Adapting cultural mixture modeling for continuous measures of knowledge and memory fluency

Yin-Yin Sarah Tan¹ · Shane T. Mueller¹

Published online: 22 October 2015
© Psychonomic Society, Inc. 2015

Abstract Previous research (e.g., cultural consensus theory (Romney, Weller, & Batchelder, *American Anthropologist*, 88, 313–338, 1986); cultural mixture modeling (Mueller & Veinott, 2008)) has used overt response patterns (i.e., responses to questionnaires and surveys) to identify whether a group shares a single coherent attitude or belief set. Yet many domains in social science have focused on implicit attitudes that are not apparent in overt responses but still may be detected via response time patterns. We propose a method for modeling response times as a mixture of Gaussians, adapting the strong-consensus model of cultural mixture modeling to model this implicit measure of knowledge strength. We report the results of two behavioral experiments and one simulation experiment that establish the usefulness of the approach, as well as some of the boundary conditions under which distinct groups of shared agreement might be recovered, even when the group identity is not known. The results reveal that the ability to recover and identify shared-belief groups depends on (1) the level of noise in the measurement, (2) the differential signals for strong versus weak attitudes, and (3) the similarity between group attitudes. Consequently, the method shows promise for identifying latent groups among a population whose overt attitudes do not differ, but whose implicit or covert attitudes or knowledge may differ.

Keywords Cultural consensus theory · Implicit memory · Finite mixture modeling

✉ Shane T. Mueller
shanem@mtu.edu

¹ Department of Cognitive and Learning Sciences, Michigan Technological University, Houghton, MI, USA

Introduction

As information systems enable distributed communities to be developed and maintained on the basis of shared affiliations, attitudes, and beliefs, distinct cultures are created that are defined by and supported by these beliefs and the communication practices that enable these communities. Although individual communities and subcultures likely have always been defined by and created by shared belief, it is also true that communities having close interaction and a shared identity can produce shared knowledge, attitudes, and beliefs. In either case, an important aspect of this culture is the shared knowledge held by the community (see Atran, Medin, & Ross, 2005; Mueller & Veinott, 2008; Sieck, Smith, & McHugh, 2007). This perspective on culture combines three different approaches, including structural/pattern, procedural, and functional methods (Cohen, 2010; Rohner, 1984), in which culture is seen as a system of shared knowledge and mental models (Fischer, 2012).

If cultural knowledge is shared knowledge, then a useful question to ask is whether some piece of knowledge, attitude, or belief is widely shared among the members of a group. If the members of a group tend to share many beliefs, then a culture of belief exists within the group, and one can justifiably characterize the culture of that group as the knowledge they share. However, sharing one piece of knowledge, attitude, or belief may not constitute a shared culture. After all, there are an infinite number of facts that everybody agrees with ($1 + 1 = 2$, Earth is the third planet from the sun, etc.). What is perhaps more compelling is finding that a set of related attitudes or beliefs could differ, but nonetheless are shared within a group. If this constellation of knowledge is consistently shared, one can reasonably argue that the group in question is a culture of shared belief. For example, a religious sect may tend to share a set of beliefs about prohibited food, and these may differ from another group's beliefs. If this were the

case, we might be able to identify group membership by the shared beliefs alone. In contrast, in some situations there may be no consistent patterns in the nature of the beliefs (e.g., the best flavor of ice cream), or the population in question may contain a mixture of several distinct beliefs (e.g., political affiliation of a group of politicians).

Cultural consensus theory (CCT; Romney, Weller, & Batchelder, 1986) was designed to assess consensus of belief within a group. That is, it can be used to identify the culturally correct set of answers, “without the answer key.” However, this theory is unable to distinguish between the different ways a consensus might fail to emerge. Mueller and Veinott (2008) described an advance of the technique (called *cultural mixture modeling*; CMM) that used an alternative statistical approach (finite mixture modeling) to identify subgroups of shared knowledge within a larger population. The multiculture General Condorcet model (Anders & Batchelder, 2012) has recently updated CCT to accommodate mixtures of belief, and these approaches have been successful at advancing both methodology and theoretical definitions of culture.

However, these methods typically use categorical or (at best) discrete ordinal responses—typically, questionnaires, surveys, or interviews. That is, they typically rely on shared knowledge that is elicited explicitly, using yes–no or agree–disagree questions, perhaps on a Likert-type scale. Researchers have become increasingly interested in examining implicit measures of knowledge and attitude. The most prominent area of this research stems from so-called social-priming effects (especially the implicit association task; Greenwald, McGhee, & Schwartz, 1998; Greenwald, Poehlman, Uhlmann, & Banaji, 2009), which was originally referred to as the “Implicit Attitude Test.” Nevertheless, many cognitive tasks have used response times (RTs) as an index of association strength, including the lexical decision task (LDT; Meyer & Schvaneveldt, 1971), the priming LDT (Meyer & Schvaneveldt, 1971; Meyer, Schvaneveldt, & Ruddy, 1975; Schvaneveldt & Meyer, 1973), and the free association task (Shimamura & Squire, 1984; Weldon & Coyote, 1996). Although RT measures are often associated with implicit knowledge, our proposed method makes no specific claims about the implicit nature of the knowledge. For example, RT differences have been used to assess privileged knowledge in memory strength and lie detection contexts (e.g., Seymour, Seifert, Shafto, & Mosmann, 2000; Verschuere, Kleinberg, & Theodoridou, 2015), and these may not be considered implicit tests of knowledge. As such, we consider our goal to be identifying aspects of knowledge fluency that are not revealed using overt measures of knowledge elicitation, and implicit knowledge may be one of those aspects.

