In a top-n task, people produce a list of items that they believe are ordered relative to a criterion, and can include any number of items in their list. We develop Thurstonian cognitive models of the individual differences and decision-making processes involved in producing top-n lists, and apply it to the problem of inferring an aggregated list from individual responses. We present 3 applications—involving predicting movie popularity, predicting the outcome of the 2014 World Cup tournament, and rating the worst U.S. presidents—using real-world data from the crowd-sourced opinion website ranker.com. The movie popularity application demonstrates the ability of the model to make relatively accurate predictions, partly through its ability to infer individual expertise. The World Cup application demonstrates the ability of the model to make accurate predictions, partly through its ability to incorporate relevant prior information, and partly through its ability to model multiple relevant behavioral data jointly. The U.S. presidents application demonstrates the ability of the model to allow for multiple latent rankings, so that a subgroup of contaminant opinions can be separated from those that reflect established historical opinion. On the basis of these applications, we argue that the psychological goals of developing models of how people produce top-n lists, and the applied goals of making accurate predictions, are usefully tackled simultaneously.

Keywords: Thurstonian models, rank data, wisdom of the crowd, Bayesian methods
very informative. The incomplete nature of top-$n$ lists, arising because people can choose how many items to include, is potentially both helpful and harmful. Shorter lists provide less explicit information, but allowing people the flexibility to terminate a list once they feel they have exhausted their knowledge may make the information they do provide more reliable and accurate.

There is a large statistical literature on methods for aggregating rankings, including top-$n$ rankings. Good overviews are provided by Marden (1995) and Lin (2010). We take the perspective that the challenge of combining the top-$n$ lists people produce is usefully tackled as a cognitive modeling problem. The ranked lists that provide the input are fundamentally behavioral data, providing an expression of cognitive knowledge produced by various memory and decision-making processes. Accordingly, in this paper, we extend the basic Thurstonian cognitive model previously developed for complete rankings (Lee et al., 2014) to the more general case of top-$n$ lists.

To demonstrate and evaluate this model, we rely on real-world data, in contrast to the laboratory controlled data that is the empirical basis of our earlier work. Our real-world data come from the crowd-sourced opinion website ranker.com, which allows people to provide top-$n$ lists for a wide variety of social, political, sporting, and other topics. Ranker receives approximately 20 million unique visitors each month, who collectively provide opinion data on tens of thousands of lists, each of which is an answer to a very specific question (e.g., “The Best Beer” or “The Best Qualities in a Person”), some of which can be thought of as predictions (e.g., “The Scariest Threats to the United States” or “The Worst NFL Teams in 2014”). The most common action on the Ranker site is to vote on whether a particular item belongs on a particular list, which is done millions of times each month. In addition, thousands of visitors create individual top-$n$ lists, either by generating lists from scratch, or by modifying existing lists.

This paper is structured around the applications of an aggregation model to the individual top-$n$ lists for three Ranker topics. The first application involves lists predicting the popularity of movies, and uses the basic top-$n$ aggregation model. The second application involves predicting the outcome of the 2014 football (soccer) World Cup. This model incorporates both ranking and voting data, and is an example of the joint modeling of multiple sources of behavioral data (Lee, 2011). It also incorporates relevant prior information for the predictions, in the form of the bracket structure of the World Cup tournament. The third application involves subjective opinion rankings of the worst U.S. presidents, and involves an extended latent mixture version of the model that allows for multiple underlying opinions.

### Movie Popularity

#### Data

The data for our first application come from the Ranker list “The Most Anticipated 2013 Films,” which asks for people’s predictions about which films will be viewed the most in U.S. theaters. The data were downloaded on 31 March, 2014, and involve a total of 28 people providing top-$n$ lists of a total of 59 movies. Because movies are released throughout a year, it was inevitable that new movies were being included in user lists during the year. About 80% of all the movies in all user lists were present in the first 4 months, however, and there was no evidence of users producing lists based on existing box-office information. Thus, we treat the user lists as predictions about movie popularity. As a benchmark, we obtained box-office attendance figures from the standard industry website http://www.imdb.com/.

