TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 4 (2012) 151–163

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2011.01175.x

Inferring Expertise in Knowledge and Prediction Ranking Tasks

Michael D. Lee, Mark Steyvers, Mindy de Young, Brent Miller

Department of Cognitive Sciences, University of California

Received 15 September 2011; received in revised form 29 October 2011; accepted 6 November 2011

Abstract

We apply a cognitive modeling approach to the problem of measuring expertise on rank ordering problems. In these problems, people must order a set of items in terms of a given criterion (e.g., ordering American holidays through the calendar year). Using a cognitive model of behavior on this problem 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 completing the ordering of items, in terms of correlation with the actual accuracy of the answers. These results apply to six general knowledge tasks, like ordering American holidays, and two prediction tasks, involving sporting and television competitions. Based on these results, we discuss the potential and limitations of using cognitive models in assessing expertise.

Keywords: Expertise; Knowledge; Prediction; Ordering problem; Ranking problem; Wisdom of crowds; Model-based measurement

1. 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 problems that people can tackle using their expertise, including estimating numerical values (e.g., ''what is the length of the Nile?''), predicting categorical future outcomes (''who will win the

Correspondence should be sent to Michael D. Lee, Department of Cognitive Sciences, University of California, Irvine, CA, 92697-5100. E-mail: mdlee@uci.edu

FIFA World Cup?''), and so on. In this article, we focus on the problem of ranking a set of given items in terms of some criterion, such as ordering a set of cities from most to least populous, or predicting the final rankings of teams in a sporting competition.

One prominent theory of expertise argues that the key requirements are discriminability and consistency (Shanteau, Weiss, Thomas, & Pounds, 2002; Weiss & Shanteau, 2003). 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 livestock judgment (Phelps & Shanteau, 1978), audit judgment, personnel hiring (see Shanteau et al., 2002), medical assessment (Williams, Haslam, & Weiss, 2008), aeronautical risk perception (Pauley, O'Hare, & Wiggins, 2009), 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. Second, they must ask the same (or very similar) questions of people repeatedly, to assess consistency, 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 article, 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. This feature is especially important because it means our approach extends naturally to prediction tasks where, by definition, there exists no ground truth at the time expertise must be assessed.

The structure of this article 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 problem 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.

2. Experiment

2.1. Participants

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

2.2. Stimuli

We used six general knowledge rank ordering problems, all with 10 items, as shown in Table 1. All involve general ''book'' knowledge and were intended to be of varying levels of difficulty for our participants and lead to individual differences in expertise. We also used two prediction problems. The first involved prediction of the order of the 32 teams in the U.S. National Football League (NFL) at the end of the 2010 season. The second involved predicting the order in which the 20 contestants in the television show ''Survivor: Nicaragua'' would be eliminated.

2.3. 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 problems. 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 eight ranking problems in a random order. Within each problem, the 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

Table 1
The six general knowledge rank ordering problems. Each involves 10 items, shown in correct order

Holidays	Landmass	Amendments	US Cities	Presidents	World Cities
New Year's	Russia	Freedom of 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

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 applied.

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, where 1 was *not confident at all* and 5 was *extremely confident*.

3. 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 for ranking 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.

3.1. Overview of model

The model is described in Fig. 1, using a simple example involving three items and two individuals. Fig. 1A shows the ''latent ground truth'' representation for the three items, represented by $\boldsymbol{\mu} = (\mu_1,\mu_2,\mu_3)$ 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 1B,C shows 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 information 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_j for the jth participant, that is applied to the distribution of all items. In Fig. 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 mental representation is based on a single sample from each item distribution, represented by the crosses in Fig. 1, where x_{ij} is the sample for

