Invariance violations and the CNI model of moral judgments

In press: *Personality and Social Psychology Bulletin*

Niels Skovgaard-Olsen

University of Freiburg

Karl Christoph Klauer

University of Freiburg

Author Note

Niels Skovgaard-Olsen, Institute of Psychology, University of Freiburg, Germany.
Karl Christoph Klauer, Institute of Psychology, University of Freiburg, Germany.

Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard.olsen@psychologie.uni-freiburg.de, n.s.olsen@gmail.com, Uni Freiburg, Psychologie, Engelbergerstraße 41, 79085 Freiburg im Breisgau, Germany).

Acknowledgment: This research was supported by a research grant (497332489) to Niels Skovgaard-Olsen from the German Research Council (DFG). We would like to further thank Jonathan Baron, the reviewers, and the editor for valuable comments on previous versions of the manuscript which helped improve the paper.

Abstract

A number of papers have applied the CNI model of moral judgments to investigate deontological and consequentialist response tendencies (Gawronski et al., 2017). A controversy has emerged concerning the methodological assumptions of the CNI model (Baron & Goodwin, 2020, 2021; Gawronski et al. 2020). In this paper, we contribute to this debate by extending the CNI paradigm with a skip option. This allows us to test an invariance assumption that the CNI model shares with prominent process-dissociation models in cognitive and social psychology (Klauer et al., 2015). Like for these models, the present experiments found violations of the invariance assumption for the CNI model. In Experiment 2, we replicate these results and selectively influence the new parameter for the skip option. In addition, structural equation modeling reveals that previous findings for the relationship between gender and the CNI parameters are completely mediated by the association of gender with primary psychopathy.

Keywords: Moral Judgment, MPT Modeling, Deontology, Utilitarianism, Individual Variation.

Introduction

Considerable parts of moral psychology have focused on the opposition between deontology and utilitarianism in sacrificial dilemmas featuring a run-away trolley (Waldmann et al., 2012). In these dilemmas, participants need to weigh the consequences (e.g., save 5 lives), which figure in utilitarian cost-benefit calculations permitting instrumental harm, against the preservation of deontological, moral principles prohibiting to kill other people intentionally.

Further methodological improvements led to the development of two models of moral judgment based on the family of multinomial processing tree (MPT) models: the process-dissociation (PD) model (Conway & Gawronski, 2013) and the CNI model (Gawronski al., 2017). These models are applied to various real-world moral dilemmas, which introduce further controls for confounds than the run-away trolley scenario, as outlined below.

In this paper, we re-examine a recent controversy concerning the latest development of the CNI model, which unfolded between Baron and Goodwin (2020, 2021) and Gawronski et al. (2020). Our focus will be on the soundness of an invariance assumption concerning the estimation of MPT parameters (like in the PD model and the CNI model), which has proved to be problematic in applications of process-dissociation models in cognitive and social psychology (Klauer et al., 2015).

Through an experiment with a large sample size (N = 486), we test this invariance assumption as applied to the CNI model and evaluate an extension of the CNI model, which avoids making the invariance assumption. Based on this model comparison, we re-examine previously reported effects concerning psychopathy and the CNI parameters via structural equation modeling to assess how effects of gender on the CNI parameters are mediated. In Experiment 2, we replicate these results showing violations of the invariance assumption for the C and N parameters. In addition, Experiment 2 introduces a manipulation of the S parameters, which our extended CNI model introduces, and shows that it is possible to selectively influence the S parameter through our manipulation.

Finally, we make a recommendation for how the CNI model should be applied in future uses based on our results.

The CNI Model

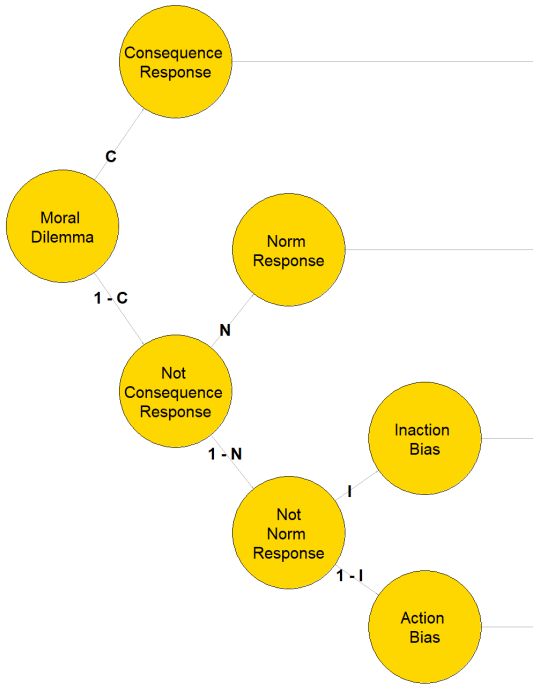
To refine the classification of norm based and consequentialist moral judgments, a process-dissociation model (Conway & Gawronski, 2013) and a multinomial processing tree (MPT) model (Gawronski et al., 2017) have been developed to disentangle factors that are not separated in the traditional, sacrificial dilemma. In Conway and Gawronski (2013), this is done by producing both congruent and incongruent conditions in which the benefits of action can be either smaller or greater than the cost of the outcome. In Gawronski et al. (2017), this takes the form of developing new stimulus materials that factorially combine action/inaction according to deontological norms and utilitarian consequences based on the insight that these two factors are confounded in the run-away trolley dilemma. Since a deontological response always requires inaction in standard sacrificial dilemmas, where victims are fixed to the tracks of a run-a-away trolley, it cannot be separated from a general response bias towards inaction. Moreover, since the utilitarian response always requires action in standard trolley dilemmas, it cannot be separated from asocial tendencies towards sacrifice.

In these improved scenarios, four conditions are created in which the benefits of the action is either greater or smaller than the costs of the outcome. In addition, the norms are manipulated to either prohibit an action to bring about the outcome (*proscriptive norm*) or to prescribe an action (*prescriptive norm*) in a situation in which some other agent plans to carry out a prohibited action. The scenarios describe realistic situations in which the sacrificial dilemma, for instance, arises in the context of a doctor treating patients. In this context, a norm-based response pattern (N) consists in a) selecting inaction, whenever actions are prohibited by deontological norms not to harm other people, and b) selecting action, whenever the action is to prevent another agent from carrying out the prohibited act. In contrast, the consequentialist

response pattern (C) consists in selecting action if and only if the beneficial consequences are greater than the detrimental consequences. Finally, a response bias towards inaction (I) consists in selecting inaction across all conditions without regard to variations in norms or the utility of the consequences.

The relative response frequencies are analyzed with a multinomial processing tree model (Batchelder & Riefer, 1999; Erdfelder et al., 2009) to characterize the processes underlying participants' selections of categorical outcomes. The processing tree contains three parameters (C, N, I) that represent the estimated probability that the observed response was based on the manipulated consequences, moral norms, or a general response bias for inaction, as illustrated in Table 1.

Table 1. CNI Model

	Proscriptive Norm		**Prescriptive Norm**	
	Benefits Greater	*Benefits Smaller*	*Benefits Greater*	*Benefits Smaller*
	Action	Inaction	Action	Inaction
	Inaction	Inaction	Action	Action
	Inaction	Inaction	Inaction	Inaction
	Action	Action	Action	Action

Note. Illustration of the CNI model based on Gawronski et al. (2017).

Based on the tree structure in Table 1, equations for action and inaction responses are formulated for each of the four CNI conditions by multiplying the parameters along a path

leading to a response and adding all paths leading to the same response. For instance, an action response in the prescriptive condition, where the benefits are greater than the costs, may either arise by reacting to the consequences [C], or by reacting to the norms given that the response is not produced by a reaction to the consequences [(1-C)×N], or by an action bias to always select action given that the response is neither produced by a reaction to the consequences nor to the norms [(1-C)×(1-N)×(1-I)]. Accordingly, p(action|prescriptive norm, benefits > costs) = C + [(1-C)×N] + [(1-C)×(1-N)×(1-I)].

Since action and inaction are complementary response options, Gawronski et al. (2017) formulate four non-redundant equations that quantify the probabilities of selecting an action and inaction, respectively, across the four CNI conditions. Modeling the responses as coming from a multinomial likelihood distribution with response probabilities given by the model equations, the three model parameters can be estimated via either maximum likelihoods methods or Bayesian statistics. It is then regularly tested whether C and N parameters differ from 0 and whether the I parameter diverges from 0.5, via 95% confidence intervals or credible intervals, respectively.

While previous studies with the traditional sacrificial dilemma have indicated a positive correlation between psychopathic traits and utilitarian sacrifices (Marshall et al., 2018), one of the interesting findings of the CNI model is that its parameters tend to be negatively correlated with psychopathic traits (Gawronski et al. 2017; Körner et al. 2020; Luke & Gawronski, 2021; Luke et al., 2021). While individuals high in psychopathy may be less opposed to the sacrifice of human life, this result indicates that they also tend to be less influenced by the difference of whether sacrifice occurs when the benefits for the greater good are larger versus smaller than the costs of the outcome. More recently, the CNI model has been further extended to permit the study of individual differences by assessing its parameters at the individual level (e.g., Kroneisen & Heck, 2020; Körner et al. 2020).

