The Roles of Knowledge and Memory in Generating Top-10 Lists

Michael D. Lee (mdlee@uci.edu)
Emily Liu (ecliu1@uci.edu)
Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA 92617 USA

Abstract

We consider the role that memory and knowledge play in the accuracy of people's generation of top-10 lists. We report data from an experiment in which people answered questions like "list the top 10 most watched TV shows in the US", with and without the help of a memory aid that provided the true top 50 items. Our analyses examine the changes in accuracy resulting from the availability of the memory aid, the patterns with which people modify their lists when the aid is provided, and the stability of individual differences in the memory and decision-making processes involved. We find clear evidence that, for those involving large number of potentially relevant items, memory retrieval plays a central role in determining the accuracy of the list. We discuss implications of these findings for the development of models for aggregating rank orders produced by people when not given the relevant items.

Keywords: top-10 lists; serial recall; memory for order; aggregating rankings

Introduction

Top-10 lists like "the top 10 most watched TV shows in the US" are ubiquitous in popular culture, and people often try to generate them on the fly. People's ability to produce these lists depends on both their knowledge and their ability to recall all of the items that might be included in the list. Inaccuracies in the list produced could be due either to a lack of knowledge about the underlying ordering, or failures to access and retrieve the relevant items despite having accurate knowledge. For example, failing to include "Sunday Night Football" in the most watched TV shows might stem from not knowing it has a large audience, or might stem from focusing on drama and other TV show genres, and forgetting to think about sport shows when generating the list. A basic research question for cognitive science is to what degree the errors are caused by deficiencies of knowledge, failures of memory, or some combination of the two.

Items in top-10 lists can be ranked according to many different criteria, such as size, value, quantity, and time, but research on human memory for order information has focused on temporal order. A standard paradigm is serial recall, where lists of (typically) random word or letter sequences are presented with the instruction to recall the items in the correct order. Standard findings include primacy and recency effects, where memory accuracy declines as a function of position in the list except for the last few items, and locality, where an item that is placed in an incorrect position is nevertheless placed nearby the original position (Estes, 1997; Farrell, 2013; Nairne, 1992). In serial reconstruction tasks (Kelley, Neath, & Surprenant, 2013) real-world knowledge of temporal ordering has been tested, such as memory for events related to September 11 (Altmann, 2003), autobiographical events (Burt, Kemp, & Conway, 2008), and the chronological order of the US presidents (Healy, Havas, & Parker, 2000; Roediger & Crowder, 1976). In addition, items can be ordered along dimensions other than time (Neath & Saint-Aubin, 2011; Lee, Steyvers, & Miller, 2014).

One important dimension of variation among serial recall tasks is the nature of the response .In some serial reconstruction tasks (e.g., Neath & Saint-Aubin, 2011), the subject has to recall the items themselves, in addition to placing the items in the correct order. Other serial reconstruction tasks (e.g., Roediger & Crowder, 1976) present the set of the to be ordered items as a memory aid to the subjec,t whose only task is to sort the items into the correct order. Clearly, the former task is more challenging because errors can arise in the process of retrieving the items as well as the ordering of the retrieved items.

This paper investigates the performance of creating top-10 lists under conditions that involve both the presence and absence of a memory aid (i.e., giving the set of to-be-ordered items). By contrasting the performance under these conditions, we assess the degree to which item access and retrieval errors contribute to performance. We also investigate the effect of the total number of relevant items on ordering performance. For some top-10 lists, there might be a few hundred potentially relevant items to choose from whereas other top-10 lists might involve thousands of relevant items. It seems reasonable to expect that, under large set size conditions, having the memory aid will benefit ordering performance because it is more likely that some items will fail be retrieved from memory and considered for inclusion in the top-10 list.

Experiment

Participants A total of 20 participants completed the control condition, and 20 different participants completed the experimental condition. All participants were recruited via Amazon Mechanical Turk, restricted to US IP addresses, and paid US$1.

Stimuli We prepared lists on 10 different topics, dealing with the popularity of TV shows and movies, the commercial success of people and companies, the populations of cities and countries, and the success of sporting teams. For each of these topics, we collected the top 50 items. Table 1 shows the top 10 items in each of the 10 topics.