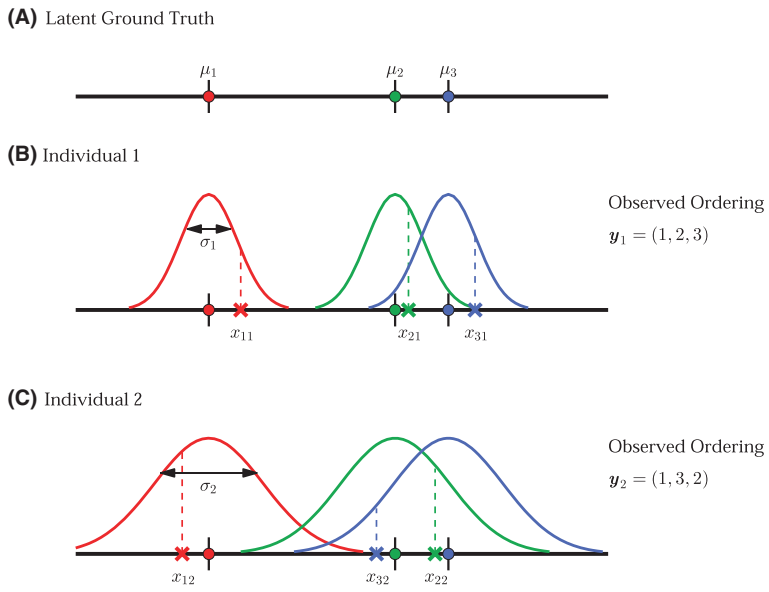

Fig. 1. Illustration of the Thurstonian model. Panel A shows the latent locations μ_1,μ_2, and μ_3 of three items to be ordered. Panels B and C show how this latent representation gives rise to mental representation and rankings for two individuals. Mental representations x_{ij} for the *j*th individual on the *i*th item are a draw from the Gaussian distribution centered on the location of the item. The ordering of these mental representation then gives the reported ranking $\mathbf{y}_j$ for the *j*th individual. Individual differences are captured by different standard deviations σ_j for the individuals.

the *i*th item and *j*th individual. The ordering produced by each individual is then based on an ordering of the mental samples. For example, Individual 1 in Fig. 1B draws sample for items that leads to the ordering (1,2,3), whereas Individual 2 in Fig. 1C 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 $\boldsymbol{\mu}$ and σ_j. In terms of the original wisdom of the crowd motivation, the most important parameter was $\boldsymbol{\mu}$, 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_j 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.

3.2. 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,

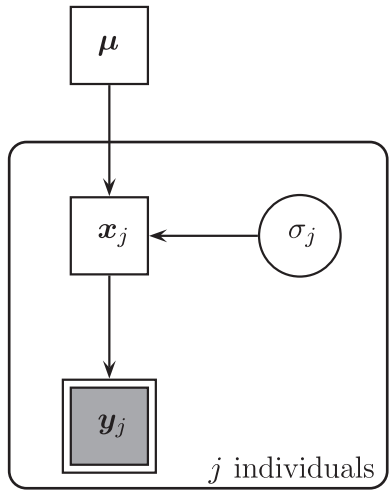

Fig. 2. Graphical model representation of the Thurstonian model, showing how latent item locations $\boldsymbol{\mu}$ combine with individual expertise σ_j to generate mental representations $\mathbf{x}_j$, on which the observed ranking data $\mathbf{y}_j$ are assumed to be based.

Kim, & Wagenmakers, 2008, for statistical and psychological introductions). 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 plate represents independent replications of the graph structure, which corresponds to individual participants in this model.

The observed data are the ordering given by the *j*th individual, denoted by the vector $\boldsymbol{y}_j$, where y_{ij} represents the item placed in the *i*th position by the individual. 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 *j*th individual samples for the *i*th item, as $x_{ij} \sim \text{Gaussian}(\mu_i,\sigma_j)$, where σ_j is the uncertainty that the *j*th 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 $\boldsymbol{y}_j = \text{Rank}(\boldsymbol{x}_j)$.

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 hyper-parameter 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_j, and μ, given the

observed orderings $\boldsymbol{y}_j$. We use Gibbs sampling to update the mental samples x_{ij}, and Metropolis-Hastings updates for σ_j and μ. Details of the MCMC inference procedure are provided in the Appendix.

4. Results

We first describe how we measure the accuracy of a rank order provided by a participant, as a ground truth assessment of his or her expertise. We then examine the correlations between this ground truth and their pre- and post-reported self-assessments, and the model-based measure.

4.1. 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 is the number of items for the problem. 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. For the 10-item general knowledge questions, this maximum is 45. The maximum τ is 496 for the NFL prediction question, and 190 for the Survivor prediction question.

