A Model-Based Approach to Measuring Expertise in Ranking Tasks

Michael D. Lee (mdlee@uci.edu)
Mark Steyvers (msteyver@uci.edu)
Mindy de Young (mdeyoung@uci.edu)
Brent J. Miller (brentm@uci.edu)
Department of Cognitive Sciences,University of California, Irvine
Irvine, CA, USA 92697-5100

Abstract

We apply a cognitive modeling approach to the problem of measuring expertise on rank ordering tasks. In these tasks, people must order a set of items in terms of a given criterion. Using a cognitive model of behavior on this task that allows for individual differences in knowledge, we are able to infer people's expertise directly from the rankings they provide. We show that our model-based measure of expertise outperforms self-report measures, taken both before and after doing the task, in terms of correlation with the actual accuracy of the answers. Based on these results, we discuss the potential and limitations of using cognitive models in assessing expertise.

Keywords: expertise, ordering task, wisdom of crowds, model-based measurement

Introduction

Understanding expertise is an important goal for cognitive science, for both theoretical and practical reasons. Theoretically, expertise is closely related to the structure of individual differences in knowledge, representation, decision-making, and a range of other cognitive capabilities (Wright & Bolderm, 1992). Practically, the ability to identify and use experts is important in a wide range of real-world settings. There are many possible tasks that people could do to provide their expertise, including estimating numerical values (e.g., "what is the length of the Nile?"), predicting categorical future outcomes ("who will win the FIFA World Cup?"), and so on. In this paper, we focus on the task of ranking a set of given items in terms of some criterion, such as ordering a set of cities from most to least populous.

One prominent theory of expertise argues that the key requirements are discriminability and consistency (e.g., Shanteau, Weiss, Thomas, & Pounds, 2002). Experts must be able to discriminate between different stimuli, and they must be able to make these discriminations reliably or consistently. Protocols for measuring expertise in terms of these two properties are well-developed, and have been applied in settings as diverse as audit judgment, livestock judgment, personnel hiring, and decision-making in the oil and gas industry (Malhotra, Lee, & Khurana, 2007). However, because these protocols need to assess discriminability and consistency, they have two features that will not work in all applied settings. First, they rely on knowing the answers to the discrimination questions, and so must have access to a ground truth. Second, they must ask the same (or very similar) questions of people repeatedly, and so are time consuming. Given these limitations, it is perhaps not surprising that expertise is often measured in simpler and cruder ways, such as by self-report.

In this paper, we approach the problem of expertise from the perspective of cognitive modeling. The basic idea is to build a model of how a number of people with different levels of expertise produce judgments or estimates that reflect their knowledge. This requires making assumptions about how individual differences in knowledge are structured, and how people apply decision-making processes to their knowledge to produce answers.

There are two key attractive properties of this approach. The first is that, if a reasonable model can be formulated, the knowledge people have can be inferred by fitting the model to their behavior. This avoids the need to rely on self-reported measures of expertise, or to use elaborate protocols to extract a measure of expertise. The cognitive model does all of the work, providing an account of task behavior that is sensitive to the latent expertise of the people who do the task.

The second attraction is that expertise is determined by making inferences about the structure of the different answers provided by individuals. This means that performance does not have to be assessed in terms of an accuracy measure relative to the ground truth. It is possible to measure the relative expertise of individuals, without already having the expertise to answer the question.

The structure of this paper is as follows. We first describe an experiment that asks people to rank order sets of items, and rate their expertise both before and after having done the ranking. We then describe a simple cognitive model of the ranking task, and use the model to infer individual differences in the precision of the knowledge each person has. In the results section, we show that this individual differences parameter provides a good measure of expertise, in the sense that it correlates well with actual performance. We also show it outperforms the self-reported measures of expertise. We conclude with some discussion of the strengths and limitations of our cognitive modeling approach to assessing expertise.

Table 1: The six rank ordering tasks. Each involves ten items, shown in correct order.

