Cultural Consensus Theory for the Evaluation of Patients' Mental Health Scores in Forensic Psychiatric Hospitals

Don van den Bergh*[1], Stefan Bogaerts[2], Marinus Spreen[3], Rob Flohr[3], Joachim Vandekerckhove[4], William H. Batchelder†[4], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam, [2]University of Tilburg, [3]NHL Stenden University of Applied Sciences, [4]University of California Irvine

Abstract

In many forensic psychiatric hospitals, 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 allows an assessment of treatment effectiveness; third, the scores may warn staff that particular patients are at high risk of turning violent, either before or after release. 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 an alternative, against often-used machine learning toolboxes. In addition, we outline the substantive conclusions afforded by the 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). R and Stan code is available at https://osf.io/jkv38/.

†Bill Batchelder passed away before the first draft of the manuscript was complete. Bill had been intimately involved in setting up the research program that forms the basis of the work reported here.

Forensic psychiatric hospitals monitor the mental health and forensic risk factors of their patients at regular intervals, typically using a method such as Routine Outcome Monitoring (de Beurs et al., 2011). A clinician, psychiatrist, or another staff member, henceforth a *rater*, scores a patient on historical, clinical, and prospective criteria. For example, a rater evaluates a patient's risk factors and behavior on a variety of criteria that relate to aggressiveness and the risk of recidivism. Such evaluations are stored so they may be used to inform future 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 often averaged as if they are exchangeable. For example, personal communication with the staff of a forensic psychiatric hospital suggested that clinicians are more lenient in their ratings than psychiatrists, but this information is not used to weigh their ratings. Furthermore, patients are analyzed in isolation, as if they are independent of one another. Any background information about patients, such as a patient's criminal record, is not accounted for and is only seen as static baseline information. In addition, uncertainty estimates of the resulting score are usually ignored.

Here we 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 forensic psychiatric hospitals, 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.[1]

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). Next, this model is expanded step by step, to allow a more sophisticated account of the data, for instance by describing multiple patients. We showcase the model in three simulation studies. First, we illustrate the benefits of this approach by analyzing two fictitious patients. Second, we show that model parameters are retrieved accurately. Third, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and against often-used machine learning toolboxes such as Random Forest (Breiman, 2001) and Boosted Regression Trees (Friedman, 2002). We highlight 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 own work.

[1] To be precise, currently available CCT models can only describe two hierarchical structures, i.e., for data of patients and raters, patients and items, or items and raters. However, existing CCT models treat the third hierarchical structure as non-hierarchical.

Cultural Consensus Theory and Three Extensions

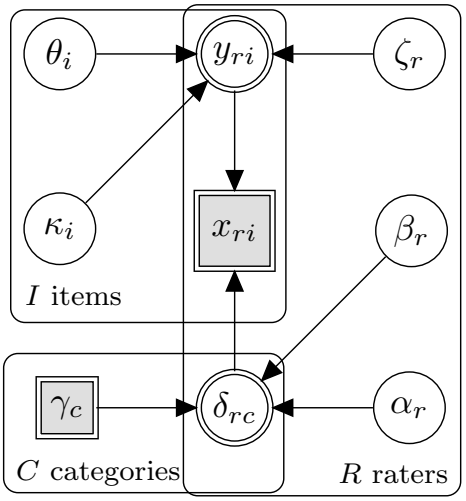
The next sections introduce Cultural Consensus Theory (CCT). First, a brief introduction to CCT is given. Next, the CCT model developed in Anders and Batchelder (2015, henceforth AB) is introduced, which serves as the simplest model for a single patient. Subsequently, we generalize the model in three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

Cultural Consensus Theory

Cultural Consensus Theory, also known as "test theory without an answer key" (Batchelder & Romney, 1988), is a statistical tool that can be used to retrieve the unknown "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 can 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 left-oriented respondents agree with the most left-oriented political statements. Note that competence and difficulty parameters are relative to the consensus and do not refer to absolute competence or difficulty. Instead competence captures the extent to which a rater evaluates according to the group consensus; likewise, difficulty captures how high a rater's competence must be to be expected to answer an item according to the group consensus. In addition, CCT models can be expanded to allow for multiple consensus truths, that is, 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 would be estimated from the data. The property of CCT models to estimate the consensus truth 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 applied 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). 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 and Table 1 provides an overview of the parameters. The LTRM captures differences among raters and items and may be viewed as the simplest model for a single patient. The rating of 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, which are briefly repeated here. There is an unknown latent shared cultural truth among the raters, which is captured by the item



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

Figure 1: Graphical model corresponding to the LTRM; a CCT model for a single patient. x_{ri} is an observed response, y_{ri} is the underlying continuous latent appraisal, θ_i is the underlying latent appraisal, and ϵ_{ri} is the appraisal error. Furthermore, κ_i and ζ_r capture the item difficulty and rater competence respectively. The unbiased thresholds are denoted γ_c, the scale and shift parameters are α_r and β_r respectively. The transformed thresholds are denoted δ_{rc}. 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.