The Invariance Assumption

Since only three MPT parameters are used to parameterize the multinomial likelihood distribution, one of the assumptions of the model is that the N parameter stays invariant across proscriptive and prescriptive norms, and that the C parameter stays invariant across the four CNI conditions. In other words, the model assumes that the strength of deontological norms is invariant to whether the norms forbid doing a questionable action (e.g., killing someone) or whether the norms prescribe interfering with the actions of someone else to prevent an action (e.g., preventing someone else in killing someone). Similarly, the model assumes that the probability of judging a questionable action with desirable consequences (e.g., saving lives) acceptable on utilitarian grounds is the same as judging the probability of the same action unacceptable on utilitarian grounds when its consequences are less desirable (e.g., averting only a minor damage). The model also assumes that the N parameter is invariant with respect to costs and benefits, but unlike the invariance across prospective and prescriptive norms, this further invariance is not deemed theoretically problematic as it directly follows from the definition of norm-consistent behavior.

The CNI model is not alone in making this type of invariance assumption. The type of process-dissociation models that is used in Conway and Gawronski (2013) as a predecessor to the CNI model similarly makes an invariance assumption in its model equations. More generally, process-dissociation models form a subset of the class of multinomial processing-tree models. In Klauer et al. (2015), it was tested empirically whether prominent instances of process-dissociation models from cognitive psychology (Stroop task, cued recall) and social psychology (racial bias in the weapon task) violated the invariance assumption. In several instances strong violations were found and it was conjectured that similar violations of the process-dissociation model of Conway and Gawronski (2013) would occur as well.

What are the consequences of violations of the invariance assumptions? As discussed by Klauer et al. (2015), such violations have the potential to compromise estimates of the model

parameters and substantive conclusions drawn from them. Moreover, traditional analyses using the CNI model, that is premised on the invariance assumptions, unfortunately do not allow one to detect such violations, nor to assess the extent of distortions that may ensue from violations of invariance.

Via the addition of proscriptive and prescriptive norms, the CNI model improves upon the process-dissociation model of Conway and Gawronski (2013). Over and above the methodological issues surrounding the invariance assumptions, these assumptions are also at the root of a recent controversy surrounding the CNI model, however.

Controversy Surrounding the CNI Model

In a recent critical exchange between Baron and Goodwin (2020, 2021) and Gawronski et al. (2020), the CNI model was criticized on several counts. Some of these points could be addressed by the rebuttal in Gawronski et al. (2020) – in particular those concerning order-effects, the interpretation of the model and its parameters – but other points still stand.

Baron and Goodwin (2020, 2021) worry that the scenarios used to apply the CNI model leave interpretational ambiguities, which may help explain the high rates of so-called "perversive responses", where participants select responses that go against both deontological and utilitarian responses in congruent conditions, where both predict action (PreGreater) or inaction (ProSmaller). They argue that this makes the CNI scenarios unsuitable for studying inaction bias.

One of the other central arguments that Baron and Goodwin (2020, 2021) make is that the reason why deontological responses have previously been investigated mainly through inaction is that deontological norms prohibiting harmful action (e.g., "first, do no harm") are stronger than norms proscribing action to do good. In fact, Kantian deontology does contain obligatory ends of developing one's own talents and to helping others, as general preconditions for pursuing our ends (see, e.g., Herman, 2011; Scanlon, 2011). But it is the duties to never treat

other people as mere means and the prohibition on self-serving action plans that cannot pass the test of universalization by the categorical imperative (e.g., plans involving deception, coercion, or direct harm), which are more often associated with Kantian ethics.

Relatedly, other researchers have studied asymmetries between a strict, duty-based system of proscriptive, moral regulation, which is focused on blame and identifying transgression versus a prescriptive regulatory system, which is focused on credit-worthy good deeds that is more desire-based and less strict (Janoff-Bulman et al. 2009). Janoff-Bulman et al. examine the distinction between these two types of moral norms as being based on an asymmetry between two motivational systems in several studies, which roughly contrast the dimensions listed in Table 2.

Table 2. Motivational Systems Underlying Moral Regulation

Prescriptive Norms	**Proscriptive Norms**
Approach	Avoidance
Positive outcomes	Negative outcomes
Activation-based	Inhibition-based
What we should do	What we should not do
Not strict	Strict
Based on either duties or desires	Duty-based
Credit-oriented	Blame-oriented
Obligation: e.g., “help others”	Obligation: e.g., “not to harm others”

Note. The table lists some of the differences between the two motivational systems that Janoff-Bulman et al. (2009) take to be underlying the divide between prescriptive norms and proscriptive norms.

Henning and Hütter (2021) refer to these studies as providing evidence of a possible contrast between proscriptive and prescriptive norms, which makes them prefer a model without prescriptive norms over the original CNI model, where instead scenarios that differ in an inaction or an action default are contrasted (see also Henning & Hütter, 2020).

Yet, the distinction between proscriptive and prescriptive norms works differently in the context of the CNI dilemma than in Janoff-Bulman et al.’s contrast between two regulatory systems. For in the former case there are always two bad outcomes and conflicting motivations for choosing between action/inaction rather than cases of univocally good outcomes. So, the

idea of a motivational system aiming at activating approach behavior towards positive end-states from Janoff-Bulman et al. (2009), and its opposition to an inhibitory system, does not directly carry over to the prescriptive norms in the context of the CNI scenarios. At the root of this difference is the fact that the CNI model mainly introduces prescriptive norms to solve the methodological problem of avoiding a possible confound between an inaction chosen from deontological reasons and a general bias towards inaction. In contrast Janoff-Bulman et al. (2009) develop a particular substantive interpretation of prescriptive norms, which they take to be more broadly based on a general distinction between two self-regulatory systems, which is found across different domains in the psychology of motivation.

In the context of the CNI implementation of prescriptive norms, a deontologically prohibited action is planned by another agent and the participant can choose to intervene to prevent this from taking place. In this way, the same action choice between two bad outcomes is presented in a configuration in which one of the actions has already been preselected by another agent. As Baron and Goodwin (2020) point out, this may lead to a weaker prescriptive norm, if this other person is a colleague or superior, as in some of the CNI scenarios, since it introduces further, unintended consequences such as the following:

> when the action is to contravene someone else's action, it has additional consequences aside from preventing the consequences of that action. It may hurt the decision maker's feelings, possibly leading him or her to take retaliatory action against the one who contravenes. It may also violate the lines of authority, thus weakening these lines for the future by discouraging those in command from taking their responsibility seriously (Baron, 1996). It may also be illegal or against the rules, and rule following likewise has a value as a precedent for future cases. (p. 424)

Accordingly, Baron and Goodwin (2020, 2021) argue that that the CNI scenarios do not succeed in keeping the relative strength of the deontological norms constant across the proscriptive and prescriptive conditions. Moreover, since unintended consequences are introduced by the way

that prescriptive norms are manipulated, Baron and Goodwin (2020, 2021) also suggest that the consequences are not held constant across the two conditions. As they say (2021: 16): "the inferential problem results both from the difference in the norms between the two alternatives presented, as well as the difference in the consequence".

Both points lead to predictions of violations of the invariance assumption of the CNI model. Baron and Goodwin's (2020, 2021) arguments most strongly suggest that the invariance assumption for parameter N should be violated so that $N_{Pro} > N_{pre}$. *A priori*, a case could, however, also be made for the converse violation with $N_{Pro} < N_{pre}$. For example, scenarios with prescriptive norms ask whether a proposed non-normative action should be thwarted and raising this very possibility of averting the action may in itself act as a clue for participants suggesting that the action is to be considered problematic and should indeed be refused. In an experiment, we set out to test the invariance assumption for parameters N and C. Unlike the possible invariance violation for the N parameters, we did not have prior expectations about the possible rank order of the four C parameter. In this case, we merely set out to test the prediction in Klauer et al. (2015) that the probability of participants judging a questionable action with desirable consequences acceptable would not in general be the same as the probability of finding the same action unacceptable when its consequences are less desirable.

Experiment 1

To investigate the invariance assumption of the CNI model, we conducted an experiment following the procedure of Klauer et al. (2015), which was used to test violations of the invariance assumption in process-dissociation models. To this end, the MPT equations of the CNI model are implemented in a Bayesian framework via hierarchical latent trait model proposed in Klauer (2010), which has also been applied to study individual variation in the context of the CNI model in Kroneisen and Heck (2020).

Since further degrees of freedom are needed to estimate separate parameters for N in the proscriptive and prescriptive condition and for C across all four CNI conditions, the model was extended via a skip option ("S"), whereby participants could opt out of selecting action/inaction in a given scenario. The MPT equations of this CNIS model are stated in Appendix A. The addition of this skip option was further motivated by reading participants' open-ended responses in Berentelg's (2020) replication study, where it was found that a sizable minority of participants complained about the exclusion of alternative courses of action in particular scenarios. Accordingly, if participants find the scenario ambiguous or the stipulation of the choice situation artificial (with neither C nor N favoring a unique choice, because information has been left out), they are permitted to skip the scenario via this extension of the CNI model.

The interpretation of the skip option is grounded in the logic of the CNI model which is our point of departure. According to that model, when consequences are activated (with probability C), the response is determined by consequences with probability 1, whether or not norms are activated and whether or not the dilemma is congruent or incongruent. When consequences are not activated (with probability 1-C), but norms are activated (with probability N), then the response is determined by norms with probability 1, whether or not the dilemma is congruent or incongruent. And thus, when consequences or norms are activated, the response is deterministically captured by all-or-none processes with consequences dominating norms.