Procedure The experiment was administered via Amazon Mechanical Turk, and programmed as a Qualtrix survey.

Table 1: The 10 list topics, and the top 10 items in each list.

US TV show audience	US brand value	US athlete income	US movie gross	NCAA basketball wins
Sunday Night Football	Apple	Floyd Mayweather	Hunger Games	Kentucky
Big Bang Theory	Microsoft	Lebron James	Iron Man 3	Kansas
The Voice	Coca-Cola	Drew Brees	Frozen	North Carolina
Modern Family	IBM	Kobe Bryant	Despicable Me 2	Duke
American Idol	Google	Tiger Woods	Man of Steel	Syracuse
The Following	McDonald's	Phil Mickelson	Monsters' University	Temple
Two-and-a-half Men	GE	Derrick Rose	Gravity	St. John's
Grey's Anatomy	Intel	Peyton Manning	The Hobbit	UCLA
NCIS	Samsung	Alex Rodriguez	Fast and Furious	Notre Dame
Football Night in America	Louis Vitton	Zach Greinke	Oz the great and powerful	Indiana
Country population	**US food chain sales**	**US city population**	**EU city population**	**Auto brand sales**
China	McDonald's	New York	London	Toyota
India	Subway	Los Angeles	Berlin	Volkswagen
United States	Starbucks	Chicago	Madrid	Ford
Indonesia	Burger King	Houston	Rome	Chevrolet
Brazil	Wendy's	Philadelphia	Paris	Nissan
Pakistan	Taco Bell	Phoenix	Bucharest	Hyundai
Nigeria	Dunkin' Donuts	San Antonio	Vienna	Honda
Bangladesh	Pizza Hut	San Diego	Hamburg	Kia
Russia	KFC	Dallas	Budapest	Renault
Japan	Applebee's	San Jose	Warsaw	Fiat

There were two between-subject experimental conditions.

In the control condition, participants were shown a memory aid consisting of the true top 50 items for the topic, ordered alphabetically, and asked to generate a top-10 list. Each topic question was explained in detail, along with instructions that emphasized there was no time limit, but not reference materials could be used. We call this the "control" list.

In the experimental condition, participants were first asked to generate a top-10 list without the aid of the true top 50 items. We call this the "before" list. After completing their list, participants were then shown the memory aid of the top 50 items. They were then asked to generate a revised top 10 list, and were able to see both their original list and the memory aid while completing this second attempt. We call this the "after" list.

Participants' answers for the control list, and the before and after lists in the experimental condition, were all completed in a set of 10 free-form text boxes in the experimental interface. Every response was post-processed to map to a standardized name for each unique item in each topic, so that, for example "LA", "LosAngles", "los angeleses", and 'los angelas'" all mapped to "Los Angeles".

Analysis

Topic Differences

Even though the topics in the experiment each had a memory aid listing the true top 50 set of items, the topics differ in the total number of items that are relevant and may be considered during recall. To assess the topic diversity or richness, we used a type-to-token ratio (TTR) analysis, which is used in lexical analysis to estimate the vocabulary knowledge of learners (Malvern & Richards, 2012). We calculated the TTR on the basis of the responses, across participants, in the experimental condition, by dividing the total number of types produced for each topic (i.e. unique items) by the number of tokens (i.e. total number of items). The TV topic led to the highest TTR (0.415) while the Auto topic led to the lowest TTR (0.140), reflecting the intuition that there are many more types of TV shows than car brands. In all of our analyses, we present the results for topics ordered from highest to lowest TTR.

Group Comparisons

Our analyses rely on measuring the difference between various lists, such as between a participant's top-10 list and the true top 10, or between the lists a participants produced before or after being provided with the memory aid. We use the *partial tau* measure, which counts the number of pairwise swaps required to change one list into another, is general enough to deal with the possibility that the sets of items in each list are not identical, and has a well-studied theoretical basis as a metric (Fagin, Kumar, Mahdian, Sivakumar, & Vee, 2006). Intuitively, partial tau is a difference that starts at zero when two lists are the same, and increases with every swap—of two items in a list, or removing an item, or adding

