Todo list

Cultural Consensus Theory for the Evaluation of Patients' Behavior in Psychiatric Detention Centers

Don van den Bergh*[1], Stefan Bogaerts[2], Marinus Spreen[3], Rob Flohr[4], Joachim Vandekerckhove[5], Mijke Rhemtulla[6], William Batchelder[5], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam, [2]University of Tilburg, [3]Mesdag Clinic, [4]Stenden University of Applied Sciences, [5]University of California Irvine, [6]University of California Davis

Abstract

In many psychiatric detention centers, patients' mental health is monitored at regular intervals. Typically, clinicians score patients using a Likert scale on multiple criteria including hostility. Having an overview of patients' scores benefits staff members in at least three ways. First, the scores may help adjust treatment to the individual patient; second, the change in scores over time allow an assessment of treatment effectiveness; third, the scores may warn staff that particular patients are at high risk of turning violent. Practical importance notwithstanding, current practices for the analysis of mental health scores are suboptimal: evaluations from different clinicians are averaged (as if the Likert scale were linear and the clinicians identical), and patients are analyzed in isolation (as if they were independent). Uncertainty estimates of the resulting score are often ignored. Here we outline a quantitative program for the analysis of mental health scores using cultural consensus theory (CCT; Anders & Batchelder, 2015). CCT models take into account the ordinal nature of the Likert scale, the individual differences among clinicians, and the possible commonalities between patients. In a simulation, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and, as a more reasonable alternative, against often-used machine learning toolboxes. In addition, we outline the substantive conclusions afforded by application of the CCT model. We end with recommendations for clinical practitioners who wish to apply CCT in their own work.

*Correspondence concerning this article should be addressed to: Don van den Bergh, University of Amsterdam, Department of Psychological Methods, Nieuwe Achtergracht 129, 1001 NK Amsterdam, The Netherlands. E-Mail should be sent to: donvdbergh@hotmail.com. This work was supported by a Research Talent grant from the Netherlands Organization of Scientic Research (NWO) awarded to EJW (???).

Psychiatric detention centers monitor the mental health of their patients at regular intervals, typically using a method such as Routine Outcome Monitoring (ROM; de Beurs et al., 2011). A clinician, psychiatrist, or another staff member, henceforth a *rater*, scores a patient on multiple criteria. For example, a rater evaluates a patient's behavior on a variety of criteria that relate to aggressiveness. Today, such evaluations are stored so they may be used later to inform decisions. The decisions informed by these ratings can vary widely. For instance, the scores may help adjust treatment to individual patients, the change in scores over time allows for an assessment of treatment effectiveness, and the scores may warn staff that particular patients are at high risk of turning violent. Moreover, these ratings are key for a quantitative approach to monitoring and forecasting patients' behavior.

Current practices for aggregating the scores are suboptimal. Evaluations from different raters are averaged as if they are exchangeable. For example, personal communication with the staff of a psychiatric detention center suggested that clinicians are more lenient in their ratings than psychiatrists, but this information is not used to weigh their ratings. Furthermore, different patients are analyzed in isolation, as if they are independent. Any information regarding a patient's criminal offense is not accounted for in a model-based manner. In addition, any uncertainty estimates of the resulting score are usually ignored.

Here, we try to address these issues using Cultural Consensus Theory (CCT; Romney, Weller, & Batchelder, 1986; Batchelder & Romney, 1988; Batchelder & Anders, 2012). The defining characteristic of CCT is that it aims to estimate the consensus knowledge shared by raters. Hence, CCT is a promising framework for analyzing data of psychiatric detention centers, where the true state of a patient is unknown and needs to be estimated from the scores given by the raters. CCT models capture individual differences between raters and items, and pool information while accounting for these differences. However, currently available CCT models can only be applied to the data of a single patient; a limitation addressed in this paper.

The focus of this paper is to outline a quantitative program for the analysis of mental health scores using CCT. First, a CCT model for ordinal data is introduced (Anders & Batchelder, 2015). Afterward, this model is expanded step by step, in order to include more characteristics of the data, such as describing multiple patients simultaneously. We showcase the model in three simulation studies. First, we show that model parameters are retrieved accurately. Second, we demonstrate how CCT could be used to monitor patients progress over time. Third, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and, as a more reasonable alternative, against often-used machine learning toolboxes such as Random Forest (Breiman, 2001) and Boosted Regression Trees (Friedman, 2002). We showcase the substantive conclusions obtained from applying the CCT model and conclude the paper with recommendations for clinical practitioners who wish to apply CCT in their work.

Cultural Consensus Theory and Three Extensions

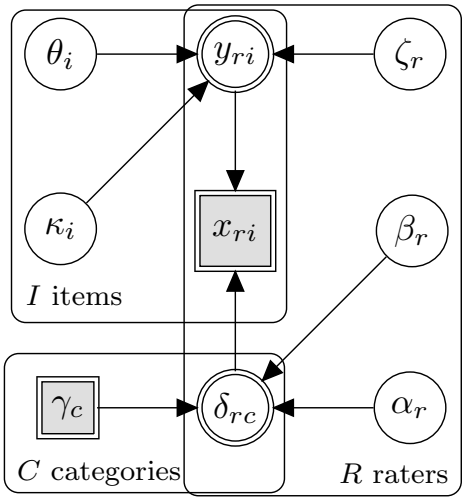
The next sections introduce Cultural Consensus Theory (CCT). First, a brief introduction to CCT is given. Afterward, the CCT model developed in Anders and Batchelder (2015, henceforth, AB) is introduced, which serves as the simplest model for a single patient. Subsequently, we generalize the model in three ways. First, the model is expanded to describe multiple patients simultaneously. Next, latent constructs are added to the model. Finally, the model is modified to include patient and rater characteristics.