#### Thurstonian Model

The key concepts in our Thurstonian model for aggregating top-$n$ lists are summarized in Figure 1. It presents an example involving a domain with five items, and the rankings pro-

---


2. Determining an appropriate benchmark to assess the aggregate list is inherently subjective, because there are multiple reasonable interpretations of what “most anticipated” means. The http://www.imdb.com/ benchmark is based on the idea that people interpret the question in terms of popularity or likelihood of viewing. An alternative interpretation would be in terms of movie quality or critical appeal. To test this alternative, we also did an analysis using the critic evaluations on the widely-used rating site http://www.rottentomatoes.com/ as a benchmark. The performance of all individuals, and all models, with respect to this benchmark was much worse. We concluded from this that the popularity interpretation is a better one, and so present only those results.
A. Latent Dimension

\[ \mu_1, \mu_2, \mu_3, \mu_4, \mu_5 \]

B. Individual 1

Observed Ranking
\[ y_1 = (1, 2, 3, 5) \]

C. Individual 2

Observed Ranking
\[ y_2 = (1, 3, 2) \]

Figure 1. Thurstonian model for top-n ranking data aggregation. Panel A shows the latent ground truth locations for five items, denoted \( \mu_1, \ldots, \mu_5 \). Panels B and C show, for two individuals, how the latent representations generate the mental samples that produce observed top-n ranking data. The mental sample \( x_{ij} \) for the \( j \)th individual on the \( i \)th item is drawn from the Gaussian distribution with mean \( \mu_i \) for the item and standard deviation \( \sigma_j \) for the individual. Only samples for the items that are included in the top-n list are drawn, and their ordering produces the reported ranking \( y_j \) for the \( j \)th individual, as listed in Panels B and C (based on Lee et al., 2014, Figure 1). See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
This model is a canonical example of the Thurstonian approach, also widely used in statistical approaches to aggregation (e.g., Böckenholt, 1992; Lin, 2010; Marden, 1995). We view the justification of the model as psychological as stemming from the special case we consider, which leads to the model parameters having natural psychological interpretations, and involving psychologically interpretable data generating processes. In particular, the model we use involves two sorts of psychologically meaningful parameters that can be inferred from top-n ranking data. The item location parameters—\(\mu_i\) for the \(i\)th item—correspond to the “true” latent locations of the items, and can be interpreted as an aggregated group representation of the order of the items. The standard deviations of the Gaussian distributions from which mental samples are drawn, given by \(\sigma_i\) for the \(j\)th individual, correspond to the “expertise” of the individual. Smaller standard deviations result in rankings that more closely follow the true latent ordering, consistent with greater expertise. These interpretations allow our applications of the canonical model to be cast in terms of stimulus and person properties, and also facilitate the psychological extensions of the model to other sorts of representations and behavioral processes we consider in the second and third case studies.

Figure 2 shows a graphical model that implements the canonical Thurstonian top-n aggregation model shown conceptually in Figure 1. Graphical models are a standard formalism for representing probabilistic models, and are especially well suited to the application of computational methods for Bayesian inference (Koller, Friedman, Getoor, & Taskar, 2007; Pearl, 1998). Lee and Wagenmakers (2013) provide an introduction to graphical modeling aimed at cognitive scientists. In a graphical model, nodes in a graph represent model parameters and data, and the graph structure indicates how the model assumes the parameters generate behavioral data. The parameters in Figure 2 are the item locations \(\mu_i\) and individual expertises \(\sigma_i\), shown by unshaded circular nodes because they are unobserved and continuous. Both are given vague priors. Together, these parameters generate the mental sample \(x_{ij}\) that the \(j\)th person draws for the \(i\)th item. These mental samples, in turn, generate the observed ranking data, \(y_{ij}\) for the rank given to the \(i\)th item by the \(j\)th person. These data are shown by a shaded square node, since they are observed and discrete. The graphical model is completed by encompassing plates used to show repetitions in the graph structure over the items and individuals.

We implemented the graphical model in JAGS (Plummer, 2003), which applies standard computational methods to provide samples from the joint posterior distribution of the model conditional on the data. The key to this implementation is the use of censoring to relate the latent mental sample variables \(x_{ij}\) to the observed rankings \(y_{ij}\). The two individuals in Figure 1 provide concrete examples of how this is done. Since individual 1 provides the data \(y_1 = (1,2,3,5)\), the mental samples must satisfy the two inequalities \(x_{11} < x_{21} < x_{31} < x_{51}\) and \(x_{41} > x_{51}\). That is, even though item 4 was not explicitly ranked, the assumptions of the model mean its mental sample must have been to the right of the mental sample for the last ranked item. Similarly, because individual 2 provides the data \(y_2 = (1,3,2)\), the mental samples must satisfy the three inequalities \(x_{12} < x_{32} < x_{22}, x_{42} > x_{22}\), and \(x_{52} > x_{22}\).

