Journal of Mathematical Psychology 61 (2014) 1–13

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Cultural consensus theory for continuous responses: A latent appraisal model for information pooling

R. Anders *, Z. Oravecz, W.H. Batchelder

Department of Cognitive Sciences, University of California, 3151 Social Science Plaza, Irvine, CA 92697-5100, United States

HIGHLIGHTS

- A new Cultural Consensus Theory (CCT) model for continuous data.
- Model-based clustering and detection of multiple cultures in the data.
- Derivation of the mathematical and statistical properties of the model.
- Hierarchical Bayesian inference for the model on real and simulated data.
- User-friendly software that facilitates application of the model.

ARTICLE INFO

Article history:
Received 17 January 2014
Received in revised form
10 May 2014

Keywords:
Cultural consensus theory
Latent class models
Latent trait models
Continuous data
Signal detection theory
Mixture modeling
Clustering

ABSTRACT

A Cultural Consensus Theory approach for continuous responses is developed, leading to a new model called the Continuous Response Model (CRM). It is a cognitive psychometric model that is applicable to consensus data, in which respondents (informants) have answered questions (items) regarding a shared knowledge or belief domain, and where a consensus (a latent set of 'true' answers applicable to the group) may exist. The model estimates the consensus answers to items, item difficulty, informant knowledge and response biases. The model can handle subcultures that have different consensuses from one another in the data, and can both detect and cluster respondents into these subcultures; it thus provides one of the first forms of model-based clustering for multicultural consensus data of the continuous response type. The model is demonstrated on both simulated and real multi-cultural data, using the hierarchical Bayesian framework for inference; two posterior predictive checks are developed to verify the central assumptions of the model; and software is provided to facilitate the application of the model by other researchers.
Published by Elsevier Inc.

1. Introduction

The purpose of the present paper is to introduce a Cultural Consensus Theory (CCT) model for continuous responses, and in tandem, supply user-friendly software that facilitates application of the model by others. CCT is a methodology conceived of in the mid 1980s (Batchelder & Romney, 1986, 1988; Romney, Weller, & Batchelder, 1986) that is applicable to *consensus data*: defined as data in which respondents (informants) have answered questions (items) regarding a shared knowledge or belief domain, and where a consensus (a set of 'true' answers applicable to the group, or culture) may exist. Exemplary forms of consensus data may consist of eyewitness testimony, probability forecasting, political polls, cultural beliefs, subjective assessment, or ideological beliefs. In order to estimate the consensus answers to items, as well as item response effects (e.g. knowledge level, response biases, item difficulty, and cultural membership) for these forms of data, CCT consists of a number of cognitive psychometric models.

The first CCT model was developed for binary data (e.g. true/false data); it is called the General Condorcet Model (GCM), and makes the assumption that the consensus truth of each item is also a binary value. This model has been widely applied in the social and behavioral sciences, especially cultural anthropology (e.g. Weller, 2007). The detection of multiple cultural truths (subcultures with differing consensuses), and cultural membership for the GCM, was developed by Anders and Batchelder (2012). An alternate assumption to that of the GCM, that continuous (fuzzy) truths in (0, 1) instead underlie binary data, was explored by Batchelder and Anders (2012). They introduced a new model for binary data that used a beta appraisal distribution to estimate these values in (0, 1); the Latent Truth Model (LTM), and these values could represent such

* Corresponding author.
E-mail addresses: andersroyce@gmail.com, andersr@uci.edu (R. Anders), zoravecz@uci.edu (Z. Oravecz), whbatche@uci.edu (W.H. Batchelder).

http://dx.doi.org/10.1016/j.jmp.2014.06.001
0022-2496/Published by Elsevier Inc.

things as a degree of truth, intensity, or a true probability. Carrying on with this assumption, Anders and Batchelder (2013) developed an extensive CCT model for ordinal data (ordered discrete options of $C \geq 2$ categories) using a Gaussian appraisal model, which links the truth locations in (0, 1) to the real line,[1] called the Latent Truth Rater Model (LTRM). The model developed herein shares much similarity to the work of the LTRM, which can likewise detect and estimate multiple cultures/clusters of informants, continuous truths, shifting and scaling response biases, informant knowledge, and item difficulty for each culture, but it is instead specified for continuous data.

The Continuous Response Model (CRM), with its application methodology and software introduced here, offers to our knowledge, the first cognitive model-based approach for cultural consensus clustering for continuous response data. The model supposes that the continuous responses of the consensus data arise from the possible effects of membership to distinct cultural consensuses, consensus knowledge, item difficulties, and cognitive response biases. A natural interval to collect continuous responses is in (0, 1); the model acts on the logit transforms of these data, to utilize the benefits that have been found with using the Gaussian modeling approach over the beta approach, as in the work of the LTRM.[2] There have been several previous efforts to develop a CCT model for continuous response data (Batchelder & Romney, 1989; Batchelder, Strashny, & Romney, 2010; France & Batchelder, 2014). However, the current model goes well beyond these efforts by extending the specification of the model to include multicultural detection, model-based clustering, and a fully hierarchical inference structure complete with Bayesian posterior predictive checks of crucial assumptions of the model. A software package that greatly facilitates its application, has also been introduced with the model.

There has also been recent progress for modeling continuous responses outside of CCT (Merkle & Steyvers, 2011; Steyvers, Wallsten, Merkle, & Turner, 2013; Turner & Steyvers, 2011; Turner, Steyvers, Merkle, Budescu, & Wallsten, 2013a), and these models have been primarily introduced for the aim of forecasting aggregation, rather than cultural consensus work. Each of these models shares some characteristics with the CRM, but may differ in regard to whether and how they model latent appraisals of truth (or error in perceiving the truth), response biases, heterogeneous item difficulty, and cultural tendencies; furthermore, none of them have been extended yet to perform model-based clustering around consensuses or cultures. This is because they are generally concerned with discovering the ground truth, whereas the CCT model is interested in discovering the cultural truth(s). Hence, the prior models will aggregate across all cultures to estimate a ground truth while the CCT approach will seek to discover the consensus truth within each culture; and thus these models are formulated differently toward their intended functions. With that, the CCT approach of the CRM becomes further differentiated: on the basis of mathematical properties of our model, we derive techniques that suggest the appropriate number of consensus truths in the data, if one assumes the CRM to be applicable, and then develop posterior predictive checks to verify if the consensus structure of the data (number of cultures) is appropriately fit by the model.

The paper consists of five sections. After the Introduction, Section 2 specifies the new model. In Section 2.2, mathematical properties of the model are developed which are then used to form important model checks. The model is formulated within the hierarchical Bayesian framework in Section 2.3, posterior predictive checks are developed in Section 3.2, and appropriate application of the model is discussed in the other parts of Section 3. Section 4 demonstrates the model and its results on both simulated and real data sets. Section 5 provides the general discussion.

2. The CRM

2.1. Specification of the model

Assume that each of N informants provides a continuous answer within (0, 1), or within an allowable range that offers an appropriate linear transform to (0, 1), to each of M items, and let the responses be the realization of a random response profile matrix $\mathbf{X} = (X_{ik})_{N\times M}$, where

$$X_{ik} = x \quad \text{iff informant } i \text{ assigns value } x \text{ to item } k. \tag{1}$$

An item's value, or measure of truth, probability, or degree may be naturally interpretative or scalable to a value in (0, 1), and the informants have latent appraisals of these values with error. In such cases, the CRM links the (0, 1) locations of these values to the real line with the logit transform, where $\text{logit}(x) = \log(x/1-x)$. Therefore, each item also has a consensus location in $(-\infty, \infty)$. The respondents have latent appraisals of these item values with a mean at the item's consensus location plus some error, which depends on the knowledge competency of the informant, and the difficulty of knowing the item. Then the observed responses are determined by the informant's response biases, which act on the latent appraisal value.

The CRM is formalized and further explained by the following axioms:

Axiom 1 (*Cultural Truths*)**.** There is a collection of $V \geq 1$ latent cultural truths, $\mathscr{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_v, \ldots, \mathbf{T}_V\}$, where $\mathbf{T}_v \in \prod_{k=1}^{M}(-\infty, \infty)$. Each informant i responds according to only one cultural truth (set of consensus locations), as $\mathbf{T}_{\Omega_i}$, where $\Omega_i \in \{1, \ldots, V\}$, and parameter $\mathbf{\Omega} = (\Omega_i)_{1\times N}$ denotes the cultural membership for each informant.

Axiom 2 (*Latent Appraisals*)**.** It is assumed that each informant draws a latent appraisal, Y_{ik}, of each $T_{\Omega_i k}$, in which $Y_{ik} = T_{\Omega_i k} + \epsilon_{ik}$. The ϵ_{ik} error variables are distributed normal with mean 0 and standard deviation σ_{ik}.

Axiom 3 (*Conditional Independence*)**.** The ϵ_{ik} are mutually stochastically independent, so that the joint distribution of the latent appraisals is given for all realizations (y_{ik}) by

$$h[(y_{ik}) \mid (T_{\Omega_i k}), (\sigma_{ik})] = \prod_{i=1}^{N}\prod_{k=1}^{M} f(y_{ik} \mid T_{\Omega_i k}, \sigma_{ik}) \tag{2}$$

where $f(y_{ik} \mid T_k, \sigma_{ik})$ is the normal distribution with mean $T_{\Omega_i k}$ and standard deviation σ_{ik}.

Axiom 4 (*Precision*)**.** There are knowledge competency parameters $\mathbf{E} = (E_i)_{1\times N}$, with $E_i > 0$, and item difficulty parameters specific to each cultural truth, $\mathscr{L} = \{\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_v, \ldots, \mathbf{\Lambda}_V\}$, where $\Lambda_v = (\lambda_{vk})_{1\times M}$, and each $\lambda_{vk} > 0$. An informant's standard appraisal error in the assessment of each $T_{\Omega_i k}$ is defined by standard deviation

$$\sigma_{ik} = E_i \lambda_{\Omega_i k}. \tag{3}$$

If all item difficulties are equal, then all λ_{vk} are set to 1.