Cultural Consensus Theory

Cultural Consensus Theory, also known as "test theory without an answer key" (Batchelder & Romney, 1988), is a statistical tool that attempts to retrieve the "truth" for an item by examining the consensus among the responses. For example, given a political questionnaire, there are no objectively correct answers. Instead, one could administer the questionnaire to left-oriented respondents and use CCT to find out what the consensus is among left oriented respondents. CCT models capture that some responders have a higher competency and will strictly answer according to the cultural consensus. Likewise, items can differ in their difficulty, i.e., the competence required to answer according to the consensus. For a political questionnaire, this implies that only extremely polarized respondents agree with the most polarized political statements. In addition, CCT models can be expanded to allow for multiple consensus truths, i.e., there can be multiple unknown truths that vary across subgroups of respondents (Anders & Batchelder, 2012). For a political questionnaire, the different consensuses (e.g., left, right, center, etc.) and respondents membership to these groups could be estimated from the data. The property of CCT models to estimate the consensus truth for an item from the data is ideal for psychiatric data, where a patient's true state is unknown and a consensus from the raters is desired. CCT models can be applies to continuous data (e.g., the LTM; Batchelder & Anders, 2012), binary data (e.g., the General Condorcet model; Batchelder & Romney, 1986), and ordinal data (AB, 2015). Since ratings are usually given on a Likert scale, we focus on a CCT model for ordinal data.

The Latent Truth Rater Model

As a starting point, consider the Latent Truth Rater Model (LTRM), a cultural consensus model for ordinal data introduced by AB. Figure 1 shows a graphical model of the LTRM. The LTRM captures differences among raters and items and may be viewed as the simplest model for a single patient. The rating given by rater r on item i is denoted x_{ri} and takes on discrete values from 1 through C. AB formalize the core ideas of the LTRM with 6 axioms. There is a latent, shared cultural truth among the raters, which is captured by the item location parameters θ_i (AB's axiom 1). Since raters are not perfect measurement instruments, they infer a noisy version of the cultural truth for each item, called



$$
\begin{aligned}
x_{ri} &\leftarrow \begin{cases} 1 \text{ if } y_{ri} \leq \delta_{r1} \\ c \text{ if } \delta_{r,c-1} < y_{ri} \leq \delta_{rc} \\ C \text{ if } y_{ri} > \delta_{r,C-1} \end{cases} \\
y_{ri} &\leftarrow \theta_i + \epsilon_{ri} \\
\gamma_c &\leftarrow \text{logit}\left(c/C\right) \\
\delta_{rc} &\leftarrow \alpha_r \gamma_c + \beta_r \\
\epsilon_{ri} &\sim \text{Logistic}\left(0,\ \kappa_i/\zeta_r\right) \\
\theta_i &\sim \text{Normal}\left(\mu_\theta,\ \sigma_\theta^2\right) \\
\kappa_i &\sim \text{Gamma}\left(\mu_\kappa^2/\sigma_\kappa^2,\ \mu_\kappa/\sigma_\kappa^2\right) \\
\alpha_r &\sim \text{Gamma}\left(\mu_{\alpha_r}^2/\sigma_{\alpha_r}^2,\ \mu_{\alpha_r}/\sigma_{\alpha_r}^2\right) \\
\beta_r &\sim \text{Normal}\left(\mu_{\beta_r},\ \sigma_{\beta_r}^2\right) \\
\zeta_r &\sim \text{Gamma}\left(\mu_{\zeta_r}^2/\sigma_{\zeta_r}^2,\ \mu_{\zeta_r}/\sigma_{\zeta_r}^2\right)
\end{aligned}
$$

Figure 1: Graphical model corresponding to the LTRM; a CCT model for a single patient. The hyper parameters are omitted from the graphical model. The group level means and standard deviations are denoted μ and σ respectively. The priors on the group level parameters are omitted. Gamma distributions are parametrized with shape and scale so that the group level parameters correspond to the mean and standard deviation of the distribution.

a latent appraisal and defined as $y_i = \theta_i + \epsilon_{ri}$, where $\epsilon_i \sim \text{Normal}\left(0,\ \zeta_r/\kappa_i\right)$ (AB's axiom 2). The appraisal error σ_i^ϵ has standard deviation κ_i which varies across items and raters. This reflects that items fluctuate in difficulty and that raters vary in competence (AB's axiom 3). Latent appraisals are assumed to be conditionally independent given the latent truth θ_i and the appraisal error ϵ_{ri} (AB's axiom 4). So far, the axioms describe a continuous latent process that underlies each observation. To translate these continuous latent appraisals to categorical responses, it is assumed that there exist $C - 1$ ordered thresholds δ_{rc}, such that each x_{ri} is generated in the following way (AB's axiom 5):

$$
x_{ri} = \begin{cases} 1 & \text{if } y_{ri} \leq \delta_{r1} \\ c & \text{if } \delta_{r,c-1} < y_{ri} \leq \delta_{rc} \\ C & \text{if } y_{ri} > \delta_{r,C-1} \end{cases}
$$