Holidays	Landmass	Amendments	US Cities	Presidents	World Cities
New Year's	Russia	Freedom speech and religion	New York	Washington	Tokyo
Martin Luther King	Canada	Right to bear arms	Los Angeles	Adams	Mexico City
President's	China	No quartering of soldiers	Chicago.	Jefferson	New York
Memorial	United States	No unreasonable searches	Houston	Monroe	Sao Paulo
Independence	Brazil	Due process	Phoenix	Jackson	Mumbai
Labor	Australia	Trial by jury	Philadelphia	Roosevelt	Delhi
Columbus	India	Civil trial by jury	San Antonio	Wilson	Shanghai
Halloween	Argentina	No cruel punishment	San Diego	Roosevelt	Kolkata
Veteran's	Kazakhstan	Right to non-specified rights	Dallas	Truman	Buenos Aires
Thanksgiving	Sudan	Power for states and people	San Jose	Eisenhower	Dhaka

Experiment

Participants

A total of 70 participants completed the experiment. Participants were undergraduate students recruited from the University of California, Irvine subject pool, and given course credit as compensation.

Stimuli

We used six rank ordering problems, all with ten items, as shown in Table 1. All involve general 'book' knowledge, and were intended to be of a varying levels of difficulty for our participants, and lead to individual differences in expertise.

Procedure

The experimental procedure involved three parts. In the first part, participants completed a pre-test self-report of their level of expertise in the general content area of each of the stimuli. This was done on a 5-point scale, simply by asking questions like "Please rate, on a scale from 1 to 5, where 1 is no knowledge and 5 is expert, your knowledge of the order of American holidays.".

In the second part, participants completed each of the six ranking questions from Table 1 in a random order. Within each problem, the ten items were presented in an initially random order, and could then be 'dragged and dropped' to any part of the list to update the order. Participants were free to move items as often as they wanted, with no time restrictions. They hit a 'submit' button once they were satisfied with their answer. No time limit was placed

The third part of the experimental procedure was completed immediately after each final ordering answer was submitted. Participants were asked to express their level of confidence in their answer, again on a 5-point scale, were 1 was 'not confident at all' and 5 was 'extremely confident'.

A Thurstonian Model of Ranking

We use a previously developed Thurstonian model of how people complete ranking tasks (Steyvers, Lee, Miller, & Hemmer, 2009). Originally, this model was developed in the context of the 'wisdom of the crowd' phenomenon as applied to order data. The basic wisdom of the crowd idea is that the average of the answers of many individuals may be as good as or better than all of the individual answers (Surowiecki, 2004). An important component in developing good group answers is weighting those individuals who know more, and so the model we use already is designed to accommodate individual differences in expertise.

We first illustrate the model intuitively, and explain how its parameters can be interpreted in terms of levels of knowledge and expertise. We then provide some more formal details, including some information about the inference procedures we used to fit the model to our data.

Overview of Model

The model is described in Figure 1, using a simple example involving three items and two individuals. Figure 1(a) shows the 'latent ground truth' representation for the three items, represented by μ_1, μ_2, and μ_3 on an interval scale. Importantly, these coordinates do not necessarily correspond to the actual ground truth, but rather represent the knowledge that is shared among individuals. Therefore, these coordinates are latent variables in the model that can be estimated on the basis of the orderings from a group of individuals.

Figure 1(b) and (c) show how these items might give rise to mental representations for two individuals. The individuals might not have precise knowledge about the exact location of each item on the interval scale due to some sort of noise or uncertainty. This mental noise might be due to a variety of sources such as encoding and retrieval errors. In the model, all these sources of noise are combined together into a single Gaussian distribution[1].