[1] Greater success was arguably achieved using the Gaussian appraisal approach over the beta appraisal approach: the inverse logit values of the truths estimated in $(-\infty, \infty)$ spanned the full interval in (0, 1) more often than the values estimated in (0, 1) by the beta, informant knowledge precision of the truth was independent of item location with the Gaussian, and the use of polychoric correlations was available to help detect and adequately fit the consensus structure of the data.

[2] Data that is not in (0, 1) can be transformed to the interval with an appropriate linear transform, e.g. financial data from \$0 to \$100,000 would be divided by 100, 000. Secondly, indeed since the model uses the Gaussian approach, it may be used on data naturally within $(-\infty, \infty)$, though as discussed later, the logit transform of data in (0, 1) may be more appropriate for the approach developed here.

Axiom 5 (*Response Biases*)**.** There are two informant bias parameters that act on each informant's latent appraisals, Y_{ik}, to arrive at the observed responses, the X_{ik}. These include a scaling bias, $\mathbf{A} = (a_i)_{1\times N}$, $a_i > 0$; and a shifting bias, $\mathbf{B} = (b_i)_{1\times N}$, $-\infty < b_i < \infty$, where

$$X_{ik} = a_i Y_{ik} + b_i. \tag{4}$$

These axioms are designed to model the continuous responses of informants that vary in competency, E_i, and response biases, a_i and b_i, to items that have different shared latent truth locations, T_{vk}, on the continuum, and potentially different item difficulties, λ_{vk}. Axiom 1 locates the item truth values, and allows for multiple cultural consensus beliefs (different sets of consensus truth locations), $\mathscr{T}$, to be involved, where each informant i belongs to only one, $v \in 1 \ldots V$, consensus truth set, in which v takes on values from Ω_i for each informant, where Ω_i is the cultural-truth membership of informant i (or cluster assignment of informant i). Multiple cultural truths may occur when the informants belong to different cultural viewpoints regarding the items, such as members of opposing political parties answering the same set of policy questions. The single culture, or simplified case of the model where $V = 1$ and all of the informants share the same cultural truth, is characteristic of the traditions in earlier CCT practices, and will be thoroughly discussed and examined later.

Axioms 2 and 3 are like those often found in CCT models as well as models in classical test theory (e.g. Lord, Novick, & Birnbaum, 1968). Axiom 2 specifies that appraisal error is normally distributed with mean zero, and Axiom 3 asserts that the appraisals are conditionally independent given the informants' cultural truth locations and the error variances of their appraisal errors. Axiom 4 sets the standard appraisal error, σ_{ik}, to depend on both the informant's competency (level of consensus knowledge in the culture E_i) and the item difficulty (difficulty of knowing that item for the culture $\lambda_{\Omega_i k}$), as shown in (3). Anders and Batchelder (2013) show that a formula like (3) can be derived from a parameterization of the Rasch model (1960) for doubly-indexed, positive-valued parameters. As with other versions of the Rasch model, there is an obvious non-identifiability since (3) is invariant under multiplying E_i and dividing $\lambda_{\Omega_i k}$ by the same positive constant. This and other identifiability issues are discussed later when the model is specified in hierarchical form.

Axiom 5 concerns each informant's response biases, or magnitude and location response tendencies on the scale. The response bias parameters a_i and b_i respectively account for magnitude (or scale) preferences: for example, a tendency to mark most values in the outer ends or middle section of the scale, and location preferences: a tendency to mark values more often on one side of the scale. These biases, a_i and b_i, are motivated by a function often used to scale bias in probability estimation in (0, 1), called the Linear in Log Odds (LLO) function (e.g. see Fox & Tversky, 1995; Gonzalez & Wu, 1999; Zhang & Maloney, 2012). The function leads to $\text{logit}(p_i) = a_i \text{logit}(p) + b_i$, where p_i is respondent i's estimate of an event with probability p. Thus the bias parameter a_i estimates the shrinkage or expansion of the latent appraisals, while the bias parameter b_i estimates a shift to the left or right.

An illustration of the model in the simplified single culture case, $\text{CRM}^{V=1}$, is provided in Fig. 1. Consider a single informant estimating the degree of radioactive danger, on a scale of (0, 1)—0 indicating none whereas 1 indicates dangerous to consume, of fish caught in the Pacific Ocean in 2013 after the 2011 Fukushima disaster. Consider three different types of fish with true degrees of radioactive danger $\mathbf{T} = 0.1, 0.3$, and 0.8. The curves in the top plot of the figure depict the probability of error in the informant's appraisals, Y_{ik}, around the consensus belief of radioactivity within each fish, T_k. The informant has a competency (or standard

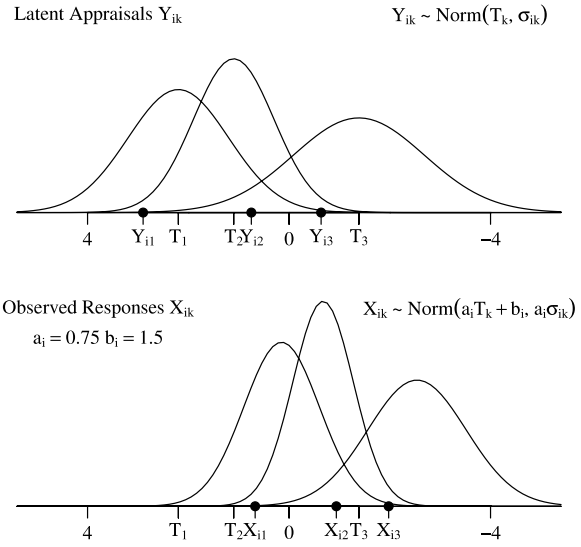

Fig. 1. An illustration of the $\text{CRM}^{V=1}$. Top plot: an individual's latent appraisals, Y_{ik}, of the true value of three different items, T_k, modeled by a normal distribution with mean T_k and standard deviation $\sigma_{ik} = E_i\lambda_k$, where $E_i = 1$ and $\lambda_k = [1, 0.8, 1.5]$. Second plot: observed responses of the same individual with response biases, a_i and b_i acting on the latent appraisals, in the form of $X_{ik} = a_i Y_{ik} + b_i$.

appraisal error) of $E_i = 1.0$, but the normal probability curves are different for each fish because of item difficulty: detecting the true radioactive danger of each fish is not equally easy, and their item difficulties are respectively, $\lambda_k = 1$ (neutral), 0.8 (more easy), 1.5 (more difficult); note that the standard appraisal error with item difficulty included is $\sigma_{ik} = E_i\lambda_k$. The bottom plot depicts the observed responses, X_{ik}, for the same informant, which are subject to his/her response biases, a_i and b_i, and the probability distribution of making various responses for this informant, is derived from the linear transformation of the Y_{ik} distribution in the form of $X_{ik} = a_i Y_{ik} + b_i$. This informant's response biases of $a_i = 0.75$ and $b_i = 1.5$ exemplify an individual who overestimates the radioactive danger of all fish ($b_i > 0$), and believes that there are few differences in the radioactive danger between fish—they are all dangerous ($a_i < 1$). This is one type of respondent, but different combinations of these response biases could capture numerous other types.

The likelihood function of the CRM is obtainable as

$$L[\mathbf{X} = (x_{ik})_{N\times M} \mid \mathscr{T}, \mathscr{L}, \mathbf{E}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Omega}] = \prod_{i=1}^{N}\prod_{k=1}^{M} f(a_i T_{\Omega_i k} + b_i, a_i\sigma_{ik} \mid E_i, \lambda_{\Omega_i k}) \tag{5}$$

where $f(\cdot)$ is the probability density function (pdf) of the normal distribution, parameterized by its mean and standard deviation. The likelihood function is derived from the property of the normal that it is closed under linear transformations. That is, a linear transformation of the normally-distributed Y_{ik} with mean $T_{\Omega_i k}$ and standard deviation σ_{ik} in the form of $a_i Y_{ik} + b_i$ is also normally-distributed with mean $a_i T_k + b_i$ and standard deviation $a_i\sigma_{ik}$.

Thus with the full model of the CRM being defined by Axioms 1–5, different or simpler cases of the model can be obtained by whether one fits data with a single culture, $V = 1$, or without item difficulty effects, in which all $\lambda_k = 1$. In the next sections, methods based on model properties are provided to distinguish the appropriate fit of the model to a given data set $\mathbf{X}$, from these various cases of the CRM.[3] Applications of these methods and the model under the different cases will be demonstrated in the paper.

[3] As the number of clusters or cultures, V, is a pre-specified number in the current model-fitting approach, techniques for assuming an appropriate number of cultures

2.2. Properties of the model

The following two properties that are developed pertain to consensus truth (Axiom 1), and appraisal precision (Axiom 4). These two properties are used to form important model checks when fitting the model to data. These two checks can respectively help determine if the number of cultures specified is appropriate for the data, and whether heterogeneous item difficulty is necessary to fit response variability across items.

2.2.1. Spearman's law property

The following property is principle to the CRM, it pertains to Axiom 1, and it is utilized in Section 3.2 to produce an essential model check for selecting the appropriate number of cultural truths to assume. A feature of the property provides that a suitably defined measure of response agreement between pairs of informants calculated over items has a simple structure; and this is a characteristic of most CCT models (e.g. see Anders & Batchelder, 2012, 2013; Batchelder & Anders, 2012). The consequence of the property is that the factor structure of the matrix of informant-by-informant correlations over the items carries information regarding the number of cultural truths relatable to the data. In this section, the similar property is shown to hold for the CRM as in other CCT models with Theorem 1. To present the theorem, a random variable that selects a random item subscript is introduced. Let K be a random variable representing item indices, with probability density:

$$Pr(K = k) = 1/M, \quad \forall k = 1, \ldots, M. \tag{6}$$

Theorem 1. *Suppose the Axioms 1–5 hold for the CRM. Then given fixed values of* $\mathscr{T}$, $\mathbf{E}$, $\mathscr{L}$, *and* $\mathbf{\Omega}$,

$$\begin{aligned} &\forall i, j = 1, \ldots, N \ni i \neq j, \\ &\rho(X_{iK}, X_{jK}) = \rho(X_{iK}, T_{\Omega_i K})\rho(X_{jK}, T_{\Omega_j K})\rho(T_{\Omega_i K}, T_{\Omega_j K}). \end{aligned} \tag{7}$$

The proof is provided in Appendix A.