In practice, the generating process of x_{ri} is probabilistic and described with an ordered logistic distribution[1], which gives:

$$P(x_{ri} \mid y_{ri}, \delta_r) = \begin{cases} 1 - \text{logit}^{-1}(y_{ri} - \delta_{r1}) & \text{if } x_{ri} = 1, \\ \text{logit}^{-1}(y_{ri} - \delta_{r,c-1}) - \text{logit}^{-1}(y_{ri} - \delta_{rc}) & \text{if } 1 < x_{ri} < C, \\ \text{logit}^{-1}(y_{ri} - \delta_{r,C-1}) & \text{if } x_{ri} = C. \end{cases}$$

The thresholds δ_{rc} accommodate the response biases of the raters. AB do so by estimating $C-1$ ordered thresholds γ and defining $\delta_{rc} = \alpha_r \gamma_c + \beta_r$ (AB's axiom 6). This translation of thresholds is called the Linear in Log Odds function and is a useful tool for capturing bias in probability estimation (Fox & Tversky, 1995; Gonzalez & Wu, 1999; Anders & Batchelder, 2015).

Figure 2 provides an intuition for how the ordered logistic distribution can model different outcomes by varying only the rater parameters. The latent appraisal y is fixed to 0, the thresholds γ are equal to $\text{logit}(c/C)$ such that $P(x_{ri} \mid y = 0, \gamma, \alpha_r = 1, \beta_r = 0)$ is uniform, and the scale α_r and shift β_r vary. In the left panel, there is no response bias, $\alpha_r = 1$ and $\beta_r = 0$, which yields a uniform distribution over the predicted Likert scores. In the right panel, an increase in response scale and shift, $\beta_r = .5$ and $\alpha_r = 2$, concentrates the predicted Likert scores around 2 and 3.

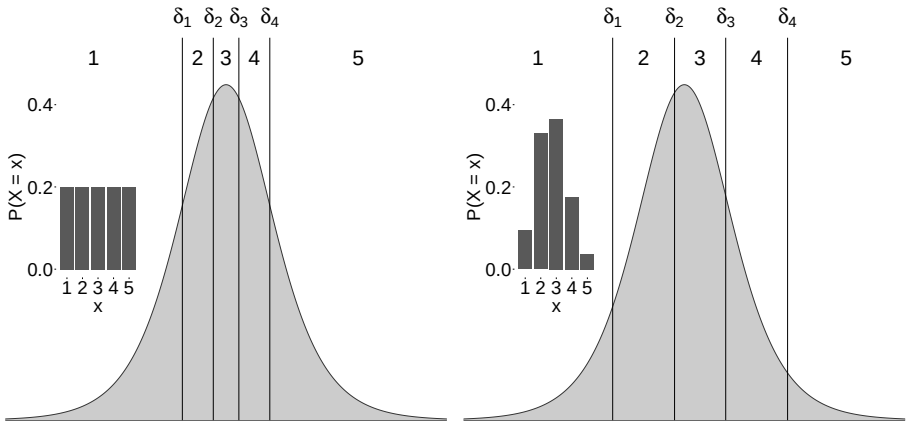


Figure 2: Ordered logistic distribution for $y_{ri} = 0$, thresholds γ equal to $\text{logit}(c/C)$, and varying rater components. The implied probability distribution over response categories is shown inside each panel. In the left panel, there is no response bias, $\alpha_r = 1$, $\beta_r = 0$. As a consequence, the distribution over the predicted Likert scores is uniform. In the right panel, the thresholds are shifted right, $\beta_r = 0.5$, and the scale increased slightly, $\alpha_r = 2$, such that the distribution over predicted Likert scores is peaked on outcomes 2 and 3.

[1]The choice for an ordered logistic distribution is arbitrary and an ordered probit distribution could also be used, as was done by Anders and Batchelder (2015).

The LTRM is a complex model and unfortunately suffers from identification issues, as was pointed out by AB already. For example, multiplying the rater competences ζ and the item difficulties κ by a constant c yields an identical variance for the appraisal distribution, since ${}^{c\zeta}/_{c\kappa} = {}^{\zeta}/_{\kappa}$. Such identification problems are avoided by restricting the mean of the respective parameters to 1 (as suggested in appendix C in AB). Another identification problems originates from estimating the thresholds individually. The number of thresholds, $C-1$, increases with the number of response options. This introduces a large number of parameters that can be difficult to estimate, in particular when some response options are not observed (i.e., when there are ceiling or floor effects). In addition, the model is only identified if the sum of thresholds is zero ($\sum_{c=1}^{C} \gamma_c = 0$; otherwise adding a constant to θ_i and δ_c yields an identical likelihood). Rather than modeling each threshold individually, we describe the thresholds using only two parameters per rater. Specifically, we model the thresholds as deviances from an initial guess, $\gamma_c = \text{logit}\,(c/C)$. This yields a set of thresholds such that if the latent appraisal is 0 then $P(x_{ri})$ is uniform. Response biases are incorporated in the same manner: $\delta_{rc} = \alpha_r \text{logit}\,(c/C) + \beta_r$. This simplification can still capture a wide variety of data sets (Selker, van den Bergh, Criss, & Wagenmakers, 2019).

Three Extensions

The LTRM as described above has many desired properties, for instance, it captures individual differences among both raters and items. However, many properties of psychiatric data are not included in the model. three sections generalize the LTRM to improve its capacity to describe the data at hand.