4.2. 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 posterior for 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 the first six columns, the two prediction problems highlighted in the last two columns, and the three measures as rows throughout. The Pearson correlations are also shown. Note that, for the self-reported measures, the goal is for higher levels of rated expertise to correspond to lower (more accurate) values of τ, and so a negative correlation means 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.

We consider first the results for the six general knowledge problems. Fig. 3 shows that they ranged in difficulty. Looking at the maximum τ needed to show results, the Holidays, Amendments, U.S. 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 important, 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

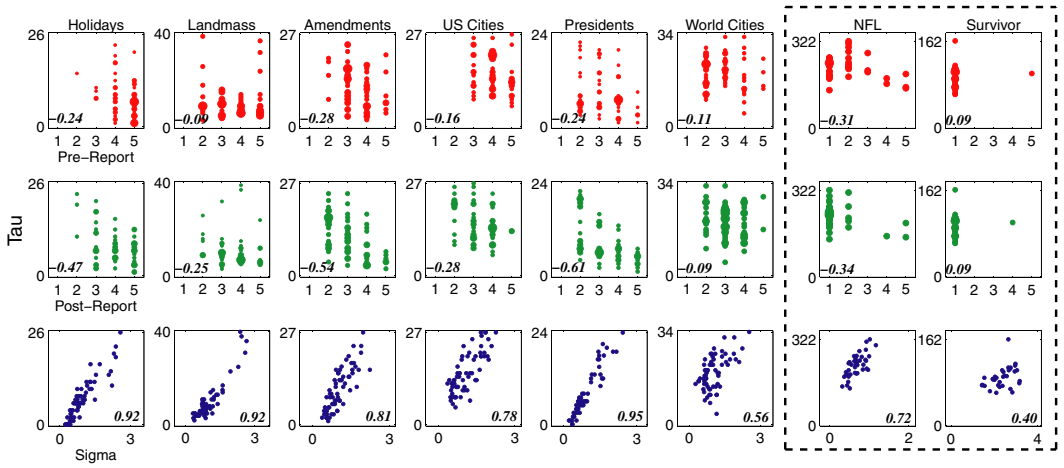

Fig. 3. Results comparing the relationship between the three measures of expertise and the accuracy of individual answers. The plots are organized with the measures in rows and the problems in columns.

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.

Perhaps most important, 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]

In terms of the prediction results, the NFL problem shows a similar pattern of results. There is a weak correlation, in the right direction, for both of the self-reported measures and the final ground truth, and this correlation is significantly improved by the model-based measure. The Survivor problem shows a slightly different pattern of results. In both the pre- and post-report measures, all but one of the participants gave the lowest possible self-rating of their expertise. This lack of variation makes correlation with the final ground truth impossible. The model-based measure manages a correlation of 0.40, which is its poorest performance over all of the eight problems, but impressive in the context of the lack of ability people apparently have to report their relative expertise.

5. 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.

5.1. 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 problem, 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 problem. 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 can be applied to prediction tasks, where there is not (yet) a ground truth. While our results only include two such problems—one involving sporting prediction, and the other involving a television competition—the results for both are encouraging. These findings are especially intruiging, because standard measures of expertise based on self-report have often been found to be unreliable predictors of forecasting accuracy (e.g., Tetlock, 2006).

Most important, the potential for real-world application to prediction problems is clear. While general knowledge can often be uncovered by means other than human judgment, prediction often fundamentally relies upon the projections of experts in business, government, sport, military, and other settings. The predictions our participants made about the performance of NFL teams are the type of predictions that need to be made in the context of sports betting, for example, and being able to identify expertise in making those predictions is an important real-world problem.

5.2. 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 needs to complete the task. How many people are required for our results to hold is an interesting question for future research. The other limitation is that the measure of expertise makes sense as a comparison between individuals and predicts their relative performance, but it does not automatically say anything about the absolute level of performance. As the results in Fig. 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 accuracy. 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 Section 1, measuring discriminability and consistency, seem likely to be better suited.