Eq. (7) states that the correlation of the responses between any two informants is decomposable into a product of three terms, namely the correlation of each informant's responses with his or her own latent truth, and the correlation between the two latent truths. While the triple correlation property behind the responses of the CRM is given in (7), note that if all informants share the same cultural truth ($V = 1$), (7) reduces to

$$\rho(X_{iK}, X_{jK}) = \rho(X_{iK}, T_K)\rho(X_{jK}, T_K), \tag{8}$$

which states that the correlation of responses between any two informants across all M items is equal to the product of their individual correlations with the truth (the shared set of consensus locations). The consequence is that by (8), only one substantive factor should be present in the informant-by-informant Pearson correlation matrix of responses over items when only one cultural truth is shared by the informants; and when multiple cultures exist by (7), then more than one substantive factor should be present.

The CCT approach for examining the factors of the informant-by-informant correlation matrix, $\mathbf{M}$, to assess property (7) in the data, involves employing standard minimum residual factor analysis (MINRES; Comrey, 1962). For a detailed explanation of this approach, see Anders and Batchelder (2012). MINRES is applicable via the fa() function from the R 'psych' package (Revelle, 2012),

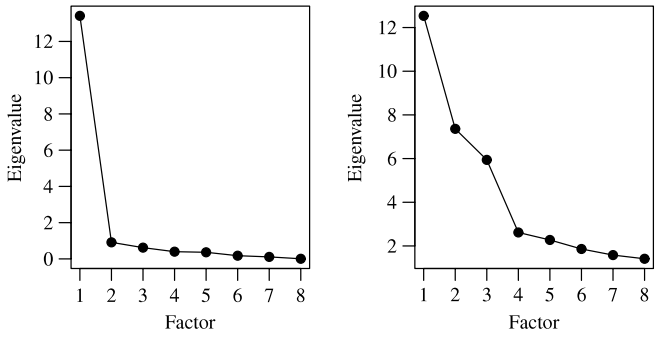

Fig. 2. Scree plots obtained from the informant–informant correlation matrix $\mathbf{M}$ of data generated by the $CRM^{V=1}$, (left) and $CRM^{V=3}$, (right) each having $N = 15 \times V$ informants by $M = 30$ items, with data simulated from the hierarchical CRM specified in Section 2.3 with homogeneous item difficulty ($\lambda_k = 1$).

which will return the factors and their corresponding eigenvalues; these eigenvalues may be used to construct a scree plot.

Using this approach on data simulated by the model, the standard patterns that indicate certain V factor solutions for the CRM are provided in the scree plots of Fig. 2. In the left plot, a one-factor solution is generated by the $CRM^{V=1}$. The plot shows the eigenvalue results from $\mathbf{M}$, which demonstrates a pattern typical of a single culture solution. The right plot of the figure contains the corresponding scree plot of data generated by the $CRM^{V=3}$, and demonstrates a strong three-culture solution. In each plot, there are V sizeable eigenvalues, where there is a large drop in the $V+1$th eigenvalue, which starts a linear decline in the eigenvalues, as its associated eigenvector and the remaining eigenvectors fit residual noise in $\mathbf{M}$. As in other multicultural extensions in CCT for the GCM and LTRM, from repeated simulation studies, the CRM also exhibits the property that generally, data generated with V number of cultures will relate to V substantial factors in the scree plot. Thus the number of substantial factors tends to relate to the number of cultural truths.[4] Furthermore as will be shown in Section 3.2, this scree plot analysis for the Spearman property in (7) can be used to develop a posterior-predictive model check for the CRM: for whether Axiom 1 is appropriately satisfied by the specification of V.

2.2.2. Informant precision and item difficulty

The following property is second in importance to that in Section 2.2.1 for the CRM, it pertains to Axiom 4, and it is utilized in Section 3.2 to produce a useful model check for the purpose of assessing the fit of response precision across items. Batchelder and Anders (2012) introduced a new statistic of the response data, $\mathbf{X} = (X_{ik})_{N\times M}$, called the Variance Dispersion Index (VDI), which reflects differences across items due to response variability:

$$\mathrm{VDI}(\mathbf{X}) = \sum_{k=1}^{M} V_k^2/M - \left(\sum_{k=1}^{M} V_k/M\right)^2 \tag{9}$$

where

$$V_k = \sum_{i=1}^{N} X_{ik}^2/N - \left(\sum_{i=1}^{N} X_{ik}/N\right)^2.$$

based on mathematical properties of the model are developed in Section 2.2; then a posterior predictive check that can successfully distinguish the model with the appropriate V value, is developed in Section 3.2. An additional posterior predictive check is also developed, which provides for the absolute model fit of a statistic related to response precision across items, and may distinguish item difficulty.

[4] However, note that this may not be the case when the additional cultures that exist are highly correlated with one another, as multicultural property (7) that results in multiple factors, reduces to (8) resulting in one factor, as the truths become highly correlated; and, when these additional cultures are small in proportion to the largest culture (and thus account for much less variance in the data).

VDI($\mathbf{X}$) is simply obtained by calculating the variance in responses over informants for each item, and then again calculating the variance of these item column variances. This VDI statistic may be used as a posterior predictive check to assess the fit of response precision, and consequently whether heterogeneous item difficulty ($\lambda_k \neq 1$ in Axiom 4) should be included in the model fit; Theorem 2 provides the rationale. Using this statistic as a posterior predictive check for item difficulty has been successfully demonstrated for other related CCT models (Anders & Batchelder, 2012, 2013; Batchelder & Anders, 2012); likewise, the present paper demonstrates the check that is appropriate for the CRM, and as a new development in CCT, expands the check to the multicultural case. Theorem 2 concerns a VDI statistic calculated instead using the variance of the latent appraisals, V_k^* in (10), rather than the manifest responses. It provides the conditional expectation of the item column appraisals in terms of the item locations, informant competencies, and item difficulty, for a single culture.

Theorem 2. *Suppose the Axioms 1–5 hold for the CRM and item difficulty is neutral ($\lambda_k = 1$). Then for fixed* $\mathbf{T}$ *and* $\mathbf{E}$*, the expected item column variances of the latent appraisals satisfy*

$$\forall k = 1, \ldots, M, \quad V_k^* = \sum_{i=1}^{N} (E_i \lambda_k)^2 = \sum_{i=1}^{N} E_i^2. \tag{10}$$

The proof is provided in Appendix B.

Eq. (10) shows that as a property of the appraisal modeling in the CRM, beyond sampling noise, the only possible source of variability in the V_k^* is due to the item difficulty parameter, λ_k. Consequently in the case of the CRM, introducing item difficulty tends to increase the VDI, and so monitoring the value of the VDI as a posterior predictive check can assist in assessing the fit of response precision across items, and whether it is necessary to include item difficulty in order to fit possible heterogeneity of response precision across items.

Note that Theorem 2 pertains to the V_k as calculated using the unobserved latent appraisals, V_k^*, rather than the observed responses which are subject to the response biases. Based on simulation studies, it has been found that a useful posterior predictive check arises from comparing the VDI of the latent appraisals generated by the model fit, to the VDI of the latent appraisals of the observed data, which can be approximated by the point estimates of the fit's response biases. The check serves for assessing absolute model fit of the appraisal variance, and makes important the appropriate recovery of both the precision parameters as well as response biases. In a large simulated data study of 200 data sets, the check has been found useful in distinguishing appropriate recovery of the appraisal variance and related parameters, as well as $\lambda_k = 1$ versus $\lambda_k \neq 1$ fits of the model; the DIC was also found to be rather effective for distinguishing the latter in cases of clear signal.

Exemplary demonstrations of using the VDI in a posterior predictive check to assess model fit of appraisal variance across items, as well as distinguish between the $\lambda_k = 1$ versus $\lambda_k \neq 1$ cases of the model, are included in the CRM analysis of both simulated and real data contained in Section 4. Furthermore, the current paper extends the VDI check to the multiple culture case, and is also the first multicultural CCT paper to include heterogeneous item difficulty for each culture. It accomplishes the VDI check extension to multiple cultures by calculating a separate VDI statistic for each component cluster, v, and clustering the real data by the mode estimated membership of each respondent, Ω_i; thus VDI($\mathbf{X}$) becomes VDI($\mathbf{X}_v$). Examples of this check are also included in Section 4.