Extension I: Multiple Patients

The first extension allows the model to describe multiple patients. Since different patients can have different mental disorders the latent truth for an item varies across patients to reflect this. Likewise, some items may be more difficult to measure, but only for some patients. Both these changes can be achieved by allowing the item truth θ_{ip} and item difficulty κ_{ip} to vary across patients. In turn, this induces that the latent appraisal y_{rip} varies across patients. We assume that the patient parameters are drawn from a group-level distribution with unknown mean and variance, for instance, the item difficulty could follow a log-normal distribution with unknown mean and variance (i.e., $\kappa_{ip} \sim \text{LogNormal}\,(\mu_\kappa, \sigma_\kappa)$).

Extension II: Latent Constructs

Often, we are not just interested in the latent truth of a single item, but also in a construct that is measured by multiple items. For instance, the latent construct aggressiveness could be measured with multiple items. To accomplish this, we introduce a latent variable η_{pl} and allow items to load on this latent variable, i.e., we introduce a factor model over the items. The relation between the latent

construct and the item consensuses is given by the regression weights λ_{il}, such that $\theta_{ip} \sim \text{Normal}\,(\lambda_{il}\eta_{lp},\, 1)$. The measurement model, i.e., which items load on what construct, is assumed to be known.

As prior distribution on the latent constructs η_{lp} we used a normal distribution with variance 1, which reflects that the variance of a latent variable is unidentified and restricted to 1. In addition, simulations showed that the estimated regressions weights and the estimated patients scores on the latent constructs exhibited label switching. For example, multiplying both the latent constructs and the regression weights by -1 yields the same prior on the item truths. To avoid label switching, we restricted the regression weights to be positive, motivated from the perspective that it is typically known which items are negatively scored (i.e., have a negative correlation with the latent construct).

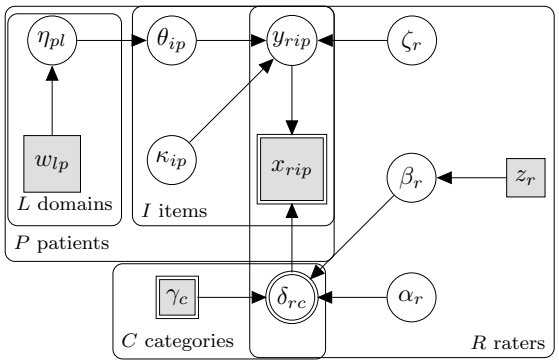
Extension III: Patient and Rater Information

The third extension adds background information about raters and patients to the LTRM. This helps the model to capture that, for instance, child molesters are less aggressive than murderers. Discrete patient and raters characteristics are captured by dividing the group level distributions into separate components for each category. For example, the hierarchical distribution over latent variables, $\eta_{pl} \sim \text{Normal}\left(\mu_{\text{Crime}_p},\, \sigma_{\text{Crime}_p}\right)$. Rater characteristics are incorporated in a similar manner, except that these would influence the group-level distributions of rater-specific parameters. This yields $\beta_r \sim \text{Normal}\,(\mu_{\text{Staff}_r},\, \sigma_{\text{Staff}_r})$. For instance, this could capture that particular groups of staff members may give more lenient ratings.

In the simulation studies, we restrict the analysis to discrete background information. However, continuous background information could also be used. Consider for instance the time a patient is committed to a detention center, Time_p. This information can be added as a regression on the mean of the group-level distribution. Thus, $\eta_{pl} \sim \text{Normal}\left(\mu_{\text{Crime}_p} + \nu\,\text{Time}_p,\, \sigma_{\text{Crime}_p}\right)$, where ν is the regression coefficient from the time a patient is committed Time_p on the mean of the group level distribution.

It is important to consider that the influence of certain background variables may differ across latent constructs. For instance, the effect of a patient's crime varies across latent constructs, allowing the model to capture that child molesters and murderers differ in aggression, but not on other latent constructs. This is accomplished by estimating an effect of a patient's crime separately for each latent construct, which implies μ_{l,Crime_p}.

To summarize, The extended LTRM first separates the rater-specific influences from the data x_{rip}, hereby accounting for different groups of raters. This results in a latent consensus for each item and patient θ_{ip}. This consensus is subsequently used as an indicator for a latent construct for all patients and constructs η_{pl}. The relation between the latent construct and the items is given by the regression weights λ_{il}, such that $\theta_{ip} \sim \text{Normal}\,(\lambda_{il}\eta_{lp},\, 1)$. The factor scores also incorporate patient specific background information, such as the crime a patient committed.



$$\theta_{ip} \sim \text{Normal}\left(\lambda_{il}\eta_{lp},\ 1\right)$$
$$\eta_{lp} \sim \text{Normal}\left(w_{lp},\ 10\right)$$
$$\lambda_{il} \sim \text{Gamma}\left(0.01,\ 0.01\right)$$
$$\beta_r \sim \text{Normal}\left(z_r,\ \sigma^2_{\beta_r}\right)$$
$$w_{pl} \sim \text{Normal}\left(0,\ 10\right)$$
$$z_r \sim \text{Normal}\left(0,\ 10\right)$$

Figure 3: Graphical model corresponding to the CCT model for multiple patients. Rater characteristics are captured by z_r and patient characteristics are captured by w_{lp}. The prior distributions for the extended LTRM were chosen to be uninformative (for positive parameters a Gamma (0.01, 0.01) was used and a Normal (0, 10) otherwise).