Such an approach may have applications, as well. For example, cultural researchers have discussed the notion of an “unpopular norm” (see Bicchieri, 2006). An unpopular norm may be an attitude in which most members of a culture knows

the culturally “correct” answer, and report that answer even if they themselves do not share the belief. This can lead to some norms becoming very resistant to change, even if most members of a culture do not support the norm, because it appears to receive strong support. It may be that implicit measures could identify lack of support, even if the explicit response were in favor of the unpopular norm. Similarly, in many contexts timing information may be available that could provide insight into sentiment or support of a notion.

In this article, we investigate the feasibility of identifying shared knowledge or attitudes using implicit measures based on response latency. After describing the basic modeling approach, we will show the results of a validation experiment in which we trained participants differentially on a common set of information. On the basis of the success (and failure) of the model to accurately discriminate groups of these data, we then report the results of a simulation study that identified some basic statistics for assessing whether shared subgroups of implicit knowledge can be recovered. Then we describe the results of a second experiment, designed on the basis of these simulations, that shows the full capabilities of our mixture modeling approach to implicit knowledge measures.

Cultural mixture modeling for continuous and implicit measures

Previous approaches to measuring shared knowledge have focused on agreement and consensus. In our previous work introducing and applying CMM (Mueller & Veinott, 2008; Sieck & Mueller, 2009; Simpkins, Sieck, Smart, & Mueller, 2009), we have typically used the “strong-consensus” model to define a community of shared belief. The strong-consensus model works by assuming that a small, fixed proportion (α) of respondents will disagree with the normative answer on any given issue. The use of this model is justified to the extent that attitudes tend to live on the extremes of a scale (an assumption that is often challenged in opinion dynamics research). In other versions of the model, we have used Gaussian mixtures to model distributions of Likert-scale responses (or, alternatively, have forced the Likert-scale responses into two categories). Neither of these approaches can be used directly for a continuous-scale dependent measure such as RTs.

The strong-consensus model of CMM requires a binary measure of agreement on a set of questions or items. The approach assumes that a fixed number of groups exist, and uses response patterns to identify the best clustering of those groups. The results of different numbers of groups are compared using the Bayesian information criterion (BIC) measure (Schwarz, 1978), and then the best model is selected. The BIC is fairly conservative and enables a trade-off between model complexity and likelihood. It is important to recognize that this process will only work well if enough items are measured to discriminate the groups.

The consensus model using RT

RTs have been used as a primary or secondary dependent measure in thousands of experimental paradigms. Modern off-the-shelf computer equipment can be measured reliably to around 1/100th of a second (see Mueller & Piper, 2014). Longer RTs can be associated with base rate in a choice decision (Weidemann & Mueller, 2008), as well as with weaker memory associations (e.g., Nelson & Shiffrin, 2013). Thus, the RT to access memory may be useful as an implicit measure of association strength.

For RTs to be useful in CMM analysis, we must be able to measure them across a set of materials. The procedure for measuring RTs must be designed so that it is sensitive to memory access retrieval or fluency (and is minimally influenced by such factors as reading time). We have developed a set of procedures for applying mixture modeling to RT patterns, and these are found in the [Appendix](#).

The model follows the approach of CMM, but it introduces a new consensus model that allows analogous consensus to be identified with RT data. Although the detail of this approach is found in the Appendix, the basic notion is that the mean RTs for particular items are modeled (after a logarithmic transformation and rescaling) as a mixture of independent Gaussians having a small fixed variance. The usefulness of this approach will be demonstrated in three experiments, which we turn to next.

Experiment 1: geographic pair-learning task

As a laboratory analogue to examining shared knowledge that differs in strength, we developed a paired-associates learning task. The goal was to develop a task using real-world knowledge, but such that differences in the verbal aspects of the knowledge varied minimally across materials, so as to reduce differences in reading times between the stimuli. Consequently, we presented geographic information (names and border outlines of countries) and asked participants to learn whether pairs were adjacent or nonadjacent. It should be recognized that although this task involves actual knowledge, it is still a somewhat artificial task and may be difficult to translate directly to an assessment of real knowledge in the domains of cognitive anthropology that people have used in the past. Regardless, it allowed us to identify conditions under which we could assess the ability of the model to recover groups we knew existed, and so was an important first step in developing the method.

Method

Participants We recruited 62 undergraduate students from Michigan Technological University (MTU) to take part in

the study for course credit. We initially collected around 30 participants who were randomly assigned to either Group 1 or Group 2. We then expanded the study with two additional training groups (3 and 4), so group membership was not completely randomly assigned.

Design We used a set of 32 country pairs to train participants with specific geographic knowledge (see Table 1). Each pair was either spatially adjacent or nonadjacent on the world map and was located within Eurasia or Africa. Four training groups were assigned, and each group was presented with the entire set of 32 country pairs. In each session, 8 pairs were presented once, 8 twice, 8 four times, and 8 eight times, for a total of 120 pairs per training session. However, each group received a different mapping of frequency to country pairs. Participants each completed four training sessions and then attended one test session within a couple of days. In the test session, each pair was shown four times (i.e., 128 trials in total), without feedback.

Procedure The first session began with reading and signing the informed consent document (approved by the MTU institutional review board). The participants were then trained on the geographic pair-learning task, implemented via the Psychology Experiment Building Language Version 0.13 (Mueller & Piper, 2014). At the beginning of the study, instructions were provided, and the 120 trials were performed. On each trial, a pair of countries was displayed on the right and left parts of the screen simultaneously (see Fig. 1). Participants had to make a decision as to whether the given pair of countries was adjacent geographically, and their RTs and accuracies were collected. After the response was made, feedback was given by showing the two countries' locations on a larger map, as well visual and audio feedback indicating whether the response was correct or incorrect. After completing four training sessions, participants later completed a test session. The test session was similar to the training, except that pairs were presented equal numbers of times and no visual or auditory feedback was given. Each training or testing session typically took around 30 min, and testing was typically spread across 2–3 days, often with back-to-back sessions on one day.