2.3. Hierarchical specification of the CRM

In this section, the CRM is extended hierarchically such that the item and respondent traits are each considered to be samples from population (or the cultural) distributions. Hierarchical cognitive modeling has recently taken a rise in practice, and the advantages of this approach are thoroughly discussed by Lee (2011). In hierarchical modeling, population distributions are specified for the regular parameters using hyperparameters. These hyperparameters are estimated from their own distributions, and can represent the central tendency and/or variability within each trait across items or respondents, which may be unique to each data set. The hierarchical specification of the CRM set by Axioms 1–5 is provided below with a following explanation.

$$\begin{aligned}
T_{vk} &\sim \text{Normal}(\mu_{T_v}, \tau_{T_v}) && \text{Item Location Values}\\
\log(\lambda_{vk}) &\sim \text{Normal}(\mu_{\lambda_v}, \tau_{\lambda_v}) && \text{Item Difficulty}\\
\log(E_i) &\sim \text{Normal}(\mu_{E_{\Omega_i}}, \tau_{E_{\Omega_i}}) && \text{Informant Competency}\\
\log(a_i) &\sim \text{Normal}(\mu_{a_{\Omega_i}}, \tau_{a_{\Omega_i}}) && \text{Informant Scaling Bias}\\
b_i &\sim \text{Normal}(\mu_{b_{\Omega_i}}, \tau_{b_{\Omega_i}}) && \text{Informant Shifting Bias}\\
\Omega_i &\sim \text{Categorical}(\pi) && \text{Group Membership}\\
\pi &\sim \text{Dirichlet}(L) && \text{Pr. of Group Membership.}
\end{aligned}$$

As the item values, T_{vk}, and shifting biases, b_i, are located on the real line, a natural population distribution is the normal distribution, parameterized with a mean and precision (inverse variance). The other model parameters, which are each located on the positive half-line: λ_{vk}, E_i, and a_i, are log-transformed to the real line and also assumed to be sampled from a normal population distribution. This approach may be particularly useful for the scalar parameters, λ_{vk} and a_i, in which equal distances on the logarithmic scale provide for proportional scaling on the positive half-line. Next, note that the three primary informant parameters remain singly-indexed by i, through a technique in which their sampling distribution is specified by their group membership, Ω_i. The probability of being in any given group is assigned a Dirichlet prior, and this allows for varying probabilities of being in any given group, of V groups; note that V is pre-specified.

In order to estimate and fit the model to data, a hierarchical Bayesian framework is employed here. Prior distributions are set for a number of hyperparameters, while others are set to fixed values in order to identify the model. The hierarchical settings used for the model are:

$$\begin{aligned}
\mu_{T_v} &\sim \text{Normal}(0, 0.25) & \tau_{T_v} &\sim \text{Gamma}(0.01, 0.01)\\
\mu_{\lambda_v} &= 0 & \tau_{\lambda_v} &\sim \text{Gamma}(0.01, 0.01)\\
\mu_{E_v} &\sim \text{Normal}(0, .01) & \tau_{E_v} &\sim \text{Gamma}(0.01, 0.01)\\
\mu_{a_v} &= 0 & \tau_{a_v} &\sim \text{Gamma}(0.01, 0.01)\\
\mu_{b_v} &= 0 & \tau_{b_v} &\sim \text{Gamma}(0.01, 0.01).
\end{aligned}$$

These prior distribution settings are suggested for data that are collected in (0, 1), and transformed using the inverse logit; for other types of applications, the researcher may want to use different prior settings. In this case, rather diffuse priors are set for each of the hyperparameters, while others are set to definite values in order to identify the model. Given that the model allows μ_{T_v} and τ_{T_v} to vary, this introduces an identification issue with the response bias hyperparameters: for example, without setting a static, neutral value for the means of the biases, μ_{a_v} could be used to fit τ_{T_v}, and correspondingly, μ_{b_v} could be used to fit μ_{T_v}; thus this issue is addressed by assigning neutral values, 1 and 0 respectively, for the mean of each bias. In addition, the Rasch model (1960) usage for item difficulty described in Section 2 also introduces an identification issue that is likewise adequately handled by setting

the mean item difficulty to neutral, $\mu_{\lambda_v} = 1$. Finally, researchers may consider adjusting the priors to suit particular aims. For example, within finite mixture model applications ($V > 1$), it has been found that informative priors on τ_{T_v} and τ_{λ_v}, such as a uniform that sets a minimum variance on the item values or difficulties, may serve beneficial; such informative priors may increase the probability for meaningful clusters to be located.

3. Hierarchical Bayesian analysis with the CRM

3.1. Applying the CRM

The model is applicable with standard inference software such as JAGS (Plummer, 2003) or Openbugs (Thomas, O'Hara, Ligges, & Sturtz, 2006). The present paper applies the model with JAGS called from the R interface by using R packages: rjags and R2jags (Plummer, 2012); the JAGS model code used for the CRM is included in Appendix C, which uses the specifications of the model in Section 2.3. In addition, software for model application with a Graphic User Interface (GUI) is available as an R package called 'CCTpack' (Anders, 2013). The analyses provided next, as well as others, have found that generally, three chains of 12,000 samples and 4000 burn-in were satisfactory specifications for obtaining adequate mixing, as assessed by trace plots and convergence diagnostics, such as appropriate $\hat{R}$ values less than 1.1 (see Gelman, Carlin, Stern, & Rubin, 2004, for an explanation of MCMC sampling terms)[5]; cases of parameter autocorrelation have been found to be well-handled by appropriate thinning, and additional samples have been found to improve the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) statistic, and occasionally improve parameter recovery.

3.2. Posterior predictive checks and model selection

In Sections 2.2.1 and 2.2.2, two properties were introduced that serve as a basis for posterior predictive checks. These properties pertain to determining if the consensus structure of the respondents is matched by the model fit (Axiom 1), and if response precision is appropriately fit by the model (Axiom 4), which may distinguish whether heterogeneous or homogeneous item difficulty should be included. For the former, the check involves plotting the series of eigenvalues of the real data and verifying whether the series of eigenvalues from many randomly-sampled posterior predictive data sets show the same pattern. As the eigenvalue posterior predictive check is known as what is called a graphical posterior predictive check (Gelman et al., 2004), it involves judgment of the researcher to determine if the check is appropriately satisfied. Fig. 4 contains examples in which the check is considered passed (third and fourth plots), as well as failed (first and second plots); and is further explained in Section 4.1.

The second posterior predictive check, for response precision across items, involves calculating the distribution of VDI statistics from the posterior predictive appraisals, and comparing it to the VDI of the latent appraisals of the real data, which is approximated by using the point estimates of the response biases. This is performed separately for each cluster retrieved. The model that better fits the response precision should result in a posterior VDI distribution in which the data VDI statistic is more likely to have arose from. As in other CCT applications of this test, one could also see if the statistic falls between a lower and upper percentile of the VDI distribution. The posterior distribution of VDI statistics from the CRM is often a right-skewed (or positively-skewed) distribution, with a mode or median usually between the 25th and 40th percentiles. In cases of the CRM's right-skewed VDI distribution, generally VDI statistics that fall between the 5th and 80th percentiles have been found to signal an appropriate model fit of the VDI; if the distribution is non-skewed, one may consider the 10th and 90th percentiles. Note that in the multicultural case ($V > 1$), the cultures of the real data are obtained by the mode posterior memberships of the model fit. The VDI check is demonstrated in Fig. 6 and in the right plots of Figs. 8 and 10; 6 and 10 are multicultural while 8 is single-cultured. Both of these checks, consensus structure and VDI, are calculated using 500 randomly-sampled data matrices from the posterior predictive data.

3.3. Generalized routine for fitting the model to data

The suggested, generalized approach for fitting the CRM to data includes the following procedure. First, the researcher should obtain the scree plot of the Pearson correlations of the data, to inspect how many cultures the scree plot suggests. If the data seem to be single-cultured, $V = 1$, then the $\text{CRM}^{V=1}_{\lambda_k=1}$ can be applied. Conversely, if the data seem to be multi-cultured, $V > 1$, then the $\text{CRM}^{V>1}_{\lambda_{vk}=1}$ can be applied, yet across a small range of V values, near the number of apparent significant factors in the scree plot.[6] Then the eigenvalue check should be run, and the model with the smallest V value that satisfies the check should be used. To check response precision fit, the VDI check can be used, as well as to distinguish between the two item difficulty cases of the model ($\lambda_{vk} = 1$ versus $\lambda_{vk} \neq 1$); the DIC can also be used, such as if both models satisfy the VDI check comparably. When the appropriate model is selected by this procedure, then it is considered suitable to examine the posterior distributions of the fit. An R package named 'CCTpack' (Anders, 2013) has been developed to facilitate application of the model, and can perform the functions above.

4. CRM application to data

In this section, the CRM is applied to both simulated and real data sets via hierarchical Bayesian inference. The simulated data application in Section 4.1 demonstrates the posterior predictive checks for distinguishing fits of the model that use different numbers of culture (V) and/or homogeneous versus heterogeneous item difficulty; it also shows the parameter recovery that one can expect using sparse data for 3 cultures (20 informants per culture, on 30 items). The real data application in Section 4.2, applies the model to two real data sets that were collected, one that pertains to consensus beliefs of probability forecasting and another to beliefs on healthy living/eating. Simulated data were produced within the logit scale space, while real data were collected within (0, 1) and then transformed to the logit scale space.

[5] In respect to the discrete parameters (the Ω_i of the $\text{CRM}^{V>1}$), inspection of their trace plots to assess that the chains have converged on similar distributions, may be more preferred than the $\hat{R}$ diagnostic; as the $\hat{R}$ diagnostic may have difficulty in properly assessing convergence for discrete parameters whose chains have a high likelihood to converge to distributions with zero or approximate zero variance (as driven by a strong signal in the data).

[6] As a characteristic of being a finite mixture model, label-switching and mixing phenomena (Stephens, 2000) are possible in the $\text{CRM}^{V>1}$, which need to be addressed prior to calculating convergence, model comparison statistics, and posterior predictive checks. For more information on handling these, see Section 3 in Anders and Batchelder (2012). The R package 'CCTpack' works to address these phenomena, should they arise.

4.1. Simulated data analysis with the CRM

In this section, the results of the CRM applied to data it has simulated are presented. First, the posterior predictive checks of the CRM are demonstrated on the data, and then parameter recovery. The data utilized in this section were simulated with $V = 3$ cultures and heterogeneous item difficulty, $\lambda_{vk} \neq 1$; using a somewhat sparse size of 20 informants per culture and 30 items. The generating hierarchical parameters were randomly drawn using uniform distributions, and were respectively: $\mu_{T_v} = [-0.47, -0.36, 1.17]$, $\tau_{T_v} = [0.76, 0.45, 0.73]$, $\tau_{\lambda_v} = [10.7, 10.1, 10.3]$, $\mu_{E_v} = [-0.44, -0.46, -0.44]$, $\tau_{E_v} = [5.9, 6.0, 5.4]$, $\tau_{a_v} = [8.9, 9.1, 8.5]$, $\tau_{b_v} = [9.6, 6.7, 6.8]$.