The model assumes that the means of these item distributions are the same for every individual, because, every individual is assumed to have access to the same infor-

---

[1]In our experiment, participants give only one ranking for each problem. Therefore, the model cannot disentangle the different sources of error related to encoding and retrieval.
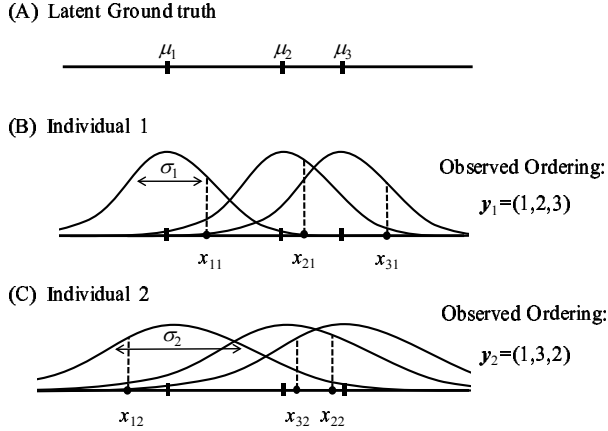
(A) Latent Ground truth

(B) Individual 1

Observed Ordering:

$\boldsymbol{y}_1 = (1,2,3)$

(C) Individual 2

Observed Ordering:

$\boldsymbol{y}_2 = (1,3,2)$

Figure 1: Illustration of the Thurstonian model.

Figure 2: Graphical representation of Thurstonian model.

mation about the objective ground truth. The widths of the distributions, however, are allowed to vary, to capture the notion of individual differences. There is a single standard deviation parameter, σ_i for the ith participant, that is applied to the distribution of all items. In Figure 1 Individual 1 is shown as having more precise item information than Individual 2, and so $\sigma_1 < \sigma_2$.

The model assumes that the realized (latent) mental representation is based on a single sample from each item distribution, represented by x in Figure 1, where x_{ij} is the sample for the ith item and jth participant. The ordering produced by each individual is then based on an ordering of the mental samples. For example, individual 1 in Figure 1(b) draws sample for items that leads to the ordering (1,2,3) whereas individual 2 draws a sample for the third item that is smaller than the sample for the second item, leading to the ordering (1,3,2). Therefore, the overlap in the item distributions can lead to errors in the orderings produced by individuals.

The key parameters in the model are μ and σ_i. In terms of the original wisdom of the crowd motivation, the most important was μ, because it represents the assumed common latent ordering individuals share. Inferring this ordering corresponds to constructing a group answer to the ranking problem. In our context of measuring expertise, however, it is the σ_i parameters that are important. These are naturally interpreted as a measure of expertise. Smaller values will lead to more consistent answers closer to the underlying ordering. Larger values will lead to more variable answers, with more possibility of deviating from the underlying ordering.

Generative Model and Inference

Figure 2 shows the Thurstonian model, as it applies to a single question, using graphical model notation (see Koller, Friedman, Getoor, & Taskar, 2007; Lee, 2008; Shiffrin, Lee, Kim, & Wagenmakers, 2008, for statistical and psychological introduction). The nodes represent variables and the graph structure is used to indicate the conditional dependencies between variables.
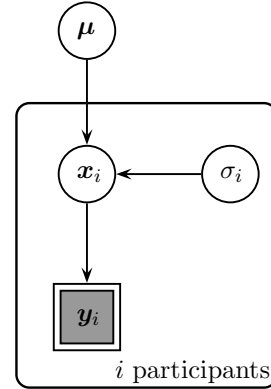
Stochastic and deterministic variables are indicated by single and double-bordered nodes, and observed data are represented by shaded nodes. The plates represent independent replications of the graph structure, which corresponds to individual participants in this model.

The observed data are the ordering given by the ith participant, denoted by the vector $\boldsymbol{y}_i$, where $\boldsymbol{y}_{ij}$ represents the item placed in the jth position by the participant.

To explain how these data are generated, the model begins with the underlying location of the items, given by the vector $\boldsymbol{\mu}$. Each individual is assumed to have access to this group-level information. To determine the order of items, the ith participant samples for the jth item, as $x_{ij} \sim \text{Gaussian}(\mu_j, \sigma_i)$, where σ_i is the uncertainty that the ith individual has about the items, and the samples x_{ij} represent the realized mental representation for the individual. The ordering for each individual is determined by the ordering of their mental samples $\boldsymbol{y}_i = \text{rank}(\boldsymbol{x}_i)$.