Only when neither consequences nor norms are activated (with probability (1-C) × (1-N)) are responses not deterministic. In this state of uncertainty, participants, metaphorically speaking, throw a loaded dice which comes up with "inaction" with probability (I) and with "action" with probability (1-I). Given that participants were explicitly instructed to use the skip option in the case of uncertainty, this state of uncertainty, reached with probability (1-C) × (1-N), is the only place in the model in which skipping can come into play. Basically, the

extension of the CNI model we present provides participants with a third face on their loaded dice, which now shows the faces "action", "inaction", and "skip". Because in the state of uncertainty, neither norms nor consequences are activated, it also makes sense that the I parameter and the S parameter do not depend upon type of dilemma, because the four CNI conditions are distinguished solely in terms of differences in norms and consequences.

Through two experiments, we test whether such invariance violations occur through the addition of the S parameter to the CNI model (see Appendix A for further details).

Method

Open Science Framework (OSF) link:

https://osf.io/569bv/

Sampling Procedures Shared by all Experiments

To reduce the dropout rate during the experiment, participants first went through three pages stating our academic affiliations, posing two SAT comprehension questions in a warm-up phase, and presenting a seriousness check asking how careful the participants would be in their responses (Reips, 2002). The following *a priori* exclusion criteria were used: not having English as native language, completing the task in less or more than the average response time $\pm$ 2 × SD, failing to answer at least one of two simple SAT comprehension questions correctly in a warm-up phase, and answering ‘not serious at all’ to the question 'how serious do you take your participation' at the beginning of the study.

Participants

The experiment was conducted over the Internet through the platform Mechanical Turk to obtain a large and demographically diverse sample. A total of 778 people finished the experiment. The participants were paid a small amount of money for their participation and sampled from USA, UK, Canada, and Australia. After applying our *a priori* exclusion criteria,

the final sample consisted of 486 participants. Mean age was 39.10 years, ranging from 19 to 76.[1] 54.11% of participants identified as male; 45.24% identified as female; 5 participants preferred not to identify with either category. 84.06 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment had a within-participants design with the following factors varying within participant: Consequence (Smaller vs. Greater) and Norm (Proscriptive vs. Prescriptive). To allow for 10 trial replications for each of the four CNI conditions, each participant in total went through 40 within-subject conditions.

Materials and Procedures

Participants were presented with the four CNI conditions across 10 scenarios adopted from Gawronski et al. (2017) and Körner et al. (2020).[2] The scenarios were modified slightly so that the proposed action was the *intended* means needed to realize the described outcome uniformly across all scenarios. In contrast, in the existing CNI stimulus materials, the manipulated consequence sometimes constituted a foreseeable consequence.[3]

To illustrate, consider the following Vaccine scenario in the ProGreater condition:

> You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because of its severe side-effects. The side-effects of the vaccine will likely cause the death of dozens of people who are not infected, but the vaccine will save hundreds of lives by preventing spread of the virus. Is it acceptable in this case to use the vaccine?

---

1 We are here ignoring one occurrence of the value '5' given that Mechanical Turk limits the participation to adults.

2 The complete list of scenarios can be found at: https://osf.io/4ny6q.

3 This distinction matters for the Doctrine of Double-Effect in deontological ethics (McIntyre, 2019).

While many of the other scenarios make the sacrifice of innocent lives a causal prerequisite for saving further lives, the death of the innocent people is here described as a side-effect, and while other scenarios provide victim counts in numbers, this scenario uses imprecise verbal descriptions. In Table 3 below, we modified both of these aspects of the scenario. In addition, we had a native speaker to read through the scenarios to suggest modifications to the English formulations.

Table 3. CNI Conditions

ProGreater	*ProSmaller*	*PreGreater*	*PreSmaller*
You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 500 lives by preventing spread of the virus.	You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 100 lives by preventing spread of the virus.	You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 500 lives by preventing spread of the virus. The plan is to use the vaccine, but you could veto this.	You are a doctor in an area that suffers from an outbreak of a highly contagious disease. Preliminary tests have shown the success of a new vaccine that is not approved by the health department of your country, because producing this vaccine is done by infecting people with the disease and then using their blood for the synthesis. You would need to infect 100 people in this way to produce enough serum. The people infected would definitely die, but the vaccine will save about 100 lives by preventing spread of the virus. The plan is to use the vaccine, but you could veto this.
Is it acceptable in this case to infect 100 people?	Is it acceptable in this case to infect 100 people?	Is it acceptable in this case to veto the infection?	Is it acceptable in this case to veto the infection?
Yes, it is acceptable vs. No, it is not acceptable vs. Skip			

Note. Example of one of the modified CNI scenarios. See https://osf.io/4ny6q for the full list of modified scenarios.

The order of the scenarios and the CNI conditions within scenarios were randomized for each participant anew.[4] Because different versions of the scenario look similar, the randomization was constrained so that different versions of the same scenario could not occur in immediate succession. Following Gawronski et al. (2017), participants were given the following instruction:

> On the following pages you will see 40 scenarios that people may come across in life. Please read them carefully. Even though some scenarios may seem similar, each scenario is different in important ways. After each scenario, you will be asked to make a judgment about whether you find the described action acceptable or inacceptable. Please note that some scenarios refer to things that may seem unpleasant to think about. This is because we are interested in people's thoughts about difficult, real-life issues.

In addition, participants were instructed that they could "skip" a moral decision for cases, where they were undecided about whether the described action was morally acceptable or inacceptable. They were also instructed that they should not make use of this option more than 10 times. Finally, some demographic questions were asked and participants' level of psychopathy was probed via Levenson et al.'s (1995) subscale for primary psychopathy in a noninstitutionalized population.

Results

For the analysis, we first fitted the original 4-parameter version of the CNIS model ($CNIS_4$) to ensure construct validity after the addition of the skip option to the CNI model. In this analysis, we test whether the CNIS model is able to replicate the mean pattern observed

[4] Garownski et al. (2017) use a pseudo-random order, which is fixed to be the same for each participant. In pilot studies, we did not find differences between this procedure and the more rigorous randomized order and chose the latter instead.

for the model parameters as well as bivariate associations with external parameters (primary psychopathy, gender) reported in Gawronski et al. (2017).

Following this analysis, a 8-parameter version of the CNIS model ($CNIS_8$) was fitted with 2 separate N parameters (N_{pro}, N_{pre}) and four separate C parameters ($C_{ProGreater}$, $C_{ProSmaller}$, $C_{PreGreater}$, $C_{PreSmaller}$). This allows us to test for violations of the invariance assumption. Finally, we extend these findings by fitting two structural equation models (SEM) to investigate whether the replicated gender effects are mediated through the association of gender and primary psychopathy.

For a Bayesian implementation of the MPT models, we followed the hierarchical extension of multinomial processing trees in Klauer (2010), which has also been implemented in the R package `TreeBUGS` (Heck et al., 2018). One of the benefits of the latent trait approach to MPT modeling proposed in Klauer (2010) is that its hierarchical structure makes it well-suited to estimate individual CNI parameters for each participant (Kroneisen & Heck, 2020). In addition, the individual MPT parameters are estimated through a multivariate normal distribution with a covariance structure that permits correlations among the individual MPT parameters, instead of stipulating *a priori* that they must be uncorrelated along the form of the Beta-MPT approach (Smith & Batchelder, 2010). We illustrate the hierarchical latent trait model of Klauer (2010) in Appendix A. The same appendix also states the MPT model equations for the extension of the CNI model with the skip parameter ("S") and explains how the invariance assumption distinguishes $CNIS_4$ from $CNIS_8$.

CNIS with Four Parameters

Figure 1 displays the distribution of the parameters estimated for each participant:

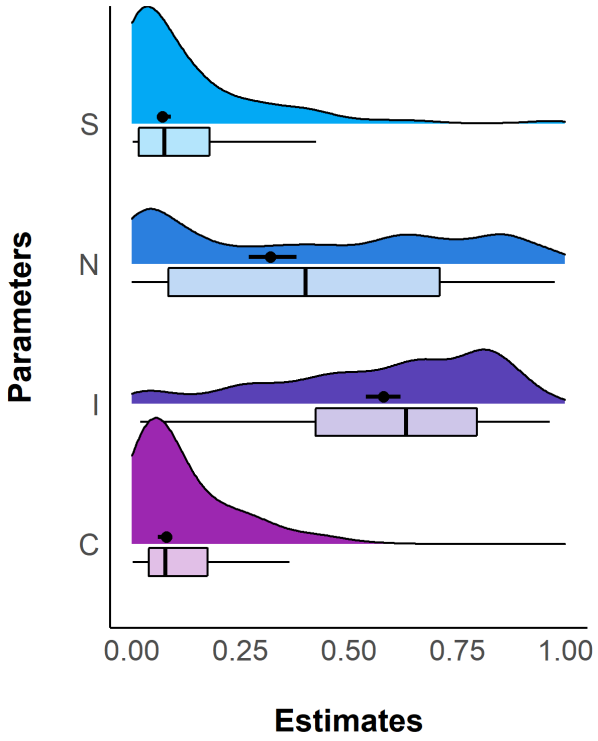


Figure 1. Distributions of the CNIS parameters estimated for each participant with boxplots indicating the quartiles of the individual estimates. The black points and lines indicate the posterior medians and 95% HDI of the group-level means.

