Revisiting N2DCG: An Empirically Grounded Reformulation of Carousel Recommendation Evaluation

Jingwei Kang*
University of Amsterdam
Amsterdam, The Netherlands
j.kang@uva.nl

Santiago de Leon-Martinez*†
Brno University of Technology
Brno, Czechia
santiago.deleon@kinit.sk

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Harrie Oosterhuis
University of Amsterdam
Amsterdam, The Netherlands
h.r.oosterhuis@uva.nl

Abstract

Carousel interfaces have been widely used in video and music streaming services, yet it remains unclear how to properly evaluate recommender systems in these two-dimensional layouts. N2DCG has been proposed to address this gap by adapting NDCG to carousel-based recommendation, but it relies on unverified assumptions borrowed from the single-list web-search setting that do not transfer well to two-dimensional carousel layouts.

We identify two substantial limitations of N2DCG: its ideal ranking, used for normalization, violates carousel constraints, and its discount function does not reflect user browsing behavior observed in empirical data. To address both limitations, we propose a reformulation of N2DCG that normalizes appropriately by respecting constraints and uses an empirically grounded discount function. We validate the proposed metric, showing that it better reflects users' empirical behavior on real-world eye-tracking data and better predicts the comparison results of carousel layouts simulated based on empirical examination patterns.

CCS Concepts

• **Information systems** → **Recommender systems**; **Search interfaces**; **Evaluation of retrieval results**.

Keywords

Carousel interface, Offline evaluation, Interface-aware RecSys

ACM Reference Format:
Jingwei Kang, Santiago de Leon-Martinez, Maarten de Rijke, and Harrie Oosterhuis. 2026. Revisiting N2DCG: An Empirically Grounded Reformulation of Carousel Recommendation Evaluation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3799682.3839894

*Both authors contributed equally to this research.
†Also with Kempelen Institute of Intelligent Technologies.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3839894

1 Introduction: New Insights in Carousels

Carousel interfaces have become ubiquitous in streaming media services like Netflix [7, 13]. They encourage user exploration by containing multiple lists or rows called *carousels*, each horizontally swipeable and organized around a single defining theme. Despite their widespread use, carousels have only recently gained attention in recommender systems research, with work on user click modeling [3, 12] and on evaluation [5, 6]. However, a recent eye-tracking study [2] and two follow-up analyses [4, 10] reveal a behavioral pattern that prior work does not account for. User examination in carousels does not follow a single global F-pattern [2, 4, 10]: the initial page shows a dual-focus F-pattern (with attention peaks at both the top-left and top-right corners), and after swiping this shifts to a mirrored F-pattern, with horizontal decay running right-to-left rather than left-to-right. This behavior is attributed to users maintaining their focus on the rightmost position after swiping. However, N2DCG [5, 6], the only evaluation metric for carousels, does not account for this behavior. Our work addresses this gap by re-examining and reformulating N2DCG.

N2DCG is a two-dimensional extension of Normalized Discounted Cumulative Gain (NDCG). NDCG evaluates a ranked list by combining item relevance with position discount. The (unnormalized) DCG of a list y with J items is defined as [1, 8, 9]:

$$DCG(y) = \sum_{j=1}^{J} g(y_j) \cdot d(j) = \sum_{j=1}^{J} \frac{g(y_j)}{\log_2(j+1)}, \tag{1}$$

where y_j denotes the item at rank j, $g(\cdot)$ represents the gain function mapping an item to its relevance score, and $d(j)$ is the position discount factor, commonly represented by a logarithm. To ensure comparability across different queries, DCG is subsequently normalized by the Ideal DCG (IDCG):

$$IDCG = \max_{y^* \in \mathcal{Y}} DCG(y^*), \quad NDCG(y) = \frac{DCG(y)}{IDCG}, \tag{2}$$

where $\mathcal{Y}$ denotes the set of all possible permutations of candidate items for the given query; IDCG represents the maximum possible DCG score for a ranking of the given items.

