Abstract

Averaging the estimates of a number of individuals has been shown to produce an estimate that is generally more accurate than those of the individuals themselves. Similarly, averaging responses from a single individual can also lead to a more accurate answer. How can we best combine estimates within and between individuals to create an accurate group estimate? We report empirical results from a general knowledge rank-ordering experiment and demonstrate that individuals that provide more consistent answers across repeated elicitations are also more accurate. We develop a consistency weighting heuristic and show that repeated elicitations within an individual can be used to improve group accuracy. We also develop a Thurstonian cognitive model which assumes a direct link between the process that explains the accuracy of an individual and response consistency and show how the model can infer accurate group answers.

Keywords: Bayesian Modeling; Rank Ordering; Knowledge; Recall; Wisdom of Crowds; Within; Expertise; Uncertainty; Coherence; Consistency.
In this paper, we focus on rank-ordering questions where the task is to rank-order a set items such as Presidents by terms in office or US cities by population size (Miller, Hemmer, Steyvers, Lee, 2009; Lee et al. 2012; Lee, Steyvers, Miller, 2014). In contrast to simple yes/no or percentage estimation question involving single quantities, rank-ordering questions involve the retrieval and coordination of many pieces of information, making it less likely that a subject can explicitly remember a previous response. In the absence of easily available episodic strategies, subjects can be asked for a second response almost immediately after their first, eliminating the need for multiple conditions and removing any question anchoring effects.

Our contribution in this paper is threefold. First, we show that the crowd within an individual effect observed by Vul and Pashler exists for rank-ordering tasks, indicating that there is a degree of statistical independence between repeated elicitations for rank-ordering judgments. Second, we demonstrate that the agreement between the first and second response is related to each subjects’ response accuracy. We present a simple consistency weighting heuristic where rank-ordering judgments from individuals that are consistent across repeated questions are given larger weight in the group average. We demonstrate that this consistency weighting heuristic significantly improves group accuracy. Finally, we introduce a new repeated-elicitation variant of a Thurstonian model for rank-ordering that has been explored elsewhere (see Steyvers et al., 2009 & Lee et al., 2012). We compare the performance of the repeated-elicitation model and the original variant, and demonstrate that accounting for the variance in an individual’s responses improves overall group aggregation performance.

**Experiment**

**Method**

The experiment was composed of 8 rank ordering questions, and an additional 3 distracter questions; the distracter questions were included to increase the delay between subject responses. Increased delay between responses has been shown previously to increase response independence and effect size (see Vul & Pashler, 2008). Subjects were 120 undergraduate students between the ages of 18 and 22 at the University of California, Irvine who were compensated with course credit.

Selection for the non-distracter questions was based on difficulty, as determined by the accuracy of subjects in previous experiments (Steyvers et al., 2009; Miller et al., 2011). Approximately one third of questions were selected for being easier (U.S. Holidays, U.S. Presidents, Book Release Dates), three for being moderately difficult (Country Landmasses, U.S. Cities, European Cities), and one two for being particularly difficult (10 Amendments, World Cities). All were general knowledge questions that subjects were likely to have had exposure to. For the distracter questions, subjects were asked to rank teams for the NFL and NBA based on what they thought their final season standing would be.

Subjects were given the eight knowledge questions in a random order, and items for each question were initially placed in random positions. Subjects were then given the distracter questions. Subjects were then prompted to give responses for the eight questions again, in the same order they appeared in the first elicitation, but with a new random initial placement of the items for each question.

All questions had a ground truth obtained from Pocket World in Figures and various online sources. An interactive interface was presented via a web browser on computer screens. Subjects were instructed to order the presented items (e.g., “Order these books by their first release date, earlist to most recent”), and responded by dragging the individual items on the screen using the computer mouse and “snapping” them into the desired locations in the ordering, as in previous experiments. Transitions between question blocks were marked by a holding page reminding subjects of the instructions for the tasks. At no point were subjects informed that they would be answering the same questions twice.