We used a flat prior for μ and a $\sigma_i \sim \text{Gamma}(\lambda, 1/\lambda)$ prior on the standard deviations, where λ is a hyperparameter that determines the variability of the noise distributions across individuals. We set $\lambda = 3$ in the current modeling, but plan to explore a more general approach where λ is given a prior, and inferred, in the future.

Although the model is straightforward as a generative process for the observed data, some aspects of inference are difficult because the observed variable $\boldsymbol{y}_j$ is a *deterministic* ranking. Yao and Böckenholt (1999), however, have developed appropriate Markov chain Monte Carlo (MCMC) methods. We used an MCMC sampling procedure that allowed us to estimate the posterior distribution over the latent variables x_{ij}, σ_i, and μ, given the observed orderings $\boldsymbol{y}_i$. We use Gibbs sampling to update the mental samples x_{ij}, and Metropolis-Hastings updates for σ_i and μ. Details of the MCMC inference procedure are provided in the appendix.

Results

We first describe how we measure the accuracy of a rank order provided by a participant, as a ground truth assess-
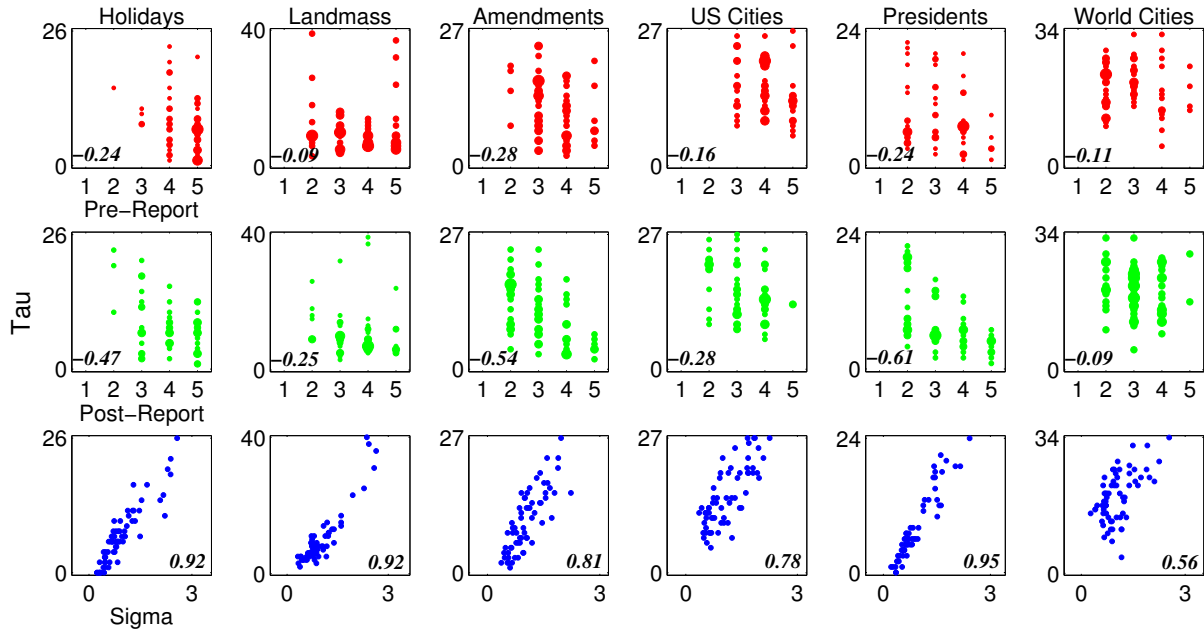
Figure 3: Results comparing the relationship between the three measures of expertise and the accuracy of individual answers. The plots are organized with the measures in rows, and the problems in columns.

ment of their expertise. We then examine the correlations between this ground truth and their pre- and post-reported self-assessments, and the model-based measure.