Implementation

The next sections illustrate the LTRM in a variety of scenarios. First, we demonstrate the benefit of applying the LTRM over analyzing the raw means in an example analysis of two fictitious patients. Second, we demonstrate that the parameters of the LTRM can be accurately recovered. Last, we compare the predictive performance of the LTRM to the unweighted mean of the observations and two machine learning toolboxes.

We estimate the parameters of the LTRM and the extended LTRM using a Bayesian approach. Therefore, we are interested in the posterior distributions of the model parameters. All models were written in Stan and approximated the posterior distributions with variational inference (Carpenter et al., 2017). We opted to use variational inference over traditional Markov chain Monte Carlo because it was computationally fast while providing similar results in terms of parameter retrieval and model predictions. All data was simulated using R (R Core Team, 2019) and Stan models were run using the R package `RStan` (Stan Development Team, 2019). R files and Stan models are available in the online appendix at `https://osf.io/jkv38/`.

Example Analysis

Here we showcase the benefits of a CCT analysis by examining results for two fictitious participants. This example demonstrates how misleading the sample mean can be. We simulated a data set of 50 patients, 10 raters, 20 items, and 5 answer categories. The items loaded on 3 different latent constructs, further referred to as aggressiveness, anxiety, and depression. A patient-specific covariate, consisting of 5 categories was added to mimic the effect of a patient's

criminal offense. Similarly, two different categories were used to imitate the different groups of raters (e.g., clinicians and psychiatrists). Next, we selected two patients whose differences in observed means were small relative to their differences in posterior means on the latent constructs. The overall mean of the observed ratings x_p was 3.51 and 3.08 for patient 1 and 2 respectively. The means for items of each construct are shown in Table 1. Analyzing these

Table 1: Means of the observed ratings for the two patients with similar mean responses yet different posterior distributions. The means are computed for each latent construct and for all scores (overall).

	Construct			
	Aggressiveness	Anxiety	Depression	Overall
Patient 1	3.86	3.04	3.65	3.51
Patient 2	3.29	3.00	2.93	3.08

patients independently by aggregating the raw observations indicates that these two patients might differ in aggressiveness and depression but not in anxiety. However, after fitting the extended LTRM to the data it becomes apparent that there is more to the data than what is shown by these averages. Using the extended LTRM, we can visualize the posterior distributions of the latent constructs for both patients, shown in Figure 4. The posterior distributions tell a

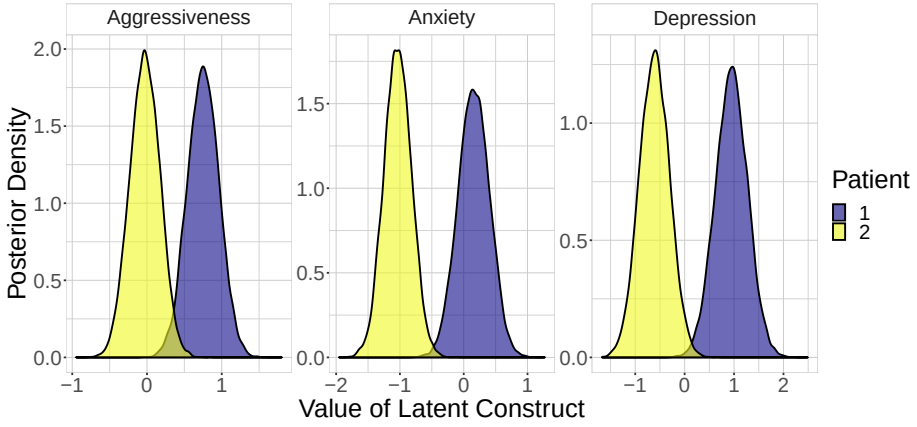


Figure 4: Approximate posterior densities for the two patients similar response pattern. The panels show different latent constructs. The posterior distributions of the patients appear to differ for Construct 1 and 2. This emphasizes that more information can be obtained from the ratings than what may be obvious from the raw scores.

different story than Table 1. Remarkably, for construct 2 where the difference in means is approximately equal, the posterior distributions differ. This difference

can be quantified by computing the probability that the posterior probability that patient 1 has a larger value on a latent trait than than patient 2. This probability is approximated by counting how often the posterior samples of a latent construct are larger for patient 1 than for patient 2. For all three constructs, the probability that patient 1 has a higher score is larger than 0.99 (Figure A.1 visualizes these probabilities).

Altogether, this example showcases that there is more relevant information in the data than just the sample averages. If we examine the parameters of the data generating model more closely, we see that there are two reasons for this discrepancy. The patients differ in item difficulty for Anxiety (1.42 for patient 1 and 0.88 for patient 2) and crime committed, which means that the population level distributions differ. In this example, all raters rated both patients. In practice, it is likely that the ratings of different patients are given by of different raters, which introduces another source of bias. The discrepancy between sample mean and posterior mean is shown for all patients in Figure A.2, which further emphasizes that the sample mean is an inadequate description of the patients scores.

Parameter Retrieval

A key step in developing a model is to assess if the model parameters can be retrieved accurately. For this purpose, we reused the data of the fictitious patients. The simulated data set consisted of 50 patients, 10 raters, 20 items, and 5 answer categories. The items loaded on 3 different latent constructs. A patient-specific covariate, consisting of 5 categories was added to mimic the effect of a patient's criminal offense. Similarly, two different categories of raters were assumed. Parameter recovery is shown in Figure 5, which graphs the values used to simulate data with against the posterior mean for each parameter.