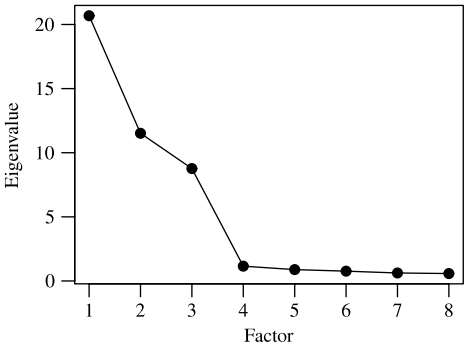

Fig. 3. Scree plot of the data simulated with the $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$.

Beginning with the scree plot of the data in Fig. 3, indeed it suggests a three culture fit, as there appears to be three substantial factors, in which the subsequent factors are significantly smaller and approximately linearly decreasing, indicating that they may primarily be fitting noise. Thus we not only apply the CRM with three cultures ($V = 3$) to the data, but also apply the model with $V = 1, 2$, and 4 latent truths to demonstrate the results.

Fig. 4 contains the cultural posterior predictive check, which pertains to selecting the appropriate specification of V, for each of these four runs. One can see that the $\text{CRM}^{V=1}$ only captures the first substantial eigenvalue of the data, the $\text{CRM}^{V=2}$ only captures the first two substantial eigenvalues of the data but not the third, and only the $\text{CRM}^{V=3}$ and $\text{CRM}^{V=4}$ successfully capture the full pattern

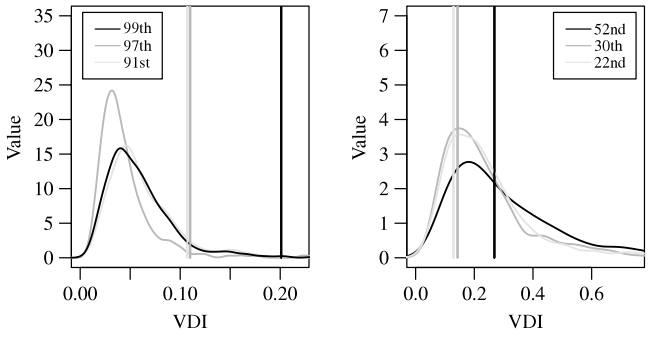

Fig. 6. The VDI posterior predictive checks for the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ (left) and $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$ (right) applied to $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$-generated data. The black, dark gray, and light gray each depict the VDI distributions of each culture, and the vertical lines are the VDI of the data clustered by the mode e_i membership parameters.

of eigenvalues.[7] In this case we know the correct specification is $V = 3$, and while the $\text{CRM}^{V=4}$ is over-specified, it simulated 3 cultures by clustering all of the informants to only 3 cultures, leaving the 4th culture empty across all samples (because of the clear signal). Since the $\text{CRM}^{V=3}$ is the simplest model that can fit the eigenvalue pattern, we proceed with it.

Next, the VDI check for item difficulty is performed to distinguish whether the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ or $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$ are appropriate for the data, and the results are contained in Fig. 6. The left plot of the figure shows that the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ poorly fits the VDI of the data for each culture, respectively at the 99th, 97th, and 91st percentiles, while the $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$ does a better job respectively at the 52nd, 30th, and 22nd percentiles; also, its distributions' central tendencies are positioned closer to the VDI values of the data than those of the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$. Thus we select the $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$. The DIC supports this decision by being lower for the $\text{CRM}^{V=3}_{\lambda_{vk}\neq 1}$ with a value of 3762.9 versus 3974.6.

With the most appropriate specifications of the model being selected by the posterior predictive checks, the parameter recovery is inspected. Fig. 5 contains five plots which respectively correspond to the recovery of the parameters T_{vk}, λ_{vk}, E_i, a_i,

[7] Whether one fits the data with item difficulty or not for each model, it did not change the result of whether it could satisfy the eigenvalue pattern.

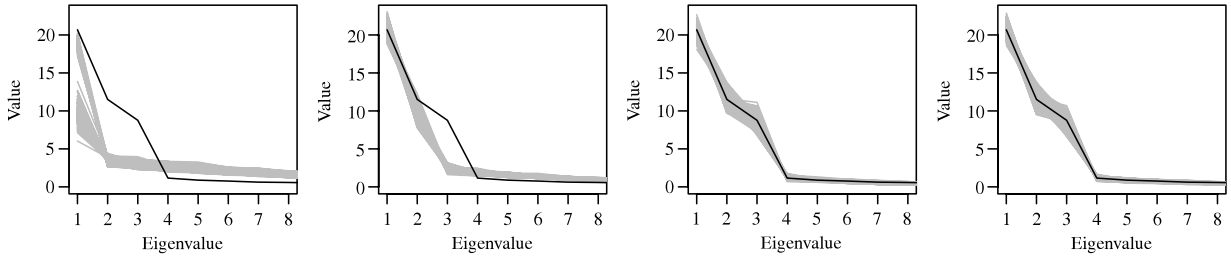

Fig. 4. Eigenvalue posterior predictive checks respectively for the $\text{CRM}^{V=1}$, $\text{CRM}^{V=2}$, $\text{CRM}^{V=3}$, and $\text{CRM}^{V=4}$ fit to the $V = 3$ simulated data. Each check is performed with 500 randomly-sampled posterior predictive data sets.

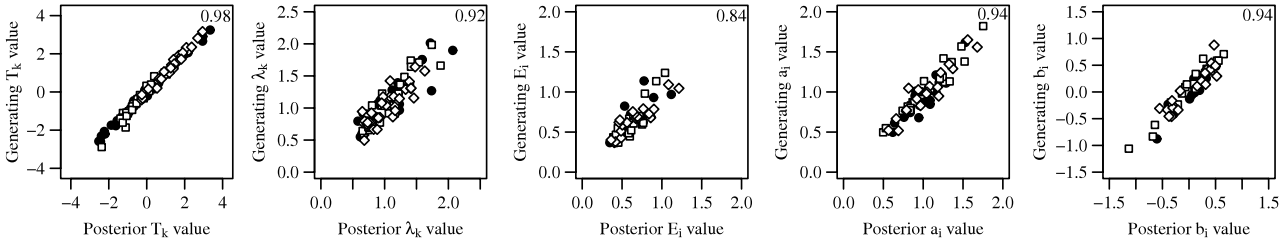

Fig. 5. Scatterplots (posterior mean value by generating value) depicting the recovery for each of the CRM parameters for $V = 3$ simulated data; circles, squares, and diamonds represent each of the 3 cultures respectively.

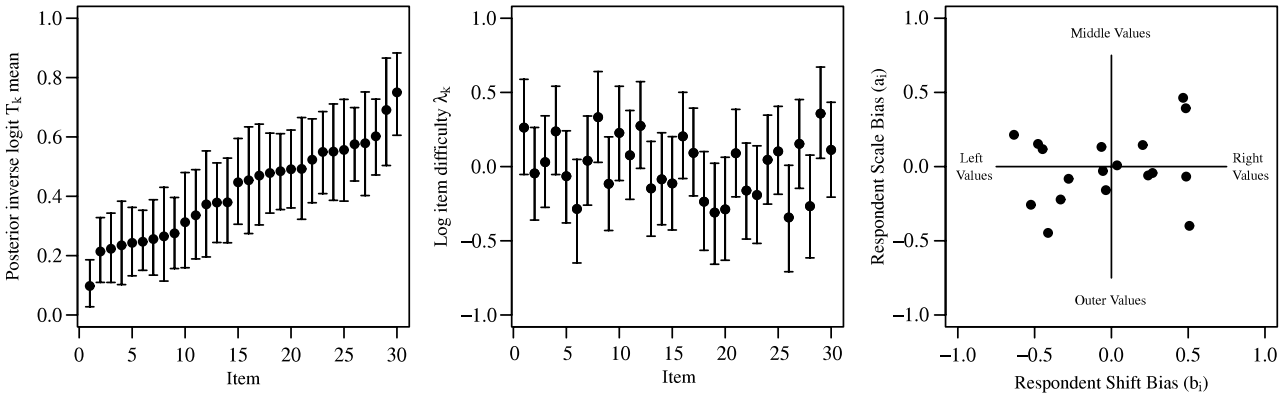

Fig. 7. Posterior mean results of the $CRM^{V=1}_{\lambda_k \neq 1}$ fit to the probability forecasting data set. The inverse logit consensus probabilities, T_k, (left) and log item difficulties, λ_k, (middle) with 95% HDI's, along with individual response biases via a_i and b_i plotted against each other for each informant (right).

b_i; the black square, white circles, and diamonds, represent the three different cultures and group memberships, related to the Ω_i parameter. The model clustered all informants correctly to their latent memberships from the simulated data; the mode of each of their estimated Ω_i values was equal to their generating Ω_i value, with very low posterior variance. For this sparse data, the recovery was strongest for the latent consensus truths T_{vk}, and response biases a_i, and b_i, with Pearson correlations between 0.94 and 0.99, while recovery was less precise for the response precision parameters, E_i and λ_k at 0.92 and 0.84.

The results demonstrated in this section are representative of larger studies performed over many data sets of similar size. In these studies, the posterior predictive checks are able to distinguish the appropriate model as demonstrated here, the 95% HDI's of each parameter overlap the generating values 90% or more of the time, and the Pearson correlations for parameter recovery are comparable across data sets that also have reasonable signal, and sufficient variance within parameters. It has been found that increases in the number of respondents and informants improve these recovery statistics. In addition, the model has been found able to recover cultures with as few as 6 members per culture, and 30 items, or 25 members and 9 items.

4.2. Real data analysis with the CRM

The CRM is demonstrated here on two collected data sets, the first pertains to forecasting via probability judgments, and the second pertains to beliefs on healthy living/eating. The suggested approach for appropriately fitting the model to data described in Section 3.3 is carried through for each real data set; and to ease interpretation of the results, the inverse logit of the T_k posterior values is reported during the analysis.