For all of the applications presented in this paper, our results are based on collecting 20,000 samples from three chains after 20,000 burn-in with 10 thinning. We checked for convergence of the chains using the standard \(\hat{R}\) statistic for the posterior predictive distribution of the data (Brooks & Gelman, 1997). When necessary, because of the very weak prior on the \(\mu_i\) location parameters, we post-processed the chains so that each \(\mu_i\) sample (i.e., the vector containing the \(\mu_i\) for all items for a sample) was zero-centered to ensure translation invariance.

**Results**

**Aggregate ranking.** A natural way of determining an aggregate ranking from the model is to order items by their posterior means for the \(\mu_i\) location parameters. Figure 3 does this for the movies, listing them from most to least popular from top to bottom, based on the means of the marginal posterior distributions shown in the

---

3 Note that we parameterize the Gaussian distribution in terms of its mean and precision.
The right-hand panel of Figure 3 summarizes the behavioral data, showing how often each movie was placed in each rank on a list by the area of circles.

The marginal posteriors shown in Figure 3 provide not only an inference about the ordering of the movies, but an expression of the uncertainty in their relative positions. It is clear, for example, that the location of movies lower on the list is much more uncertain, consistent with the fact they are included in many fewer individuals’ rankings.

**Model and individual performance.** To assess the accuracy of the rankings provided by individual people, and by the model, we use the partial tau measure. This is a generalization of the standard Kendall’s tau measure that compares complete rank orders by counting the number of pairwise swaps required to change one list into another (Marden, 1995). The partial tau measure allows for the possibility that some items are tied in a partial ranking. This generalization can be applied to top-\( n \) lists by listing the \( n \) ranked items as usual, and then considering the remaining items in the domain that are not ranked as tied in the \((n + 1)\)th position. Computationally, for every pair of items, the partial tau measure accrues a penalty of 1 if that pair is present in a different order in the two partial rankings, and a penalty of \( 0 \leq p \leq 1 \) if they are ordered in one ranking but tied in the other. Intuitively, partial tau is a difference or error measure that starts at zero when two lists are the same, and increases as the two lists differ in their rankings, or one list fails to rank items that the other list has ranked. The partial tau measure has a sound and well-studied theoretical basis as a metric (Fagin, Kumar, Mahdian, Sivakumar, & Vee, 2006), and we use the default choice \( p = 1/2 \).

Figure 4 shows the partial tau measure of performance for each individual, measured against the benchmark box-office ranking, as a distribution of stick figures. The chance distribution for full rankings of all the movies is shown by the broken curve. The Thurstonian model partial tau of 329 is shown by a circular marker. It is clear that this is better performance than that achieved by the best individual, who had a partial tau of 383.5, demonstrating a so-called “strong wisdom of the crowd” effect (Surowiecki, 2004). Figure 4 also shows it is slightly better than the partial tau of 339 for the aggregated ranking found by a standard statistical Borda count method, which sums the ranks of every item over all lists, giving each list equal weight, and orders the items according to their totals (Marden, 1995).

**Expertise predictions.** Besides inferring an aggregate ranking of the movies, the Thurstonian model makes inferences about the expertise of each individual. In particular, the model provides an inference about the accuracy of individuals. The scatter-plot shown in the upper-right of
Figure 4 tests these inferences by showing the relationship over all individuals between the posterior expectation of the expertise parameter, and actual performance, as measured by partial tau, for each person. There is a strong positive correlation of $r = .72$, showing that people inferred by the model to have higher expertise did generally provide more accurate rankings. This correlation between inferred expertise and observed performance mirrors those found for complete rankings by Lee et al. (2014). It is worth emphasizing that this result is genuinely the confirmation of a prediction. The order of the movies inferred by the model is based entirely on the behavioral ranking data of people, and the model never has access to the ground truth benchmark box-office order against which its prediction is measured.