Results

Data preprocessing and the impact of frequency To remove trials on which the RT was unlikely to indicate true familiarity, we removed the 65 trials (out of more than 36,000) whose RTs were less than 400 ms (and whose mean accuracy was 47%) and the 171 trials whose RTs were greater than 10,000 ms. In addition, two participants were excluded from further data analysis due to low accuracy (<80%). Although these participants showed frequency-based effects on time, we wanted to

Table 1 Country pairs used in both experiments

Adjacent-Eurasia	Adjacent-Africa	Nonadjacent: Within	Nonadjacent: Cross
Albania–Macedonia	Angola–Zambia	Iraq–Lebanon	Algeria–Cyprus
Belarus–Poland	Djibouti–Eritrea	Liechtenstein–Montenegro	Benin–Turkmenistan
Bulgaria–Serbia	Ethiopia–Somalia	Kuwait–Oman	Lesotho–Armenia
Estonia–Latvia	Gabon–Cameroon	Lithuania–Afghanistan	Libya–Israel
Hungary–Ukraine	Mali–Guinea	Syria–Yemen	Niger–Turkey
Iran–Azerbaijan	Mozambique–Swaziland	Liberia–Botswana	Tanzania–Slovakia
Romania–Moldova	Nigeria–Chad	Namibia–Ghana	Togo–Georgia
Slovenia–Austria	Senegal–Mauritania	Sudan–Kenya	Uganda–Jordan

demonstrate that the method would work even for participants who had learned the information to a relatively high criterion.

Multiple studies have shown training frequency to be associated with faster RTs and higher accuracies (Becker, 1979; Lupker & Pexman, 2010; Neely, 1991; Nelson & Shiffrin, 2013; Scarborough, Cortese, & Scarborough, 1977). Our results were aligned with this frequency trends for both mean RTs and accuracies at the time of testing. The accuracies ranged from 87% to 99% for the infrequent versus frequent training pairs (Fig. 2). As training progressed, RTs became faster and accuracy higher, but the impact of frequency was maintained along both dimensions. An analysis of variance (ANOVA; treating participant code as a randomized factor) showed that both the mean RT and mean accuracy depended significantly on the training condition during the test session [$\log(\text{RT})$: $F(3, 177) = 80, p < .0001$; accuracy: $F(3, 177) = 32, p < .0001$].

Mixture modeling It is important to recognize that a consensus-based mixture model similar to the ones proposed by Mueller and Veinott (2008) would be unable to distinguish these small differences in accuracy, because they require a consensus near 1.0 and 0.0, and accuracies

for the least-frequent items were above .8. Consequently, we completed the rescaling procedure described in the [Appendix](#) and fit mixtures of Gaussians to predict the mean scaled RTs.

The classification results of CMM showed that, without using participants' training conditions as a predictor, Group 1 and Group 2 (totaling 30 participants) were identified perfectly (i.e., the model identified two groups that were identical to the training groups). However, when two additional groups were incorporated, none of the other combinations could be correctly identified by the model (results shown in Fig. 3), with the accuracy of the model declining to 60%–70% over different mixtures.

Discussion

The implicit mixture model perfectly identified two initial groups that were designed to be fairly distinct. However, when other groups were examined in this study, the model could not produce effective, satisfactory results. One of the reasons for this can be seen in Table 2, which shows the correlations between the designs (frequency of presentation) of each pair of training groups. This shows that the only pair that could be

Fig. 1 Screenshot of a training trial (left), with correct and incorrect feedback maps. Testing trials were identical to the training trials, but no feedback was given.

Fig. 2 Speed-accuracy results across the four training sessions and the test session in Experiment 1. Error bars indicate ± 1 standard deviations of the participants' means

distinguished was highly negatively correlated ($-.704$). In contrast, groups that were positively correlated (G1–G3, G3–G4) were not distinguishable, as were pairs that were close to independent (G1–G4, G2–G4) or those with a moderately negative correlation (G2–G3).

To investigate whether the failures of the model stemmed from the correlation between training frequencies, we performed a simulation experiment in which we created simulated data from a simple learning model and assessed whether the model could recover the original training groups under different levels of noise.

Experiment 2: simulation study

Experiment 1 indicated that groups of shared knowledge could be identified on the basis of implicit measures of knowledge, but only when the strengths of that knowledge were highly negatively correlated. In Experiment 2, our aim was to determine whether this finding is a fundamental limitation of the method, or if the training could have been recovered if the RT measures were made with less noise. Consequently, in this experiment we tested the same models and study conditions as in Experiment 1, except that here we used a simple stochastic model to simulate RTs based on learning.

Method

We generated experiments of the same size and design as Experiment 1 (15 participants per simulation group, 32 stimulus pairs), with individual RTs (measured in seconds) taking the following form:

$$RT = \alpha + \frac{\beta}{F} + \varepsilon \quad (1)$$

In this formula, the nondecision time $\alpha = .3$, a scaling factor $\beta = .3$, F represents the country pair frequency ratios (as 1, 2, 4, or 8), and ε indicates the RT noise distribution. The values

of α and β were chosen to produce mean RTs across conditions that roughly corresponded to the observed data, but no systematic attempt to fit the model to the data was made, as that was not the purpose of this experiment. We parameterized the simulation of noise level ε with a one-parameter log-normal distribution, with Z indicating a standard normal distribution, and σ indicating a noise level:

$$\varepsilon \approx \exp(\sigma Z), \quad (2)$$

Each of the four training groups in Experiment 1 was simulated to create a complete experiment. The same data analysis processes were completed on the simulated data, and the implicit CMM model was used to identify distinct subgroups under a variety of conditions. This simulation was performed 200 times for $\sigma = 0.02, 0.1$, and 0.2 .