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 a latent appraisal and defined as $y_i = \theta_i + \epsilon_{ri}$, where $\epsilon_{ri} \sim \text{Logistic}\left(0,\, \zeta_r/\kappa_i\right)$ (AB's axiom 2). The logistic density with location l and scale s is defined as

$$
\text{Logistic}\left(x;\, l,\, s\right) = \frac{\exp\left(-\frac{x-l}{s}\right)}{s\left(1+\exp\left(-\frac{x-l}{s}\right)\right)^2} \quad \text{where } s > 0.
$$

The scale of the logistic distribution for the latent appraisals consists of two components. Differences in item difficulty are captured by κ_i and differences in rater competence are captured by ζ_r (AB's axiom 3). The ratio of item difficulty over rater competence is the variance of the latent appraisal. For example, if an item is difficult then the variance of the latent appraisals is high, which leads to a spread-out probability distribution over observed ratings. Likewise, if the rater competence is high, then the variance of the latent appraisals is low and the probability distribution over observed ratings is concentrated. Latent appraisals y_{ri} are assumed to be conditionally independent given the latent truth θ_i, the item difficulty κ_{ip}, and the rater competence ζ_r (i.e., their joint distribution can be factored into a product of univariate distributions that only depend on the three aforementioned parameters; 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 deterministically in the following way (AB's axiom 5):

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

where $c = 1, \ldots, C$. The appraisal y_{ri} is latent and thus we consider the probability that an appraisal falls between two thresholds to obtain the probability of an observed score. This makes the generating process of x_{ri} probabilistic and described by an ordered logistic distribution[2], which gives:

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

where $F(x) = (1 + \mathrm{e}^{-x})^{-1}$, the cumulative distribution function of the standard logistic distribution. 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). Specifically, the scale parameter α concentrates the thresholds closer together or farther apart, and thus can yield a flat probability distribution over outcomes or a peaked one. The shift parameter β moves all thresholds up and down relative to the item location and thus captures the fact that some raters give higher overall ratings than others.

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.

The LTRM is a complex model and unfortunately suffers from identification issues, as AB already pointed out. 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 problem 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

[2]The choice for an ordered logistic distribution is arbitrary and an ordered probit distribution could also be used, as was done by AB. We use a logistic distribution rather than a normal distribution because its cumulative distribution function has an analytic expression.

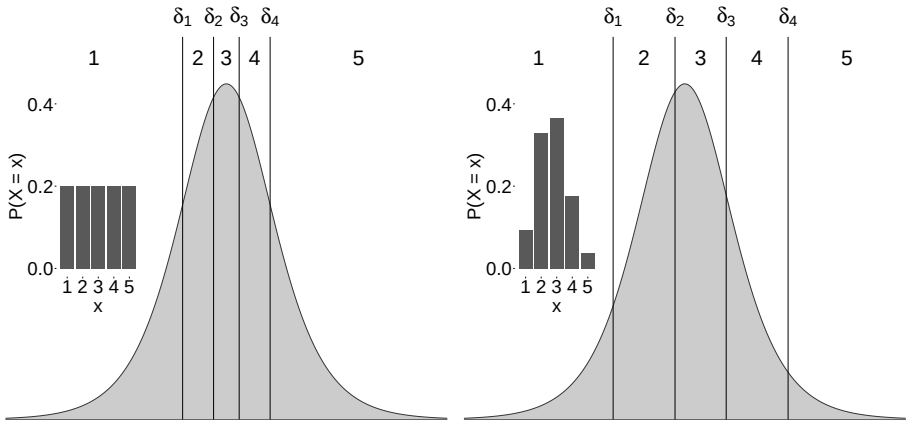


Figure 2: The ordered logistic distribution relates latent appraisals y_{ri} to response categories $\gamma = 1, \ldots, 5$ via category thresholds $\delta_1, \ldots, \delta_4$. The implied probability distribution over response categories is shown inside each panel. In the left panel, there is no response 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. In both panels, the item location parameter θ_i is 0.

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 desirable properties; for instance, it captures individual differences among both raters and items. However, many properties of psychiatric data are not captured by the model. Three extensions 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 instead of a single patient. Since even patients with the same disorder can have different item scores, the latent truth for an item varies across patients to reflect this. Likewise, it can be more difficult for raters to answer specific items according to the consensus, but only for some patients. To describe parameters that vary across patients we introduce the subscript p for patient. Both these changes can

Table 1: Overview of the parameters in the LTRM. The first column indicates the parameter, the second the parameter bounds, and the third provides the definition or prior distribution of that parameter. The last column provides a brief description of the parameter.