Standard NDCG assumes a 1D ranked list, whereas carousel interfaces present items in a 2D layout. To account for row and column positions, 2D Discounted Cumulative Gain (2DCG) extends DCG to an $I \times J$ grid [5, 6]. Let $y_{i,j}$ denote the item located at the

i-th row and j-th column in a 2D arrangement y:

$$2DCG(y) = \sum_{i=1}^{I} \sum_{j=1}^{J} g(y_{i,j}) \cdot d(i,j), \tag{3}$$

where I denotes the number of rows (e.g., carousels) and J the number of items in each row. The gain term $g(\cdot)$ remains the same as in DCG, whilst the 2D discount term $d(i,j)$ models the combined effect of visual attention and interaction effort. The design of $d(i,j)$ is the key modeling choice for carousels. Ferrari Dacrema et al. [6] propose two formulations:

$$d_t(i,j) = \frac{1}{\log_2(\alpha i + \beta j)}, \; d_a(i,j) = \frac{1}{\log_2(\alpha i + \beta j + \gamma n_h + \lambda n_v)}. \tag{4}$$

Here, α and β penalize lower row positions i and farther-right column positions j, respectively, reflecting F-pattern behavior. For d_a, γ and λ additionally penalize the number of horizontal swipes n_h and vertical scrolls n_v required to reveal an item.

Similar to NDCG, 2DCG is normalized by the Ideal 2DCG (I2DCG) to obtain a comparable metric bounded between 0 and 1:

$$I2DCG = \max_{y^* \in \mathcal{Y}_{2D}} 2DCG(y^*), \quad N2DCG(y) = \frac{2DCG(y)}{I2DCG}, \tag{5}$$

where I2DCG represents the maximum possible 2DCG score achieved by mapping the most relevant items to grid coordinates with the highest position discount factors. Here, $\mathcal{Y}_{2D}$ denotes the set of all possible 2D permutations of candidate items. Ferrari Dacrema et al. [6] express this mathematically as: the ideal arrangement y^* ensures $g(y^*_{i,j}) \geq g(y^*_{i',j'})$ for any pair of positions where $d(i,j) > d(i',j')$.

Building on this foundation, we address the limitations of N2DCG through the following contributions:

- We reformulate I2DCG to account for categorical constraints necessary for carousel interfaces and their evaluation.
- We identify and correct a failure of the original N2DCG discount factor to model mirrored F-pattern examination behavior.
- We design and evaluate four novel position-discount formulations; among these, the row-page discount aligns most closely with empirical examination frequencies and most reliably predicts user preferences between simulated carousel layouts.

2 Limitations of the Original N2DCG

2.1 Ignoring Categorical Constraints

The original N2DCG formulations ignore a key property of carousel interfaces: each carousel is, by design, organized around a single category or theme, and all items within it are drawn from that category. I2DCG, as defined above, treats all $I \times J$ positions as a single global pool and assigns items solely to maximize 2DCG, with no awareness of these categorical constraints. The resulting I2DCG values come from layouts that no valid carousel could realize, and dividing by them lowers the N2DCG scores of valid layouts.

Table 1 illustrates this flaw using an example 2×3 carousel. When relevant items belong to different categories, the unconstrained I2DCG simply places all of them into the top positions, producing an invalid layout that no carousel interface could display. In the example, the optimal valid layout reaches a 2DCG of 2.2, while the unconstrained optimal is 2.7. As a result, the valid layouts can only reach a maximum N2DCG of 0.81. This causes inconsistencies in evaluation as the actual upper bound of N2DCG is lower than 1.

Table 1: A 2×3 carousel interface with row discounts $(1.0, 0.9, 0.8)$ and $(0.3, 0.2, 0.1)$. Candidates span two categories, $A = \{a_1, a_2, a_3\}$ and $B = \{b_1, b_2, b_3\}$. Relevance is binary: a_1, a_2, b_1 are relevant, a_3, b_2, b_3 are non-relevant. The unconstrained ideal mixes categories within rows, inflating the N2DCG denominator.

Case	Layout	2DCG
Optimal valid layout	$\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix}$	2.2
Optimal unconstrained layout	$\begin{pmatrix} a_1 & a_2 & b_1 \\ a_3 & b_2 & b_3 \end{pmatrix}$	2.7
Maximum N2DCG among valid layouts (2.2/2.7)		0.81

2.2 Mismodeling Mirrored F-Pattern Behavior

A position discount function does not need to exactly match empirical examination frequencies. For evaluation, it is often sufficient for the discount to preserve the correct qualitative trend of user attention. However, even this weaker requirement is not met by the original N2DCG discount.