All parameters are retrieved adequately. An exception is the item difficulty κ, which estimates appear more variable as the true item difficulty increases. However, the spread in posterior means of the item difficulty is similar to that in Figure 6 of AB. Also, the posterior means of μ_β are approximately equal to the sample means of β_r split by the rater specific covariates. These sample means differ somewhat from the true means since there are 5 observations per mean.

Although it is good to know when the parameters of the extended LTRM can be recovered, it may be more useful to know when the data are not informative enough to use the extended LTRM. This is likely the case when there are few items and raters. Exact numbers, however, may vary depending on the specific situation at hand. For most purposes, it is straightforward to adjust the number of raters, items, and patients, and then repeat the simulation. Parameter recovery for the LTRM of BA is shown in Figure B.1

Predictive Performance

Here, we compare the predictive performance of the LTRM to that of the sample mode and, as a more informative comparison, to Random Forest and Boosted

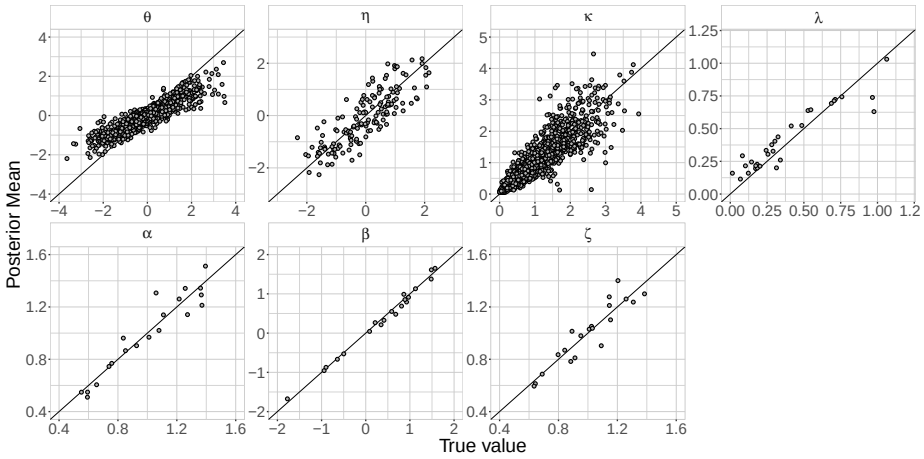


Figure 5: True value used for data simulation (x-axis) and posterior mean of that parameter (y-axis), for all parameters of the LTRM. Above each panel is indicated which parameter is shown.

Regression Trees (Boosting). Random Forest and Boosting analyses were done using the R packages `ranger` and `gbm` respectively (Wright & Ziegler, 2017; Greenwell, Boehmke, Cunningham, & Developers, 2019). We used the default settings for the hyperparameters in both R packages.

We simulated two data sets each consisting of 20 raters, 30 items, 50 patients, and thus in total 30,000 observations. The first data set represented a dense design, where all raters scored all patients. The second data set represented a sparse design, where each rater scored 10 patients. Raters were assigned to patients so that the number of obtained scores was about equal for all patients. To simulate this, we first simulated a dense data set and subsequently removed a score if the raters that did not rate a patient. This remaining data set consisted of 6,000 observations. Next, both data sets were split into a training set (80%) and a test set (20%). The performance of the four methods was evaluated by training the models on the training set and using the trained model to predict the outcomes for a test set. For the LTRM, we used the mean of the posterior predictive distribution as a point-prediction.[2] Predictions for Random Forest and Boosting were obtained by taking the majority vote of the trained classification trees. We used the observed mode of all observations for the same rater, item, or patient (i.e., to predict x_{rip} we used mode$(x_{-r,ip}, x_{r,-i,p}, x_{ri,-p})$).[3]

Don: Misschien kunnen we deze paragraaf ook gewoon skippen

[2] In this particular example, model predictions could also be interpreted as imputing missing values. If these are regarded as missing observations rather than predictions, they should be modeled as unknown discrete parameters of the model (Ch. 8; Gelman et al., 2014). That way, uncertainty about these missing observations is propagated into the parameters. Although we did not sample the missing observations from the joint posterior distribution, the code in the online appendix does show how to do this.

[3] We used the mode rather than the commonly used mean, which predictions lie outside of the ordinal scale.

A technical difference between this approach and the previous simulation is that model predictions are effectively treated as missing data. In terms of implementation, a complete data set, where every patient is scored on each item by all raters, is represented as an array of dimensions $R \times I \times P$ (and possibly a dimension for measurement occasion). However, it is easier to handle an incomplete data set by representing the data as a matrix of $(R \times I \times P)$ rows and 4 columns (i.e., long format).[4] In this format, the first column indicates the outcome, the second the rater, the third the item, the fourth the patient, and subsequent columns indicate rater and patient covariates.

We quantified prediction error by calculating the confusion matrix, a frequency table of predicted versus hold-out responses where the diagonal indicates the number of correct responses. Prediction accuracy is defined as the number of correct responses divided by the total number of responses.

Table 2: Prediction accuracy for the LTRM, Random Forest, Boosting, and the sample mode. The LTRM outperforms all other methods, but Random Forest, and Boosting perform remarkably well. Since the data are simulated the choices for the simulation setting are arbitrary, and different settings could yield a very accurate of very inaccurate predictive performance. Therefore, the absolute prediction error cannot be interpreted and only a relative comparison should be made.