Ground Truth Accuracy

To evaluate the performance of participants, we measured the distance between their provided order, and the correct orders given in Table 1. A commonly used distance metric for orderings is Kendall's τ, which counts the number of adjacent pairwise disagreements between orderings. Values of τ range from $0 \leq \tau \leq n\,(n-1)/2$, where $n = 10$ is the number of items. A value of zero means the ordering is exactly right, and 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 45.

Relationship Between Expertise and Accuracy

Figure 3 presents the relationship between the three measures of expertise—pre-reported expertise, post-reported confidence, and the mean of the σ parameter inferred in the Thurstonian model—and the τ measures of accuracy. In each plot, a point corresponds to a participant. The plots are organized with the six problems in columns, and the three measures as rows. The Pearson correlations are also shown. Note that, for the self-reported measures, the goal is for higher levels of rated expertise should correspond to lower (more accurate) values of τ, and so a negative correlation would mean the measure was effective. For the model-based σ measure, smaller values correspond to higher expertise, and so a positive correlation means the measure is effective.

Figure 3 shows that the six different problems ranged in difficulty. Looking at the maximum τ needed to show results, the Holidays, Amendments, US Cities and Presidents questions were more accurately answered than the Landmass and World Cities questions. This finding accords with our intuitions about the difficulty of the topic domains and the experience of our participant pool.

More importantly, there is a clear pattern, for all six problems, in the way the three expertise measures relate to accuracy. The correlations are generally in the right direction, but small in absolute size, for the pre-reported expertise. They continue to be in the right direction, and have larger absolute values, for the post-reported confidence measure of expertise. But correlations are in the right direction, and strongest, for the model-based σ measure of expertise.

Perhaps most importantly, it is also clear that the model-based measure improves upon the self-reported measures. It achieves, for all but the world cities problem, an impressively high level of correlation with accuracy. With correlations around 0.9, the σ measure of expertise explains about 80% of the variance between people in their accuracy in completing the rank orderings.[2]

---

[2] A legitimate concern is that the correlations for the Thurstonian model benefit from σ being continuous, whereas the pre- and post-report measures are binned. To check this, we also calculated correlations for the Thurstonian model using 5 binned values of σ, and found correlations of 0.88, 0.88, 0.80, 0.77, 0.92 and 0.54 for the six problems in the order shown in Figure 3. While slightly reduced, these correlations clearly support the same conclusions.

Discussion

We first discuss the advantages of the modeling approach we have explored for measuring expertise, then acknowledge some of its limitations, before finally mentioning some possible extensions.

Advantages

Our results could be used to make a strong case for the assessment of expertise, at least in the context of rank order questions, using the Thurstonian model. We have shown that by having a group of participants complete the ordering task, the model can infer an interpretable measure of expertise that correlates highly with the actual accuracy of the answers.

One attractive feature of this approach is that it does not require self-ratings of expertise. It simply requires people to do the ordering task. Our results indicate that the model-based measure is much more useful than self-reported assessments taken before doing the task, focusing on general domain knowledge, or confidence ratings done after having done the task, focusing on the specific answer provided.

An even more attractive feature of the modeling approach is that it does not require access to the ground truth to assess expertise. We used ground truth accuracies to assess whether the measured expertise was useful, but we did not need the τ values to estimate the σ measures themselves. The model-based expertise emerges from the patterns of agreement and disagreement across the participants, under the assumption there is some fixed (but unknown) ground truth, as per the wisdom of the crowd origins of the model.

A natural consequence is that the approach developed here could be applied to *prediction* tasks, where there is not (yet) a ground truth. For example, we could ask people to predict the end-of-season rankings of sports teams, and potentially use the model to assess their expertise ahead of time. If the model-based approach continues to perform well with prediction, it would be especially valuable, since standard measures of expertise based on self-report are have often been found to be unreliable predictors of forecasting accuracy (e.g., Tetlock, 2006).