Results

The best-fitting results for any single model run might or might not have the same number of groups as the original. Thus, to provide a basic accuracy measure, we needed to develop a measure that was fair to solutions that were partially correct. For example, for a mixture of two groups, three groups might be obtained, but if all of the members of one group were placed in one cluster, and the members of the second group were divided into two different clusters, this would be a better solution than one in which the three obtained groups were each equally divided between the true groups. To evaluate each simulation, we examined each true group and found the obtained group that had the most members classified together, and divided by the number of members in the true group. This was done for each group in the mixture and for both the true groups and the model-produced groups, and these values were averaged to produce an accuracy score. The lower bound of these scores is theoretically close to $1/(\text{number of groups})$.

Fig. 3 Classification results from Experiment 1. Each panel shows the maximum-likelihood model obtained via expectation–maximization optimization using two groups, along with the ideal classification results. Only the mixture of Groups 1 and 2 was accurately recovered in this experiment

The simulation results show (Fig. 4) that CMM can successfully identify all mixtures when the noise level is as low as $\sigma = 0.02$. When the noise rose to $\sigma = 0.1$, highly negatively correlated groups (e.g., Groups 1 and 2) could typically be identified correctly, but other groups were not recovered as accurately as more groups were mixed together. This corresponds roughly to the results of our Experiment 1. As noise grew larger, to $\sigma = 0.2$, none of the mixtures were recovered satisfactorily.

Overall, this simulation experiment revealed that our proposed method might be able to distinguish groups—even those whose training designs were moderately positively correlated—if the noise level was low enough relative to their true difference in training. Two ways in which this difference could increase would be to use more trials during testing, to obtain a more reliable measure of the mean RT per participant, and to increase the disparity between frequent and rare stimuli, to attempt to increase the mean RT differences between rare

Table 2 Correlations among the frequency of presentation of individual pairs across the four training groups

	G1	G2	G3	G4
G1	1.000	−.704	.391	.078
G2	−.704	1.000	−.217	.096
G3	.391	−.217	1.000	.270
G4	.078	.096	.270	1.000

Only Groups 1 and 2, whose designs were highly negatively correlated, were discriminated via the model.

and frequent stimuli. In Experiment 3, we looked at both of these options simultaneously.

Experiment 3: modified geographic pair-learning task

The simulation in Experiment 2 confirmed that the statistical methods we employed can discriminate groups of shared knowledge using implicit knowledge measures. Furthermore, the results of the simulation suggested that the failure of the method for Experiment 1 likely stemmed from large noise relative to the difference in memory strength induced by our training approach. Consequently, in Experiment 3 we modified the method used in Experiment 1 so as to both reduce noise and increase the differential memory strength.

Method

Participants A total of 48 students were recruited from the psychological participant pool in MTU. Students were randomly assigned to the same training as Groups 1, 2, and 4 of Experiment 1 (each with around 15 participants).

Procedure The task that we presented to participants was identical to that in Experiment 1, except that the training frequencies of paired countries were increased from 1:2:4:8 to 1:3:9:27 and that more test trials were added (from four to six trials per pair) as a means to obtain a stable RT measure. Because of the different numbers of trials involved in a single counterbalancing (40 vs. 15), the training was collapsed into two 320-trial blocks and one 192-trial test block.

Results

As in Experiment 1, we removed trials that were either faster than 400 ms (235 trials of 42,827; accuracy .48) or slower than 10,000 ms (915 trials). With the same criteria as in Experiment 1, four out of 48 participants were eliminated due to their low accuracy after training ($<.8$). Similar to Experiment 1, the results exhibited effects of training frequency on both accuracy and RTs (see Fig. 5). Within the testing block, in which feedback was not given, RTs were initially slower than at the end of the training block, but they decreased to be faster than those in the training block (mean RTs = 1,860, 1,412, and 1,256 ms, respectively, for consecutive 64-trial blocks, in comparison to 1,378 ms for the last 64 trials of training). Without feedback, accuracy dipped slightly from the end of training, but it did not change during the test session (for consecutive 64-trial blocks, accuracy was .894, .896, and .893, in comparison to .95 for the last block of training). Repeated measures ANOVAs on both log(RT) and accuracy during the test session, treating participant as a randomized factor, found that both DVs depended reliably on frequency [log(RT): $F(3, 147) = 106, p < .001$; accuracy: $F(3, 147) = 100.2, p < .001$].

Mixture modeling on the transformed RT means We used the same transforms and mixture models as in Experiment 1, comparing models of size 1 through 5, with 200 random

Fig. 4 Cultural mixture modeling classification results of the data simulation. Each group combination was simulated under three different noise levels ($\sigma = 0.02, 0.1$, and 0.2)

Fig. 5 Speed–accuracy results for the two training sessions and the one test session of Experiment 3. Error bars indicate ± 1 standard deviations of the participant means

starting configurations of each. The results appear as the first model in Table 3. Here, the BIC statistic identified three

groups as the best model, and all 44 participants were correctly classified into their training groups. Separate models (not

Table 3 Clustering results of the proposed model (Model 1) and of several alternative clustering approaches