Method	Dense	Sparse
LTRM	0.52	0.46
Sample Mode	0.43	0.43
Random Forest	0.42	0.42
Boosting	0.39	0.40

Given that the data were generated by the LTRM, it comes as no surprise that it predicts more accurately than the other methods. However, even though data generated from the LTRM is likely a gross simplification of reality, the results show that black box machine learning methods perform somewhat adequately. This is somewhat surprising because the data at hand are ill-suited for black-box machine learning methods, as these have difficulty capturing the hierarchical structure of the data which contains most of the information (but see Hajjem, Bellavance, & Larocque, 2014). Instead, if a lot of background information about patients and raters is available, this could likely improve their performance. However, machine learning methods do not provide interpretable models, which may be undesirable in practice because it makes it difficulty to substantiate decisions.

[4] For a Stan implementation the long format is even required as missing values must be handled explicitly, unlike for other software e.g., JAGS.

Discussion

In this paper, we extended the Cultural Consensus model developed by Anders and Batchelder (2015) to be suited for data often encountered in psychiatric detention centers. The original model was suited for data from a single patient and we extended this to multiple patients, latent constructs, and patient and rater specific covariates. The benefit of this approach is that we can obtain estimates for e.g., a patients aggressiveness while accounting for rater bias, item-specific measurement error, and a patient's criminal offense.

Although the LTRM provided better predictions than black-box machine learning approaches, this is likely because the data were simulated from the LTRM. It seems more reasonable that an optimal method for prediction would combine results from the LTRM with some machine learning approach. For example, augmenting a Random forest model with features based on psychological theories resulted in improved predictions of human decisions (Plonsky et al., 2019; Plonsky, Erev, Hazan, & Tennenholtz, 2017). However, machine learning approaches, despite their predictive power, result in uninterpretable models which may be undesirable in psychiatric practice where decisions need to be motivated and possibly defended (e.g., when releasing a patient).

Ideally, patients are monitored over a period of time and data from multiple measurement occasions is obtained. Next, the LTRM is applied to analyze the data. Rather than applying the LTRM repeatedly to data from individual measurement occasions, all observations should be analyzed simultaneously. That way, a patient's progress could be monitored over time and predictions for the future time points could be obtained along with uncertainty estimates. To extend the LTRM to incorporate time varying components is straightforward, but the exact properties of the time varying components may depend on the data at hand. For example, one can imagine that the factor scores of a patient vary over time. A model to describe these changes would be a dynamic factor model (Molenaar, 1985; Forni, Hallin, Lippi, & Reichlin, 2000). However, if patients are only rated, say, every six months then a rigorous time series model is likely unfit due to the small number of time points. Instead, simply estimating the difference between consecutive time points with an intercept may suffice.

Limitations

In the LTRM, we assumed that the factor structure is known. In practice however, this need not be the case. Estimating the factor structure from the data is possible, however, such an endeavor would shift the focus of the LTRM to model selection rather than assessing the progress of patients. Another more flexible approach is to view the latent true scores of the items as a network and estimate the relations among the items (but see Epskamp, Kruis, & Marsman, 2017 for possible drawbacks).

Since the posterior distributions were approximated with variational inference, it is possible that the obtained posterior distribution are biased. In general, these biases rarely affect the estimated posterior means, but the posterior

variance can be underestimated (Blei, Kucukelbir, & McAuliffe, 2017). As a consequence, uncertainty estimates may be too narrow. To alleviate this, it is relatively trivial to modify the Stan code in the appendix to use MCMC instead of variational inference (e.g., in the code in the appendix change `vb(model)` to `sampling(model)` to use MCMC). However, note that MCMC algorithms for the models discussed run for hours to obtain a reasonable number of posterior samples, whereas variational inference finishes after several minutes.

Recommendations for clinical practice

To successfully apply the LTRM model in practice, the data should meet several minimum requirements. This is to obtain some minimum quality

Evaluations should be recorded and stored long-term (e.g., multiple years). Ideally, ratings are obtained with high frequency. Differences between raters should be minimized, for instance by training staff. Additional information about patients, such as the reason of incarceration, should be added to the model.

To sum up, we extended the Latent Truth Rater model (LTRM) introduced by Anders and Batchelder (2015) a model that is suitable for data typical in psychiatric detention centers. The model accounts for individual differences between raters, items, and patients. We demonstrated that the extended LTRM can provide more information about the data at hand than the raw means for two fictitious patients. In addition, we have shown that the parameters of the extended LTRM can be adequately retrieved and that the LTRM outperforms the observed mode and several machine learning toolboxes in terms of predictive power. Finally, we have provided recommendations for clinical practitioners who wish to apply the LTRM in practice. Altogether, we hope the extended LTRM improves current practices for analyzing the scores in psychiatric detention centers.

References

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

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*(5), 316–332.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In G. O. E. B. Grofman (Ed.), *Information pooling and group decision making: proceedings of the second university of california irvine conference on political economy* (pp. 103–112). JAI Press Greenwich, CN.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32. Retrieved from https://www.jstatsoft.org/v076/i01 doi: 10.18637/jss.v076.i01

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J. A., Giltay, E. J., van Noorden, M. S., ... Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, *18*(1), 1–12.

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PloS one*, *12*(6), e0179891.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, *82*(4), 540–554.