In real carousel interfaces, users often navigate within a carousel through horizontal swiping, which creates page boundaries within each carousel. Eye-tracking evidence suggests that after users swipe to a new page, their attention may reset toward the right side of the newly visible page, so that on later pages attention decreases from right to left as a mirrored F-pattern [2, 10], as illustrated in Figure 1(b). Because the N2DCG discount monotonically penalizes larger column positions, it instead predicts attention that keeps decreasing from left to right across the entire carousel (Figure 1(a)), the opposite of the observed trend on the second and third pages. Thus, the original position discount captures the wrong direction of attention change after horizontal swiping.

3 Proposal: Category-Aware N2DCG

We propose a category-aware variant of N2DCG in which the ideal layout is constrained to respect the category structure of the carousel interface. Let $\mathcal{Y}^{\text{cat}}_{2D} \subset \mathcal{Y}_{2D}$ denote the subset of 2D layouts in which all items in a row belong to the same category:

$$\mathcal{Y}^{\text{cat}}_{2D} = \left\{ y \in \mathcal{Y}_{2D} \;\middle|\; \forall\, i, j, j',\ \text{cat}(y_{i,j}) = \text{cat}(y_{i,j'}) \right\}, \tag{6}$$

where $\text{cat}(\cdot)$ denotes the category of an item. Since each category corresponds to exactly one carousel, $\mathcal{Y}^{\text{cat}}_{2D}$ consists precisely of the layouts obtained by (i) assigning category-specific carousels to display rows and (ii) ranking items within each carousel; items from different categories never share a row. The category-aware $\text{I2DCG}_{\text{cat}}$ is the maximum 2DCG over this restricted space:

$$\text{I2DCG}_{\text{cat}} = \max_{y^* \in \mathcal{Y}^{\text{cat}}_{2D}} 2\text{DCG}(y^*), \tag{7}$$

and the category-aware $\text{N2DCG}_{\text{cat}}$ is defined as $\frac{2\text{DCG}}{\text{I2DCG}_{\text{cat}}}$.

In our setting, computing $\text{I2DCG}_{\text{cat}}$ reduces to assigning category-specific carousels to display rows. Given a category-specific carousel and a display row, the optimal within-carousel ranking is obtained

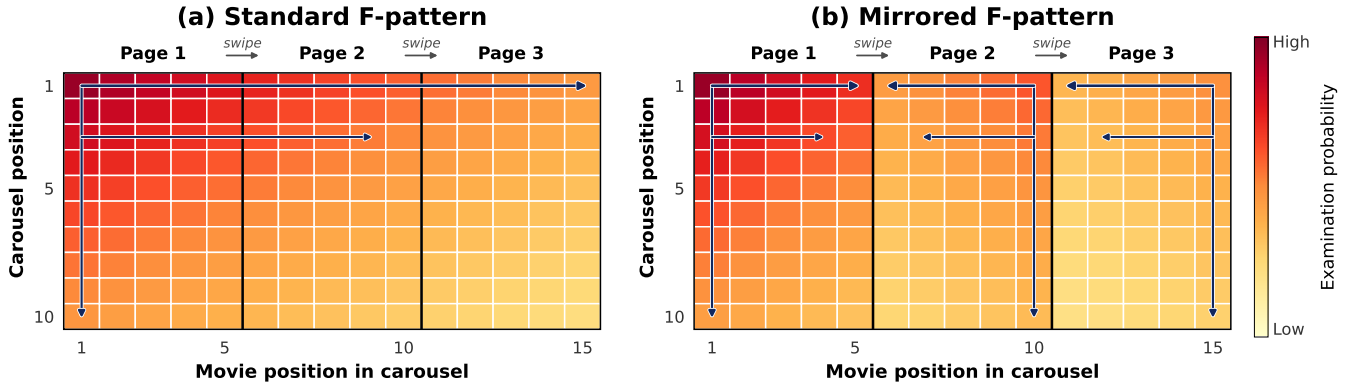


Figure 1: Schematic comparison of the assumed (a) and observed (b) attention patterns in a carousel interface. Darker cells indicate higher examination probability; arrows indicate the direction of attention decay.

by placing items in decreasing relevance order onto that row's decreasing-discount positions. This yields a score for each carousel–row pair. The remaining problem is to map category-specific carousels to display rows to maximize the total 2DCG, which can be solved using assignment algorithms such as the Hungarian algorithm [11].