As Figure 1 shows, the 95 % HDI[5] for the posterior medians of the C and N parameters exclude zero, and a general bias towards inaction is found, since the 95 % HDI of the posterior median of the I parameter excludes .5.

Since published work reports bivariate correlations, and the first goal is to replicate previous results, we here plot bivariate correlations between the C, N, I parameters and primary psychopathy (P), self-reported gender (G) with ‘male’ encoded as 1 and ‘female’ encoded as 0, and total response time (T):

[5] A HDI interval is an interval of the posterior distribution where all points within the interval have a higher probability density than points outside it.

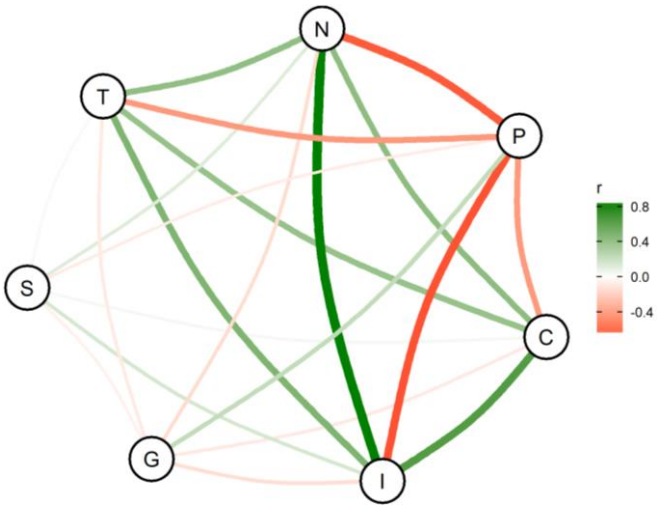


Figure 2. Bivariate associations between model parameters and external variables. 'G' = self-reported gender (excluding five participants, who preferred not to respond), 'P' = primary psychopathy, 'T' = total response time to complete the items, 'S' = skip.

As Figure 2 shows, negative bivariate associations between the CNI parameters and psychopathy were found (r_{PC} = -.44, 95% HDI [-.50, -.36]; r_{PI} = -.70, 95% HDI [-.74, -.65]; r_{PN} = -.67, 95% HDI [-.72, -.62]). In addition, Figure 2 shows that male participants scored higher on primary psychopathy (r_{GP} = .23, 95% HDI [.14, .31]) and that male participants scored lower than females on both the N (r_{GN} = -.15, 95% HDI [-.23, -.06]) and I parameter (r_{GI} = -.14, 95% HDI [-.23, -.05]).

These results replicate the findings in Gawronski et al. (2017) while adding the S parameter to the CNI model. Below we will use structural equational modeling (SEM) to further analyze mediation relationships in these results. But first we need to find out whether the invariance assumption is violated in the CNI model by contrasting the present model with a 8-parameter version.

CNIS with Eight Parameters

Next, we tested the invariance assumption by fitting separate N and C parameters in a 8-parameter version of the CNIS model. The parameter estimates are displayed in Figure 3 below.

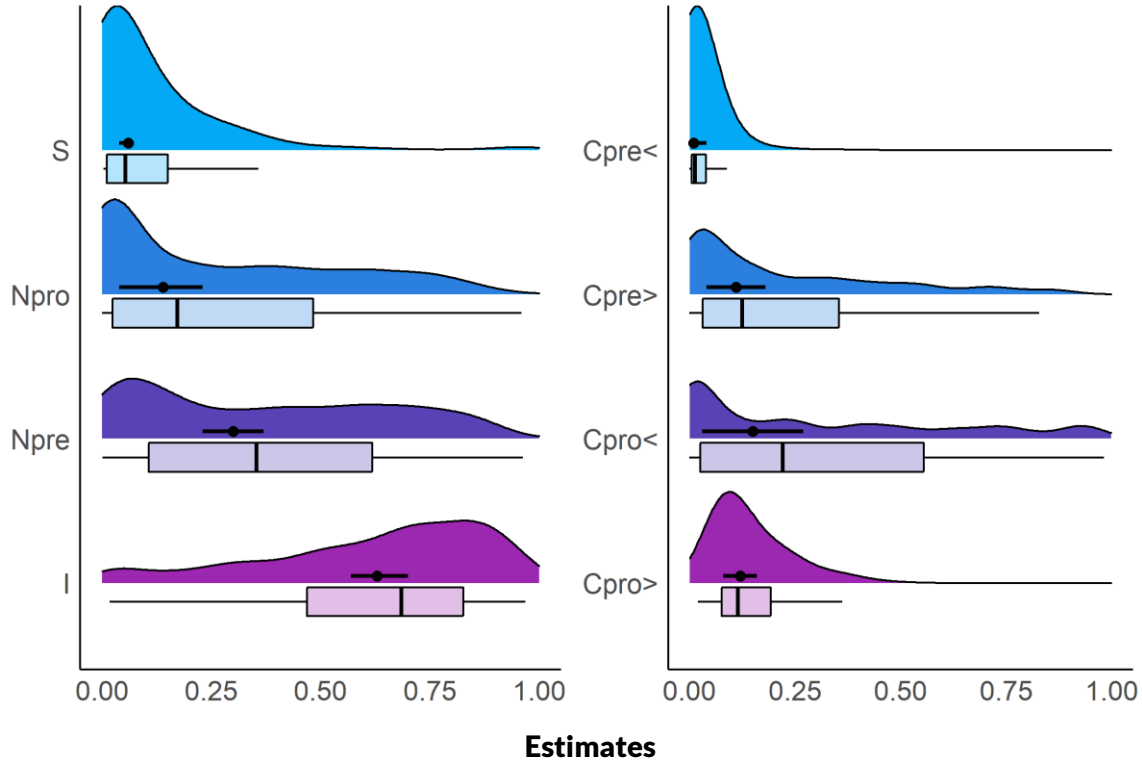


Figure 3. Distributions of the CNIS parameters estimated for each participant with boxplots indicating the quartiles of the individual estimates. The black points and lines indicate the posterior medians and 95% HDI of the group-level means. 'Cpro>' = $C_{ProGreater}$, 'Cpro<' = $C_{ProSmaller}$, 'Cpre>' = $C_{PreGreater}$, 'Cpre<' = $C_{PreSmaller}$.

To test possible violations of the invariance assumption that Npro = Npre and that $C_{proGreater} = C_{proSmaller} = C_{preGreater} = C_{preSmaller}$, we analyzed contrasts of pairs of parameters on the probability scale by identifying whether the 95% HDI intervals for the difference between the two parameters included zero. Credible differences were found for Npre > Npro ($\Delta^{N}_{pro\text{-}pre}$ = -0.16, 95% HDI [-0.27, -0.04]), $C_{proGreater} > C_{preSmaller}$ ($\Delta^{C}_{proGreater\text{-}preSmaller}$ = 0.11, 95% HDI [0.06, 0.15]), $C_{proSmaller} > C_{preSmaller}$ ($\Delta^{C}_{proSmaller\text{-}preSmaller}$ = 0.14, 95% HDI [0.02, 0.26]), and $C_{preGreater} > C_{preSmaller}$ ($\Delta^{C}_{preGreater\text{-}preSmaller}$ = 0.10, 95% HDI [0.02, 0.18]).

Finally, we compared the two models in terms of their expected out-of-sample predictive accuracy via information criteria and found $CNIS_8$ to be the better fitting model, as shown in Table 4 below.

Table 4. Model Comparison

	WAIC	LOOIC	Δelpd (SE)	p_{T1}	p_{T2}
$CNIS_4$	9435.8	9764.2	-79.62 (14.77)	< .0001	< .0001
$CNIS_8$	9230.0	9605.0	--	.03	< .01

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. Note that information criteria can take both positive and negative values and that the lowest value on the real line still indicates best fit. The test statistics T_1 and T_2 represent Bayesian *p* values and are based on the posterior predictive model checks in Klauer (2010).

In addition, model fit was assessed with the posterior-predicted *p* values based on T_1 and T_2 posterior model checks proposed in Klauer (2010). T_1 measures the adequacy of the models in capturing the mean observed outcome frequencies (aggregated across persons). T_2 measures the adequacy of the models in capturing the variability (variances and covariances) among the observed response frequencies (computed across persons). The proportion with which T_i(observed) < T_i(predicted) are given by Bayesian *p* values. A small *p* value for these test statistics indicates that the posterior predictive distribution of the model fails to capture an aspect of the data. It was found for both the aggregate outcome frequencies and the variability across individuals that both models failed to capture aspects of the data. This is not unusual for large data sets such as the present, but the comparison also shows that $CNIS_8$ performed better than $CNIS_4$.

Structural Equational Modeling

Next, we fitted two structural equation models with the R-package `blavaan` (Merkle & Rosseel, 2018) based on the winning $CNIS_8$ model.[6] Structural equation modeling (SEM) is a generalization of regression models used for causal inference in statistics, which models the covariance matrix. Some of its benefits are to permit the estimation of direct and indirect effects of explanatory variables as well as imposing conditional independence constraints

[6] We performed the same SEM analysis on $CNIS_4$, which produced the same qualitative results. Further details can be found in the supplementary materials or on the OSF project page: https://osf.io/569bv/

from a causal model (Kline, 2016; Shipley, 2016). In the context of our study, we used this statistal tool to investigate mediation effects on the relationship between the CNI model parameters and external variables like gender, primary psychopathy, and response time.