Parameter	Domain	Definition/ prior	Meaning
y_{ri}	$\mathbb{R}$	$\theta_i + \epsilon_{ri}$	Appraisal of rater r on item i.
γ_c	$[0,\ 1]$	logit (c/C)	Unbiased thresholds for outcome c.
δ_{rc}	$\mathbb{R}$	$\alpha_r\gamma_c + \beta_r$	Transformed thresholds for rater r on outcome c.
ϵ_{ri}	$\mathbb{R}$	Logistic $(0,\ \kappa_i/\zeta_r)$	Residual of appraisal.
θ_i	$\mathbb{R}$	Normal $\left(\mu_\theta,\ \sigma^2_\theta\right)$	Location of item i.
κ_i	$\mathbb{R}^+$	Gamma $\left(\mu^2_\kappa/\sigma^2_\kappa,\ \mu_\kappa/\sigma^2_\kappa\right)$	Difficulty of item i.
α_r	$\mathbb{R}^+$	Gamma $\left(\mu^2_{\alpha_r}/\sigma^2_{\alpha_r},\ \mu_{\alpha_r}/\sigma^2_{\alpha_r}\right)$	Scale-bias of rater r.
β_r	$\mathbb{R}$	Normal $\left(\mu_{\beta_r},\ \sigma^2_{\beta_r}\right)$	Shift-bias of rater r.
ζ_r	$\mathbb{R}^+$	Gamma $\left(\mu^2_{\zeta_r}/\sigma^2_{\zeta_r},\ \mu_{\zeta_r}/\sigma^2_{\zeta_r}\right)$	Competence of rater r.

be achieved by allowing the item truth θ_{ip} and item difficulty κ_{ip} to vary across patients, so that the latent appraisal y_{rip} varies across patients. Note that item difficulty is no longer specific to items, but also captures the interaction between patients and items. Modeling this interaction is useful when, for example, a patient barely cooperates with a question about his or her feelings; as a result, it is hard to score this item according to the consensus, but only for this patient. As in Figure 1, 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 gamma distribution with unknown mean and variance (i.e., $\kappa_{ip} \sim$ Gamma $\left(\mu^2_\kappa/\sigma^2_\kappa,\ \mu_\kappa/\sigma^2_\kappa\right)$).

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 allow the model to measure constructs, we introduce a latent variable η_{pl} that represents the score of patient p on latent variable l. Items can load on different latent variables, which introduces a factor model over the items. The relation between the latent construct l and the item consensus i is given by the factor loading λ_{il}, such that $\theta_{ip} \sim$ Normal $\left(\lambda_{il}\eta_{pl},\ \sigma^2_{\eta_p}\right)$. The measurement model, i.e., which items load on what latent construct, is assumed to be known.

As prior distribution on the latent constructs η_{pl} we used a normal distribution with mean 0 and variance 1, which reflects that the mean and variance of a latent variable are typically unidentified. 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 η and the factor loadings λ by -1 yields the same distribution

over the item truths. To avoid label switching, we restricted the factor loadings to be positive. Since we assume the factor structure to be approximately known, items that will have a negative loading on the latent construct can be reverse-scored. Here, approximately implies that if an item loads on a scale, we know whether it correlates positively or negatively with the scale although the magnitude is unknown.

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, patients with a pedophilic disorder are typically less aggressive than murderers. Discrete patient characteristics, such as criminal record, and rater characteristics are captured by introducing separate parameters of the group-level distributions for each level of the discrete characteristic. For example, the mean of the group-level distribution of the aggressiveness scale is estimated separately for murderers and patients with a pedophilic disorder. More formally, background information is represented by a categorical indicator w_p that takes on values 1 through D for each patient p. The group-level distribution for factor scores then becomes $\eta_{pl} \sim \text{Normal}\left(\mu_{w_p l}, \sigma_{w_p l}\right)$.

Rater characteristics are denoted z_r and are incorporated in similar manner. Rater characteristics influence the group-level distributions of rater-specific parameters, which yields $\beta_r \sim \text{Normal}\left(\mu_{z_r}, \sigma_{z_r}\right)$. For instance, this could capture that clinicians give more lenient ratings than psychiatrists.

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 psychiatric hospital, Time_p. This information can be added as a regression on the mean of the group-level distribution. Thus, $\eta_{pl} \sim \text{Normal}\left(\mu_{w_p l} + \nu\, \text{Time}_p, \sigma_{w_p l}\right)$, 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 background variables can differ across latent constructs. For instance, the effect of a patient's crime varies across latent constructs, allowing the model to capture that patients with a pedophilic disorder and murderers differ in aggression, but not on depression. This is accomplished by estimating the effect of a patient's crime separately for each latent construct.

Figure 3 graphically summarizes the extended LTRM and Table 2 provides an overview of the parameters. 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 factor loadings λ_{il}, such that $\theta_{ip} \sim \text{Normal}\left(\lambda_{il}\eta_{pl}, 1\right)$. The factor scores also incorporate patient-specific background information, such as the crime a patient committed.