Model	Log-Likelihood	BIC Value	# of Groups	Cramér's V	Cluster ID	SD Estimate	Training Group 1	Training Group 2	Training Group 4
1. All trials, <i>SD</i> fixed, BIC solution	−304.7	980	3	1.000	1	0.1	15	0	0
					2	0.1	0	15	0
					3	0.1	0	0	14
2. All trials, <i>SD</i> estimated, BIC criterion	−194.8	900.5	4	.904	1	0.041	9	0	0
					2	0.066	0	14	0
					3	0.052	1	0	11
					4	0.212	4	1	4
3. Correct trials, <i>SD</i> fixed, BIC solution, <i>NA</i> = 1.0	−546	1,588	4	.663	1	0.1	9	0	0
					2	0.1	0	5	0
					3	0.1	0	0	4
					4	0.1	5	10	11
4. Correct trials, <i>SD</i> fixed, BIC solution, <i>NA</i> = 2.0	−649.9	1,795.6	4	.919	1	0.1	13	0	0
					2	0.1	0	13	0
					3	0.1	0	0	8
					4	0.1	1	2	7
5. Scaled drift rate, BIC solution, <i>SD</i> estimated	−1,461	3,303	3	.732	1	0.587	11	0	8
					2	0.501	0	15	0
					3	0.202	3	0	7
6. Drift rate, BIC solution, <i>SD</i> fixed	−1,487	3,345	3	.901	1	0.5	13	0	1
					2	0.5	0	14	0
					3	0.5	1	1	14
7. Accuracy, binomial model*	−179.4	479.9	1	NA	1	*	14	15	15
8. Accuracy, strong-consensus model	−234.1	589.4	1	NA	1	0.05	14	15	15

In each case, the model was selected among alternatives on the basis of having the smallest BIC value. Cramér's (1946) *V* is a categorical correspondence/correlation index based on the chi-squared statistic. *Binomial model estimates accuracy for each question, and the *SD* is calculated according to binomial formula.

Fig. 6 Comparison between training frequency and the estimated parameters across the three training groups. Best-fitting regression lines show fairly similar parameter estimates

shown) comparing mixtures of two groups at a time produced similar classification results. This confirmed the simulation results of Experiment 2 in a behavioral study, and shows that the methods described here are able to identify groups of shared knowledge using implicit measure of knowledge strength.

To examine the results of this experiment more closely, recall that each group in the model produces a parameter estimate for the mean scaled RT of each country pair. We expected the recovered parameters to be negatively correlated with the frequency of training, and Fig. 6 shows that these estimates were highly negatively correlated with the training frequency (as would be expected if more training leads to shorter RTs). A more detailed analysis of each group, predicting the learned parameter on the basis of $\ln(\text{training frequency})$, as we assumed in the model used in Experiment 2, showed a reasonable account of the learning effect and fairly similar parameter estimates across the three learning groups.

Consequently, this model served as a proof of concept that with sufficient data and proper transformation, a Gaussian fixed-variance mixture model can successfully identify distinct training groups, even when training conditions are not orthogonal, based on log-transformed mean RTs. Yet, to produce this model, a number of choices were made in the modeling. In the next section, we explore some alternative models resulting from different choices about the assumptions and procedures. Together, these help illustrate some of the boundary conditions of the process and provide additional insight into the group training structure. We display the results of the baseline model and alternate models in Table 3, which shows the method, group membership, estimates of variance, and a categorical correspondence/correlation measure known as Cramér's V (Cramér, 1946), computed with the lsr package in R (Navarro, 2015).

Alternative models for identifying clusters

The present set of studies illustrate that with appropriate transformation and modeling, we can identify clusters of shared knowledge on the basis of memory strength and fluency, even when the overt responses are typically correct and would not allow for accurate identification of group memberships. However, a number of related approaches make different assumptions or estimate different values from the data that might provide additional insight into either the method or the underlying data set. Thus, we examined several of these to help understand the potential alternative models.

Estimating variance directly

The proposed model follows the strong-consensus model, in which we define the meaning of a consensus a priori for a particular data type. For accuracy data, Mueller and Veinott (2008) typically used a small fixed proportion of disagreement (e.g., 10%), meaning that a group of individuals can share a set of ideas as long as an individual tends to disagree less than 10% of the time. For RTs, the analogue is using fixed-variance Gaussians once the distributions have been scaled to their mean values, so that the standard deviation of the Gaussian represents the coefficient of variation of a group's RT. As an alternative, that standard deviation can be estimated directly. During the expectation–maximization process, the variance and covariance of the Gaussian can be estimated directly. We will consider a model in which each group of people is modeled with a single variance parameter (i.e., variances are identical for all questions), which adds just one parameter per group, for $K(N + 1)$ total parameters (where K is the number of groups and N is the number of items). More complex models could be considered, as well, with individual variance and covariance terms

estimated directly from the data, but this could increase the number of parameters to $2KN$ (for individual variance estimates) or KNN (if the covariance was also estimated).

The result of such an analysis is shown as Model 2 in Table 3. Here, instead of three groups, the method identified four, one corresponding roughly to each training group, and a fourth catch-all group with relatively large variance that contained individuals who did not fit well into any of the other groups. Interestingly, the estimated standard deviations of the remaining intact groups are smaller than our fixed 0.1 value. Consequently, this approach produced results similar to those of Model 1, with the added benefit of identifying some ill-fitting participants that did not provide strong enough evidence for membership in the identified groups.

RT of correct trials

As we discuss in the Appendix, one may choose to apply the approach to either mean RTs or the mean RTs of correct responses. In our study, most participants had one or more items that they had learned incorrectly and that thus had no correct responses during the test phase (since feedback was not given in this phase, it was not uncommon to be incorrect on all response trials for a pair). This makes examining correct RTs difficult, since the likelihood computation and parameter estimation are not well behaved with missing data. To address this problem, we replaced any missing values with a placeholder estimate—a value of either 1.0 (the scaled mean value; Model 3) or 2.0 (greater than 93% of the means; Model 4). These models each identified four-group solutions using the BIC criterion, with results akin to those of Model 2. Importantly, when missing values were imputed with the 1.0 mean value, three small accurate groups were identified, consisting of about half of the participant population; the rest were placed in a larger catch-all group; when these values were replaced with relatively greater values, the classification improved. This indicates that even without considering incorrect learning, the process is reasonably successful, but a nontrivial amount of the process is driven by the longer RTs from less-well-trained items.