**Results**

**Assessing Accuracy** Performance was measured relative to the ground truth using Kendall’s tau distance τ. This metric is used to count the number of pair-wise disagreements between the reconstructed and correct ordering (lower is better). The larger the distance, the more dissimilar the two orderings are. Values of τ range from: $0 \leq \tau \leq N(N-1)/2$, where N is the number of items in the order (ten for all of our questions). A value of zero means the ordering is exactly right, a value of one means that the ordering is correct except for two neighboring items being transposed, and so on up to the maximum possible value of forty-five (indicating that the list is completely reversed). An average score of 22.5 is expected for random performance.

**Averaged Responses** We first evaluated whether or not averaging the responses within each individual reduced the error relative to the individual responses, indicating statistically independent error of the sort observed in the simple recall tasks of Vul and Pashler (2008). Table 1 shows the median Kendall’s tau distance for individual rank-ordering problems for the first and second response as well as the combined first and second response using the Borda aggregation method (see modeling section for Borda details). Subjects’ error on the first and second responses were not significantly different, on average, and varied according to question difficulty. The averaged first and second responses of each subject (combined column in Table 1) was less erroneous than the first and second responses – t(120)=2.16, p<.05 and t(120)=2.87, p<.01 respectively – replicating the findings of Vul and Pashler (2008) for rank ordering tasks.
Response Consistency and Accuracy If subject response consistency is correlated to the precision of an individual’s knowledge, then multiple independent responses should be further apart from each other the less knowledgeable a subject is. We quantified (inverse) response consistency as the Kendall’s tau distance between subjects’ responses. Subjects with a larger distance between their first and second judgment should show a higher tau distance to the ground truth. Figure 1 illustrates this relationship separately for the first and second response. The correlation between each subject’s response disagreement, and the error of their first and second responses, is $\rho = .51$ and $\rho = .55$ respectively. This correlation is observed not only across all questions, but also for each individual question. The correlation between response disagreement and accuracy appears to scale linearly with overall subject accuracy for the problem.

Modeling

While averaging across a given individual’s responses yields answers that are more accurate, the improvement is far smaller than averaging two responses across subjects (Miller et al., 2011). Given a large number of subjects, it is unclear whether repeated elicitations would improve group responses aggregation if they are merely treated as extra subjects. Can within-subject response consistency be integrated into a between-subject aggregation model to improve overall accuracy? To test this, we evaluate two models – a heuristic approach based on Borda aggregation method and a Thurstonian cognitive model of subject behavior.

Borda Aggregation

In order to assess if incorporating within-subject response consistency can improve between-subject estimates for rank ordering tasks, we used a modified version of Borda count aggregation that incorporates subject weighting. Borda aggregation is a representative aggregation heuristic that has been used widely elsewhere (see Miller et al., 2009). In traditional Borda count aggregation, all items are assigned points based upon their location in a given response: 1 point for being in position 1, 2 points for being in position 2, up to 10 points for a list of 10 items. In a standard Borda aggregation method, the points are added across all rank-orderings provided by subjects and the items are ordered according to the sum totals for each item. In our modified Borda aggregation method, we add a weighting factor for each individual subject in order to upweight subjects that are more consistent. Specifically, we calculate the point total $s_k$ for each item $k \in \{1, \ldots, K\}$ by:

$$s_k = \sum_{j=1}^{N} r_{j,k} w_j$$

where $r_{j,k}$ is the rank of item $k$ for subject $j \in \{1, \ldots, N\}$, and $w_j$ is the weight given to subject $j$. As in a standard Borda method, the sums of these points for each item are then ranked from smallest to largest to determine the final Borda aggregate rank ordering.

For an unweighted aggregate rank-ordering, the subject weights were set to the same value for all participants. We used this as the baseline for comparison. For the aggregate rank-ordering weighted by response consistency, we use the inverse of the tau disagreement between the first and second rank-ordering:

$$w_j = 1/(\tau_j + 1)$$

where we add one to the distance in order to avoid zero division. Therefore, the rank-orderings of participants with larger response consistency have a stronger influence on the aggregate rank ordering.