Experiment 1: Forecasting

In this study, $N = 18$ informants rated the likelihood of $M = 30$ events to occur using a continuous slider response format that indicated probabilities from 0 to 1, and allowed for increments as small as 0.01.[8] The respondents consisted of University of California, Irvine students, whose ages ranged from 18 to 32 with a mean age of 21.2 and a median age of 20. The survey tool used to collect data was from SurveyPlanet,[9] and the students received credit for completing the survey. The items of the survey are provided in Appendix D.

[8] For the analysis, responses at exactly 0 and 1 were replaced with 0.001 and 0.999 respectively. This likewise done for the health data analyzed in the next section.

[9] Accessible at the time of this paper here: http://surveyplanet.com.

The scree plot from the data, depicted by the black line in the first plot of Fig. 8, strongly suggests a single-factor design with a large eigenvalue followed by a sharp decline to a second factor, which begins a linearly decreasing trend. Thus the $CRM^{V=1}_{\lambda_k=1}$ and $CRM^{V=1}_{\lambda_k \neq 1}$ were fit to the data and compared.

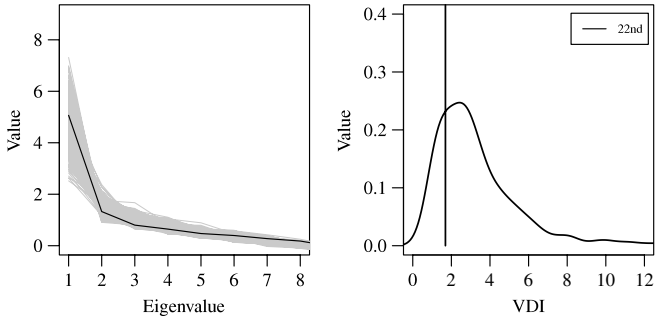

Fig. 8. The posterior predictive checks for the $CRM^{V=1}_{\lambda_k \neq 1}$ fit to the probability forecasting data set.

The analysis of the data is begun with an inspection of the posterior predictive checks for appropriate model fit of each application. Both models were found to satisfy the eigenvalue check. However, the $CRM^{V=1}_{\lambda_k \neq 1}$ satisfied the VDI check markedly better with the VDI statistic at the mode of the predictive distribution at the 22nd percentile, while the $CRM^{V=1}_{\lambda_k=1}$ had a much further distance at the 80th percentile; the DIC was also lower for the former at 2028.2 versus 2048.4. These successful eigenvalue and VDI checks for the $CRM^{V=1}_{\lambda_k \neq 1}$ are respectively shown in the two plots of Fig. 8. Given the better fit of the $CRM_{\lambda_k \neq 1}$, its posterior results are examined.

The left plot of Fig. 7 contains the posterior mean inverse logit T_k, which is interpreted as the consensus beliefs of the true probability for each event occur, sorted from lowest to highest, with 95% HDI's. In the figure, one can see that the model estimated very few of the items to be extremely probable or improbable (near 0 or 1). It was found that the least probable events pertained to a soon-to-arrive major world war/terrorist attack, or cure for a serious health illness, such as Alzheimer's or HIV, to be implemented. The most probable events pertained to greater income disparity between the rich and poor, greater income taxes on average income, and same-sex marriage becoming federally recognized by the end of president Obama's term in the United States.

The posterior mean item difficulty values, λ_k, with 95% HDI's are contained in the middle plot of Fig. 7, and are in the same ordering of the T_k. One can see that judging the consensus probability of

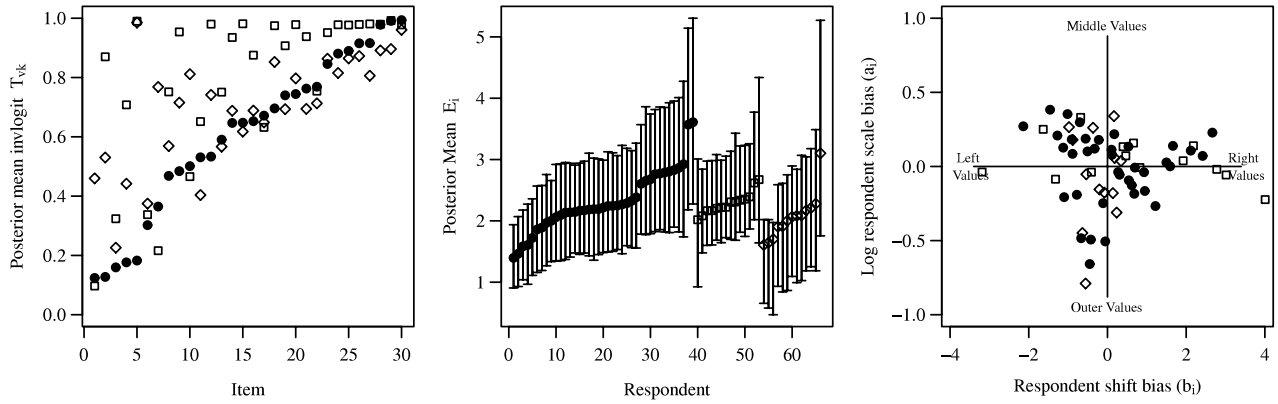

Fig. 9. Posterior mean results of the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ fit to the health beliefs data set. The inverse logit degree of importances, T_{vk}, (left), informant competencies, E_i, with 95% HDI's (middle), and individual response biases via a_i and b_i plotted against each other for each informant (right). Black circles, white squares, and diamonds, respectively denote the three separate cultures' truths, and the informant memberships to each truth.

various items was more difficult for some items than others. For example, item 26 pertaining to US firearm restrictions with mean $T_k = 0.58$ was one of the easiest to know at mean log $\lambda_k = -0.34$ while item 29, the consensus belief about same-sex marriage in the US with mean $T_k = 0.69$, was one of the hardest to know with mean log $\lambda_k = 0.36$; the standard deviation of these log mean λ_k values across items, was 0.21.

The right plot in Fig. 7 shows the measurement of each informant's individual bias profile (a_i plotted by b_i). The quadrants depict the extent to which each rater prefers using positions on the scale, such as outer and right probability values, versus middle and left probability values, for example. As for the informants' competency on the task, the standard appraisal errors ranged from 1.2 to 2.4 with mean 1.5 and standard deviation 0.28.

Experiment 2: Health beliefs

In this study, $N = 66$ respondents rated the degree of importance of $M = 30$ lifestyle/diet items for healthy living. Responses were collected using tools from SurveyGizmo,[10] which involved a continuous slider response format with values from 0 to 100%, with the label "Percent Important for Good Health". A link to the survey was posted online to gather responses. Five percent of the respondents were between the ages of 18 and 24, 70% between 25 and 34, 10% between 35 and 44, and the remaining 15% ranged up to 70 years old. The items of the survey are provided in Appendix E.

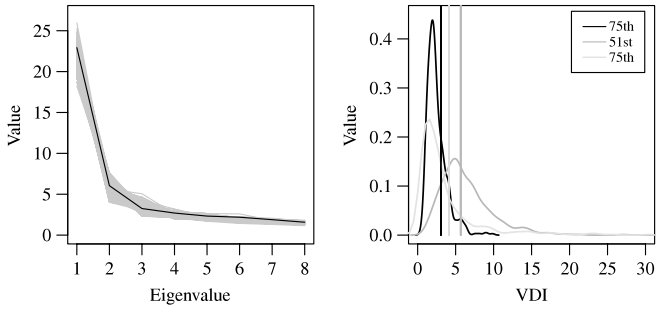

Fig. 10. The eigenvalue (left) and VDI (right) posterior predictive checks for the $\text{CRM}^{V=2}_{\lambda_{vk}=1}$ fit to the health beliefs data set. The black and dark grey each depict the VDI distributions of each culture, and the vertical lines are the VDI of the data clustered by the mode e_i membership parameters, which are respectively at the 77th and 28th percentiles of the corresponding VDI distributions.

[10] Accessible at the time of this paper here: http://surveygizmo.com.

The scree plot of the data, from the black line in the left plot of Fig. 10, at first glace gives sign of two cultures, in which the second culture may be small in number to the first, or its truth is highly correlated to the other, since the magnitude of the second factor is relatively small to the first, this property of the factors was discussed in Section 3.1. The $\text{CRM}^{V=1}$, $\text{CRM}^{V=2}$, and $\text{CRM}^{V=3}$ were each applied to the data, and their performance on the posterior predictive checks was evaluated. In summary, the $\text{CRM}^{V=2}$ and $\text{CRM}^{V=3}$ satisfied the consensus posterior predictive check while the $\text{CRM}^{V=1}$ (DIC 9044.3) failed the check. Then the VDI check was failed by the $\text{CRM}^{V=2}_{\lambda_{vk}=1}$ (DIC 9001.2), but not by the $\text{CRM}^{V=2}_{\lambda_{vk}\neq 1}$ (DIC 9447.4) and $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ (DIC 8996.0). However, the 2nd factor of the consensus posterior predictive check was consistently underestimated by the $\text{CRM}^{V=2}_{\lambda_{vk}\neq 1}$ while the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ satisfied the factors the most appropriately; and it appears there is a third meaningful factor in the plot, albeit very small in proportion, and not grossly larger than the immediately proceeding ones. Thus as the $\text{CRM}^{V=3}_{\lambda_{vk}=1}$ satisfied both checks, as depicted in Fig. 10, adequately over the other fits, it was chosen for analysis. In this case, three clusters explain the scree plot, each of which have substantially correlated truths: Pearson r's between 0.56 and 0.68, and two of which are rather small in group size proportion, having only 13 or 14 out of 66 total respondents.