Combining time and accuracy: inverse efficiency

The subsequent models imply that part of the classification accuracy is driven by factors related to the time of error trials. As a consequence, it might be useful to combine timing and accuracy into a composite score. One simple way of combining time and accuracy is to use so-called *inverse efficiency* (IE), in which the mean correct RT is adjusted by dividing by the accuracy. IE was developed by Townsend and Ashby (1978, 1983) to measure energy consumption over time and the power of

energy systems, and it is popularly used as a method to summarize both RTs and accuracy simultaneously in cognitive psychology. One interpretation of IE (see Mueller, Simpkins, Anno, Fallon, Price, & McClellan, 2011) is that it provides a total work throughput for the system, assuming zero error recovery costs and fixed-time completion costs: A task that takes 1 min with 100 accuracy is equivalent to a task that takes 40 s with 2/3 chance of succeeding on each try, since the latter on average will take 60 s to complete. However, some research suggests that IE scores should not be the only independent measure regarded in experimental psychology, because the ratio used will produce large variances that make them difficult to interpret (Bruyer & Brysbaert, 2011), and this is especially true for situations such as ours, with relatively few trials per condition and some low accuracy (Akhtar & Enns, 1989). In fact, the existence of conditions in which participants mislearned the answers impedes the ability to estimate IE (because there is no baseline correct RT to adjust, and an accuracy of 0 would inflate this value to infinity). The results were either uninteresting or inconclusive, so we chose not to report any specific models using this approach. Instead, we considered a model-based approach in which we used a simple diffusion model to estimate information accumulation.

Combining time and accuracy: diffusion drift rate

This alternative approach utilized drift rates produced by the EZ-diffusion model (Wagenmakers, van der Maas, Dolan, & Grasman, 2008), which is a simplified application of the Ratcliff diffusion model (Ratcliff, 1978). The EZ-diffusion model takes the error rate and RT mean and variance as inputs and computes a drift rate, response boundary, and nondecision time as outputs. The EZ-diffusion model can address the issues of RT and accuracy simultaneously and is prevalently applied in psychological areas such as color and brightness stimulus discrimination, recognition and prospective memory, and lexical decision (Voss et al., 2013; Wagenmakers et al., 2008).

To obtain the drift rates, we estimated EZ-diffusion model parameters based on Wagenmakers et al.'s (2008) implementation. However, this estimation fails when accuracy is either 1.0 or 0.0, so we replaced these values arbitrarily with .99 and .01, respectively. As with the earlier measures, we then normalized the drift rate for each individual by dividing by the mean absolute value of all estimated drift rates. Because of the existence of negative drift rates, this means that the normalization is somewhat different from the normalization performed on the RTs, so the baseline coefficient of variation of 0.1 was no longer appropriate. We thus present two models:

Model 5, in which the drift rate for each group was estimated on the basis of group membership, and Model 6, which fixed the group standard deviation at a level similar to the mean standard deviation of Model 5. Model 5 produced one group that consisted of Group 2, one mostly of Group 4, and a third mixing Groups 1 and 4. In contrast, the fixed-standard-deviation Model 6 produced classification results fairly close to the training group membership. Thus, this shows that even using a simple estimate of drift rate, reasonable solutions can be produced using the Gaussian CMM approach.

Using accuracy alone: cultural consensus modeling

Given that error trials appear to impact the model, it is important to show that we were not simply using the error-impacted RT as a proxy for accuracy, and that a traditional cultural consensus model would fail to identify multiple distinct groups. Accuracy was modeled using a binomial distribution, with the accuracy for each item either estimated directly from the group or coming from the most popular response and forcing agreement within a small bound (the strong-consensus model). We examined both models (shown as Models 7 and 8 in Table 3), and both converged to a single group. Thus, the error trials alone were not enough to enable classification into distinct groups; rather, the successful classification achieved in Models 1–6 was driven by the impact of slowing on less familiar information.

General discussion

In this article, we have introduced a statistical method based on CMM to identify cultural groups on the basis of implicit measures of knowledge using RT patterns. The results revealed that the ability to recover and identify shared belief groups depends on (1) the level of noise in the measurement, (2) the differential signals for strong versus weak attitudes, and (3) the similarity between group attitudes.

This model is applicable to behavioral and cognitive tasks that use latency measures, such as the implicit association task, LDT, priming LDT, free association task, and many visual attention tasks. The basic benefits of the approach to social priming and related paradigms is that one can classify a set of participants into distinct groups of shared beliefs or attitudes prior to taking their demographic background into account, and without knowing what types of patterns particular test items may elicit.

Previous research had established cultural differences between Westerners and Easterners on a number of RT measures in tasks involving visual attention (e.g., Boduroglu, Shah, & Nisbett, 2009; Masuda & Nisbett,

2001). The present methodological advances give the opportunity to assess whether consistent patterns emerge across a broad set of such tests, and how such groups (if they exist) are related to East–West cultures as well as other factors (such as education, language, and bilingualism) that tend to be confounded with cultural differences.

To our knowledge, this study is the first of its kind that has shown how RT measures may be used as a means for assessing cultural consensus. Yet the basic approach could be useful for a number of different dependent measures beyond latency/RT, including brain activation, cerebral blood flow, skin conductance measures (often used for lie detection), voice stress, and numerous measures that are increasingly available as the outputs of sensors and machine-learning algorithms. The present work is a first step on the path to exploring how a number of such measures, in isolation or combination, can help to identify shared cultures of belief.

