# 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]??

**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

*Correspondence concerning this article should be addressed to:
Don van den Bergh
University of Amsterdam, Department of Psychological Methods
Postbus 15906, 1001 NK Amsterdam, The Netherlands
E-Mail should be sent to: donvdbergh@hotmail.com.

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.

Psychiatric detention centers monitor the mental health of their patients at regular intervals. 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. Next, these ratings of patients' mental health are used for multiple purposes. For instance, the scores may help adjust treatment to individual patients; second, the change in scores over time allows for an assessment of treatment effectiveness; third, the scores may warn staff that particular patients are at high risk of turning violent. Moreover, these ratings are key to 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.

Don: quote opzoeken over non exchangeability van raters/ items.

An appropriate model for these data captures individual differences among the patients, raters, and items. Cultural Consensus Theory (CCT) is an ideal starting point for such a model, as CCT is designed to pool information from different raters and items (Romney, Weller, & Batchelder, 1986; Batchelder & Romney, 1988; Batchelder & Anders, 2012). CCT models describe raters and items using hierarchical structures and thereby accommodates their individual differences. A model based on CCT would allow psychiatric detention centers to assess the effectiveness of their treatments and the mental health of their patients.

Don: Iets over goals en benefits!

Here we 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. 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 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

The next sections introduce Cultural Consensus Theory (CCT). A brief introduction to CCT is given. Afterward, the CCT model developed in (Anders &

Batchelder, 2015) is introduced, which may serve as the simplest model for a single patient. Subsequently, this model is expanded in two steps. First, the model is expanded to describe multiple patients. Next, the model is modified to include patient and rater characteristics.

Don: Is het leuk om een korte historische context the geven? Persoonlijk boeit dat mij meestal niet maar sommige lezers zouden het leuk kunnen vinden.

Historically, CCT used in context of questionnaires where the 'true' answers are unknown and estimated from the data. Examples: "unknown answer key", political questionnaire.

As a starting point, consider the Latent Truth Rater Model cultural consensus model for ordinal data (LTRM) as introduced by Anders and Batchelder (2015). We use a model for ordinal data as ratings are generally given on a Likert scale, but CCT can be applied to continuous data as well (e.g., the LTM; Batchelder & Anders, 2012). A graphical model of the LTRM is shown in Figure 1. The LTRM captures differences among raters and items and may be viewed as the simplest model for a single patient.

Don: Definieer Ordered Logistic als een kansverdeling met locatie, scale en thresholds als parameters.

The rating given by rater $r$ on item $i$ is denoted $x_{ri}$ and takes on discrete values from 1 through $C$. Anders and Batchelder (2015) 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 $T_i$ (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 $y_{ri} = T_i + \epsilon_{ri}$, where $\epsilon_{ri} \sim \text{Normal}\,(0,\, \sigma^{\epsilon}_{ri})$ (Axiom 2). The standard deviation of the appraisal error $\sigma^{\epsilon}_{ri}$ reflects that raters differ in their precision, captured by $\zeta_r$, and that items differ in their difficulty, captured by $\lambda_r$. These two components define the standard deviation, $\sigma^{\epsilon}_{ri} = \kappa_r/\zeta_r$ (Axiom 3). Latent appraisals are assumed to be conditionally independent given the latent truth $T_i$ and the appraisal error $\epsilon_{ri}$ (Axiom 4). So far, the axioms describe a latent process that underlies each observation. To translate these latent appraisals to observed categorical responses, it is assumed that there exist $C-1$ ordered thresholds $\delta_{rc}$, such that each $x_{ri}$ is generated in the following way (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