The left plot of Fig. 9 contains the posterior mean inverse logit item values for each of the three cultures recovered, and are ordered by the majority culture's beliefs (black circles). This majority culture seems to generally find the items less important for health on average, than the other cultures. Some of its lowest T_k value beliefs: item 5 on avoiding GMOs, and item 2 on whether one should periodically detoxify the body, are in largest disagreement with the other two cultures, which find these much more important. Instead, the majority culture's top three important beliefs for healthy living involve eating fruits and vegetables frequently (item 28), exercising regularly (item 29), and not smoking (item 30). Then, the square culture may be generalized as a more hyper health-conscious culture than the majority culture, is highly against GMOs, and is more against animal products for good health. In contrast, the diamond culture selectively emphasizes certain actions for health, and instead favors animal products for good health, yet is also highly against GMOs. On a post-survey question regarding how concerned the informants are about healthy living, those clustered in majority culture 1 (black circles) were the least concerned with mean 77% and median 80%, in culture 2 (white squares) were the most concerned with mean 87% and median 93%, and in culture 3 (white diamonds) were between the two with mean 83% and median 86%.

The middle plot of Fig. 9 contains the clustering of informants estimated by the model. Here the clustering is presented by the mode membership, Ω_i, that each informant had. The plot includes the standard appraisal errors, E_i, of each informant, sorted from lowest to highest and separated by cluster. Clusters 1 and 2 had comparable average standard appraisal errors, and Cluster 3 had the lowest average standard appraisal errors at $\bar{\mathbf{E}}_v = [2.30, 2.31, 2.06]$, $\text{sd}(\mathbf{E_v}) = [0.5, 0.2, 0.4]$, generally these standard errors are higher on average by 0.7 than in the forecasting data. The level of competency for these differently-grouped informants may also be used to understand how well they fit into the cluster they were assigned to. Thirty-nine informants were clustered into culture 1, 14 were clustered into culture 2, and 13 were clustered into culture 3. Less than half of the informants clustered in culture 3 were US residents, while more than half of the informants clustered into the other cultures were US residents.

The right plot of Fig. 9 provides the individual response bias behaviors of the informants, with posterior mean a_i and b_i, and their cultural membership, from mode Ω_i. The quadrants depict the extent to which the members of each culture prefer different locations of the scale. One can see that members of the 2nd culture had some of the strongest shifting biases (preferring one side of the scale), while some members of the 1st and 3rd culture had some of the strongest re-scaling biases (preferring outer values). While there were strong outlying biases to prefer outer values, there were not strong outlying biases to prefer middle values. On average, the least extreme biases were among culture 3.

5. Discussion

A CCT approach was developed for continuous response data, leading to the CRM. The model can be applied to informant-by-item continuous response data, where each informant is assumed to share consensus knowledge about the location of each item on a latent item trait. The model estimates culturally-shared item values, T_{vk}, the culturally-shared difficulty of knowing each item, λ_{vk}, each informant's knowledge competency within their culture, E_i, and the response bias tendencies of each informant, a_i and b_i. The model is one of the first to handle multicultural continuous data, it can detect V number of subgroups (cultures) and cluster informants into each group, as denoted by the membership parameter, Ω_i. For each culture, different population parameters are estimated for both the items and informants.

The CCT approach for the continuous data case includes a generalized method for making cultural measurements from continuous data, which involves fitting the CRM and assessing whether the fit is appropriate. The methodology is based on an exploration of the properties of the model, as well as key assumptions of CCT: a most notable one is that the set of informants is either composed of one group which shares a cultural truth, or of subgroups that share different cultural truths. The methodology includes the assessment of model fit with two useful posterior predictive checks, which are based on mathematical properties of the model. Specifically, the eigenvalue posterior predictive check serves to verify whether the model, with a specified number of possible cultures, predicts data with a consensus structure similar to the actual data, and the VDI check assesses fit of the response precision, which may also be used to determine whether the model should account for heterogeneous item difficulty.

The CRM was applied to both simulated and real data sets, and was demonstrated to recover the parameters of simulated data well, even when working with sparse data for the tri-cultural case, and provide interesting and plausible interpretations of the real data. The CRM was demonstrated as a powerful tool that is versatile to provide an analysis for a number of different applications, such as probability forecasting, and beliefs for healthy living, though many more varieties are possible. The T_{vk} values represent the consensus belief for culture v, the λ_{vk} can reveal which items are the most difficult to judge accurately for each culture, the E_i can reveal which informants are more knowledgeable, a_i and b_i reflect each informant's response biases, and Ω_i gives the cultural clustering of the informants.

While the fitting and estimation approaches for the CRM developed here have shown to be effective and useful, the exploration of additional techniques to fit the model, such as Differential Evolution MCMC (DE-MCMC, see Ter Braak, 2006; Turner, Sederberg, Brown, & Steyvers, 2013b) or ways to better optimize the current framework (e.g., reparameterization), are of interest to be explored in future work; as there is room for improvement in the current approach, and in further refining mixing quality. As a finite mixture model that seeks to assign clusters, appropriately fitting the CRM becomes increasingly difficult as one increases the number clusters, V, and/or adds item difficulty $\lambda_{vk} \neq 1$. In addition, one could develop an algorithm that samples between separate CRM models specified at each V level, and estimates the most appropriate V, rather than using the posterior predictive check approach. Nonetheless, the current work has proven to be a large advance in CCT model-based clustering for continuous data.

The CCT tradition, and approach developed here for continuous responses, is specialized and appropriate for consensus situations to which culture is relevant or sought to be observed. The present paper considerably developed the CCT approach for continuous data, by establishing a new model that is specialized for the aims of CCT, and which can account for multicultural data. In addition, the approach includes two important posterior predictive model checks that address issues which are of importance to CCT: assessing appropriate fit of consensus structure, and response precision across items. A GUI was developed to facilitate model application, and the calculation of these posterior predictive checks. All in all, this work has shown to be a valuable addition to CCT, and its applicability to continuous data.

Acknowledgments

Work on this paper was supported by a grant to the second author from the Army Research Office (ARO), as well as an appointment administered by the Oak Ridge Institute for Science and Education (ORISE) to work with Dr. Michael Young at the U.S. Air Force Research Laboratory in Wright Patterson Air Force Base.

Appendix A. Spearman law property with proof

Theorem 1. *Suppose the Axioms 1–5 hold for the CRM. Then given fixed values of* $\mathscr{T}$, $\mathbf{E}$, $\mathscr{L}$, *and* $\mathbf{\Omega}$: $\forall i, j = 1, \ldots, N \ni i \neq j$,

$$\rho(X_{iK}, X_{jK}) = \rho(X_{iK}, T_{\Omega_i K})\rho(X_{jK}, T_{\Omega_j K})\rho(T_{\Omega_i K}, T_{\Omega_j K}), \tag{A.1}$$

where K is a random variable representing item indices, with probability density: $Pr(K = k) = 1/M,\ \forall k = 1, \ldots, M$.

Proof. Note that $Y_{ik} = T_{\Omega_i k} + \epsilon_{ik}$ for the CRM, and that $\forall i, k$, $E(\epsilon_{ik}) = 0$. Further, by Axiom 3, conditional independence requires that all of the $(\epsilon_{ik})_{i=1}^{N}$ are conditionally independent for fixed $\mathscr{T}$, $\mathbf{E}$, $\mathscr{L}$, and $\mathbf{\Omega}$. From these assumptions, the terms in (A.1) can be calculated as follows. First note that with the matrix of latent appraisals, $\mathbf{Y} = (Y_{ik})_{N\times M}$, the correlation between two informants over items is

$$\begin{aligned} rho(Y_{iK}, Y_{jK}) &= \frac{\text{Cov}(Y_{iK}, Y_{jK})}{\sqrt{\text{Var}(Y_{iK})\text{Var}(Y_{jK})}} \\ &= \frac{E(Y_{iK}Y_{jK}) - E(Y_{iK})E(Y_{jK})}{\sqrt{\text{Var}(Y_{iK})\text{Var}(Y_{jK})}}. \end{aligned} \tag{A.2}$$

Next, note that

$$E(Y_{iK}Y_{jK}) = E_K[E(Y_{iK}Y_{jK} \mid K)]$$
$$= \frac{1}{M}\sum_{k=1}^{M} E[(T_{ik}+\epsilon_{ik})(T_{jk}+\epsilon_{jk})]. \quad (A.3)$$

Now from the zero mean and conditional independence properties of the error random variables, it is clear that (A.3) becomes

$$E(Y_{iK}Y_{jK}) = \frac{1}{M}\sum_{k=1}^{M} T_{ik}T_{jk}.$$

Similarly other aspects of (A.2) can be computed, such as

$$E(Y_{iK}) = \frac{1}{M}\sum_{k=1}^{M} E(T_{ik}+\epsilon_{ik}) = E(T_{iK}),$$

so the result is

$$\rho(Y_{iK}, Y_{jK}) = \frac{\mathrm{Cov}(T_{iK}, T_{jK})}{\sqrt{\mathrm{Var}(Y_{iK})\mathrm{Var}(Y_{jK})}}. \quad (A.4)$$

Next consider the terms on the right side of (A.1). Using the same methods, it is easy to calculate

$$\rho(Y_{iK}, T_{iK}) = \frac{\sum_{k=1}^{M} T_{ik}^2/M - E^2(T_{iK})}{\sqrt{\mathrm{Var}(Y_{iK})\mathrm{Var}(T_{iK})}} = \sqrt{\frac{\mathrm{Var}(T_{iK})}{\mathrm{Var}(Y_{iK})}},$$

$$\rho(Y_{jK}, T_{jK}) = \frac{\sum_{k=1}^{M} T_{jk}^2/M - E^2(T_{jK})}{\sqrt{\mathrm{Var}(Y_{jK})\mathrm{Var}(T_{jK})}} = \sqrt{\frac{\mathrm{Var}(T_{jK})}{\mathrm{Var}(Y_{jK})}},$$

where the computational formula for the variance of a random variable, $\mathrm{Var}(X) = E(X^2) - E^2(X)$, is used. Finally, the third correlation is obtained as

$$\rho(T_{iK}, T_{jK}) = \frac{\mathrm{Cov}(T_{iK}, T_{jK})}{\sqrt{\mathrm{Var}(T_{iK})\mathrm{Var}(T_{jK})}}.$$