Predicting the World Cup

Data

The data for our second application come from the ranker.com list “Who Will Win the
The data were downloaded on 19 October 2014, and involve a total of 31 people providing top-n lists of a total of 32 teams in the World Cup, and an unknown number of users providing totals of 5,667 up votes and 11,202 down votes distributed across the 32 teams. All of the lists and votes were completed before the start of the World Cup tournament. One person included Uzbekistan in their list, despite this team not qualifying for the World Cup. Their list was truncated at the item before Uzbekistan. We obtained benchmark (non-corrupted) information about the tournament, including the matches played and their results, from the governing body website http://www.fifa.com/.

Model

To model both the ranking and voting data, we extended the canonical Thurstonian model to produce the graphical model shown in Figure 5. The latent locations continue to generate ranking behavior, modulated by individual expertise, as before. The same latent location parameters, however, now also generate voting behavior, formalized as $v_j$ up-votes for the jth team out of a total of $n_j$ votes. The cognitive process used to generate voting behavior assumes there is a logistic link function that maps latent item locations to up-voting probabilities. This logistic function is parameterized by a slope $\alpha$, which is given a prior distribution that constrains teams with better latent strengths to correspond to higher probabilities of up-voting. The number of up-votes then naturally follows a binomial distribution with respect to this probability and the total number of votes.

The nature of the World Cup competition provides additional relevant information that, given the goal of making accurate predictions, should be combined with the information provided by people’s rankings and votes. In particular, the competition has a bracket structure, in which the 32 teams are divided into eight groups of four, and an initial round of group play determines the top two teams from each group, who progress to the elimination stages of the tournament. In this way, the bracket structure provides the prior knowledge that, for each of the eight groups, exactly two teams will finish in the top 16, and the other two will finish in the bottom 16. Our Bayesian ap-
approach allows this information—which, unlike the ranking and voting behavioral data, is inherently nonpsychological, and comes from the logical structure of the problem—to be incorporated naturally into the joint prior over the $\mu_i$ location parameters. Computationally, we achieved this by applying the constraint that, for the four teams in every group, two of the location parameters were negative, and the other two were positive. These constraints insure that, for any posterior sample of the $\mu_i$ parameters for all teams, exactly two from each group are in the top 16 positions and the other two are in the bottom 16 positions. This feature of the model is denoted by the $\beta$ constraint in the specification of the prior for the $\mu_i$ parameters in Figure 5. We acknowledge that the structure of the tournament provides additional constraints, involving which teams can play each other in the knock-out stages. Conceptually, this information can also be incorporated in the joint prior over the $\mu_i$ parameters, but we were unable to find or develop a method for applying these additional constraints in practice.

**Results**

**Aggregate ranking.** Figure 6 shows the inferred aggregate ranking of teams, listing the teams from predicted first to last place from top to bottom. The left-hand panel shows the voting data, and the right-hand panel shows the ranking data. The influence of the prior imposed by the bracket structure is evident in this aggregate ranking. There are several teams, such as Uruguay and Chile, ranked in relatively low positions in the list given the behavioral data, because they were drawn in difficult groups with at least two other strong teams. Chile, for example, was in the same group as the Netherlands and Spain. On the other hand, Switzerland is an example of a team placed in a relatively high position in the aggregate list given its rankings, because it was drawn in a less difficult group.

**Model and individual performance.** The elimination stages include round of 16 matches followed by quarter- and semi-finals, before the final and a third-place playoff. Thus, overall, the World Cup produces a partial order of the teams. Rankings 1, 2, 3, and 4 are determined exactly, but the four teams that lost quarterfinals are tied for fifth place, the eight teams that lost in the round of 16 are tied for sixth place, and the 16 teams eliminated after group play are tied for last place. The partial tau metric is general enough to handle this partial ordering as a ground truth, and can compare the final tournament standings to both complete and top-$n$ rankings.

Figure 7 shows the partial tau measure of performance for each individual, as a distribution of stick figures. The chance distribution for complete random orderings of all 32 teams is shown by the broken curve. Circular markers show the performance of a number of particular rankings. Because the World Cup tournament does not produce a complete ranking of the 32 teams—all of the 16 teams eliminated after the group stage, for example, effectively finish equal last—the best possible partial tau is 77, and will be achieved by any ranking that is completely consistent with the final tournament standings. The partial tau for the aggregate ranking produced by the model is 147. This is better than all of the individuals, and better than any reduced version of the model formed by removing any combination of the ranking, voting, or bracket information. The model predictions

![Figure 5. Graphical model formalization of the joint Thurstonian model of top-$n$ rankings and voting behavior, including a bracket structure constraint on the joint prior distribution of the latent location parameters.](image-url)
outperform the ordering produced by the Borda count method, which has a partial tau of 174. It also outperforms two external points of comparison, given by the predictions published by the betting site betfair.com, which has a partial tau of 187, and by the predictions published by the site fivethirtyeight.com, which has a partial tau of 174.