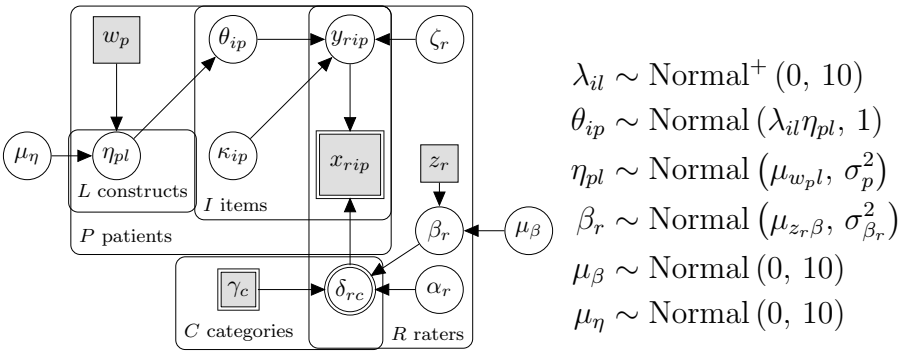


Figure 3: Graphical model corresponding to the CCT model for multiple patients. The data x_{rip}, latent appraisals y_{ri}, latent truths θ_{ip}, and item difficulty κ_{ip} now vary across patients, as indicated by the subscript p. Furthermore, for raters, competence, scale and shift are captured by ζ_r, α_r, and β_r respectively. The unbiased and biased thresholds are denoted γ_c and δ_{rc}. Rater and patient covariates are represented by z_r and w_{pl}, whereas their effects are captured by the vectors μ_β and μ_η respectively. The prior distributions are shown on the right for modified parameters. Priors not shown can be found in Figure 1. The prior distributions for the extended LTRM were chosen to be weakly informative.

Implementation

The next sections illustrate the LTRM in a variety of scenarios. First, we demonstrate the benefit of the LTRM over 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 were 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 patients that are part of a sample of 50 fictitious patients. 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 latent constructs, further referred to as aggressiveness, anxiety, and de-

Table 2: Overview of the parameters in the extended LTRM. The first column indicates the parameter, the second the parameter bounds, and the third provides the definition or prior distribution of that parameter. The last column provides a brief description of the parameter.

Parameter	Domain	Definition/ prior	Meaning
y_{rip}	$\mathbb{R}$	$\theta_{ip} + \epsilon_{ri}$	Appraisal of rater r on item i and patient p.
γ_c	$[0,\ 1]$	logit (c/C)	Unbiased thresholds for outcome c.
δ_{rc}	$\mathbb{R}$	$\alpha_r\gamma_c + \beta_r$	Transformed thresholds for rater r on outcome c.
ϵ_{rip}	$\mathbb{R}$	Logistic $(0,\ \kappa_{ip}/\zeta_r)$	Residual of appraisal.
θ_{ip}	$\mathbb{R}$	Normal $(\lambda_{il}\eta_{pl},\ 1)$	Location of item i for patient p.
κ_{ip}	$\mathbb{R}^+$	Gamma $\left(\mu_\kappa^2/\sigma_\kappa^2,\ \mu_\kappa/\sigma_\kappa^2\right)$	Difficulty of item i for patient p.
α_r	$\mathbb{R}^+$	Gamma $\left(\mu_{\alpha_r}^2/\sigma_{\alpha_r}^2,\ \mu_{\alpha_r}/\sigma_{\alpha_r}^2\right)$	Scale-bias of rater r.
β_r	$\mathbb{R}$	Normal $\left(\mu_{z_r\beta},\ \sigma_{\beta_r}^2\right)$	Shift-bias of rater r.
ζ_r	$\mathbb{R}^+$	Gamma $\left(\mu_{\zeta_r}^2/\sigma_{\zeta_r}^2,\ \mu_{\zeta_r}/\sigma_{\zeta_r}^2\right)$	Competence of rater r.
λ_{il}	$\mathbb{R}^+$	Normal$^+$ $(0,\ 10)$	loading of θ_{ip} on factor η_{pl}.
η_{pl}	$\mathbb{R}$	Normal $\left(\mu_{w_pl},\ \sigma_p^2\right)$	latent construct underlying θ_{ip}.
μ_β	$\mathbb{R}$	Normal $(0,\ 10)$	Rater group effects.
μ_η	$\mathbb{R}$	Normal $(0,\ 10)$	
w_p	$\{0, 1, \dots\}$	Data	Indicator variable that groups patients.
z_r	$\{0, 1, \dots\}$	Data	Indicator variable that groups raters.

pression. A patient-specific covariate consisting of 5 categories was added to mimic the effect of a patient's criminal offense. Similarly, two categories were of raters (e.g., clinicians and psychiatrists) were simulated. Next, we selected two patients whose differences in observed means were small relative to their differences in posterior means on the latent constructs. The means for items of each construct are shown in Table 3. The aggregates of the raw scores suggests

Table 3: Raw means of the observed ratings for the two patients with similar mean responses. The standard errors of the means are shown in parentheses. The means and standard errors are computed for each scale.