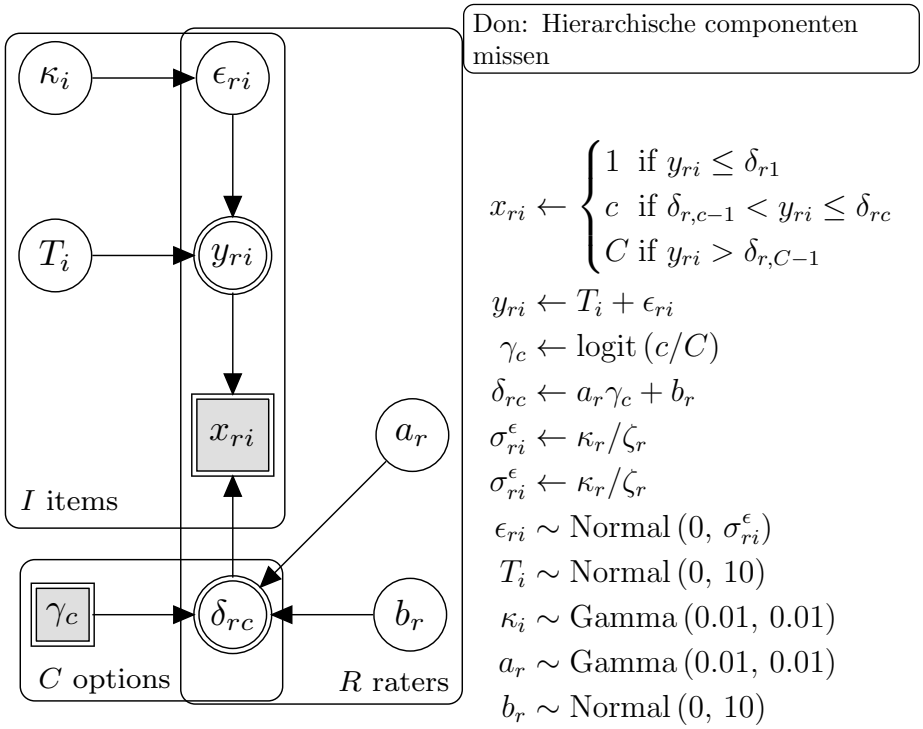



Figure 1: Graphical model corresponding to the LTRM; a CCT model for a single patient. Gray nodes are observed whereas white nodes are unobserved parameters. Square nodes are discrete and circular nodes are continuous. A double border implies that a node is ???.

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 $\delta_{rc}$ are not fixed across raters but accommodate the response biases of the raters. Anders and Batchelder (2015) do so by estimating $C-1$ ordered thresholds $\gamma$ and defining $\delta_{rc} = a_c\gamma_c + b_c$ (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).

A technical difficulty of estimating $C-1$ is that the number of thresholds $\gamma_c$ increases with the number of response options. This introduces a large number

[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).

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 thresholds sum to zero (otherwise adding a constant to $T_i$ and $\gamma_c$ yields an identical likelihood). Rather than modeling each threshold individually, 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: $\delta_{rc} = a_r \text{logit}\,(c/C) + b_r$. This simplification can still capture a wide variety of data sets (Selker, van den Bergh, Criss, & Wagenmakers, 2019).

An intuition for how the ordered logistic distribution can model different data sets is given in Figure 2. In the left panel, there is no response bias, $a_r = 1$ and $b_r = 0$, which yields a uniform distribution over the outcomes. In the right panel, an increase in response scale, $a_r = 2$, concentrates the responses around the middle outcomes. An increase in response shift, $b_r = .5$ moves the average response towards lower categories. The latent appraisal is the same in both panels.

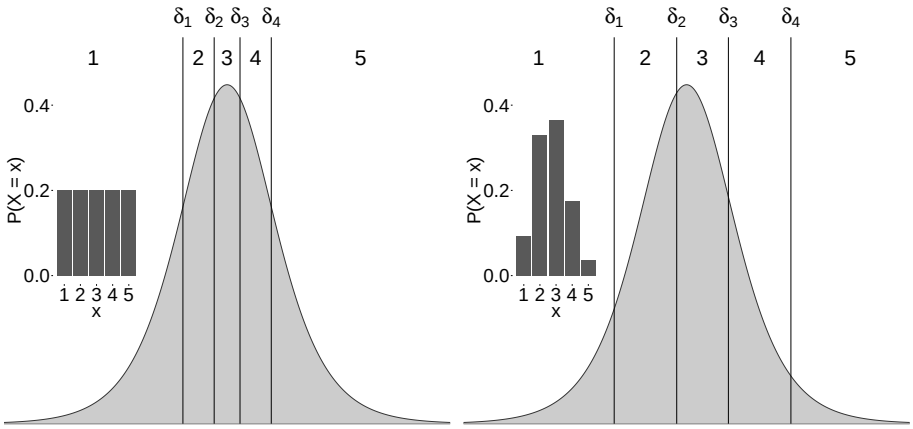



Figure 2: Ordered Logistic Distribution for $y_{ri} = 0$ and varying thresholds. The implied probability distribution over categories is shown inside each panel. In the left panel, there is no response bias, $\alpha_r = 1$, $\beta_r = 0$. As a consequence, the distribution over outcomes is uniform. In the right panel, the thresholds are shifted right and the scale increased slightly, $\alpha_r = 2$, $\beta_r = 0.5$. The distribution over outcomes is peaked on 2 and 3.