For clarity, this computation is only used for normalization during evaluation. We are not proposing to use the Hungarian algorithm in production to create recommendations for actual users.

4 Proposal: Mirrored F-Pattern N2DCG

To capture both mirrored F-pattern attention and the interaction effort of accessing carousel items, we propose four formulations of the discount function $d(i, j)$ at row i and column j. We first modify the position discount to reflect attention shift after a swipe, and then explore three ways to incorporate user action penalties into the base discount.

4.1 Mirrored F-Pattern Positional Discount

The original position discount d_t in Eq. 4 models F-pattern position bias through a two-dimensional logarithmic decay. To account for attention shift after horizontal swipes, we modify this decay:

$$d_{\text{mir}}(i, j) = \frac{1}{\log_2(\alpha i + \beta \tilde{j})} \tag{8}$$

Let δ denote the number of items visible per page within a carousel. After a swipe, a new page of δ items is displayed, and the user's attention is naturally drawn first to the newly displayed items on the right side, and subsequently shifts toward the left. We capture this by replacing the original column index j with a re-indexed position $\tilde{j}$:

$$\tilde{j} = \begin{cases} j, & j \le \delta, \\ \delta \left\lfloor \dfrac{j-1}{\delta} \right\rfloor + (\delta - ((j-1) \bmod \delta)), & j > \delta. \end{cases} \tag{9}$$

On the first page ($j \le \delta$) the indexing is unchanged; on each subsequent page, the order within the page is mirrored, so the rightmost position carries the smallest $\tilde{j}$ and therefore the highest discount.

4.2 Incorporating User Action Penalties

Accessing items that are not currently visible requires additional swiping. Let n_h and n_v denote the number of horizontal and vertical swipes required to reveal the item at position (i, j). We compare three strategies for penalizing these actions.

Additive Swipe Penalties. Following d_a in Eq. 4, the swipe cost is added inside the logarithmic denominator with weight parameters γ and λ:

$$d_{\text{add}}(i, j) = \frac{1}{\log_2(\alpha i + \beta \tilde{j} + \gamma n_h + \lambda n_v)}. \tag{10}$$

Multiplicative Swipe Penalties. The swipe cost is instead applied multiplicatively. Let $\eta, \theta \in (0, 1)$ denote the penalty factors applied per horizontal and vertical swipe, respectively:

$$d_{\text{mul}}(i, j) = \frac{1}{\log_2(\alpha i + \beta \tilde{j})} \cdot \eta^{n_h} \cdot \theta^{n_v}. \tag{11}$$

Row-Page Discount. Instead of penalizing every swipe, this variant only penalizes items that require at least one horizontal swipe, combined with geometric decay across rows. Let $\mu \in (0, 1)$ be the page penalty and let $\nu \in (0, 1)$ be the row decay factor:

$$d_{\text{RPD}}(i, j) = \frac{1}{\log_2(\alpha i + \beta \tilde{j})} \cdot \mu^{\mathbf{1}[j > \delta]} \cdot \nu^{i-1}, \tag{12}$$

where $\mathbf{1}[j > \delta]$ equals 1 when the item is not on the first page and 0 otherwise. Intuitively, this formulation treats the *first* horizontal swipe as costly but does not penalize subsequent pages further, and also applies a multiplicative decay for each carousel row a user must enter to reach the item.

5 Experiments

This section details our experiments using the RecGaze eye-tracking dataset to estimate the parameters of our proposed discount functions and systematically validates the robustness of the reformulated N2DCG metric through simulated carousel layouts. Our code is available at https://github.com/jkang98/carousel-metric.

RecGaze Dataset. To date, RecGaze [2] is the *first and only* publicly available eye-tracking dataset designed for carousel recommendation. It provides a unique empirical basis for modeling user examination behavior in carousels. We focus on the *free-browsing*