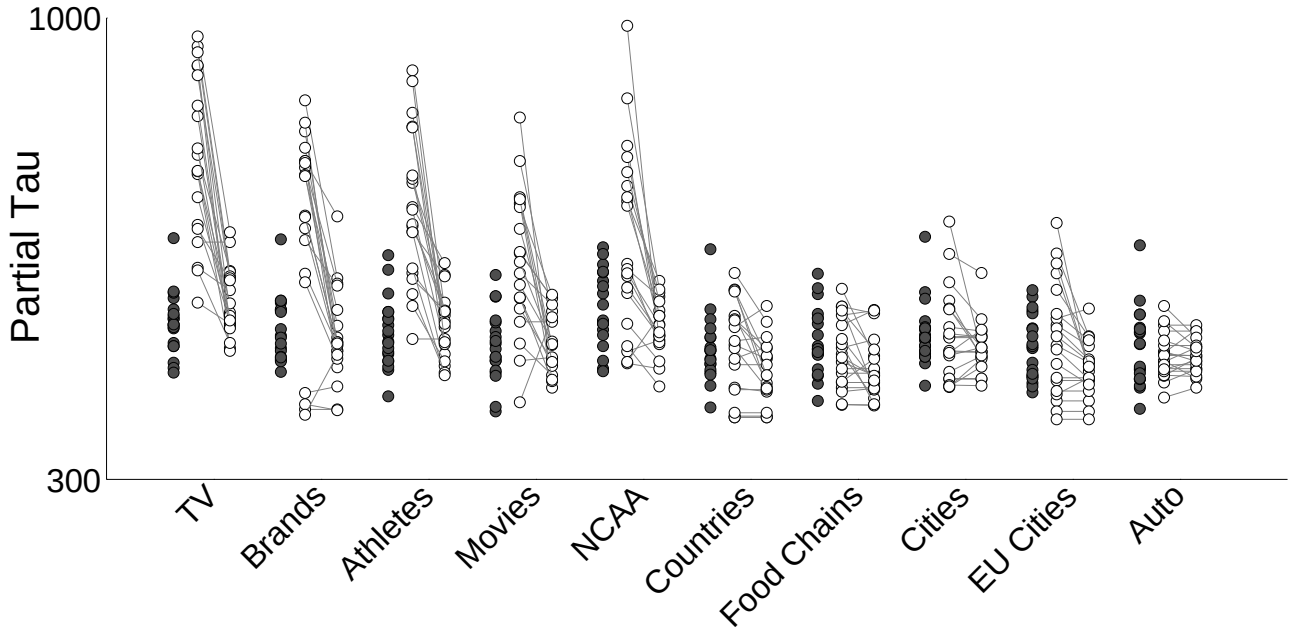

Figure 1: Partial tau measures of list accuracy for every participant in every experimental condition. Topic questions are shown from left to right, ordered by decreasing type-to-token ratios. Performance in the control condition is shown by the black circles, with each circle representing one participant. Performance in the experimental condition is shown by white dots, connected for the same participant, with performance before and after the memory aid shown from left to right.

a new item—that is needed to convert one list to the other.

Figure 1 shows the partial tau measure for every list completed in both of the experimental conditions.[1] The topics are ordered from left to right in terms of decreasing type-to-token ratio. For each topic, the partial tau for the lists in the control condition are shown by black circles. The partial taus for the same participant in the experimental condition are shown as white circles, joined by a line, moving from their before list to their after list from left to right.

To measure the size of change in accuracy between sets of lists, we measured the effect size of the change in partial tau (Cohen, 1992). To test for the sameness or difference of accuracy, we estimated Bayes factors (Kass & Raftery, 1995) using the methods developed by Rouder, Speckman, Sun, Morey, and Iverson (2009).[2] Bayes factors have the advantage of being expressed as easily interpreted likelihood ratios, and of being able to express evidence in favor of the null hypothesis of sameness. Table 2 lists the effect sizes and Bayes factors for every topic question comparing the three possible pairings of list-generating experimental conditions. The three list comparisons presented graphically in Figure 1 and statistically in Table 2 allow inferences about the roles of memory and knowledge in the accuracy of the generated lists.