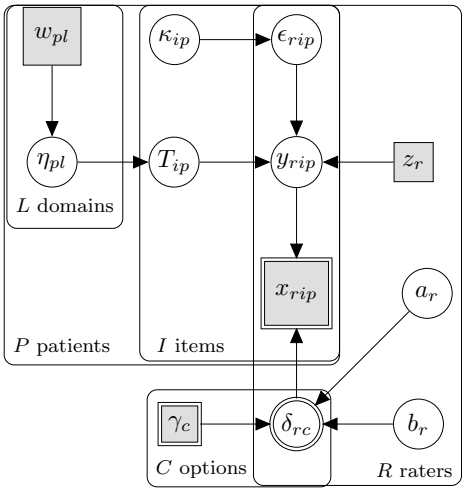
The first extension is to allow the model to describe multiple patients. This can be achieved by allowing all item specific parameters to vary across patients ($T_{ip}, \kappa_{ip}$, which induces $\epsilon_{rip}, y_{rip}, x_{rip}$). This implies that the latent cultural truth of an item and the difficulty of an item vary across patients. A second extension is that we are often not just interested in the latent truth of a single item, but also in a construct that is measured by multiple items. This is essentially a factor model, where a latent construct (e.g., aggressiveness) is

measured through multiple items. For this particular application, however, the items may contain rater bias which is simultaneously estimated and corrected for. Formally, there are $L$ constructs captured by latent variables which are indicated by some items. The measurement model, i.e., which items load on what construct, is assumed to be known.

Don: Als we het delict in de prior stoppen, dan moet het zo zijn dat elke moordenaar aggressiever is dan elke kindermisbruiker, ongeacht of een moordenaar al 30 jaar vast zit. Dat is wel erg resrictief.

A third extension is to add additional information about raters and patients to the model. This would facilitate the model to capture that, for instance, child molesters are less aggressive than murderers. Such patient and raters characteristics could be perceived as covariates and their influence could be captured through a regression. For instance, the crime of a patient can be regressed on each latent variable, allowing the model to capture that child molesters and murderers differ in aggression, but not in say ... . Rater characteristics could be regressed directly on the latent appraisal. Adding these three components to the model results in the graphical model shown in Figure 3.

Don: ik heb eigenlijk geen idee wat ze allemaal meten




$$
\begin{aligned}
T_{ip} &\sim \text{Normal}\left(\lambda_{il}\eta_{lp},\ 1\right) \\
\eta_{lp} &\sim \text{Normal}\left(w_{pl},\ 10\right) \\
w_{pl} &\sim \text{Normal}\left(0,\ 10\right) \\
b_r &\sim \text{Normal}\left(z_r,\ 10\right) \\
z_r &\sim \text{Normal}\left(0,\ 10\right)
\end{aligned}
$$

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 $w_p$.

A technical difficulty of the LTRM is that some parameters are poorly identified. In the specification of (Anders & Batchelder, 2015) another rater specific parameter was introduced,

For instance, given a positive real numbers $c$, $\kappa_r/\zeta_r = (c\kappa_r)/(c\zeta_r)$. Hence, different parameters values may yield an identical likelihood. Such identification problems are avoided by restricting the mean of a parameter vector to 0 whenever necessary.

# Simulation Results

The sections below illustrate the LTRM for a variety of scenarios. First, we demonstrate that the parameters of the LTRM can be accurately retrieved. Afterward, we show how the LTRM could monitor the progress of a patient across measurement occasions. Finally, we compare the predictive performance of the LTRM to the unweighted mean of the observations and two machine learning toolboxes.

In all simulations, data was simulated using R (R Core Team, 2019) and the posterior samples were obtained using Stan (Carpenter et al., 2017). Machine learning analyses were done using ref1 and ref2. R files and Stan models are available in the online appendix at *osflink*.

Don: Fill in refs.

## Parameter Retrieval

Before interpreting a model, it is key to assess if its parameters can be accurately retrieved. For this purpose, we simulated a data set of 50 patients, 30 raters, 20 items, and 5 answer categories. A patient specific covariate containing 5 levels was added to mimic the effect of criminal offense. Similarly, three different categories of raters were assumed, which was also captured as a covariate. Parameter retrieval is shown in Figure 4 and shows that most parameters are retrieved without any difficulties. Most parameters are accurately retrieved. The spread in the posterior means of $\kappa$ is similar to that in Anders and Batchelder (2015).