It is interesting to note that the inferences about individual expertise are less useful in this application. The scatter-plot in Figure 7 shows a weakened correlation between the posterior expectation of the expertise parameter and partial tau. The correlation of $r = .21$ is smaller than for the movie application, and for the many applications to complete lists presented in Lee et al. (2014). We suspect the key difference in this application is that people’s rankings violated the bracket structure of the tournament. For example, people placed three teams from the same group inside their top 16 a total of 31 times, and there were two occasions in which all four teams from a bracket were placed inside the top 16. A total of 17 out of the 31 people (55%) produced at least one such violation in their top-$n$ rankings. These individual violations, contrasted with the use of the bracket structure information by the model, creates a disconnect between people’s answers and the model’s aggregate answer. We suspect this corrupts the regularities on which the effective inference of expertise, as explained by Lee et al. (2014, see especially Figure 7), is based.

**Groups of Political Opinions**

**Data**

The data for our third application come from the Ranker list “The Worst U.S. Presidents.”\(^6\) The data were downloaded on 19 of April 2014, and involved a total of 99 people providing top-$n$ lists of 46 items.

\(^6\) http://www.ranker.com/crowdranked-list/the-10-worst-u-s-presidents.
Model

These data reflect political opinions, and do not have an objective ground truth against which an inferred aggregate list can be assessed, although there are many polls of historians and other experts that attempt to rank the U.S. presidents (e.g., Schlesinger, 1997). Based on the obvious intuition that there might be more than one political opinion underlying the ranking of the presidents, we developed an extended Thurstonian model that allows for the possibility of multiple latent rankings. This extended graphical model is shown in Figure 8. There is now a plate for different groups, corresponding to different latent rankings, and indicator variables $z_i$ that indicates to which group the $i$th person belongs.

The model with latent groups is closely related to models developed in Cultural Consensus Theory (CCT; Romney, Batchelder, & Weller, 1987). In CCT, the goal is to model underlying beliefs without access to a ground truth or “answer key,” and it is often assumed there are cultural subgroups with different beliefs. The goal of inferring beliefs is closely related to our goal of aggregating knowledge, and the lack of an objective ground truth is consistent with our focus on modeling the observed behavioral data. The latent mixture modeling formalism we use is also consistent with recent implementations of CCT models (Anders & Batchelder, 2012; Karabatsos & Batchelder, 2003; Oravecz, Anders, & Batchelder, 2015). CCT typically does not deal with ranking data in general, nor top-$n$ data in particular, and one way of understanding the model in Figure 8 is as an extension of CCT to these sorts of data.

Results

Group membership. Figure 9 shows the inferences about group membership for all 99 people, under the assumption there are $G = 2$ groups. The posterior mean for the $z_i$ indicator variables measures the certainty with which each person is placed in one or the other group. It is clear that the vast majority of people are confidently classified, with everybody having a level of certainty of at least 70% for their most likely group, and most being very confidently assigned. A posterior predictive analysis (Gelman, Carlin, Stern, & Rubin, 2004) showed correlations between observed and model-based posterior predictive distributions of 0.625,

![Figure 7. Partial tau performance for the World Cup application. Stick figures show the distribution of people’s performance. Circular markers show the performance of the Thurstonian model with and without the vote and bracket information, the Borda count, and the fivethirtyeight and betfair predictions. Also shown is the best possible partial tau of 77, corresponding to a prediction matching the tournament result. The broken curve shows the chance distribution of partial tau for full rankings of the 32 teams. The inset scatter-plot shows the posterior mean of the expertise parameter $\sigma$ against the partial tau for each individual.](image-url)
0.709, and 0.705 for one-, two-, and three-group models, respectively. Conceptually, it would be desirable to estimate Bayes factors between models assuming different numbers of groups, but this is a difficult computational problem. Taken together, the certainty of group membership and posterior predictive correlation results heuristically suggest the assumption of two latent groups is a useful one. In addition, roughly equal numbers of people are assigned to each group, suggesting both are important for understanding the data.