Figure 2 shows the aggregation results. As we found previously (Miller et al., 2009), unweighted Borda

Table 1: Subject response error (Kendall’s tau) across individual rank-ordering problems.

<table>
<thead>
<tr>
<th>Problems</th>
<th>1st Error</th>
<th>2nd Error</th>
<th>Combined Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landmass</td>
<td>9</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Holidays</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Presidents</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Books</td>
<td>11</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Euro Cities</td>
<td>15</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>US Cities</td>
<td>16</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>World Cities</td>
<td>21</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>10 Amendments</td>
<td>16</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td><strong>12.5</strong></td>
<td><strong>12.7</strong></td>
<td><strong>11.9</strong></td>
</tr>
</tbody>
</table>
aggregation outperforms the average subject for all eight questions. Additionally, the weighted Borda model performs as well as or better than the unweighted model for all but two of the questions. The weighted Borda model performed worse for the Presidents question because most subjects performed so well that weighting over-penalized the many subjects with near-correct responses. Similarly, the model performed poorly on the European Cities question because there were so few subjects that performed well. Aggregation for the unweighted Borda model was performed across both trials so as not to give the weighted model the advantage of extra subject responses. This superior performance in reconstructing the ground truth ordering demonstrates that response consistency can be used to improve group accuracy for rank ordering tasks. Next we explore whether a cognitive model of the rank-ordering task can better describe subject behavior and more accurately reconstruct the ground truth.

Thurstonian Model

Given that subject response consistency is clearly related to accuracy in rank-ordering tasks, what kind of mechanism might be responsible for this observed behavior? We developed a probabilistic model based upon a Thurstonian approach. In a Thurstonian representation, the latent ground truth ordering for a specific problem is represented by coordinates on an interval scale. As Figure 3a illustrates, each item $k$ is represented as a latent coordinate $\mu_k$ on an interval dimension. Note that this represents not the actual ground truth but the latent truth as perceived by a group of individuals. The one-dimensional representation of items is appropriate as all problems in our study involve one-dimensional relative judgments (e.g. the size of items and the timing of events).

Each individual $i$ is assumed to have access to all of the ground truth latent coordinates $\mu$, but without precise knowledge about their exact locations. This uncertainty is represented with normal distributions that are centered on the shared latent ground truth locations and with a subject-level $\sigma_i$ that represents the uncertainty of the individual about the item locations. Note that for a given subject, all items have the same standard deviation which is a strong assumption but simplifies the model considerably.

As Figure 3b shows, the subject draws mental samples from these item distributions. Repeated elicitations are modeled simply by repeating the sampling process which leads to a new set of samples. The rank-ordering produced by a subject is then based on the order of the mental samples.

As illustrated in Figure 3c, different subjects can have different uncertainty $\sigma_i$, and this influences not only the response accuracy but also the response consistency. For example, the larger uncertainty associated with the subject illustrated in Figure 3c leads to more transposition errors in the mental samples associated with a given response – it becomes more likely that samples of nearby distributions are out of order (relative to the latent ground truth) which lowers accuracy. In addition, the larger uncertainty also leads to increased differences in orderings between different responses. Therefore, the model assumes an inherent connection between response consistency and accuracy – they are both driven by a latent parameter $\sigma_i$ that represents the (inverse) expertise level of a subject for a particular

Figure 2: Aggregation performance of unweighted and weighted Borda aggregation across first and second responses, compared to the average subject performance.

Figure 3: Illustration of the Thurstonian Model for repeated elicitations. (a) The latent ground truth is represented as a set of coordinates on an interval scale (b) Uncertainty about the latent ground truth is represented by Gaussian noise and responses are created by sampling latent values from each item distribution (c) Example of a subject with larger uncertainty about the ground truth and larger variability in the item samples across the first and second response
This multiple-elicitation model is different from previous Thurstonian models that we have presented, where subjects only give a single response per question (Steyvers et al., 2009; Miller & Steyvers, 2011). This extended model allows us to examine whether accuracy and response consistency can be described with the same underlying mechanism.