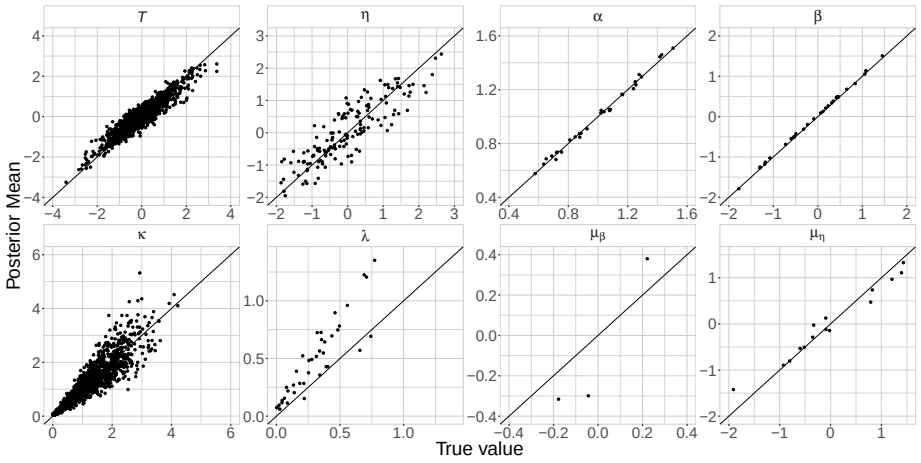



Figure 4: text

## Progress Monitoring

Don: Een probleem is dat dit model niet geidentificeerd is tenzij we sommige dingen op 0 zetten (zoals het gemiddelde van de factor scores). Dus het observeren van "progress" kan eigenlijk niet op populatie niveau.

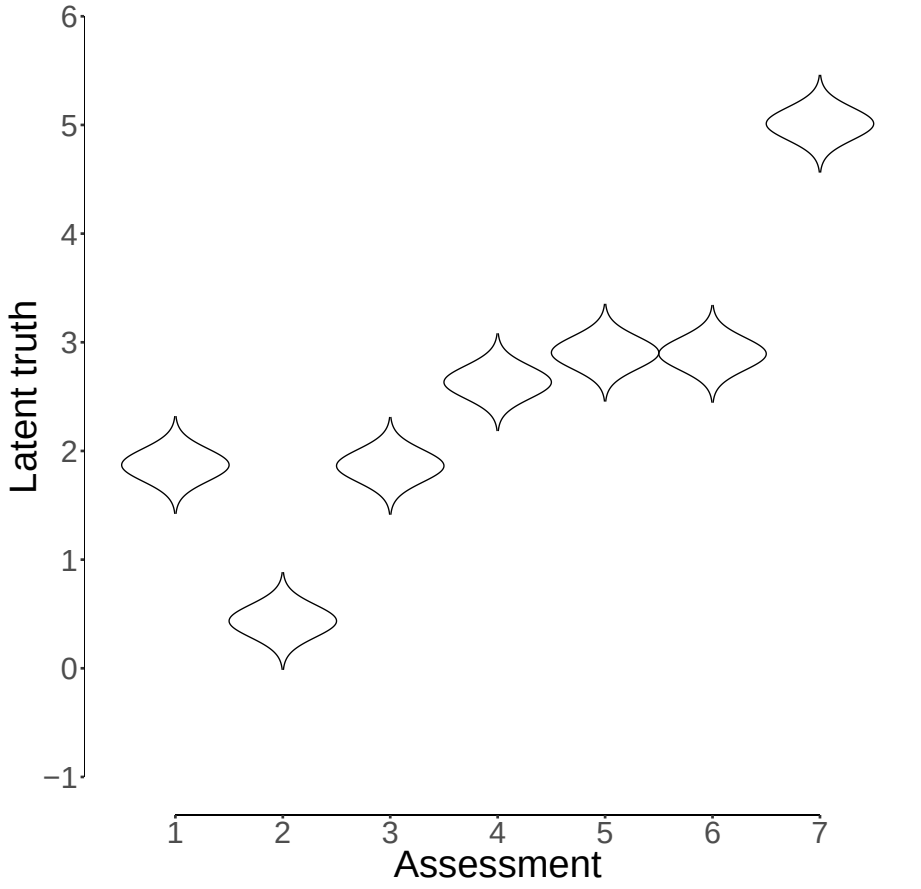



Figure 5: Example of how a patient's progress could be monitored across measurement occasions.