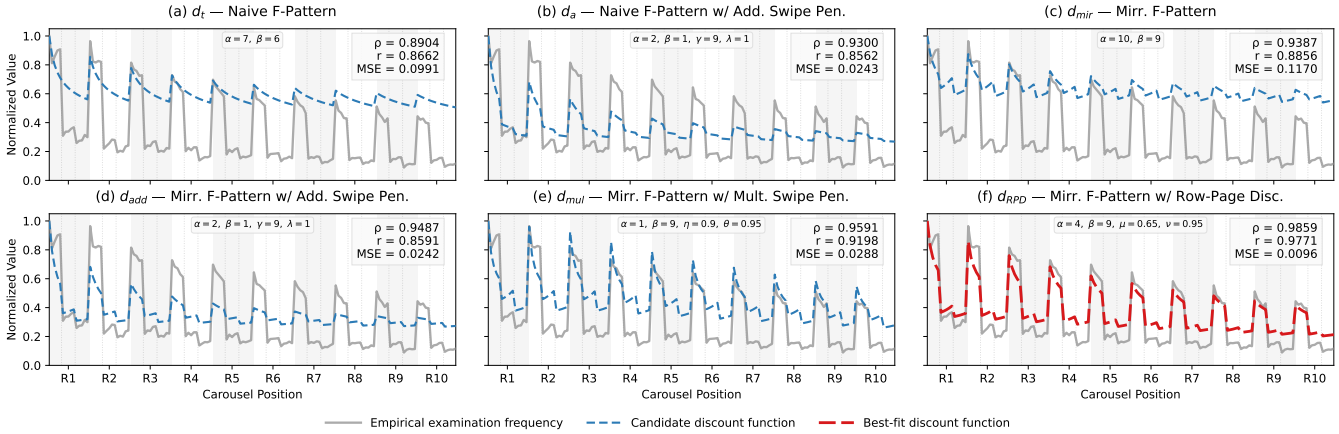


Figure 2: Fitted parameters and test-set performance (ρ, r, MSE) of each discount function are reported. Both the empirical examination frequencies and the computed discounts are max-normalized to $[0, 1]$ before plotting and MSE evaluation, so the comparison reflects how well each formulation captures the *relative* decay of user attention rather than its absolute magnitude.

task of RecGaze because it best reflects natural exploration behavior in carousel interfaces. In this task, each screen contains ten genre-based carousels, with fifteen movies in each carousel divided into three pages of five movies, of which the first page is initially presented and the last two can be seen by horizontally swiping once or twice, respectively. The first three carousels are initially visible with the remaining requiring vertical scrolling. This layout allows us to model both vertical carousel positions and horizontal within-carousel positions.

Experimental Setup. We use fixations as an observable proxy for item examination. Following the RecGaze preprocessing, an item is considered examined if at least one valid fixation falls within its area of interest. From the dataset, we estimate the examination probability at position (i, j) with the observed frequency of examination:

$$P(E_{i,j} = 1) \approx \hat{P}(E_{i,j} = 1) := \frac{N(E_{i,j} = 1)}{N_{\text{total}}}, \tag{13}$$

where $N(E_{i,j} = 1)$ denotes the number of screens in which the item displayed at position (i, j) was examined and N_{total} the total number of screens. These empirical frequencies serve two purposes: they are the targets for fitting the parameters of each candidate discount function, and they provide the examination-based ground truth position weights in our simulation study below.

To compare different design choices for the discount function, we perform a train/test split based on the location of the participants. The training set is used to estimate the parameters of the discount function, while the test set is used to evaluate whether the learned function generalizes to a different user group. In our experiment, we use data from 61 participants in Bratislava, Slovakia for training and 26 participants in Amsterdam, the Netherlands for testing; after removing screens with accidental clicks, this yields 1665 train and 710 test screens, each containing one clicked item.

Discount Function Parameter Estimation. We constrain the parameter search space based on the interpretation of each discount function. Across all variants, α and β control the weights of the vertical carousel position and the horizontal item position inside the logarithm. For additive action-based variants, γ and λ weight the numbers of horizontal swipes and vertical scrolls. Since these parameters weight integer-valued positions or actions, we restrict them to positive integers. In the multiplicative action-based variant, η, θ control horizontal/vertical action decay, and μ, ν control page/row decay; all are constrained to $(0, 1)$ as they directly scale the discount value. To determine the optimal parameters, we conduct a grid search, varying α, β, γ, and λ over $[1, 10]$ in steps of 1, and η, θ, μ, and ν over $(0, 1)$ in steps of 0.05.

For each discount variant, we fit the optimal parameters on the training set by maximizing the Spearman's rank correlation coefficient (ρ). The goal is to identify the parameter setting that accurately preserves the ordinal ranking of the empirical examination frequencies observed at each position, ensuring that positions with higher expected attention correctly receive larger discount.