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*(3), 585–603.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd ed.)*. Boca Raton (FL): Chapman & Hall/CRC.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. (2019). gbm: Generalized boosted regression models. Retrieved from https://CRAN.R-project.org/package=gbm (R package version 2.1.5)

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., ... others (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. In *Thirty-first AAAI conference on artificial intelligence.*

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338.

Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019, May 08). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*. Retrieved from https://doi.org/10.3758/s13428-019-01231-3 doi: 10.3758/s13428-019-01231-3

Stan Development Team. (2019). *RStan: the R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.19.2)

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi: 10.18637/jss.v077.i01

A Example Analysis

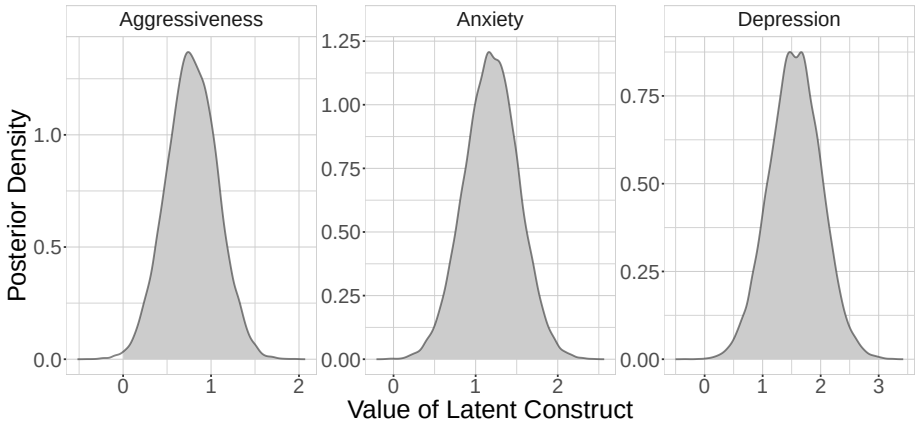


Figure A.1: Approximate posterior densities for the two patients similar response pattern. The panels show different latent constructs. The posterior distributions of the patients appear to differ for Construct 1 and 2. This emphasizes that more information can be obtained from the ratings than what may be obvious from the raw scores.

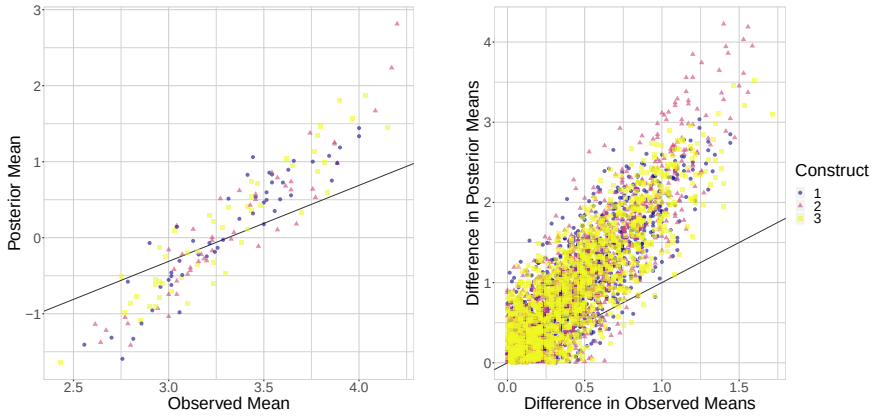


Figure A.2: The left panel plots the means of the observed ratings against the posterior means of the latent variables. The right panel shows for each combination of patients i, j the absolute difference in means, $|\hat{x}_i - \hat{x}_j|$, against the absolute difference in posterior means of the latent variables, $|\hat{\eta}_i - \hat{\eta}_j|$. Note that in the left panel, there is a difference in intercept because the responses are on a scale from 1 to 5, whereas the latent variables are assumed to have a mean of 0. The large spread in the right panel again demonstrates that there is more to the data than what the sample mean tell us.

B Parameter Recovery

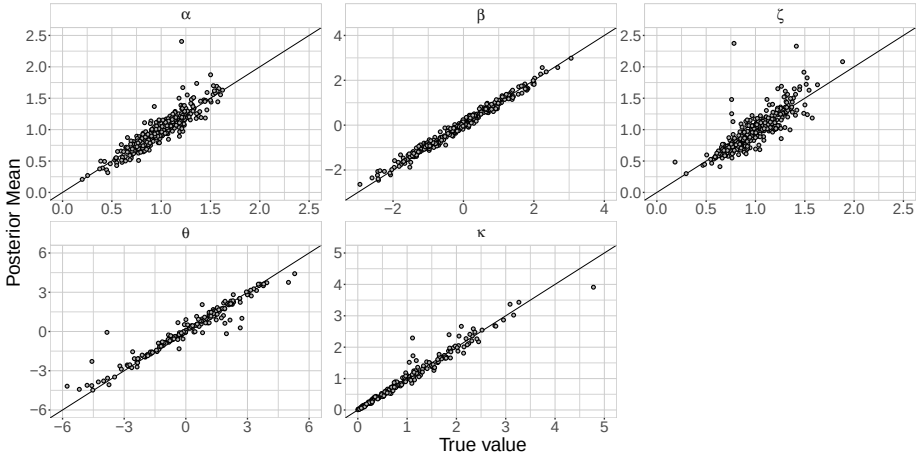


Figure B.1: Parameter recovery for the Latent Truth Rater model displayed in Figure 1. The data set consisted of 1 patient, 200 items, and 300 raters. Items had 5 possible outcomes.