Before vs After One way to assess the role of the memory aid is to compare the accuracy of the before and after lists in the experimental condition. The natural interpretation is that the list produced after being shown the memory aid is inaccurate because of deficiencies in knowledge, whereas the list produced before the aid is inaccurate because of both gaps in knowledge and failures in memory. This implies the change in partial tau from the before to after lists is a measure of the benefits of not having to rely on memory. The effect sizes and Bayes factors in Table 2 show that, for the last five topics, with lower TTR values, there is little evidence of a large improvement. This suggests that participants were able to retrieve the cities, countries, cars, and fast food items that needed to be assessed. For the first five topics, however, with relatively high TTR values, there is a large improvement in the accuracy of lists generated with the memory aid. This suggests the memory aid was useful, providing lists of movies, shows, brands, people, and teams about which participants were knowledgeable, but which were not easily retrieved from memory.

Control vs Before A second way to assess the role of memory is to compare the accuracy of the control lists to the before lists. The control lists benefit from the memory aid, whereas the before lists do not, and both are the first attempt a partici-

[1] Except for two outliers removed from the experimental condition. These corresponded to a single participant in the movies topic, and a single participant in the countries topic, who provided meaningless responses in their original lists.

[2] We used the `http://pcl.missouri.edu/bayesfactor` two-sample test applet to compare control lists to experimental lists, and the one-sample applet to compare the before and after lists. The Bayes factors we report use the JZS Cauchy prior option with the default effect size scale provided.

Table 2: Effect sizes and Bayes factors comparing the partial tau in the before vs after, before vs control, and after vs control lists for each topic question. A Bayes factor in favor of the alternative hypothesis of different distributions is listed as B_{10}. A Bayes factor in favor of the null hypothesis of the same distribution is listed as B_{01}.

	After vs Before		Before vs Control		After vs Control	
Topic	Effect Size	Bayes Factor	Effect Size	Bayes Factor	Effect Size	Bayes Factor
US TV show audience	$\delta_{b-a} = 2.3$	$B_{10} > 100$	$\delta_{b-c} = 2.7$	$B_{10} > 100$	$\delta_{a-c} = -0.6$	$B_{10} = 1.1$
US brand value	$\delta_{b-a} = 1.3$	$B_{10} > 100$	$\delta_{b-c} = 1.3$	$B_{10} = 91$	$\delta_{a-c} = 0.2$	$B_{01} = 2.8$
US athlete income	$\delta_{b-a} = 2.0$	$B_{10} > 100$	$\delta_{b-c} = 2.1$	$B_{10} > 100$	$\delta_{a-c} = -0.2$	$B_{01} = 2.4$
US movie gross	$\delta_{b-a} = 1.5$	$B_{10} = 48$	$\delta_{b-c} = 1.5$	$B_{10} > 100$	$\delta_{a-c} = -0.2$	$B_{01} = 3.0$
NCAA basketball wins	$\delta_{b-a} = 1.3$	$B_{10} > 100$	$\delta_{b-c} = 1.0$	$B_{10} = 9.5$	$\delta_{a-c} = 0.5$	$B_{01} = 1.4$
Country population	$\delta_{b-a} = 0.6$	$B_{10} = 1.5$	$\delta_{b-c} = 0.0$	$B_{01} = 3.2$	$\delta_{a-c} = 0.7$	$B_{10} = 2.1$
US food chain sales	$\delta_{b-a} = 0.5$	$B_{10} = 2.4$	$\delta_{b-c} = -0.3$	$B_{01} = 2.0$	$\delta_{a-c} = 0.9$	$B_{10} = 8.0$
US city population	$\delta_{b-a} = 0.3$	$B_{01} = 1.7$	$\delta_{b-c} = 0.0$	$B_{01} = 3.2$	$\delta_{a-c} = 0.5$	$B_{01} = 1.1$
EU city population	$\delta_{b-a} = 0.7$	$B_{10} = 9.3$	$\delta_{b-c} = 0.2$	$B_{01} = 2.7$	$\delta_{a-c} = 0.7$	$B_{10} = 3.3$
Auto brand sales	$\delta_{b-a} = 0.0$	$B_{01} = 4.2$	$\delta_{b-c} = -0.2$	$B_{01} = 2.7$	$\delta_{a-c} = 0.2$	$B_{01} = 2.6$

pant makes. The effect sizes and Bayes factors again show a division between the first five and second five topics. There is strong evidence that the control lists are different from, and more accurate than, the before lists, for these topics (although the Bayes factor of 9.5 for the NCAA topic is probably better regarded as "evidence" than "strong evidence"). The effect sizes for the second five topics are very small, and the Bayes factors favor the hypothesis of the same group accuracy. Once again, these results suggest the memory aid was helpful for the first five topics, but not for the second five. In this way, the comparison of control vs before lists supports the findings of the before vs after list comparison, from a complementary perspective that is between- rather than within-participants, and controlling for the number of attempts a participant has made to generate the list.