Appendix: Applying CMM to RT measures

To begin, we assume that we measured RTs across a set of questions, items, or item categories (columns) and respondents (rows), and that the overt responses were mostly the same (i.e., 80% or more responses were in the same direction). If the actual response patterns deviated substantially, then the strong-consensus model could be used to identify clusters of overt belief, but we are assuming that the deviations from the group modal response on any individual question were relatively small, and so we needed to use RT patterns to identify shared knowledge. One must decide whether to examine the mean RT of just the correct responses or of all responses (correct and incorrect). Because, in our case, errors probably indicated a failure to encode the geographical fact, it made sense to

Table 4 Raw response time (RT) patterns for five participants across four questions in a hypothetical experiment using RT as an implicit measure of knowledge access

	Q1	Q2	Q3	Q4
P1	609	757	405	514
P2	620	432	525	757
P3	1,037	1,105	819	970
P4	723	453	623	815
P5	1,147	873	1,044	1,233

Table 5 Scaled RT scores remove between-participant differences in mean RT

	Q1	Q2	Q3	Q4	Mean
P1	1.094	1.360	0.728	0.923	1.026
P2	1.085	0.756	0.919	1.325	1.022
P3	1.062	1.131	0.838	0.993	1.006
P4	1.132	0.709	0.976	1.276	1.023
P5	1.076	0.819	0.980	1.157	1.008
Mean	1.090	0.955	0.888	1.135	1.017

use correct RTs. In other cases, in which errors might stem from either lower retrieval fluency or other decision processes, it might make sense to use the mean of all responses. In our alternative models section, we showed that, for our data set, either model worked equally well.

Table 4 shows example RT patterns that might be seen. Here, each row represents a different participant, and each column represents a different question. Possibly, each value might be computed as the mean RT for a set of items sharing the same question class.

Even so, there are typically large but consistent between-participants variations in the RTs. If the raw RTs were used to identify clusters of belief strength, there would be a risk that the methods would simply identify slow versus fast responders, and not people with differential associations. In Table 4, Participants 3 and 5 have long RTs relative to the rest, and even though their RTs have a correlation of $-.31$, a model using raw RTs might place them in the same group. Consequently, our first step is to normalize the RTs for each participant. We do this by computing the exponentiated mean of the $\log(\text{RT})$ ($\exp\{\text{mean}[\log(x)]\}$) for each participant, and dividing each RT by the resulting mean for each participant. Each participant's data thus have an average close to 1.0 (Table 5).

Table 6 Here, the values in Table 5 are rescaled so that each question has a mean value of 1.0

	Q1	Q2	Q3	Q4	Mean
P1	1.004	1.424	0.819	0.814	1.015
P2	0.996	0.792	1.035	1.168	0.998
P3	0.974	1.184	0.944	0.875	0.994
P4	1.039	0.743	1.098	1.124	1.001
P5	0.988	0.858	1.103	1.019	0.992
Mean	1.000	1.000	1.000	1.000	1.000

Finally, we rescale each question so its mean is 1.0, by dividing by its mean. This step serves to allow easier interpretation of the parameter estimates, because to the extent a group's parameter deviates from 1.0, it indicates that the group is consistently above or below average (see Table 6).

Now, by careful examination, some patterns emerge. Looking at Q2, we can see that P2, P4, and P5 are substantially faster than average, whereas P1 and P3 are slower. Similarly, for Q4, P2, P4, and P5 are substantially slower than the other two. This pattern indicates that two distinct and incompatible patterns of belief strength exist in the sampled population.

An analogue to the strong-consensus model for RTs is to model the deviations from the mean of any group as a Gaussian distribution, with all responses having the same variance, which is either predetermined (essentially a coefficient of variation that is deemed to be the upper limit of group's variability), estimated as a single parameter across all questions, or estimated individually for each question. Of course, even more parameters could be identified, including the covariance between RT distributions for each question, but additional parameters can be costly, because they represent more complex models that are better able to fit arbitrary data and may require substantial data to make worthwhile. Consequently, in the present approach, we use the simplest reasonable model analogous to the strong-consensus model used for overt responses: We fix the variance of the Gaussians to be a small value σ (we have typically used around $\sigma = 0.1$). Because of the transformations, this value essentially represents the average coefficient of variation allowed within a group. If a group tends to have larger variability than this, it will be beneficial to break the group into two distinct groups, provided that the new groups can account for the data.

The model was implemented via a custom likelihood model using the flexmix package in R (Leisch, 2004) to perform inference via the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977) in the R statistical computing language, Version 2.15 (R Development Core Team, 2014). We selected the best-fitting and simplest model using the BIC metric (Schwarz, 1978). The flexmix package provides a number of tools for developing and analyzing the results of a wide variety of mixture models, and its documentation can be consulted for specific examples of how the present models were implemented. Code for the specific custom flexmix driver (a Gaussian strong-consensus model in which the sigma parameter controls the fixed standard deviation) is presented below.

```

MyNormalSC <- function (formula = .~., sigma=.1, diagonal = TRUE)
{
  z <- new("FLXMC", weighted = TRUE, formula = formula, dist = "mvnorm",
          name = "model-based Gaussian clustering")
  z@defineComponent <- expression({
    logLik <- function(x, y) mvtnorm::dmvnorm(y, mean = center,
                                              sigma = diag(cov, length(center)),
log = TRUE)
    predict <- function(x, ...) matrix(center, nrow = nrow(x),
                                       ncol = length(center), byrow = TRUE)
    new("FLXcomponent", parameters = list(center = center,
                                          cov = cov), df = df, logLik = logLik,
predict = predict)
  })
  z@fit <- function(x, y, w, ...) {
    center <- colSums(y * w, na.rm=T) / sum(w)
    decentered <- [ t(t(y*w) - center))

    cov3 <- sigma
    para <- list(center=center, cov=cov3, df=ncol(y))
    with(para, eval(z@defineComponent))
  }
  z
}