When these three correlations are multiplied, the result is

$$\rho(Y_{iK}, Y_{jK}) = \rho(Y_{iK}, T_{\Omega_i K})\rho(Y_{jK}, T_{\Omega_j K})\rho(T_{\Omega_i K}, T_{\Omega_j K}). \quad (A.5)$$

Next note that $X_{ik} = a_i Y_{ik} + b_i$. Then using the property that the correlation coefficient ρ is invariant under linear transformations, (A.1) is equivalent to (A.5). □

While the triple correlation property behind the latent appraisals of the CRM$^{V>1}$ is given in (A.1), note that if all informants share the same cultural truth ($V = 1$), (A.1) reduces to

$$\rho(X_{iK}, X_{jK}) = \rho(X_{iK}, T_K)\rho(X_{jK}, T_K). \quad (A.6)$$

Appendix B. Item difficulty property with proof

Theorem 2. *Suppose the Axioms 1–5 hold for the CRM and item difficulty is neutral* ($\lambda_k = 1$). *Then for fixed* **T** *and* **E**,

$$\forall k = 1, \ldots, M, \quad V_k^* = \sum_{i=1}^{N}(E_i\lambda_k)^2 = \sum_{i=1}^{N} E_i^2.$$

Proof. First, using conditional independence, for any item k:

$$V_k^* = \mathrm{Var}\left(\sum_{i=1}^{N} Y_{ik}\right) = \sum_{i=1}^{N}\mathrm{Var}(Y_{ik})$$
$$= \sum_{i=1}^{N}\mathrm{Var}(T_k+\epsilon_{ik}) = \sum_{i=1}^{N}\mathrm{Var}(\epsilon_{ik}).$$

From (2), $\mathrm{Var}(\epsilon_{ik}) = \sigma_{ik}^2$, and under the assumption that $\lambda_k = 1$ in (3),

$$V_k^* = \sum_{i=1}^{N} E_i^2,$$

and this is the same for all items k. □

Appendix C. JAGS model code

```
model{
   for (i in 1:n){
      for (k in 1:m){
Y[i,k] ~ dnorm((a[i]*T[k,Om[i]])+b[i],
              pow(a[i]*E[i]*lam[k,Om[i]],-2)) }}
#Parameters
   for (i in 1:n){
      Om[i] ~ dcat(pi)
      Elog[i]~ dnorm(pow(Emu[Om[i]],Etau[Om[i]])
      E[i] <- exp(Elog[i])
      alog[i] ~ dnorm(atau[Om[i]],atau[Om[i]])
      a[i] <- exp(alog[i])
      b[i] ~ dnorm(bmu[Om[i]],btau[Om[i]]) }
   for (k in 1:m){
    for (v in 1:V){
     T[k,v] ~ dnorm(Tmu[v],Ttau[v])
     lamlog[k,v] ~ dnorm(lammu[v],lamtau[v])
     lam[k,v] <- exp(lamlog[k,v])
     }}

     pi[1:V] ~ ddirch(L)

#Hyperparameters
 for (v in 1:V){
  Tmu[v] ~ dnorm(0,0.25)
  Ttau[v] ~ dgamma(0.01,0.01)
  lammu[v] <- 0
  lamtau[v] ~ dgamma(0.01,0.01)
  Emu[v] ~ dgamma(0.01,0.01)
  Etau[v] ~ dgamma(0.01,0.01)
  amu[v] <- 0
  atau[v] ~ dgamma(0.01,0.01)
  bmu[v] <- 0
  btau[v] ~ dgamma(0.01,0.01)
  L[v] <- 1
 }
}
```

Appendix D. Forecasting questionnaire

These items are ordered according to Fig. 7. Please rate the probability (0% to 100% probable) . . .

1. . . . that a major world war will occur by 2016.
2. . . . that a terrorist attack will occur in the US next year.
3. . . . that a cure for Alzheimer's disease will be implemented by 2016.
4. . . . that a cure for HIV will be implemented by 2016.
5. . . . that a major disease epidemic will occur in the US next year.
6. . . . of a major destructive earthquake occurring in California next year.
7. . . . that the US will make immigration laws more strict during Obama's term.
8. . . . that the average rate of HIV transmission will increase next year.
9. . . . that a major tsunami will hit one of the US coasts within the next 4 years.

10. . . . that national average crime rates will increase next year.
11. . . . that a woman will be elected US president next term.
12. . . . that humans will successfully visit Mars by year 2020.
13. . . . that the number of birth defects will increase next year.
14. . . . that the average worth of the US dollar will be higher next year.
15. . . . that unemployment in the US will be lower next year.
16. . . . that the growth rate of the national US debt will decrease next year.
17. . . . that the national average of pollutants or pesticides in foods will increase next year.
18. . . . that the general US stock market will be better next year than in the previous year.
19. . . . that the proportion of people diagnosed with mental diseases will be higher next year.
20. . . . that the national average of pollutants in tap water will increase next year.
21. . . . that the average rate of cancers in American people will increase next year.
22. . . . that the next elected US president will be a Republican.
23. . . . that US laws on genetically-modified foods will become more strict by 2015.
24. . . . that the average number of kids per household in the US will decrease next year.
25. . . . that the average worth of coastal homes in the US will increase next year.
26. . . . that the federal government will place greater restrictions on firearms by the end of Obama's term (2017).
27. . . . of a destructive hurricane occurring in Florida next year.
28. . . . that the next US president will increase income taxes for people with average incomes.
29. . . . that same-sex marriage will be federally recognized by the end of Obama's term (2017).
30. . . . that income gaps between the lower 25% and upper 75% will increase next year.

Appendix E. Healthy living/eating questionnaire

These items are ordered according to Fig. 9. Please rate how important it is (0%–100% important) to do the following things for good health.

1. Drinking animal milks.
2. Periodically detoxifying the body of food contaminants and pollutants.
3. Drinking plant-based milks.
4. Balancing acids/bases: trying to create a more "alkaline" state of the body.
5. Avoiding genetically-modified food (GMOs).
6. Taking vitamin or mineral supplements.
7. Eating protein from animal sources.
8. Lightly cooking foods to preserve more nutrients.
9. Eating organic foods as opposed to non-organic foods.
10. Showering daily.
11. Having a preference for eating raw vegetables/fruits (as opposed to cooked).
12. Avoiding foods with preservatives.
13. Minimizing salt in-take.
14. Avoiding sitting at a computer or with other technology for long periods of time.
15. Using natural sweeteners/sugars rather than artificial sweeteners/sugars.
16. Incorporating social activities into your weekly routine.
17. Avoiding foods associated with high cholesterol.
18. Avoiding artificial food ingredients.
19. Minimizing sugar in-take.
20. Only taking over-the-counter medicines when they are truly necessary.
21. Developing/maintaining a healthy posture when doing daily activities (sitting, exercising, driving).
22. Avoiding foods with large amounts of saturated fats.
23. Engaging in challenging and enjoyable activities.
24. Avoiding excessive alcohol consumption.
25. Having a positive disposition (or positive attitude).
26. Reducing stress and anxiety.
27. Avoiding processed snack foods (chips, cookies, bars, etc.).
28. Eating fruits and vegetables frequently.
29. Exercising regularly.
30. Not smoking.

References

Anders, R. (2013). CCTpack: cultural consensus theory applications to data. *R package version 0.9*.

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal for Mathematical Psychology*, *56*, 452–469.

Anders, R., & Batchelder, W. H. (2013). Cultural consensus theory for the ordinal data case. *Psychometrika*.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general condorcet model for dichotomous choice situations. In B. Grofman, & G. Owen (Eds.), *Information pooling and group decision making: proceedings of the second University of California Irvine conference on political economy* (pp. 103–112). Greenwich, Conn: JAI Press.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92.

Batchelder, W. H., & Romney, A. K. (1989). New results in test theory without an answer key. In Roskam (Ed.), *Mathematical psychology in progress* (pp. 229–248). Heidelberg, Germany: Springer-Verlag.

Batchelder, W. H., Strashny, A., & Romney, A. K. (2010). Cultural consensus theory: aggregating continuous responses in a finite interval. In S. K. Chai, J. J. Salerno, & P. L. Mabry (Eds.), *Social computing, behavioral modeling, and prediction 2010* (pp. 98–107). New York: Springer-Verlag.

Comrey, A. L. (1962). The minimum residual method of factor analysis. *Psychological Reports*, *11*, 15–18.

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*, 585–603.

France, S. L., & Batchelder, W. H. (2014). Maximum likelihood item easiness models for test theory without an answer key. *Educational and Psychological Measurement*, 0013164414527448.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (second ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*, 129–166.

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

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*: *Vol. 47*. Reading, MA: Addison-Wesley.

Merkle, E. C., & Steyvers, M. (2011). A psychological model for aggregating judgments of magnitude. In *Social computing, behavioral-cultural modeling and prediction* (pp. 236–243). Springer.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling.

Plummer, M. (2012). Rjags: Bayesian graphic models using mcmc. R package version 3.2.0 (http://CRAN.R-project.org/package=rjags).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institute.

Revelle, W. (2012). *Psych: procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University, R package version 1.2.1.

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.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, *6*, 583–640.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*, 795–809.

Steyvers, M., Wallsten, T. S., Merkle, E. C., & Turner, B. M. (2013). Evaluating probabilistic forecasts with bayesian signal detection models. *Risk Analysis*.

Ter Braak, C. J. (2006). A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*, 12–17.

Turner, B., & Steyvers, M. (2011). A wisdom of the crowd approach to forecasting. In *2nd NIPS workshop on computational social science and the wisdom of crowds*.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2013). Forecast aggregation via recalibration. In *Machine learning* (pp. 1–29).

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368.

Weller, S. W. (2007). Cultural consensus theory: applications and frequently asked questions. *Field Methods*, *19*, 339–368.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action and cognition. *Frontiers in Neuroscience*, *6*.