To conduct this analysis, we compared two SEM models. These two models differed on whether direct paths were included from gender to the CNI parameters (SEM_1) or whether the effect of gender was completely mediated through the effect of primary psychopathy on the CNI parameters (SEM_2), as displayed in Figure 4.

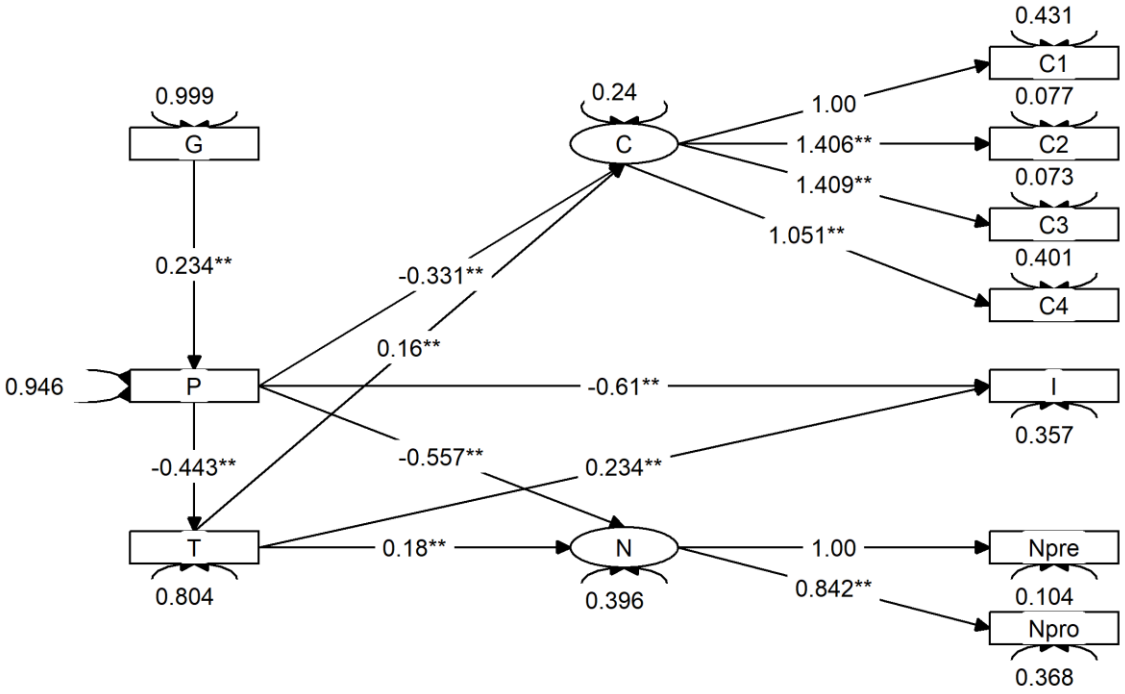


Figure 4. SEM_2 model. Path coefficients indicate posterior medians and are marked with '**', if the 95% HDI interval does not include 0. 'G' = self-reported gender with 'male' = 1 and 'female' = 0 (excluding five participants, who preferred not to respond), 'P' = Primary psychopathy, 'T' = total response time for all the items. Both 'P' and 'T' were scaled to take values between 0 and 1 before fitting the model to prevent large differences in the variances of the different parameters. Direct and indirect effects are encoded via arrows. The loops indicate variances. The covariances have been left out to simplify the graph. Latent variables are marked with circles. The scales of the latent variables were fixed by setting the first path coefficient equal to 1.0. All the variables were z-transformed before fitting the structural equation models to ensure that their variance were of the same order of magnitude. 'C1' = $C_{ProGreater}$, 'C2' = $C_{ProSmaller}$, 'C3' = $C_{PreGreater}$, 'C4' = $C_{PreSmaller}$.

The C and N parameters in Figure 4 are estimated latent variables, which are measured via C1-C4 and Npre/Npro, respectively. The violations of the invariance assumption can be read off from the differences in the standardized path coefficients from the latent C and N parameters to the C1-C4 and Npre/Npro parameters.

Table 5 shows a model comparison between SEM_1 and SEM_2 as well as a mediation analysis of SEM_1, which shows why including direct paths from gender to the CNI parameters does not improve the fit of the model. This in turn creates a slight preference for the SEM_2 model displayed in Figure 4. A feature of SEM_2 is that its underlying directed acyclic graph (DAG) entails the following conditional independencies, which imply that the partial correlations between gender and the CNI parameters are zero, when controlling for the influence of primary psychopathy.

C _||_ G | P G _||_ I | P G _||_ N | P G _||_ T | P

In words: gender is independent of the C, N, and I parameters when conditioning on primary psychopathy.

Table 5. SEM Models

	Model Comparison			
	WAIC	*LOOIC*	*Δelpd (SE)*	R^2
SEM_1	8634.017	8639.281	-3.84 (1.98)	C=.43, N=.52, I=.61
SEM_2	8629.955	8631.595	--	C=.43, N=.52, I=.61

	Mediation Analysis based on SEM_1			
	Direct Path G → X	*Indirect Path* G → P → X	*Total Effect: Direct + Indirect*	*Proportion Mediated*
C	$\tilde{x}$ = -.02 [-.07, .03]	$\tilde{x}$ = -.08 [-.11, -.04]	$\tilde{x}$ = -.10 [-.15, -.04]	0.832
N	$\tilde{x}$ = .03 [-.03, .08]	$\tilde{x}$ = -.13 [-.18, -.08]	$\tilde{x}$ = -.11 [-.18, -.03]	1.456
I	$\tilde{x}$ = .02 [-.03, .08]	$\tilde{x}$ = -.14 [-.20, -.09]	$\tilde{x}$ = -.12 [-.20, -.05]	1.271

Note. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. Note that information criteria can take both positive and negative values and that the lowest value on the real line still indicates best fit. The proportion mediated can take values larger than one in cases where the direct and indirect effects are of opposite signs, as here. The square brackets indicate 95% HDI.

The results in Table 5 show that the negative correlations between gender and N and I that are found in the bivariate correlations (which indicate that male participants scored lower than females) are completely mediated by the effect of gender on primary psychopathy.

A further advantage of structural equation modeling is that it gives a principled way of identifying the minimal adjustment set of covariates that need to be controlled for to avoid spurious correlations (Pearl, 2009; Kline, 2016). Applying the graphical criteria from Pearl (2009) on the underlying DAG in SEM_2, we thus find that associations between the CNI

parameters and response time need to control for primary psychopathy to avoid spurious correlations, because P acts as a common cause on the T and CNI parameters in Figure 4.

In Appendix C, SEM models that include the S parameter are contrasted and an integrative model that includes both $CNIS_8$ and the best fitting SEM model is fitted to the data. It is found that the main results of the analysis above are replicated, when this SEM model and $CNIS_8$ are combined into one integrative model to permit the propagation of uncertainty from the estimated CNIS parameter to the structural equation analysis.

Discussion

To test the construct validity of the CNIS model, a 4-parameter version was first fitted to the data, and it was tested whether known patterns of means for the CNI parameters and known relations of the CNI parameters to primary psychopathy and gender could be replicated. Like previous work, we found evidence for parameters N and C to be substantially larger than zero, and for the inaction parameter to exhibit a credible bias towards inaction ($I > .5$). In addition, it was found that the C and N parameters were negatively associated with primary psychopathy and that male participants scored lower on the N and I parameters than females. This result replicates previous work reporting similar gender effects and negative associations between the model parameters and primary psychopathy with mixed findings concerning a negative association with the C parameter across studies (Gawronski et al. 2017; Körner et al. 2020; Luke & Gawronski, 2021; Luke et al., 2021).

In a second analysis, a 8-parameter version of the CNIS model was fitted to the data and it was found that credible differences between the N and C parameters emerged. This indicates a violation of the invariance assumption. In a further exploratory analysis, we investigated whether the invariance assumption was violated within each item. The item specific estimates of the 8 MPT parameters are reported in Appendix B, and it is found that violations of the invariance assumption occur almost within every scenario tested.

As explained above (Section “The Invariance Assumption”), violations of invariance have the potential to compromise estimates of the model parameters by introducing systematic bias and to invalidate substantial conclusions drawn from them. Comparing the parameter estimates and results pattern for the 4- and 8-parameter versions of the CNIS model suggests that the consequences in terms of substantive conclusions were relatively minor for the present data: Both models yielded roughly similar overall patterns of mean parameter estimates and correlational results. This need of course not be the case for other data sets and situations; there is simply no way to tell unless the model is extended as exemplified here to allow one to estimate separate N and C parameters.