Inferred rankings. Figure 10 summarizes the two rankings inferred by the model. Each panel corresponds to one of the subgroups, and the first 15 presidents in each of these subgroups is shown, listed from worst to best. The first group naturally interpreted as a “Republican” group that lists recent Democrats—Barack Obama, Jimmy Carter, and Bill Clinton—as the worst presidents. The second group lists George W. Bush as the worst president, and then includes a mix of Democrats and Republicans, many of whom are consistently regarded by the polls of historians and experts as among the worst presidents (e.g., Schlesinger, 1997).

One way of conceiving the results in Figure 10 is that Group 1 acts as a “contaminant” group, identifying those people with rankings that correspond to one political perspective, and removing their rankings from the inference about the aggregate ranking in Group 2. If the original model in Figure 2 is applied to the data, the single latent ranking that is inferred is essentially a corrupted version of the Group 2

---

**Figure 8.** Graphical model formalization of the Thurstonian model of top-\(n\) rankings, assuming multiple latent groups with different rankings.

**Figure 9.** Assignment of people to two latent groups of political opinions about worst U.S. presidents. The posterior expectation of the \(z_i\) indicator variable for each person is shown, ordered from the person most likely to belong to Group 1 to the person most likely to belong to Group 2.

**Figure 10.** Summarizes the two rankings inferred by the model. Each panel corresponds to one of the subgroups, and the first 15 presidents in each of these subgroups is shown, listed from worst to best.
Discussion

The basic goal of aggregation is to combine information across people to create accurate or otherwise useful group answers. Depending on the nature of the problem, this aggregation potentially involves two complementary components. One is signal-amplification, and arises if there is a common signal that each individual perceives in an imperfect way, such as the number of jelly beans in a jar, or the 10 longest rivers in the world. In these cases, effective aggregation processes naturally amplify the common signal and cancel out the idiosyncratic noise, bias, or errors in individual estimates. A second component is jigsaw-completion, and arises if there are multiple parts to an answer, and different individuals have access to different parts. Ranking the highest earning athletes might be best done by combining the ranks of individual experts in boxing, golf, football, and other relevant sports.
In this paper, we tackled the problem of combining top-\(n\) rankings, which generally requires both components. In the movie popularity and World Cup applications, there is a ground truth that must be predicted, and this serves as the common signal that is sought by aggregation. In all three applications, the nature of the top-\(n\) task means that most people do not rank most items, and so an aggregated list needs to be formed by combining items across individual lists. The Thurstonian model we developed naturally incorporates both the signal-amplification and jigsaw-completion components of aggregation. It makes simple assumptions about how the ranking of items are represented, how individuals differ in their representation, and how the latent ordering and individual expertise combine to produce top-\(n\) rankings. The specification of a single probabilistic process that maps a latent ordering of the items to observed individual rankings automatically enables jigsaw-completion, because it formalizes how likely every possible latent ordering is to have produced each individual’s list. As emphasized in the movie popularity application, the relative expertise of individuals is often accurately inferred from the patterns of agreement between their rankings, and the assumptions of the model then naturally lead to the rankings of the more expert individuals being up-weighted in determining the aggregate ranking.

The applications presented in this paper all have dual goals as evaluations of cognitive models of top-\(n\) rankings, and applications of that cognitive model to analyze real-world data. We think these are symbiotic goals, and pursuing both simultaneously leads to challenges and advantages that are not encountered by pursuing either in isolation. If our goals were entirely focused on using data from Ranker lists to make predictions, it is likely discriminative machine-learning algorithms would be preferred to the probabilistic models of cognition and generative inference methods we used (Lasserre, Bishop, & Minka, 2006). It is also possible that purely statistical approaches, of the type surveyed by Lin (2010), could perform as well or better on the data we consider. These approaches, however, would forgo the useful capabilities of the modeling that come from the focus on cognitive variables and Bayesian inference. As one example of the benefits of the psychological approach, the ability to make inferences about individual-level expertise in the movie application comes from relying on a parameterization of the canonical model—with an independent \(\sigma\) per person—that can be interpreted as their expertise. As a second example, the joint modeling of multiple behavioral data achieved by the extended model used in the World Cup application is a natural psychological goal. It is central to many modeling efforts in psychology, such as models of decision-making that make predictions about both choices and response times (e.g., Brown & Heathcote, 2008; Ratcliff & McKoon, 2008; Vickers, 1979). The World Cup application showed the benefit of jointly modeling both ranking and voting data to improve predictions. This application also highlighted the ability to include the bracket structure information, using Bayesian priors naturally as a means of expressing relevant existing knowledge (Jaynes, 2003). Ultimately, it would be possible to incorporate analogous up-weighting schemes to mimic expertise, some sort of combination process to mimic joint modeling, and some sort of constraint satisfaction extension to mimic priors in a machine learning or statistical algorithm. Such algorithms might display better predictive performance, but it is clear the motivation for these capabilities come from the goal of understanding the cognitive parameters and processes involved, and the use of Bayesian methods.