Another basic limitation of our approach is that it relies on assuming there is one underlying truth, and people have knowledge of this truth that, while inaccurate, is not systematically distorted. If the knowledge that most people use to provide rankings is fundamentally wrong, or if there are multiple different justified answers, it is unlikely our approach will be effective. Systematic error could arise in practice if there is a widely held erroneous belief. Multiple truths could arise in practice if, for example, different cultures have different justifiable beliefs, as in Cultural Consensus Theory (Romney, Batchelder, & Weller, 1987). We think both of these issues could potentially be addressed with more complicated cognitive models than the one assumed in Fig. 2, using hierarchical models to capture systematic distortions, and mixture models to accommodate multiple truths (Lee, 2011). But these extensions remain a challenge for future work.

5.3. 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 as 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 problems (e.g., if we had believed there was expertise for city populations, linking the U.S. 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.

6. Conclusion

In this article, we have developed and demonstrated a model-based approach to measuring expertise for rank ordering problems. The approach simply requires people to complete the problem on which their expertise is sought, with parameter inference then automatically providing the measure of expertise. The method was shown to work extremely well, on both general knowledge and prediction problems. It allowed the inference of expertise measures

that correlated strongly with the actual accuracy of people's performance, and providing significantly better information that two self-reported measures.

Notes

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.
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 discrete. To check this, we also calculated correlations for the Thurstonian model using five binned values of σ and found correlations of 0.88, 0.88, 0.80, 0.77, 0.92, 0.54, 0.67, and 0.42 for the eight problems, in the order shown left to right in Fig. 3. These correlations are only slightly different from those shown, and they support the same conclusions.

Acknowledgments

M.d.Y. acknowledges the support of UROP and SURP funding from the University of California, Irvine. M.D.L. and M.S. acknowledge support from the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20059. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

This article is based on a paper presented at the 2011 Annual Conference of the Cognitive Science Society. That paper was awarded the Computational Modeling Prize for best Applied Cognition paper at the Conference.

References

Klementiev, A., Roth, D., Small, K., & Titov, I. (2009). Unsupervised rank aggregation with domain-specific expertise. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence* (pp. 1101–1106). Menlo Park, CA: AAAI Press.

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* (pp. 13–55). 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.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

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.

Pauley, K., O'Hare, D., & Wiggins, M. (2009). Measuring expertise in weather-related aeronautical risk perception: The validity of the Cochran-Weiss-Shanteau (CWS) index. *International Journal of Aviation Psychology*, *19*, 201–216.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance*, *21*, 209–219.

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *Americal Behavioral Scientist*, *31*, 163–177.

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*, *2*, 1785–1793.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

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

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors*, *45*, 104–116.

Williams, C. A., Haslam, R. A., & Weiss, D. J. (2008). The Cochran-Weiss-Shanteau performance index as an indicator of upper limb risk assessment expertise. *Ergonomics*, *51*, 1219–1237.

Wright, G., & Bolderm, F. (1992). *Expertise and decision support*. New York: 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_{ij} conditional on all other variables. Using Bayes rule and the conditional independencies in the model, this distribution can be evaluated by

$$p(x_{ij} \mid \mu_i, \sigma_j, \boldsymbol{y}_j, \boldsymbol{x}_{-ij}) \propto p(\boldsymbol{y}_j \mid \boldsymbol{x}_j) p(x_{ij} \mid \mu_i, \sigma_j), \tag{1}$$

where $\boldsymbol{x}_{-ij}$ refers to all samples $\boldsymbol{x}$ for individual j except the sample for the ith item. The distribution $p(x_{ij} \mid \mu_i, \sigma_j)$ has a Gaussian distribution and $p(\boldsymbol{y}_j \mid \boldsymbol{x}_j)$ is

$$p(\boldsymbol{y}_j \mid \boldsymbol{x}_j) = \begin{cases} 1 & \text{if } \boldsymbol{y}_j = \text{Rank}(\boldsymbol{x}_j) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Taken together, the sampling distribution for x_{ij} conditional on all other variables can be evaluated by

$$x_{ij} \mid \mu_i, \sigma_j, x_{lj}, x_{uj} \sim \text{TruncGauss}_{x_{lj}, x_{uj}}(\mu_i, \sigma_j). \tag{3}$$