Experiment 2

The goal of Experiment 2 was to replicate the findings of Experiment 1 concerning violations of the invariance assumption for the C and N parameters. In addition, Experiment 2 introduced a manipulation aimed at the S parameter to test whether it was possible to selectively influence the S parameter in an experimental comparison. Finally, Experiment 2 added a third model to the model comparison, which estimates four S parameters (one for each of the four CNI conditions). We added this third model to test in a model comparison which of the following two models fits the data best: 1) a model that avoids the invariance assumption for the C and N parameters ($CNIS_8$), or 2) a model that avoids the invariance assumption for the S parameter ($CNIS_7$). Limited by the degrees of freedom in our data, we in this way tested which of the different invariance assumptions led to the worse fit of the data: an invariance assumption in the C and N parameters or an invariance assumption in the S parameter.

Method

OSF link:

https://osf.io/569bv/

Participants

Unless otherwise noticed, Experiment 2 followed the design, sampling procedure, and materials of Experiment 1. A total of 1124 people finished the experiment. After applying our *a priori* exclusion criteria, the final sample consisted of 1040 participants. Mean age was 40.97 years, ranging from 19 to 78. 45.05% of participants identified as male; 53.60% identified as female; 14 participants preferred not to identify with either category.[7] 69.26 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment had a mixed design. Between-participants, the factor "Skip Anchor" (10% Skip Anchor vs. 25 % Skip Anchor) was varied. Within participant, the same two factors were varied as in Experiment 1 (Consequence and Norm) with 10 trial replications, thus resulting in 40 within-participant conditions.

Materials and Procedures

To manipulate the size of the S parameter in a between-participants comparison, two different instructions for how to use the Skip response were shown to the participants. Both groups were instructed to use the Skip option for cases where they were undecided about whether the described action was morally acceptable or inacceptable. The two groups differed in that one group (N = 538) was cautioned about the possibility of false positives and provided with an anchor that we typically observe that at most 10% of the responses consist of Skip-responses.

> Please only use this option when undecided. It is better if you answer action or inaction than completely skip a decision! As a guideline, for a typical scenario, we observe that at most 10% of the responses are skip options.

[7] For the analyses below, we focus on the 1026 participants who did identify with either male or female.

The second group (N = 502) was cautioned about the possibility of false negatives and provided with an anchor that we typically observe that at least 25% of the responses consist of Skip-responses.

> Please always use this option when undecided. It is better if you skip a decision than mistakenly answer action or inaction! As a guideline, for a typical scenario, we observe that at least 25% of the responses are skip options.

In both cases, we were careful not to introduce a count that sets an upper limit of the number of skip responses to ensure that each trial had the same probability of activating a skip response.

Results

As an initial manipulation check, it was found that 80.30% of the participants made use of the skip option in the 25% anchor condition and that 12.84% of the responses were skip responses. In contrast, 45.82% of the participants made use of the skip option in the 10% anchor condition and it was found that 3.89% of the responses were skip responses. It was found that the proportions of the three outcomes differed significantly across the two conditions, $\chi^2(2) = 1073$, $p < .0001$.

To better investigate the influence of the anchor manipulation in the context of the CNI model, the following models were contrasted in a model comparison:

> **CNIS$_4$**: The original CNI model with the S parameter added. Model with three invariance assumptions: C-invariance, N-invariance, and S-invariance.
>
> **CNIS$_7$**: The model builds on CNIS$_4$ but avoids the S-invariance assumption by estimating four S parameters (one for each of the four CNI conditions).
>
> **CNIS$_8$**: The model builds on CNIS$_4$ but avoids the C-invariance and N-invariance assumptions by estimating four C parameters (one for each of the four CNI conditions) and 2 N parameters (one for proscriptive norms and one for prescriptive norms).

For each of the two between-participants conditions, these models were fitted separately. Like in Experiment 1, we quantify the fit of the model in terms of how low an absolute value the information criteria, LOOIC and WAIC, have on the real line. In addition, in Table 6 we test whether there are statistically significant misfits of the models as indicated by the posterior predictive checks (T_1, T_2) proposed in Klauer (2010).

Table 6. Model Comparison

	LOOIC	Δelpd	SE	WAIC	Weight	p_{T1}	p_{T2}
			10% Anchor				
CNIS$_4$	9239.1	-83.3	15.4	8975.0	.00	< .0001	< .0001
CNIS$_7$	9133.1	-30.3	13.4	8845.0	.32	< .0001	< .0001
CNIS$_8$	9072.4	0	--	8775.8	.68	.052	< .01
			25% Anchor				
CNIS$_4$	12659.1	-333.04	26.49	12412.5	.00	< .0001	< .0001
CNIS$_7$	12168.0	-87.50	17.48	11894.6	.16	.097	< .0001
CNIS$_8$	11993.0	0	--	11703.3	.85	.13	< .02

Note. 'elpd' = expected log predictive density. elpd is a measure of out-of-sample predictive adequacy. LOOIC = -2*elpd. The weights are stacking weights based on LOOIC. Note that information criteria can take both positive and negative values and that the lowest value on the real line still indicates best fit.

As the comparison in Table 6 shows, CNIS$_8$ performed best for both anchor conditions in light of both the trade-off between fit and parsimony measured by the information criteria (LOOIC and WAIC). It was, moreover, found that a model that makes all three invariance assumptions (CNIS$_4$) is incapable of capturing the mean observed outcome frequencies across both the 25% and 10% anchor conditions (T_1). Similarly, it was found that a model that makes the C and N invariance assumptions (CNIS$_7$) was only able to capture the mean observed outcome frequencies in the 25% anchor condition. In contrast, a model (CNIS$_8$) that avoids the invariance assumption for both the C and the N parameter was found to be capable of passing this posterior predictive check, across both anchor conditions. Yet, none of the models was capable of capturing the variability across individuals as quantified by the T_2 posterior predictive check in each of the between-participants conditions. We attribute this result to the large sample sizes that these conditions had.

Next, Figure 5 displays the parameter estimates of $CNIS_8$ fitted separately to each of the two conditions.

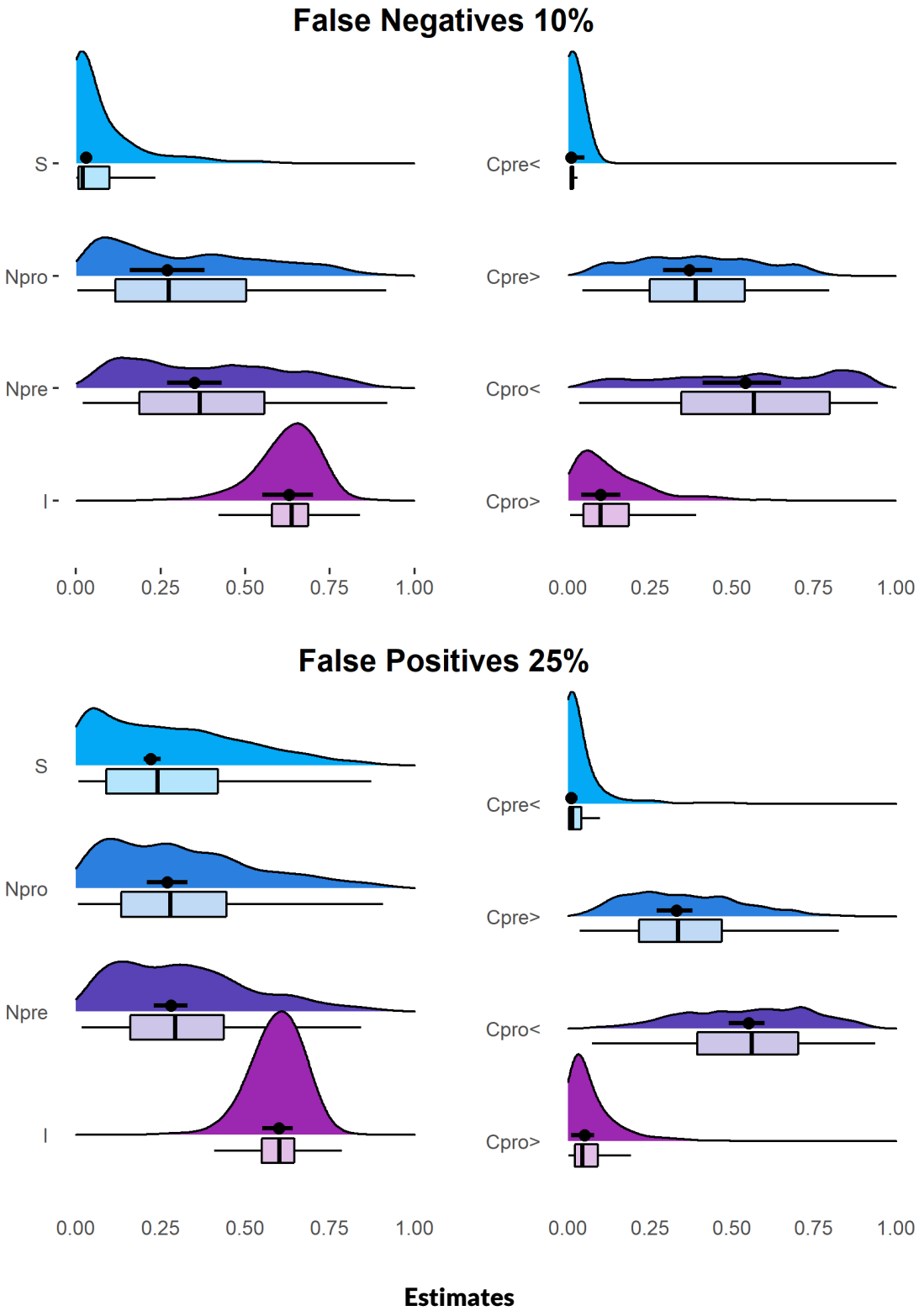


Figure 5. Distributions of the CNIS parameters across the two between-participants conditions. The parameters were estimated for each participant with boxplots indicating the quartiles of the individual

estimates. The black points and lines indicate the posterior medians and 95% HDI of the group-level means. 'Cpro>' = $C_{ProGreater}$, 'Cpro<' = $C_{ProSmaller}$, 'Cpre>' = $C_{PreGreater}$, 'Cpre<' = $C_{PreSmaller}$.