```

An example application of the model is shown below. Here, `normed` is a matrix containing the mean RT, drift rate, or other dependent measure, in which each

participant is a row and each column is an item. Values should be scaled and normalized prior to applying the model.

```

library(flexmix)
out <- stepFlexmix(normed~1, model=MyNormalSC(sigma=.1),

                  k=1:5,          ##The number of groups to assess

                  nrep=200,       ##Number of E-M models per group size

                  control=list(iter.max=500,minprior=0.1))

plot(out)
model <- GetModel(out, "BIC") ##Obtain BIC best fit model
p <- parameters(model)  ##get parameter matrix; one column per group
cl <- clusters(model)   ##get cluster membership for each row of data matrix
table(cl, trainingcond) ##assuming trainingcond specifies the experimental condition

```

References

- Akhtar, N., & Enns, J. T. (1989). Relations between covert orienting and filtering in the development of visual attention. *Journal of Experimental Child Psychology*, 48, 315–334.
- Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, 56, 452–469. doi:10.1016/j.jmp.2013.01.004
- Atran, S., Medin, D. L., & Ross, N. O. (2005). The cultural mind: Environmental decision making and cultural modeling within and across populations. *Psychological Review*, 112, 744–776. doi:10.1037/0033-295X.112.4.744
- Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 252–259. doi:10.1037/0096-1523.5.2.252

- Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge, UK: Cambridge University Press.
- Boduroglu, A., Shah, P., & Nisbett, R. E. (2009). Cultural differences in allocation of attention in visual information processing. *Journal of Cross-Cultural Psychology*, 40, 349–360.
- Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, 51, 5–13. doi:10.5334/pb-51-1-5
- Cohen, A. B. (2010). Just how many different forms of culture are there? *American Psychologist*, 65, 59–61. doi:10.1037/a0017793
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39, 1–38.
- Fischer, R. (2012). Intersubjective culture: Indeed intersubjective or yet another form of subjective assessment? *Swiss Journal of Psychology*, 71, 13–20. doi:10.1024/1421-0185/a000067
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464–1480. doi:10.1037/0022-3514.74.6.1464
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97, 17–41. doi:10.1037/a0015575
- Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8), 1–18.
- Lupker, S. J., & Pexman, P. M. (2010). Making things difficult in lexical decision: The impact of pseudohomophones and transposed-letter nonwords on frequency and semantic priming effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1267–1289. doi:10.1037/a0020125
- Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81, 922–934.
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227–234. doi:10.1037/h0031564
- Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Loci of contextual effects on visual word recognition. In P. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 98–118). London, UK: Academic Press.
- Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL test battery. *Journal of Neuroscience Methods*, 222, 250–259.
- Mueller, S. T., Simpkins, B., Anno, G., Fallon, C. K., Price, O., & McClellan, G. E. (2011). Adapting the task–taxon–task methodology to model the impacts of chemical protective gear. *Computational and Mathematical Organizational Theory*, 17, 251–271. doi:10.1007/s10588-011-9093-7
- Mueller, S. T., & Veinott, E. S. (2008). Cultural mixture modeling: Identifying cultural consensus (and disagreement) using finite mixture modeling. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 64–70). Austin, TX: Cognitive Science Society.
- Navarro, D. J. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5)*. Adelaide, Australia: University of Adelaide.
- Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336). Hillsdale, NJ: Erlbaum.
- Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, 120, 356–394. doi:10.1037/a0032020
- R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108. doi:10.1037/0033-295X.85.2.59
- Rohner, R. P. (1984). Toward a conception of culture for cross-cultural psychology. *Journal of Cross-Cultural Psychology*, 15, 111–138.
- Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313–338.
- Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 1–17. doi:10.1037/0096-1523.3.1.1
- Schvaneveldt, R. W., & Meyer, D. E. (1973). Retrieval and comparison processes in semantic memory. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 395–409). New York, NY: Academic Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess “guilty knowledge.”. *Journal of Applied Psychology*, 85, 30–37.
- Shimamura, A. P., & Squire, L. R. (1984). Paired-associate learning and priming effects in amnesia: A neuropsychological study. *Journal of Experimental Psychology: General*, 113, 556–570. doi:10.1037/0096-3445.113.4.556
- Sieck, W. R., & Mueller, S. T. (2009, February). *Cultural variations in collaborative decision making: Driven by beliefs or social norms?* Paper presented at the International Workshop on Intercultural Collaboration, Palo Alto, CA.
- Sieck, W. R., Smith, J. L., & McHugh, A. P. (2007). Cross-national comparison of team competency values. In *Proceedings of the Human Factors and Ergonomics Society, 51st Annual Meeting* (pp. 268–272). Thousand Oaks, CA: Sage.
- Simpkins, B., Sieck, W., Smart, P., & Mueller, S. (2009, September). *Idea propagation in social networks: The role of “cognitive advantage.”* Paper presented at the 1st ITA Workshop on Network-Enabled Cognition, Raleigh, NC.
- Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. *Cognitive Theory*, 3, 200–239.
- Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge, UK: Cambridge University Press.
- Verschuere, B., Kleinberg, B., & Theodoridou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition*, 4, 59–65.
- Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in associative and categorical priming: A diffusion model analysis. *Journal of Experimental Psychology: General*, 142(2), 536.
- Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, 15, 1229–1235. doi:10.3758/PBR.15.6.1229
- Weidemann, C. T. & Mueller, S. T. (2008). Decision noise may mask criterion shifts: Reply to Balakrishnan and MacDonald (2008). *Psychonomic Bulletin & Review*, 15, 1031–1034.
- Weldon, M. S., & Coyote, K. C. (1996). Failure to find the picture superiority effect in implicit conceptual memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 670–686. doi:10.1037/0278-7393.22.3.670