Discount Fit to Examination Data. Using the learned parameters, we evaluate each discount function against the empirical examination frequencies observed in the held-out test set. Figure 2 shows the results for different discount functions. The gray curve represents the empirical examination frequency from eye-tracking data, and the dashed curves denote the fitted discount functions. Our best-performing formulation d_{RPD} is highlighted in red.

The two naive F-pattern-based functions, d_t and d_a, roughly capture the overall decreasing trend, but fail to reflect the repeated drops and rises caused by pagination; adding user-action information, d_a, improves the overall fit, but the pagination problem persists. The four mirrored F-pattern-based functions fit the empirical curve better, especially in capturing the repeated pagination pattern. Within this family, implementing user actions multiplicatively (d_{mul} and d_{RPD}) produces closer match. The best result is achieved by combining the mirrored F-pattern with the row-page discount d_{RPD}, which yields the highest Spearman's ρ and Pearson's r correlations and the lowest MSE after max normalization. Accordingly, the next experiment compares the original and reformulated 2DCG using d_a and d_{RPD}, respectively.

Metric Evaluation on Simulated Carousels. To systematically evaluate the reliability of our reformulated metric, we conduct a simulation experiment with 20,000 independent trials. Each trial samples a set of category-specific carousels: each carousel corresponds to a category and contains candidate items with fixed relevance labels. We then construct two complete 10×15 carousel layouts, denoted as A and B, by independently permuting the vertical order of the carousels and the horizontal order of items within each carousel. Therefore, the only difference between the two layouts is how carousels and items are arranged. This experimental design evaluates only the position discount functions: because both layouts contain the same candidate items, any ideal normalizer would be identical for A and B and cancel in the pairwise comparison. The category-aware $\text{I2DCG}_{\text{cat}}$ concerns the definition of the ideal layout rather than the modeling of user attention, and is therefore not the subject of this validation.

To obtain an empirical reference, we compute examination-based ground-truth scores, $N2DCG_A^{\text{exam}}$ and $N2DCG_B^{\text{exam}}$, as the category-aware $\text{N2DCG}_{\text{cat}}$ with its position discounts replaced by the empirical examination frequencies $\hat{P}(E_{i,j} = 1)$ (Eq. 13), estimated on the held-out test set. The resulting score reflects the expected relevance that users actually examine. We then measure the difference between the two layouts as $\Delta = \left|N2DCG_A^{\text{exam}} - N2DCG_B^{\text{exam}}\right|$. A near-zero Δ indicates that the two layouts are difficult to distinguish, while a larger Δ means that one layout is clearly better than the other in terms of actual user browsing behavior.

Finally, we score both layouts using the unnormalized 2DCG with the original discount d_a and with our reformulated discount d_{RPD}. A metric is considered correct in a trial if its pairwise preference matches the preference given by the empirical ground truth. By comparing the accuracy under different Δ thresholds, we can evaluate the robustness of each metric. Table 2 reports accuracy under both binary and graded relevance (1–5 scale) settings. Different thresholds are used for the ground-truth difference Δ. Here, ***Corr.*** denotes the percentage of trials in which the original 2DCG makes an incorrect pairwise decision, while the reformulated 2DCG makes the correct one. ***Both Err.*** denotes the percentage of trials in which both metrics make incorrect decisions. The results show that the reformulated 2DCG consistently achieves higher accuracy than the original metric. When the ground-truth difference is very obvious ($\Delta \geq 0.10$), the original 2DCG still misjudges 1.3% (binary) / 1.8% (graded) of trials.

Lessons Learned. Experiments show the row-page discount d_{RPD} best captures user attention and preferences in carousel interfaces. Thus, our reformulated 2DCG with our best found parameters is:

$$2DCG_{\text{ref}}(y) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{g(y_{i,j})}{\log_2(4i + 9\tilde{j})} \cdot 0.65^{\mathbf{1}[j>\delta]} \cdot 0.95^{i-1} \qquad (14)$$

We reformulate $\text{I2DCG}_{\text{cat}}$ and $\text{N2DCG}_{\text{cat}}$ correspondingly.

6 Conclusion

We identified two limitations of N2DCG—the neglect of categorical constraints and the misalignment with actual browsing behavior in carousel interfaces—and proposed a reformulation addressing both. Validation using empirical examination frequencies and simulated pairwise comparisons demonstrates the effectiveness of our reformulation. This suggests that our reformulated N2DCG is a better choice to evaluate carousel recommendation. Future work should extend this metric-based evaluation into behavior-aware simulator-based evaluation, enabling carousel policies to be evaluated under explicit models of user browsing and choice behavior.