We apply Bayesian estimation techniques to infer the group representation from individual orderings. Figure 4 shows the Thurstonian model for a single question across subjects using graphical model notation (see Koller, Friedman, Getoor, & Taskar, 2007; Shiffrin, Lee, Kim, & Wagenmakers, 2008, for statistical and psychological introductions). Each node represents a model variable, and the graph structure is used to indicate the conditional dependencies between these variables. Stochastic and deterministic variables are indicated by single-and double-bordered nodes (\(\mu\), \(\sigma\), \(x\) and \(y\) respectively), and observed data are represented by a shaded node (\(y\)). The plate represents independent replications of the graph structure, which corresponds to multiple elicitations from each individual \(i\) and across individuals for each question \(j\).

To explain how these data are generated, the model begins with the underlying ground truth location of the items, given by the vector \(\mu\). The latent ground truth \(\mu\) is given a flat prior such that all item locations are equally likely a priori. Each individual has an associated uncertainty parameter \(\sigma_j \sim \text{Gamma}(\lambda_0, 1/\lambda)\) where \(\lambda\) is a hyper-parameter that determines the variability of the expertise levels across individuals. We set \(\lambda = 3\) in the current model.

To determine the order of items for the \(i\)th repetition, the \(j\)th individual samples a location \(x_{ijk}\) for each item \(k\) where \(x_{ijk} \sim \text{Normal}(\mu_k, \sigma_j)\). The sample \(x_{ijk}\) represents the realized mental representation for the individual at that particular time. The ordering for each individual is determined by the ordering of all of their mental samples \(y_{ij} = \text{Rank}(x_{ij})\).

While the generative model is relatively straightforward, the inference is challenging because the observed data \(y_{ij}\) is a deterministic ranking. We utilized MCMC procedures originally developed by Yao and Böckenholt (1999), which allowed us to estimate the posterior distribution over the latent variables \(x_{ijk}\), \(\sigma_j\), and \(\mu\) given the observed orderings \(y_{ij}\). We use Gibbs sampling to update the mental samples \(x_{ijk}\), and Metropolis-Hastings updates for \(\sigma_j\) and \(\mu\).

Figure 5 shows the accuracy of three aggregation models, and demonstrates that the repeated elicitation Thurstonian model performed best overall. It outperformed the weighted Borda model and also outperformed a Thurstonian model that is given both the first and second response of participants but treats the second responses as coming from a new set of participants. Additionally, the repeated elicitation Thurstonian model matched or exceeded other models’ performance for each individual question.

The advantage of the repeated elicitation Thurstonian model over the Thurstonian model where the first and second responses are not linked to the same subject is not due to the fact that it has access to additional response information (it uses the same set of subject responses), but because the model simultaneously infers a subject’s uncertainty based upon their disagreement with other subjects and their disagreement with themselves. In this way, we have some confidence in the Thurstonian representation of individual-level uncertainty for subject item recall, both as a generative model and as a means of yielding more accurate group estimates for rank ordering tasks.

**Conclusions**

In this paper, we have shown that repeated elicitations for general knowledge rank-ordering tasks exhibit statistically independent error, and the variance of that error is correlated to the accuracy of subject responses for easy and difficult questions. Additionally, we have shown that this response consistency can be used to improve group
aggregate accuracy in reconstructing the ground truth answer for rank ordering knowledge tasks. These findings might also be applicable to tasks that do not have a known ground truth, as we have discussed elsewhere (Lee et al., 2012). Finally, we introduced a cognitive model of rank-order judgement wherein a subject-level uncertainty parameter accounted for both subject response accuracy and response consistency, and found that it was best able to capture subject behavior and reconstruct the original ground truth ordering for each of our questions. This lends credence to the idea of a combined probabilistic mechanism for consistency and accuracy underlying the subject behavior observed in these complex knowledge recall tasks.

**References**