## Predictive Performance

Here, the predictive performance of the LTRM is compared to that of the mean and, as a more informative comparison, to Random Forest and Boosted Regression Trees. Consider the following scenario: Given one measurement occasion for a single patient, the goal is to predict the behavior of that patient for the

next day. It is assumed that the patient remains unchanged from one day to the next, i.e., the generating model remains the same. The performance of the three methods is evaluated by training the models on the first measurement occasion and using the trained model to predict the outcomes for the next day. To compare the accuracy of the methods, 100 data sets were simulated for the next day. For each data set, the root-mean-squared-error is computed between simulated and predicted observations.

# Discussion

A common application of CCT is testing for multiple consensus truths.

Model extensions, covariate for years of admission.

## Recommendations for clinical practice

To successfully apply the LTRM model in practice, several minimum requirements should be met. 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. Additional information, such as the reason of incarceration, should be added to the model.

## Limitations

Aanbeveling voor de praktijk

Bijhouden van evaluaties (scores + rater), liefst met hoge frequentie. Reden waarom iemand opgesloten is (reason of incarceration). zo min mogelijk missing not at random. "handig": training in invullen om verschillen tussen raters te minimalizeren. (V)AR component?

# References


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. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92.

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*.

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.

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

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

## Parameter Recovery

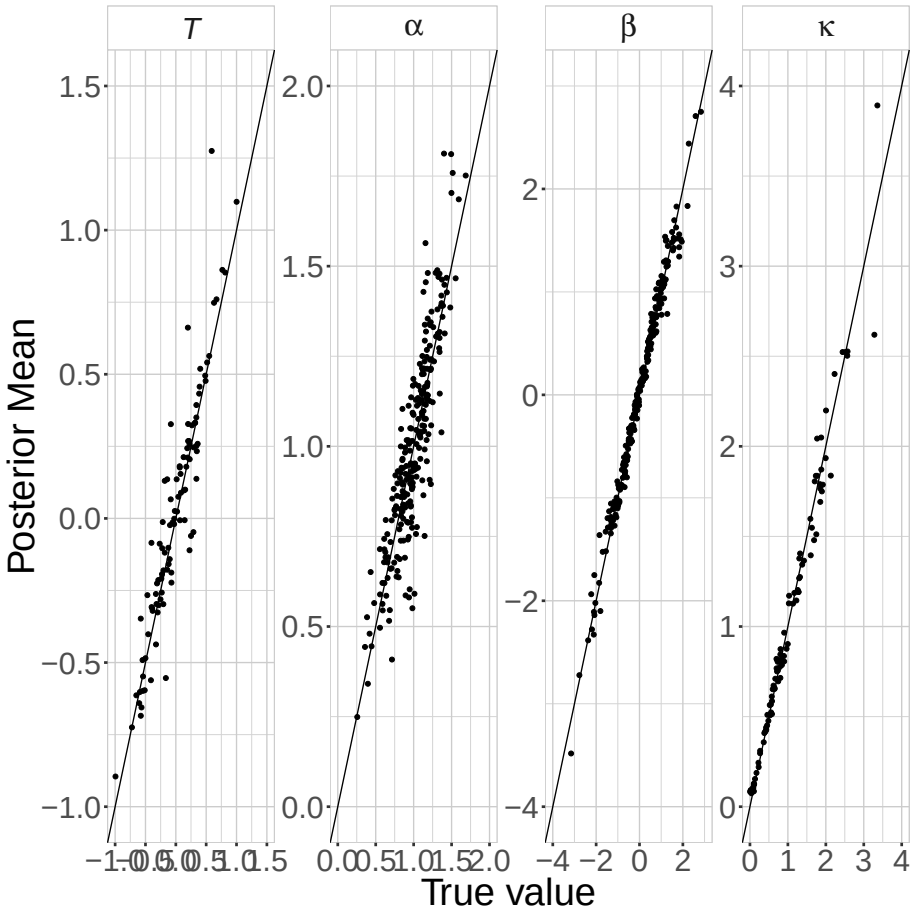



Figure 6: Parameter recovery for the model displayed in Figure 1