The sampling distribution is the Truncated Gaussian with the lower and upper bounds determined by x_{lj} and x_{uj}, respectively. The values x_{lj} and x_{uj} are based on the next smallest and largest values from $\boldsymbol{x}_j$ relative to x_{ij}. Specifically, if $\pi(i)$ denotes the rank given to item i and $\pi^{-1}(i)$ denotes the item assigned to rank i, $l = \pi^{-1}(\pi(i) - 1)$, and $u = \pi^{-1}(\pi(i) + 1)$. We also define $x_{lj} = -\infty$ when $\pi(i) = 1$, and $x_{uj} = \infty$, when $\pi(i) = N$. With these bounds, it is guaranteed that the samples satisfy Eq. 2 and the ordering of samples $\boldsymbol{x}_j$ is consistent with the observed ordering $\boldsymbol{y}_j$ for the jth individual.

We update the group means $\boldsymbol{\mu}$ using a Metropolis Hastings step. We sample a new mean μ_i from a proposal distribution $Q(\mu_i' \mid \mu_i)$ and accept the new value with probability

$$\min\left(1, \frac{p(\mu_i'|\boldsymbol{x}_{i\cdot}, \boldsymbol{\sigma})}{p(\mu_i|\boldsymbol{x}_{i\cdot}, \boldsymbol{\sigma})} \frac{Q(\mu_i|\mu_i')}{Q(\mu_i'|\mu_i)}\right). \tag{4}$$

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

$$\begin{aligned} \frac{p(\mu_i'|\boldsymbol{x}_{i\cdot}, \boldsymbol{\sigma})}{p(\mu_i|\boldsymbol{x}_{i\cdot}, \boldsymbol{\sigma})} &= \prod_j \frac{p(x_{ij}|\mu_i', \sigma_j)}{p(x_{ij}|\mu_i, \sigma_j)} \\ &= \exp\left(-\frac{1}{2}\sum_j \frac{(x_{ij}-\mu_i')^2 - (x_{ij}-\mu_i)^2}{\sigma_j^2}\right). \end{aligned} \tag{5}$$

For the proposal distribution, we use a Gaussian distribution with mean equal to the current mean, $Q(\mu_i'|\mu_i) \sim \text{Gaussian}(\mu_i, \zeta)$, where the standard deviation ζ controls the step size of the adjustments in μ_i.

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

$$\min\left(1, \frac{p(\sigma_j'|\boldsymbol{x}_{\cdot j}, \boldsymbol{\mu})}{p(\sigma_j|\boldsymbol{x}_{\cdot j}, \boldsymbol{\mu})} \frac{Q(\sigma_j|\sigma_j')}{Q(\sigma_j'|\sigma_j)}\right). \tag{6}$$

Using Bayes rule, the first ratio can be simplified to

$$\begin{aligned} \frac{p\left(\sigma_j'|\boldsymbol{x}_{\cdot j}, \boldsymbol{\mu}\right)}{p\left(\sigma_j|\boldsymbol{x}_{\cdot j}, \boldsymbol{\mu}\right)} &= \frac{p\left(\sigma_j'|\lambda\right)}{p\left(\sigma_j|\lambda\right)} \prod_i \frac{p\left(x_{ij}|\sigma_j', \mu_i\right)}{p\left(x_{ij}|\sigma_j, \mu_i\right)} \\ &= \frac{p\left(\sigma_j'|\lambda\right)}{p\left(\sigma_j|\lambda\right)} \exp\left(-\frac{1}{2}\frac{\sigma_j^2 - \sigma_j'^2}{\sigma_j^2 \sigma_j'^2}\sum_i \left(x_{ij}-\mu_i\right)^2\right). \end{aligned} \tag{7}$$

We use a Gamma proposal distribution with a mean set to the current value of $\sigma_j, Q(\sigma_j'|\sigma_j) \sim \text{Gamma}(\sigma_j v, 1/v)$, and a precision parameter v.

For the MCMC sampling procedure, the proposal distribution parameters were $\zeta = 0.1, v = 20$, to give approximately an acceptance probability of 0.5. We started each chain with randomly initialized values. In a single iteration, we used Eqs. (3), (4), and (6) to sample new values in the vectors $\boldsymbol{x}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$, 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 eight chains and collected 160 samples.