Limitations

A basic property of the approach we have presented is that it involves assessing the relative expertise for a large group of people. There are two inherent limitations with this.

One is that a possibly quite large number of participants need to complete the task. The other limitation is that the measure of expertise makes sense as a comparison between individuals, and predicts their relative performance, but does not automatically say anything about the absolute level of performance. As the results in Figure 3 show, the relationship between σ and τ is well correlated, but with different slopes and intercepts. This means we cannot equate an inferred σ value for the expertise of an individual with a predicted τ level of ac-

curacy. We can merely say which individuals are more accurate.

For this reason, our approach is best suited to real-world problems where the goal is to be able to find the most expert individuals from a large pool. If more precise statements about levels of accuracy are important the sorts of protocols we mentioned in the Introduction, measuring discriminability and consistency, seem likely to be better suited.

Extensions

Our current results are specific to rank ordering tasks, but the basic approach could be applied to other sorts of tasks for expressing knowledge and expertise. One obvious possibility is estimation tasks, in which people have to give values for quantities (Merkle & Steyvers, 2011). It should also be possible to develop suitable models for tasks, such as multiple choice questions, where the answers are discrete and nominally scaled.

Our analysis considered each problem independent of the others, which seems reasonable as a starting point. However, if there was reason to believe a domain-level expertise might exist for a set of related questions (e.g., if we had believed there was expertise for city populations, linking the US and World Cities questions), that assumption could be incorporated into the model. The basic idea would be to create a hierarchical model, with a single σ for each participant that applied to all of the relevant problems in the domain (e.g., Klementiev, Roth, Small, & Titov, 2009). Usually, when hierarchical assumptions are reasonable, they improve inference, leading to better estimates of parameters from fewer data. As such, this is an interesting possibility worth exploring, both to test the theoretical assumption of domain-level expertise, and to make the measurement of expertise more efficient in practical applications.

Conclusion

In this paper, we have developed and demonstrated a model-based approach to measuring expertise for rank ordering tasks. The approach simply requires people to complete the task on which their expertise is sought, with parameter inference then automatically providing the measure of expertise. The method was shown to work extremely well, giving expertise measures that correlated strongly with the actual accuracy of people's performance, and providing significantly better information that two self-reported measures.

Acknowledgments

References

Klementiev, A., Roth, D., Small, K., & Titov, I. (2009). Unsupervised rank aggregation with domain-specific expertise. *IJCAI*, 1101–1106.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Malhotra, V., Lee, M. D., & Khurana, A. K. (2007). Domain experts influence decision quality: Towards a robust method for their identification. *Journal of Petroleum Science and Engineering*, *57*, 181–194.

Merkle, E. C., & Steyvers, M. (2011). A psychological model for aggregating judgments of magnitude. *Lecture Notes in Computer Science*, *6589*, 236–243.

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operations Research*, *136*, 253–263.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Steyvers, M., Lee, M. D., Miller, B., & Hemmer, P. (2009). The wisdom of crowds in the recollection of order information. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1785–1793).

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Random House.

Tetlock, P. E. (2006). *Expert political judgment*. Princeton University Press.

Wright, G., & Bolderm, F. (1992). *Expertise and decision support*. Plenum Press.

Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, *52*, 79–92.

Appendix: MCMC Details

In the first Gibbs sampling step, we sample a value for each $x_{i,j}$ conditional on all other variables. Using Bayes rule and the conditional independencies in the model, this distribution can be evaluated by

$$P(x_{ij}|\mu_j,\sigma_i,y_i,x_{i\bar{j}}) \propto P(y_i \mid x_i)P(x_{i,j} \mid \mu_j,\sigma_i) \quad (1)$$

where $x_{i\bar{j}}$ refers to all samples x for individual i except the sample for the jth item. The distribution $P(x_{i,j}|\mu_j,\sigma_i)$ is Normally distributed and $P(y_i|x_i)$ is

$$P(y_i \mid x_i) = \begin{cases} 1 & \text{if } y_i = f(x_i) \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where $f(\cdot)$ is the ranking function. Taken together, the sampling distribution for x_{ij} conditional on all other variables can be evaluated by:

$$x_{ij} \mid \mu_j,\sigma_i,x_{il},x_{iu} \sim \text{TN}_{x_{il},x_{iu}}(\mu_j,\sigma_i). \quad (3)$$