Finally, contrasts in the mean estimates of the parameters of the winning model were investigated across the 10% and 25% anchor conditions (see Table 7).

Table 7. Contrasts in the Parameters of $CNIS_8$

Contrast	$\tilde{\Delta}$	95% HDI
$C_1^{25\%} - C_1^{10\%}$	-.05	[-.12, .02]
$C_2^{25\%} - C_2^{10\%}$	.008	[-.11, .16]
$C_3^{25\%} - C_3^{10\%}$	-.04	[-.14, .05]
$C_4^{25\%} - C_4^{10\%}$	-.001	[-.04, .03]
$I^{25\%} - I^{10\%}$	-.04	[-.13, .06]
$N_{pre}^{25\%} - N_{pre}^{10\%}$	-.07	[-.16, .03]
$N_{pro}^{25\%} - N_{pro}^{10\%}$	-.006	[-.12, .13]
$S^{25\%} - S^{10\%}$	.19	[.16, .22]

Note. The contrasts were calculated based on the differences in 10,000 random posterior draws of the mean CNIS parameters of the $CNIS_8$ model in the 25% and 10% anchor conditions.

As Table 7 shows, the anchor manipulation had a selective influence on the S parameter, since this was the only model parameter for which the 95% HDI did not include 0. In accordance with prior expectations, the mean of the S parameter was higher in the anchor 25% condition, when participants were cautioned against false negatives and informed that on average 25% of the responses tended to be skip-responses.

Discussion

In Experiment 2, the result from Experiment 1 that the $CNIS_8$ model outperforms the $CNIS_4$ model was replicated in two new between-participants conditions. In both experiments, it was found that a model that estimates separate C and N parameters for the CNI conditions performs better than a model that sets these equal.

To encounter the potential criticism that the $CNIS_8$ model trades two problematic invariance assumptions (concerning the C and N parameters) for a new invariance assumption

(concerning the S parameter), a further model was included in the model comparison in Experiment 2, which estimates separate S parameters for each of the four CNI conditions. In a model comparison, it was found that this new model ($CNIS_7$) also performs worse than the $CNIS_8$ model. Our results thus indicate that enforcing the invariance assumption for the C and N parameters leads to severe misfit with the data and that this finding is replicable.

Next, we investigated whether the introduction of a between-participants manipulation of both the severity of false positives and false negatives and the size of an anchor of how often other participants made use of the skip option would have a selective influence on the S parameter. By investigating contrast effects based on the $CNIS_8$ model, it was found that of all its model parameters, the 95% HDI interval only excluded credible effects of a zero contrast for the S parameter. It was thus found that indeed the effect of the anchor and error type manipulation was circumscribed to the S parameter, as a validation of its psychological interpretation.

General Discussion

The CNI model (Gawronski et al., 2017) has advanced the computational modeling of moral judgments by systematically pairing factors that are normally confounded in traditional research on moral judgment via Trolley-type dilemmas. Using multinomial processing trees and scenarios with four contrast cases, the CNI model attempts to dissociate adherence to utilitarianism and deontology in participants' case judgments.

At the same time, the model is surrounded by controversy concerning its underlying assumptions and their implications for moral psychology (see e.g., Baron & Goodwin, 2020, 2021; Gawronski et al., 2020). Part of the latter controversy implicitly concerns an invariance assumption made by process-dissociation models and related MPT models alike, which has been found problematic in other domains of psychology in Klauer et al. (2015). For estimating adherence to Utilitarianism and Deontology, the CNI model assumes that the

probability of judging a questionable action with desirable consequences (e.g., saving lives) acceptable on utilitarian grounds is the same as judging the probability of the same action unacceptable on utilitarian grounds when its consequences are less desirable (e.g., averting only a minor damage). Similarly, the model assumes that the strength of deontological norms is invariant to whether the norms forbid doing a questionable action (e.g., killing someone) versus whether the norms prescribe interfering with the actions of someone else to prevent an action (e.g., preventing someone else in killing someone).

To investigate these invariance assumptions, we compared two hierarchical Bayesian implementations of the CNI model in two experiments. The models differ in whether they assume different or the same parameters for utilitarian and deontological judgments in these contrast cases. What enabled the estimation of the parameters of the CNI model without the invariance assumption was extending the CNI paradigm with a skip option and the CNI model by a S parameter ("skip"). It was found through a model comparison in both experiments that the extended 8 parameter version of the CNIS model, which does not make the invariance assumption, outperformed the 4 parameter version, which differs from it solely by making the invariance assumption (Tables 4 and 6).

While previous controversy surrounding the CNI model suggests that the invariance assumption would be violated, Baron and Goodwin (2020, 2021) strongly predict that such violations would take the form of $N_{Pro} > N_{pre}$. In contrast, the data show that the violations go in the opposite direction: $N_{Pro} < N_{pre}$. We offered a speculative account for why N_{pre} might be larger than N_{Pro} above based on the idea that presenting the possibility of overwriting the action of another agent pragmatically implicates that the action is to be considered problematic and should indeed be refused. Further violations of the invariance assumption occurred with respect to the C parameter, where it was found that the posterior median of $C_{PreSmaller}$ approaches zero and is reliably smaller than the posterior median of the C parameters in all other conditions. For PreSmaller scenarios, consequentialist choices imply judging

refusals to act inacceptable. We suspect that the double negation uniquely implied in understanding and making this particular choice leads to it being adopted only infrequently and thus to the depressed $C_{PreSmaller}$ parameter. Taken together, the extended CNIS model provides (a) a methodological tool for estimating the CNI parameters without the need for the problematic invariance assumption and (b) suggests interesting new hypotheses (e.g., possible roles for pragmatic implicatures and double negation) and thereby opens avenues for future research when violations of invariance are found.

The distributions of the individual C parameters in Figure 3 moreover indicate that the variance for the C parameter in the congruent conditions ($C_{ProSmaller}$, $C_{PreGreater}$) is larger than the variance for the C parameter in the incongruent conditions ($C_{ProGreater}$, $C_{PreSmaller}$). We refrain from interpreting this finding substantively, however, because it may have to do with the amount of statistical information that is available for estimating the different parameters, and hence the estimation uncertainty expressed in the variances, that may differ between the conflict scenarios and the congruent scenarios.

To rule out the possibility that this new 8 parameter version merely traded two problematic invariance assumptions concerning the C and N parameters for a new invariance assumption concerning the S parameter, a further model was included in the model comparison in Experiment 2. This further 7 parameter model enforced the invariance assumption for the C and N parameters but avoided it for the S parameter by estimating four separate S parameters, one for each of the four CNI conditions. Compared with the performance of the 8 parameter version, it was found that the 7 parameter model lead to a worse fit of the data (Table 6). It was thus found that the C and N invariance assumptions uniquely contribute to the misfit of the four parameter version.

To validate the psychological interpretation of the S parameter as representing a process that is activated if participants reach a state of stochastic uncertainty in the absence of either a norm or a consequence response, Experiment 2 introduced a manipulation targeting

the S parameter. In a between-participants comparison, it was investigated whether both emphasizing the severity of false positives and false negatives and providing anchors (10% vs. 25%) of the likelihood with which other participants produced a skip response on a given trial would selectively influence the S parameter. By investigating contrast effects of the best fitting model of Experiment 2, it was confirmed that indeed this manipulation had a selective influence only on the S parameter (Table 7).

Finally, based on a structural equation analysis of $CNIS_8$, we were able to show that the previously reported gender effects on the CNI parameters (e.g., Gawronski et al. 2017) were completely mediated by the association of gender with primary psychopathy (see Table 5 and Figure 4).

An additional finding in Figure 4 is that longer total response time is positively associated with the CNI parameters. Accordingly, a consequentialist response pattern, sensitivity to norms, and an inaction bias have a higher probability for participants who spend more time on the task. In contrast, primary psychopathy is found to be negatively associated with both total response time and the I parameter. This indicates that participants who score higher on primary psychopathy have a higher probability of spending less time on the task and having an action bias. In contrast, the negative associations between primary psychopathy and the C and N parameters indicate that participants who score higher on primary psychopathy are less sensitive to the effect of norms (proscriptive vs. prescriptive) and to whether the outcomes benefit the greater good (greater benefit vs. smaller benefit), in line with previous results (Gawronski et al. 2017; Körner et al. 2020; Luke & Gawronski, 2021).