After vs Control The third comparison relates the control list to the after list. In both cases, the memory aid is available, and so this comparison considers the possible benefit of making multiple attempts to generate a list. Table 2 shows that the effect sizes are generally small, and the Bayes factors generally provide no strong evidence for either sameness or difference. Only for the food chain topic is there some evidence the lists generated in the after condition are more accurate than those generated in the control condition. A reasonable overall conclusion is that, once the memory aid is presented, participants produce similarly accurate lists, regardless of whether they made an initial attempt without the memory aid.

Individual Changes

Figure 2 presents an analysis of individual differences in performance for participants in the experimental condition, made possible by the within-subject design. The left-hand panel relates to the first five topic questions with larger TTR values, and the right-hand panel relates to the five question with smaller TTR values. The white circles show the average initial partial tau of each participant before the memory aid, and the average change in their partial tau for the list generated after the memory aid, for those topics. The same measures of performance on the five individual topics making up the average are connected to the circle by lines. In this way, Figure 2 shows the relationship between initial performance and the improvement in performance across participants, for the two qualitatively different types of list topics.

It is clear there is a strong positive correlation between initial performance and improvement, with participants who generally were inaccurate with lists before the memory aid improving the most. This relationship is especially strong for the first five topics, where the memory aid is most helpful, but is evident in both topic sets. It is also clear that every participant improved on average, since improvement is always a positive difference in the partial tau measure. There are only a few individual topic and participant combinations in which lists became less accurate after the presentation of the memory aid, as is also clear from Figure 1. It is the case, however, that for the second five topics, a number of participants do not improve on average when the memory aid is presented. This suggests that they have already remembered the relevant items, and their inaccuracies are caused by imperfect knowledge. Finally, Figure 2 permits some suggestive conclusions about the stability of individual differences across the topics. There is some evidence that the individual questions are near the averages—showing larger inter- than intra-individual differences—suggestion that there are consistent individual differences.

Item Accuracy and Change

Figure 3 presents an analysis of how items are changed in the experimental condition, as participants generate a revised list after they have been provided with the memory aid. The item positions 1–10 in both the before and after lists are shown by nodes. The overall pattern of change in items is shown in the main panel on the left, with the thickness of each line indicat-

