACM, 2024, pp. 110–113.

[5] O. Jeunen, B. Goethals, Top-K contextual bandits with equity of exposure, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, ACM, 2021, pp. 310–319.

[6] B. Rahdari, P. Brusilovsky, Under the hood of carousels: Investigating user engagement and navigation effort in multi-list recommender systems, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25, ACM, 2025, pp. 112–124.

[7] G. Marchionini, Information Seeking in Electronic Environments, Cambridge University Press, 1995.

[8] C. He, D. Parra, K. Verbert, Interactive recommender systems: A survey of the state of the art and future research directions, Human-Computer Interaction 31 (2016) 145–188.

[9] D. Parra, P. Brusilovsky, User-controllable personalization: A case study with setfusion, International Journal of Human-Computer Studies 78 (2015) 43–67.

[10] B. Rahdari, P. Brusilovsky, A. J. Sabet, Controlling personalized recommendations in two dimensions with a carousel-based interface, in: Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS '21), volume 2948 of *CEUR Workshop Proceedings*, 2021, pp. 112–122.

[11] R. Hu, P. Pu, Enhancing recommendation diversity with organization interfaces, in: Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI '11, ACM, 2011, pp. 347–350.

[12] D. Jannach, S. Naveed, M. Jugovac, User control in recommender systems: Overview and interaction challenges, in: E-Commerce and Web Technologies, volume 278 of *Lecture Notes in Business Information Processing*, Springer, 2016, pp. 3–17.

[13] P. Brusilovsky, Adaptive hypermedia, User Modeling and User-Adapted Interaction 11 (2001) 87–110.

[14] S. de Leon-Martinez, B. Rahdari, R. Móro, P. Brusilovsky, M. Bieliková, Interface-aware recommender systems, in: Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '26, ACM, 2026, pp. 668–671.

[15] B. Rahdari, Human-AI Collaboration in Search and Recommendation: Navigation, Personalization and Evaluation of Carousel-Based Interfaces, Ph.D. thesis, University of Pittsburgh, 2025.

[16] C. Trattner, Y.-L. Lin, D. Parra, Z. Yue, W. Real, P. Brusilovsky, Evaluating tag-based information access in image collections, in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, ACM, 2012, pp. 113–122.

[17] D. Dimitrov, D. Helic, C. Trattner, M. Strohmaier, Tag-based navigation and visualization, in: Social Information Access: Systems and Technologies, Springer, 2018, pp. 535–575.
[18] W. Bendada, G. Salha-Galvan, T. Bontempelli, Carousel personalization in music streaming apps with contextual bandits, in: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, ACM, 2020, pp. 420–425.
[19] M. Ferrari Dacrema, P. Cremonesi, Optimizing the selection of recommendation carousels with quantum computing, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, ACM, 2021, pp. 691 – 696.
[20] B. Ermiş, P. Ernst, Y. Stein, G. Zappella, Learning to rank in the position based model with bandit feedback, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, ACM, 2020, pp. 1195–1204.
[21] Y. Xi, J. Lin, W. Liu, X. Dai, W. Zhang, R. Zhang, R. Tang, Y. Yu, A bird's-eye view of reranking: From list level to page level, in: Proceedings of the 16th ACM International Conference on Web Search and Data Mining, WSDM '23, ACM, 2023, pp. 1075–1083.
[22] B. Rahdari, B. Kveton, P. Brusilovsky, From ranked lists to carousels: A carousel click model, arXiv preprint arXiv:2209.13426 (2022).
[23] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, B. Carterette, Offline evaluation to make decisions about playlist recommendation algorithms, in: Proceedings of the 15th ACM WSDM, WSDM '19, ACM, 2019, pp. 420–428.
[24] N. Felicioni, M. Ferrari Dacrema, P. Cremonesi, A methodology for the offline evaluation of recommender systems in a user interface with multiple carousels, in: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 10–15. URL: https://doi.org/10.1145/3450614.3461680. doi:10.1145/3450614.3461680.
[25] N. Felicioni, M. Ferrari Dacrema, P. Cremonesi, Measuring the ranking quality of recommendations in a two-dimensional carousel setting, in: Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS '21), volume 2948 of *CEUR Workshop Proceedings*, 2021, pp. 98–111.
[26] S. de Leon-Martinez, Understanding user behavior in carousel recommendation systems for click modeling and learning to rank, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM '24, ACM, 2024, pp. 1064–1065.
[27] B. Rahdari, P. Brusilovsky, B. Kveton, Towards simulation-based evaluation of recommender systems with carousel interfaces, ACM Transactions on Recommender Systems 2 (2024) 1–25.
[28] L. A. Granka, T. Joachims, G. Gay, Eye-tracking analysis of user behavior in WWW search, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, ACM, 2004, pp. 478–479.
[29] B. Rahdari, P. Brusilovsky, A. J. Sabet, Connecting students with research advisors through user-controlled recommendation, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, ACM, 2021, pp. 745–748.
[30] A. Starke, E. Asotic, C. Trattner, "serving each user": Supporting different eating goals through a multi-list recommender interface, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, ACM, 2021, pp. 124–132.
[31] A. D. Starke, E. Asotic, C. Trattner, E. J. Van Loo, Examining the user evaluation of multi-list recommender interfaces in the context of healthy recipe choices, ACM Transactions on Recommender Systems 1 (2023) 1–31.
[32] B. Rahdari, Food.com - enhanced recipes with images, 2026. URL: https://www.kaggle.com/ds/1436959. doi:10.34740/KAGGLE/DS/1436959.
[33] S. A. Keaton, Rational-experiential inventory-40 (REI-40), in: The Sourcebook of Listening Research: Methodology and Measures, Wiley, 2017, pp. 517–523.
[34] W. E. Hick, On the rate of gain of information, Quarterly Journal of Experimental Psychology 4 (1952) 11–26.