	Construct		
	Aggressiveness	Anxiety	Depression
Patient 1	3.86 *(0.14)*	3.04 *(0.19)*	3.65 *(0.18)*
Patient 2	3.29 *(0.17)*	3.00 *(0.18)*	2.93 *(0.21)*

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 distribu-

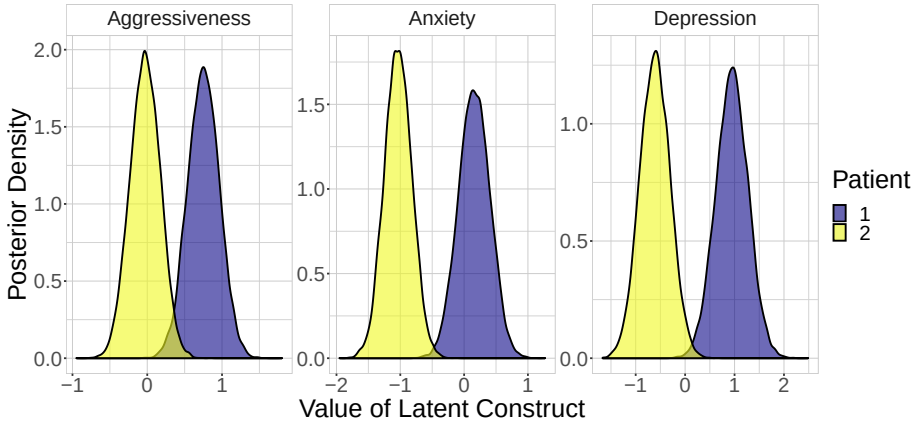


Figure 4: Approximate posterior densities for the two patients with similar response patterns. The panels show different latent constructs. The posterior distributions suggest these patients differ on all three latent constructs, unlike what the raw means in Table 3 would suggest. This demonstrates that more information can be obtained from the ratings than what may be obvious from the raw scores.

tions tell a different story than Table 3. Remarkably, for Anxiety where the raw means are approximately equal, the posterior distributions differ. This difference can be quantified by computing the posterior probability that patient 1 has a larger value on a latent trait 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 shows that there is more information in the data than what the averages convey. Examining the parameters of the data generating model more closely reveals two reasons for this discrepancy. The first reason is that the item difficulty parameter κ differed among the patients for the anxiety items (the average item difficulty for anxiety was 1.42 for patient 1 and 0.88 for patient 2). The second reason is that the fictitious patients differed in background information, that is, they committed different crimes. This means that the population level distributions for the latent constructs differ for these patients.

In this example, all raters rated both patients. In practice, the ratings of different patients are likely given by different raters, which introduces a third source of bias. The discrepancy between the 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.

Naturally, the sample mean need not always perform this poorly. The more

the data from different raters, items, and patients are exchangeable, the closer the predictions of the LTRM will be to that of the sample mean.

Parameter Retrieval

A key step in developing a model is to assess if the model parameters can be retrieved accurately. For this purpose, we simulated data as in the previous example; 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. These simulation settings resemble data sets often obtained in clinical practice (e.g., Kamphuis, Dijk, Spreen, & Lancel, 2014). Figure 5 displays the true values against the posterior means for each parameter. Details and code to replicate the simulation can be found at `https://osf.io/jkv38/`.

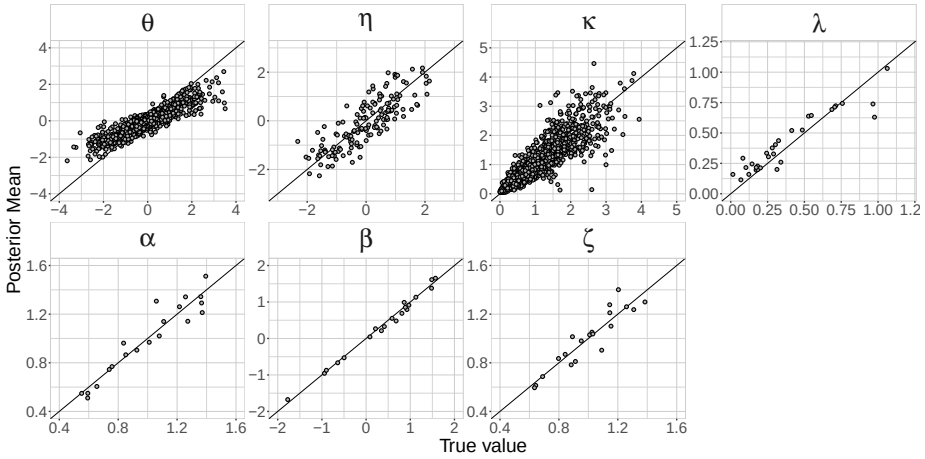


Figure 5: True value used for data simulation (x-axis) and posterior mean of that parameter (y-axis), for all parameters of the LTRM. Above each panel is indicated which parameter is shown.

All parameters are retrieved adequately. An exception is the item difficulty κ whose estimates appear more variable as the true item difficulty increases. The spread in posterior means of the item difficulty is similar to that in Figure 6 in AB. The item truths θ seem underestimated as their absolute magnitude increases. The hierarchical structure of the extended LTRM likely shrinks the item truths towards the mean. Typically, there is more shrinkage if the values of the parameter are larger, as is the case here. The bias in the item truths does not appear to influence the retrieval of any other parameters, for example, the latent construct scores η are retrieved accurately.

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 sufficiently informative to apply 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. As an exercise, we also recovered the parameters for AB's LTRM. Code for the simulation is available in the online appendix and parameter recovery is shown in Figure B.1.