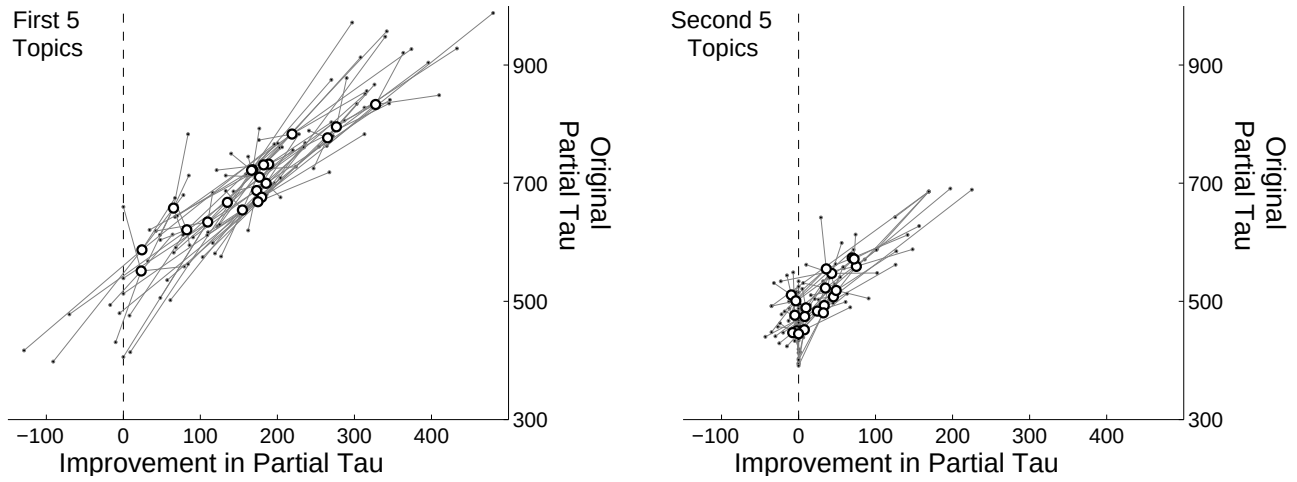

Figure 2: Individual differences in initial performance and improvement for participants in the experimental condition, considered separately for the first five topic questions (left-hand panel) and the second five topic questions (right-hand panel). The large white circles show the average performance of each participant initially (in the before condition), and their improvement after the memory aid (the change in their performance from the before to after condition). Performance on the individual topic questions that comprise these averages are shown by connecting lines.

ing to the number of times an item in a position in the before list was kept in the same position, moved to a different position in the list, or removed from the list, as denoted by the "-" node. Also shown are the frequency with which items added to the after list, coming from the "+" node.[3] One clear pattern from is that items are rarely changed to other positions in the list. They are either kept at their original position, or removed from the list entirely. The main panel in Figure 3 shows very few swaps to other positions. As might be expected, items that were originally in lower positions in the list are more often removed, and items that are introduced to the list after the memory aid are more often placed in lower positions.

The sub-panels in Figure 3 show that there are clear differences in how items are changed in the first five topics where the memory aid plays a major role, when compared to the second five, where the memory aid is less important. The differences are intuitive, with many more items added and removed in the first five topic lists, and both these insertions and deletions sometimes taking place high on the lists. The lists for the second five topics, in contrast, involve fewer changes, and especially fewer additions and removals of items.

Discussion

Our motivating question was how memory for items and knowledge of their properties interact in producing top-10 lists. The experimental design we used involves conditions that do not require retrieval of the items to be ordered, because a memory aid is presented listing the relevant items. It also involves a condition in which no memory aid is presented, and people must generate a set of items and order them. Constrasting performance across these sorts of conditions allows the relative contribution of memory retrieval and memory for the properties of items in determining the overall accuracy of top-10 lists to be assessed. We found that failures to generate or retrieve relevant sets of items does play a significant role for some topic domains, characterized by having high type-to-token ratios. These are essentially topic domains where people include many different items in their lists.

An interesting challenge for models of serial reconstructions, such as SIMPLE (Brown, Neath, & Chater, 2007; Kelley et al., 2013; Lee & Pooley, 2013), is to account for the differences in accuracy across domains with different type-to-token ratios. Whlle SIMPLE naturally accommodates memory for order with respect to different criteria, additional theorizing and model development is needed to model the memory retrieval processes that generate the items to be ordered.

More broadly on the modeling front, our empirical results have implications for aggregation models that combine rankings across subjects. In previous research (Lee et al., 2014), we demonstrated that a cognitive probabilistic model performed well in producing aggregate rankings that were often close to the ground truth and performed better than most of the individuals. The ranking data used for the cognitive model consisted of a full ranking of all the relevant items. An important goal for future work is to extend the cognitive model to aggregate top-10 lists. While there are standard algorithmic approaches to aggregating top-n lists (e.g., Marden, 1995), these methods typically ignore human memory retrieval errors or deficiencies in producing the list. Therefore, if a subject did not place a particular item in the top-10, the assumption is that item would have to have a rank higher than 10, which might unnecessarily shift the item in the group

[3] In the main panel, only change transitions that occurred at least 5 times are shown. This threshold is lowered to 3 for the sub-panels.