In a complementary way, the real-world data we use, and the applied problems we consider, help further the basic cognitive science goals of understanding how people represent knowledge about orders and express that knowledge in top-\(n\) lists. The data from Ranker lists cover a wide variety of topic areas, and involve users with large individual differences. This variety, and the lack of restrictive experimental control in data collection, mean the data considered in this paper raise many cognitive modeling challenges that remain to be addressed. One challenge is to improve the way in which individual differences are modeled. The current model does this through a single individual-level measure of general expertise, and through the use of two latent groups in the U.S. presidents application. These are simple approximations to the true individual differences that likely exist. Presumably different people give different emphasis to different genres depending on their movie-going preferences, and it is extremely likely
the World Cup predictions are biased by the country each person supports. These intuitions suggest that a more complete model of individual differences should allow for a more theoretically motivated prior than the uniform one currently used, individual-by-item interactions, structured latent groups of both people and items, or some combination of all of these possibilities.

A second challenge is to develop a more complete cognitive model of the top-\(n\) data by modeling termination processes. The current model does not include any psychological parameters or processes that account for why different people produce lists of different lengths, even though it is a basic task characteristic. There are existing theories and models that deal with the related topic of modeling search termination (e.g., Lee, Newell, & Vandekerckhove, 2014) that could be applied. A more complete model could, for example, parameterize the caution of different people when producing top-\(n\) lists for different topics, and include processes that—perhaps by interacting with expertise—make predictions about how many items are included in that person’s list. Even more broadly, there are a number of standard statistical methods, besides the Thurstonian one pursued here, that could form the basis for alternative psychological models. These include seminal models like Mallow’s model (Lu & Boutilier, 2011), the Bradley-Terry model (Caron & Doucet, 2012), and Plackett-Luce models (Guiver & Snelson, 2009), as well as recent generalizations and extensions (e.g., Qin, Geng, & Liu, 2010).

A final cognitive modeling challenge posed by the current data involves incorporating memory processes. One reason different people rank different items relates to whether or not they retrieve the items from memory. Lee, Liu, and Steyvers (2015) present empirical evidence, from a controlled study involving the generation of top-10 lists, that people often fail to include some items in their lists because of limitations of memory rather than knowledge about the items. It seems very likely this is true for at least the movie popularity application, which included many unique movies only sparsely ranked by individuals. A more complete cognitive model of top-\(n\) lists in cases like these would involve interacting memory and decision-making processes, and would have potentially strong implications for the inferred aggregate lists. If a model allows for the possibility that a person does not include an item because of memory failure, the absence of that item on their list does not necessarily mean they believe the item has a low rank. This possibility is not afforded by the current model, and could lead to very different aggregate rankings for some sets of top-\(n\) lists. The alternatives to the Thurstonian model mentioned above, with their different assumptions about the processes used to generate lists, could prove especially valuable for including memory and termination processes in cognitive models.

Addressing these cognitive modeling challenges is a natural direction for future work. People are able to produce top-\(n\) lists easily for all sorts of topics, reflecting their ubiquity as expressions of basic knowledge. Understanding how people produce these lists helps understand how people represent knowledge, how they make estimates, judgments, and decisions based on that knowledge, and what the nature of the individual differences in these representations and processes are. Our expectation is that, the more completely and accurately we are able to model the individual differences and cognitive processes that generated the data, the better the aggregated rankings inferred by the model will perform. Good cognitive models of top-\(n\) lists have rich potential for application in combining human knowledge, enabling the measurement of the expertise, caution, and other basic properties of individuals, and making accurate predictions about orders.

References


Received September 10, 2015
Revision received December 22, 2015
Accepted February 8, 2016