Predictive Performance

Here, we compare the predictive performance of the LTRM to that of the sample mode, the sample median, the sample mean rounded towards the nearest integer, and, as a more informative comparison, to Random Forest and Boosted Regression Trees (Boosting). The sample mode is the most often observed outcome (since the data are discrete). Random Forest and Boosting analyses were done using the R packages `ranger` and `gbm` respectively (Wright & Ziegler, 2017; Greenwell, Boehmke, & Cunningham, 2019). We used the default settings for the hyperparameters in both R packages.

Here we briefly introduce Random Forest (Breiman, 2001) and gradient boosted regression trees (Friedman, 2001). Random Forest and Boosting are tree-based machine learning methods that learn from a training data set in order to predict out-of-sample observations. Both methods can be used across a wide range of applications. The methods make no parametric assumptions and their predictions tend to generalize extremely well to new observations.[3] However, both Random Forest and Boosting also have downsides. Both models are so-called black boxes, that is, their parameters are statistically unidentified and do not have a meaningful interpretation. So although their predictions are often on point, they cannot answer the *how* or the *why* of the phenomena they predict. Furthermore, these models cannot be simulated from and they do not provide uncertainty estimates.

We simulated two data sets each consisting of 20 raters, 30 items, 50 patients, and thus in total 30,000 observations. Patients were scored on a 5-point Likert scale. 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, which mimics the practically plausible situation where ratings of different patients are given by different raters. Raters were pseudo-randomly assigned to patients so that the number of obtained scores was about equal for all patients. To simulate a sparse data set, we first simulated a dense data set and subsequently removed a score if the raters that did not rate a patient. This remaining sparse 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 and the extended LTRM, we used the mode of the posterior predictive distribution as a point-prediction.[4] Predictions for Random Forest

[3] On *kaggle*, an online platform for machine learning competitions, Random Forest and Boosting are among the most successful machine learning techniques, see `https://www.kaggle.com/bigfatdata/what-algorithms-are-most-successful-on-kaggle`.

[4] 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.

and Boosting were obtained by taking the majority vote of the trained classification trees. For the observed sample mean, median, and mode we used all observations for the same rater, item, or patient.[5]

We quantified predictive performance by computing the confusion matrix between observations in the test set and predicted values; a contingency table with correct predictions on the diagonal. Prediction accuracy is defined as the proportion of correct predictions.

Table 4: Prediction accuracy for the Extended LTRM, the LTRM, Random Forest, Boosting, the sample mean, the sample median, and the sample mode. The LTRM outperforms all other methods, but Random Forest and Boosting perform worse than the sample mode. Since the data are simulated the choices for the simulation settings are somewhat arbitrary, and different settings could yield a very accurate or very inaccurate predictive performance (e.g., by adjusting item difficulty and rater competence). Therefore, the absolute prediction error cannot be interpreted and only a relative comparison should be made. Since there were 5 possible outcomes, an accuracy of 0.2 corresponds to chance performance.

Method	Dense	Sparse
Extended-LTRM	0.52	0.41
LTRM	0.46	0.34
Sample Mode	0.43	0.33
Random Forest	0.42	0.36
Boosting	0.41	0.35
Sample Median	0.33	0.32
Sample Mean	0.23	0.27

Given that the data were generated by the Extended LTRM, it comes as no surprise that it predicts more accurately than the other methods. However, even though data generated from the Extended 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 difficult to substantiate decisions.

[5]Predictions for the mode, median, and mean are obtained in the following manner. Let a negative subscript refer to all observations except that particular one, e.g., $x_{-r,ip}$ refers to $x_{1,ip}, x_{2,ip}, \ldots, x_{r-1,ip}, x_{r+1,ip}, \ldots, x_{R,ip}$; all observations for item i and patient p but not observation rip. Then predictions for the mode, median, or mean are obtained by taking respectively the mode, median, or mean of $x_{-r,ip}$, $x_{r,-i,p}$, and $x_{ri,-p}$.

Discussion

In this paper, we extended the Cultural Consensus model developed by Anders and Batchelder (2015) to apply to mental health scores of patients in forensic psychiatric hospitals. 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, for example, a patients aggressiveness while accounting for rater bias, item-specific measurement error, and the nature of a patient's previous criminal offense. We have shown in a simulation that the parameters of the extended LTRM can be retrieved accurately.

Although the LTRM provided better predictions than black-box machine learning approaches, this is likely because the data were simulated from the LTRM. In practice, it might be advantageous to combine the results from the LTRM with a machine learning method, as this may improve prediction accuracy. For example, augmenting a Random Forest model with features based on psychological theories resulted in a model with better predictions of human decisions than naive machine learning models and models based on psychological theories alone (Plonsky et al., 2019; Plonsky, Erev, Hazan, & Tennenholtz, 2017). However, machine learning approaches, despite their predictive power, may result in uninterpretable models which may be undesirable in psychiatric practice where decisions need to be motivated and possibly defended (e.g., when determining whether a treatment is effective or when deciding if a patient should be released). In addition, the LTRM provides richer information. For example, clinicians or psychiatrists may want to know if they rate very leniently or not. On the other hand, management might be interested in what covariates determine, for instance, the aggressiveness of patients.