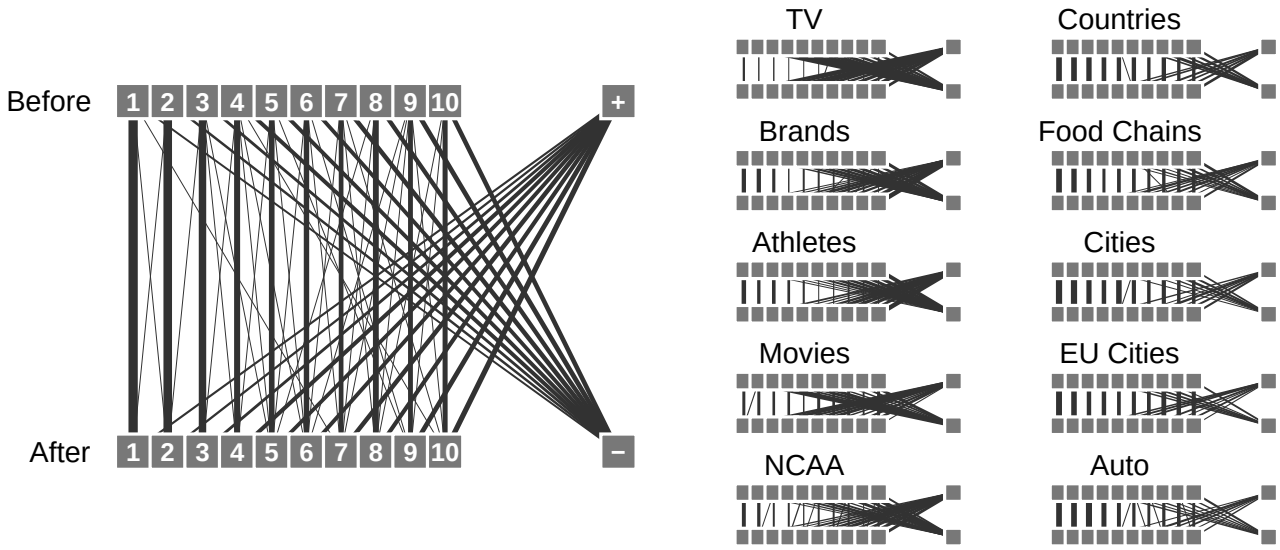

Figure 3: The pattern of changes in items in the list generated before and after the memory aid in the experimental condition, including removed and added items. The overall pattern is shown in the main panel on the left, while the sub-panels on the right show the patterns for each individual list.

average. In a cognitive modeling approach, however, the possibility of memory errors can be considered especially when other subjects consistently place that item in their top 10. We expect that our empirical results–which make clear the important role of memory in the sorts of lists people produce—will place important constraints on the design of such aggregation models.

References

Altmann, E. M. (2003). Reconstructing the serial order of events: A case study of September 11, 2001. *Applied Cognitive Psychology*, *17*, 1067–1080.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(1), 539–576.

Burt, C. D. B., Kemp, S., & Conway, M. (2008). Ordering the components of autobiographical events. *Acta Psychologica*, *127*, 36–45.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, *104*, 148–169.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2006). Comparing partial rankings. *SIAM Journal of Discrete Mathematics*, *3*, 628–648.

Farrell, S. (2013). Serial order memory, computational perspectives. In H. Pashler (Ed.), *Encyclopedia of the mind.* Thousand Oaks, CA: Sage.

Healy, A. F., Havas, D. A., & Parker, J. T. (2000). Comparing serial position effects in semantic and episodic memory using reconstruction of order tasks. *Journal of Memory and Language*, *42*, 147–167.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Kelley, M. R., Neath, I., & Surprenant, A. M. (2013). Three more semantic serial position functions and a SIMPLE explanation. *Memory & Cognition*, *41*, 600–610.

Lee, M. D., & Pooley, J. P. (2013). Correcting the SIMPLE model of free recall. *Psychological Review*, *120*, 293–296.

Lee, M. D., Steyvers, M., & Miller, B. J. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE.*, *9*, 1–9.

Malvern, D. D., & Richards, B. (2012). Measures of lexical richness. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Blackwell/Wiley.

Marden, J. I. (1995). *Analyzing and modeling rank data*. Chapman & Hall.

Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, *3*, 199–202.

Neath, I., & Saint-Aubin, J. (2011). Further evidence that similar principles govern recall from episodic and semantic memory: The Canadian prime ministerial serial position function. *Canadian Journal of Experimental Psychology*, *65*, 77–83.

Roediger, H. L., III, & Crowder, R. G. (1976). A serial position effect in recall of United States presidents. *Bulletin of the Psychonomic Society*, *8*, 275–278.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.