Table 2: Accuracy comparison between the original and reformulated 2DCG under different ground-truth differences.

Setting	Diff. $\Delta \geq$	Orig. Acc.	Our Acc.	Corr.	Both Err.
Binary	0.00	82.9%	**93.2%**	11.4%	5.7%
	0.01	85.7%	**96.4%**	11.0%	3.2%
	0.02	88.4%	**98.3%**	10.0%	1.5%
	0.05	94.8%	**99.9%**	5.2%	0.1%
	0.10	98.7%	**100.0%**	1.3%	0.0%
Graded	0.00	84.9%	**93.9%**	10.4%	4.8%
	0.01	87.4%	**96.7%**	9.9%	2.7%
	0.02	89.4%	**98.2%**	9.1%	1.5%
	0.05	94.3%	**99.8%**	5.5%	0.2%
	0.10	98.2%	**100.0%**	1.8%	0.0%

Acknowledgments

This work is supported by the Dutch Research Council (NWO) under grants VI.Veni.222.269, 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006; by the European Union under grant agreement No. 101201510 (UNITE); and by the Eyes4ICU project, funded by the European Union under the Horizon Europe Marie Skłodowska-Curie Actions, grant agreement No. 101072410. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

GenAI Usage Disclosure

We used generative AI tools to assist with language polishing and with code cleaning and optimization during manuscript preparation. The authors carefully reviewed all generated content and take full responsibility for the final manuscript.

References

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 89–96. doi:10.1145/1102351.1102363

[2] Santiago de Leon-Martinez, Jingwei Kang, Robert Moro, Maarten de Rijke, Branislav Kveton, Harrie Oosterhuis, and Maria Bielikova. 2025. RecGaze: The First Eye Tracking and User Interaction Dataset for Carousel Interfaces. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3702–3711. doi:10.1145/3726302.3730301

[3] Santiago de Leon-Martinez, Robert Moro, Branislav Kveton, and Maria Bielikova. 2026. From Latent to Observable Position-Based Click Models in Carousel Interfaces. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Republic of Korea) *(KDD '26)*. Association for Computing Machinery, New York, NY, USA, 767–778. doi:10.1145/3770855.3818211

[4] Santiago de Leon-Martinez, Robert Moro, Branislav Kveton, and Maria Bielikova. 2026. Riding the Carousel: The First Extensive Eye Tracking Analysis of Browsing Behavior in Carousel Recommenders. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)*. Association for Computing Machinery, New York, NY, USA, 2120–2130. doi:10.1145/3742413.3789166

[5] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. 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* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 10–15. doi:10.1145/3450614.3461680

[6] Maurizio Ferrari Dacrema, Nicolò Felicioni, and Paolo Cremonesi. 2022. Offline Evaluation of Recommender Systems in a User Interface With Multiple Carousels. *Frontiers in Big Data* 5 (2022). doi:10.3389/fdata.2022.910030

[7] Carlos A. Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (Dec. 2016), 19 pages. doi:10.1145/2843948

[8] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 41–48. doi:10.1145/345508.345545

[9] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. doi:10.1145/582415.582418

[10] Jingwei Kang, Maarten de Rijke, and Harrie Oosterhuis. 2026. Following the Eye-Tracking Evidence: Established Web-Search Assumptions Fail in Carousel Interfaces. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Australia) *(SIGIR '26)*. Association for Computing Machinery, New York, NY, USA, 734–744. doi:10.1145/3805712.3809657

[11] Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955), 83–97. doi:10.1002/nav.3800020109

[12] Behnam Rahdari, Peter Brusilovsky, and Branislav Kveton. 2024. Towards Simulation-Based Evaluation of Recommender Systems with Carousel Interfaces. *ACM Trans. Recomm. Syst.* 2, 1, Article 9 (March 2024), 25 pages. doi:10.1145/3643709

[13] Lequn Wang, Jiangwei Pan, and Linas Baltrunas. 2026. GenPage: Towards End-to-End Generative Homepage Construction at Netflix. doi:10.48550/arXiv.2606.31031 arXiv:2606.31031 [cs.IR]