Ideally, patients are monitored over some time and data from multiple measurement occasions is obtained and analyzed using the extended LTRM. Rather than applying the LTRM repeatedly to data from individual measurement occasions, all observations should be analyzed simultaneously. That way, a patient's progress may 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 conceptually straightforward, but the exact properties of the time-varying components should depend on the data at hand. For example, one can imagine that the factor scores of a patient vary over time as described by a dynamic factor model (Molenaar, 1985; Forni, Hallin, Lippi, & Reichlin, 2000). However, when patients are rated only rarely – say every six months – then the application of a sophisticated time series model is not feasible. Instead, simply estimating the difference between consecutive time points with an intercept may suffice. For these reasons, we did not explore a time series extension of the LTRM.

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, although such an endeavor shifts the focus of the LTRM to model selection rather than assessing the progress of patients. Furthermore, we ensured that the factor structure is identified by fixing all loadings to be

positive. Strictly speaking, this restriction is stronger than needed to ensure that the model is identified. An alternative way is to fit the model without constraints and afterward relabel such that a factor solution that corresponds to one posterior mode is obtained (e.g., Erosheva & Curtis, 2017). Another more flexible approach is to view the latent true scores of the items as a network rather than a latent variable model and estimate the relations among the items (but see Epskamp, Kruis, & Marsman, 2017 for possible drawbacks).

The generalizability of the reported results is limited due to the small number of simulations. However, we do not believe there is much reason for concern as the characteristics of the simulated data were chosen to mimic those in practice.

Since the posterior distributions were approximated with variational inference, the obtained posterior distributions may be 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 intervals may be too narrow. To alleviate this problem, it is relatively straightforward 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.

In the extended LTRM, extreme location parameters θ_{ip} are underestimated (e.g., see Figure 5). From a Bayesian perspective, there is little to worry about. Given the priors and the data, the posterior follows automatically. From a frequentist perspective, this bias may be worrying. This bias can be mitigated in several ways. However, we want to stress that addressing bias should be considered in light of the decisions made based on the estimates. Furthermore, bias should not be considered in isolation of the bias-variance tradeoff, that is, reducing the bias may increase the variance of an estimator, which harms generalization. For example, one straightforward approach to reduce bias is to tune the prior to minimize shrinkage. On the other hand, there are many success stories of shrinkage, Stein's paradox being a well-known example (Efron & Morris, 1977). Rather than interpreting point estimates one could instead consider the uncertainty intervals, assuming these have frequentist coverage (e.g., given enough data points or by using a procedure similar to Yu and Hoff (2018) or Hoff and Yu (2019)).

Recommendations for clinical practice

To successfully apply the extended LTRM in practice, the data should meet several minimum requirements. For instance, it should be recorded which rater gave what rating, and patient and rater covariates should contain as few missing observations as possible. Furthermore, although the model accounts for differences between raters, it is best to minimize these differences, for instance through clear scoring instructions. Minimizing differences between raters ensures that rater bias is minimal and helps to ensure validity. In addition, there should be overlap among (groups of) raters and the patients they score. That is, patients should be scored by multiple raters in such a way that there are no isolated groups of raters and patients, where one group of raters only rates one group of patients and another group of raters rates a different group of patients. A lack of overlap between two groups complicates a comparison between raters

and patients between them. A lack of overlap can be avoided by having rater 1 score patients 1 through 5, having rater 2 score patients 3 through 7, etc. Additional information about patients should be added to the model, such as the reason for incarceration. That should help the extended LTRM to distinguish between groups of patients that differ on these covariates. This also holds for the raters; if certain background variables are suspected of causing rater bias then these should be included in the model.

An important step in applying any model is assessing its fit to the data. There are at least two options for doing so with the extended LTRM. First, a traditional approach is to take the residuals of the extended LTRM and examine these for any leftover structure. As in linear models, there should be no structure in the residuals if the model accurately describes the data. Second, one could compare the predictive performance of the LTRM to that of a machine learning toolbox (e.g., Random Forest or Boosting). The data set is split into a training set and a validation set. Subsequently, the models are fitted to the training set and are evaluated on the validation set. This provides an idea of how much fit is lost by using a parametric model (the extended LTRM) as opposed to a nonparametric alternative (a machine learning toolbox).

Conclusion

We extended the Latent Truth Rater model (LTRM) introduced by Anders and Batchelder (2015) to a model that can be applied to patients' mental health scores in forensic psychiatric hospitals. 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 believe the extended LTRM constitutes a promising approach for the analysis of mental health scores in forensic psychiatric hospitals.

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.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

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

Erosheva, E. A., & Curtis, S. M. (2017). Dealing with reflection invariance in bayesian factor analysis. *psychometrika*, *82*(2), 295–307.

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. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 1189–1232.

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. (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.