The sampling distribution is the truncated Normal with the lower and upper bounds determined by x_{il} and x_{iu} respectively. The values x_{il} and x_{iu} are based on the next smallest and largest values from x_i relative to x_{ij}. Specifically, if $\pi(j)$ denotes the rank given to item j and $\pi^{-1}(j)$ denotes the item assigned to rank j, $l = \pi^{(-1)}(\pi(j)-1)$, and $u = \pi^{(-1)}(\pi(j)+1)$. We also define $x_{i,l} = -\infty$ when $\pi(j) = 1$, and $x_{i,u} = \infty$, when $\pi(j) = N$. With these bounds, it is guaranteed that the samples satisfy Equation (2) and the ordering of samples x_i is consistent with the observed ordering y_i for the ith individual.

We update the group means μ using Metropolis Hastings. We sample a new mean μ_j from a proposal distribution $Q(\mu'_j \mid \mu_j)$ and accept the new value with probability

$$\min\left(1, \frac{P(\mu'_j \mid x_{i\cdot},\sigma)}{P(\mu_j \mid x_{i\cdot},\sigma)}\frac{Q(\mu_j \mid \mu'_j)}{Q(\mu'_j \mid \mu_j)}\right). \quad (4)$$

With Bayes rule and a uniform prior on μ_j, the first ratio can be simplified to

$$\frac{P(\mu'_j|x_{\cdot j},\sigma)}{P(\mu_j|x_{\cdot j},\sigma)} = \prod_i \frac{P(x_{ij}|\mu'_j,\sigma_i)}{P(x_{ij}|\mu_j,\sigma_j)} \quad (5)$$

which has a natural expansion in terms of an exponential sum. For the proposal distribution, we use a Normal distribution with mean equal to the current mean, $Q(\mu'\_j \mid \mu_j) \sim \text{N}(\mu_j,\zeta)$, where the standard deviation ζ controls the step size of the adjustments in μ_j.

We update the standard deviations for each individual σ_i using another Metropolis Hastings step. We sample a new standard deviation σ_i from a proposal distribution $Q(\sigma'_i \mid \sigma_i)$ and accept the new value with probability

$$\min\left(1, \frac{P(\sigma'_i|x_{i,\cdot},\mu)}{P(\sigma_i|x_{i,\cdot},\mu)}\frac{Q(\sigma_i \mid \sigma'_i)}{Q(\sigma'_i \mid \sigma_i)}\right). \quad (6)$$

Using Bayes rule, the first ratio can be simplified to

$$\frac{P(\sigma'_i|x_{i\cdot},\mu)}{P(\sigma_i|x_{i\cdot},\mu)} = \frac{P(\sigma'_i|\lambda)}{P(\sigma_i|\lambda)}\prod_i \frac{P(x_{i,j}|\sigma'_i,\mu_j)}{P(x_{i,j} \mid \sigma_i,\mu_j)} \quad (7)$$

which also has an exponential sum expansion. We use a Gamma proposal distribution with a mean set to the current value of σ_i, $Q(\sigma'_i|\sigma_i) \sim \text{Gamma}(\sigma_i\nu,1/\nu)$, and a precision parameter ν.

For the MCMC sampling procedure, the proposal distribution parameters were $\zeta = 0.1$, $\nu = 20$, to approximately give an acceptance probability of 0.5. We started each chain with randomly initialized values. In a single iteration, we used Equations (3), (4), and (6) to sample new values in the vectors x, μ, and σ respectively. Each chain was continued for 500 iterations, and samples were taken after 300 iterations with an interval of 10 iterations. In total, we ran 8 chains and collected 160 samples.