Hoff, P., & Yu, C. (2019). Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, *13*(1), 94–119.

Kamphuis, J., Dijk, D.-J., Spreen, M., & Lancel, M. (2014). The relation between poor sleep, impulsivity and aggression in forensic psychiatric patients. *Physiology & Behavior*, *123*, 168–173.

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

Yu, C., & Hoff, P. D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, *105*(2), 319–335.

A Example Analysis

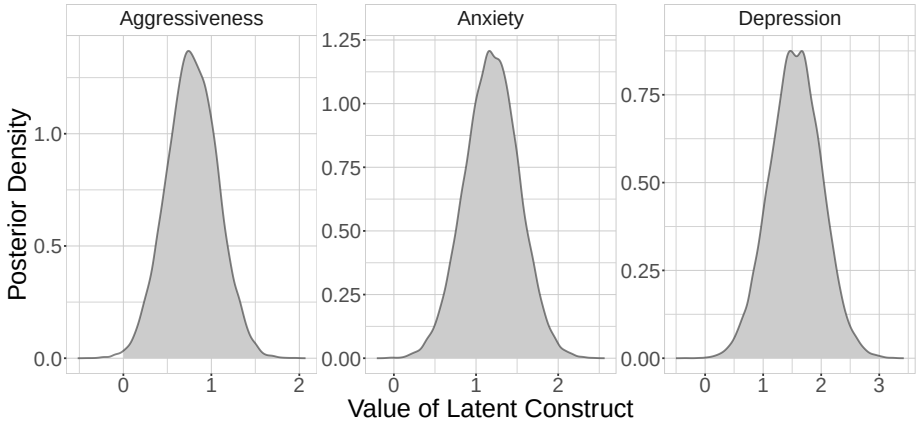


Figure A.1: Approximate posterior densities for the differences in latent constructs of two fictitious patients with response pattern. The probability that the difference is larger than 0 is above 0.99 for all constructs.

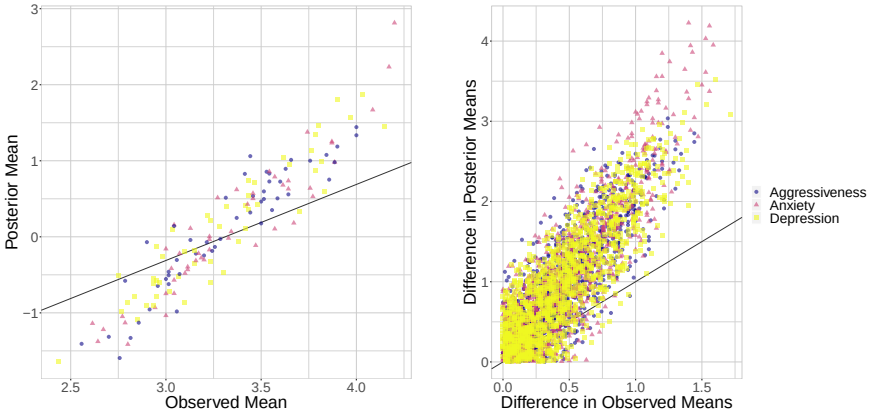


Figure A.2: The left panel plots the means of the observed ratings against the posterior means of the latent variables. The right panel shows for each combination of patients i, j the absolute difference in means, $|\hat{x}_i - \hat{x}_j|$, against the absolute difference in posterior means of the latent 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 demonstrates that the sample mean is an unreliable indicator of the truth underlying the data.

B Parameter Recovery

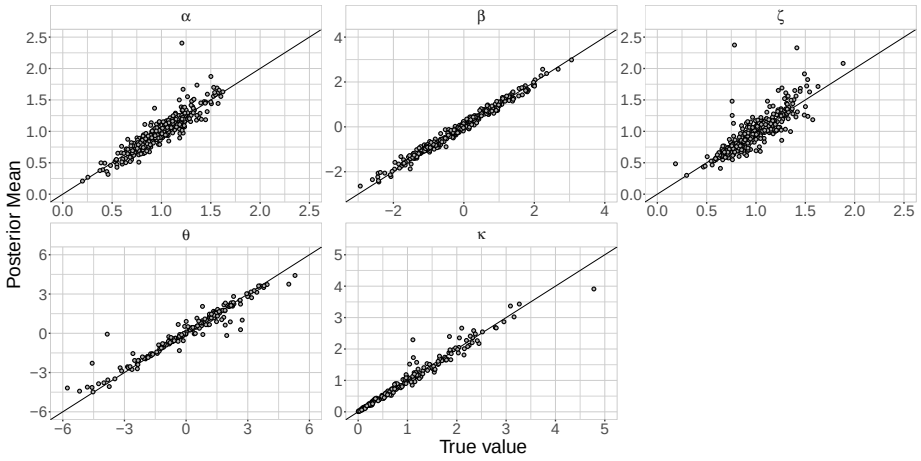


Figure B.1: Parameter recovery for the Latent Truth Rater model displayed in Figure 1. The data set consisted of 1 patient, 200 items, and 300 raters. Items had 5 possible outcomes.