That the total response time was positively associated with all of the C, N, and I parameters indicates that in the context of the CNI scenarios, neither a norm-based response pattern nor a bias towards inaction is the result of a rapid, automatic response. In contrast, previous work on the dual process theory of moral judgment has assumed that utilitarian judgments were produced by controlled cognitive comparisons of costs and benefits while

deontological responses in sacrificial dilemmas were based on automatic, emotional responses (Greene et al., 2001, 2004). Other authors have argued that deontological judgments are the result of participants' efforts to arrive at coherence by satisfying often conflicting constraints concerning rights and duties in their common-sense moral reasoning (Holyoak & Powell, 2016). While the former view would have predicted a negative association of the N parameter with total response time, the latter view is consistent with our finding of a positive association.

Other studies have found discrepancies between the temporal predictions of the dual process theory of moral judgments and response time data (e.g., Baron et al., 2012; Koop, 2013). Using the CNI model, Gawronski et al. (2017) were able to qualify earlier results by Greene et al. (2008) reporting a selective influence of cognitive load on utilitarian responses, which were obtained using the ProGreater condition only, where deontological responses and an inaction bias coincide. When including all four CNI conditions, which permit the separate estimation of each process, Gawronski et al. (2017) found that the effect of cognitive load was restricted to increasing the inaction bias rather than accentuating an automatic, emotional response underlying deontological responses. Note that like our results, the results in Gawronski et al. (2017, studies 2a, 2b) were obtained using response time data from online studies. However, given that previous studies have found that findings from cognitive psychology involving response times can be replicated in online studies (see, e.g., Semmelmann & Weigelt, 2017); we do not consider this a limitation.[8] However, as one

[8] Since these studies moreover use large sample sizes, it is to be expected that effects of noise will be mitigated. In this context it is also worth pointing out that as part of our exclusion criteria, we used both average response time $\pm$ 2 × SD and comprehension question in two initial high hurdle SAT questions, where participants were also required to read a lot of text to produce accurate responses. The data we analyse thus come from participants who are not outliers in their response time and who have demonstrated that they can adequately process a dense text passage to find the correct answers.

reviewer points out, future studies may be interested in investigating relationships between response times and additional covariates like participants' age or conscientiousness.

Figure 4 shows that the C1-C4 and Npre/Npro parameters of the $CNIS_8$ can be used as a measurement model for two latent C and N parameters. From this structural equation model, the violations of the invariance assumption can be read off from the differences in the standardized path coefficients from the latent C and N parameters to the parameters of the $CNIS_8$ model that they are measured by. This in turn shows that the violation of the invariance assumption does not only take the form of an additive shift to the means but can also be found in different path coefficients and thereby in the correlations between the different measures. The possibility of fitting a structural equation model with C and N parameters as latent variables, which takes the violations of the invariance assumption into account, shows that it possible to specify a model that fits the CNI model's intended use while addressing the methodological skepticism raised by Baron and Goodwin (2020, 2021) and others.

Conclusion

Implicit in a recent controversy concerning the CNI model of moral judgment (Gawronski et al., 2017) lies a problematic invariance assumption that process-dissociation and related multinomial processing-tree models make in their applications in different areas of psychology (Klauer et al., 2015). By extending the CNI model with a skip option, we implemented a version of the CNI model which avoided making the invariance assumption. This allowed us to test the invariance assumptions built into the CNI model, which were found to be violated both for the C and the N parameters in two experiments. Across two experiments, we obtained evidence of invariance violations in the CNI model and found that a 8-parameter version which avoids these invariance assumptions provided a better fit.

In light of these results, we recommend that future use of the CNI model adds this further S parameter and follows the 8-parameter version that we presented (Appendix A). In

Experiment 2, we showed that it was possible to selectively influence this new S parameter through a behavioral manipulation. Through structural equation modeling, we further analyzed mediation effects on the role of psychopathy on the CNI parameters and extended previous findings which were primarily based on bivariate correlations. It was found that the previously reported effect of gender on the CNI parameters is completely mediated by the association of gender with primary psychopathy.

References

Baron, J. (1996). Do no harm. In D. M. Messick, & A. E. Tenbrunsel (Eds.), *Codes of conduct: Behavioral research into business ethics* (pp. 197–213). New York: Russell Sage Foundation.

Baron, J. & Goodwin, G. P. (2020). Consequences, norms, and inaction: A comment. *Judgment and Decision Making, 15*(3), 421–442.

Baron, J. & Goodwin, G. J. (2021). Consequences, norms, and inaction: Response to Gawronski et al.. *Judgment and Decision Making*, *16*(2), 566-595.

Baron, J., Gürçay, B., Moore, A. B., Starcke, K. (2012). Use of a Rasch model to predict response times to utilitarian dilemmas. *Synthese, 189*, 107-117.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86.

Berentelg, M. (2020). Multinomial Modeling of Moral Dilemma Judgment: A Replication Study. Retrieved in November 2021 from https://osf.io/mb32t/.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*, 216–235.

Erdfelder, E., Auer, T., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift fur Psychologie / Journal of Psychology*, *217*, 108–124.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hutter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*, 343–376.

Gawronski, B., Conway, P., Hütter, M., Luke, D.M., Armstrong, J., & Friesdorf, R. (2020). On the validity of the CNI model of moral decision-making: Reply to Baron and Goodwin (2020). *Judgment and Decision Making, 15*(6), 1054-1072.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment, *Cognition*, *107*, 1144-1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Heck, D.W., Arnold, N.R. & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50***,** 264–284.

Hennig, M., & Hütter, M. (2021). Consequences, norms, or willingness to interfere: A proCNI model analyses of the foreign language effect in moral dilemma judgment. *Journal of Experimental Social Psychology, 95, 104148.*

Hennig, M., & Hütter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. *Journal of Personality and Social Psychology, 118*, 22-56.

Herman, B (2011). A Mismacth of Methods. In Scheffler, S. (Eds.), *On What Matters,*

Volume 2 (pp. 83-115). Oxford: Oxford University Press.

Holyoak, K. J. & Powell, D. (2016). Deontological coherence*: A* framework *for* commonsense moral reasoning. *Psychol Bull., 142*(11), 1179-1203

Janoff-Bulman, R., Sheikh, S. and Hepp, S. (2009). Proscriptive Versus Prescriptive Morality: Two Faces of Moral Regulation. *Journal of Personality and Social Psychology, 96*(3), 521-537.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*, 131–164.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.

Klauer, K. C. (2010). Hierarchical Multinomial Processing Tree Models: A Latent-Trait Approach. *Psychometrika, 75*(1), 70-98.

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General, 144*(1), 198–221.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4. Edition). New York: The Guildford Press.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making*, *8*(5), 527-539.

Körner, A., Deutsch, R., and Gawronski, B. (2020). Using the CNI Model to Investigate Individual Differences in Moral Judgments. *Personality and Social Psychology Bulletin*, 1-16.

Kroneisen, M., & Heck, D. W. (2020). Interindividual Differences in the Sensitivity for Consequences, Moral Norms, and Preferences for Inaction: Relating Basic Personality Traits to the CNI Model. *Pers Soc Psychol Bull.*, *46*(7), 1013-1026

Lee, M. D., and Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Levenson, M. R., Kiehl, K. A., Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population, *Journal of personality and Social Psychology*, *68*, 151-158.

Luke, D. M. & Gawronski, B. (2021). Psychopathy and Moral Dilemma Judgments: A CNI Model Analysis of Personal and Perceived Societal Standards. *Social Cognition*, *39*(1), 41-58.

Luke, D. M., Neumann, C.S., Gawronski, B. (2021). Psychopathy and Moral-Dilemma Judgment: An Analysis Using the Four-Factor Model of Psychopathy and the CNI Model of Moral Decision-Making. *Clinical Psychological Science,* 1-17. doi:10.1177/21677026211043862

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2013). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*(1), 205–235. https://doi.org/10.1007/s11336-013-9374-9

McIntyre, A. (2019). Doctrine of Double Effect. In: The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.). Retrieved in November 2021 from https://plato.stanford.edu/archives/spr2019/entries/double-effect/.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of Statistical Software*, *85*(4), 1–30.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2th Ed.). Cambridge: Cambridge University Press.

Scanlon, T. M. (2011). How I Am Not a Kantian. In Scheffler, S. (Eds.), *On What Matters,*

Volume 2 (pp. 116-139). Oxford: Oxford University Press.

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: reaction time effects in cognitive experiments. *Behavior Research Methods, 49*, 1241–1260.

Shipley, B. (2016). *Cause and Correlation in Biology*. Cambridge: Cambridge University Press.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology, 54*, 167–183.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). Oxford University Press.

Appendix A: The CNIS Model

Model Equations of the CNIS Model

The model equations for the 8-parameter version of the CNIS model ($CNIS_8$) are:

$$P(\text{action}|\text{ProGreater}) = C_1 + (1-C_1) \times (1-N_{pro}) \times (1-S) \times (1-I)$$

$$P(\text{inaction}|\text{ProGreater}) = (1-C_1) \times N_{pro} + (1-C_1) \times (1-N_{pro}) \times (1-S) \times I$$

$$P(\text{skip}|\text{ProGreater}) = (1-C_1) \times (1-N_{pro}) \times S$$

$$P(\text{action}|\text{ProSmaller}) = (1-C_2) \times (1-N_{pro}) \times (1-S) \times (1-I)$$

$$P(\text{inaction}|\text{ProSmaller}) = C_2 + (1-C_2) \times N_{pro} + (1-C_2) \times (1-N_{pro}) \times (1-S) \times I$$

$$P(\text{skip}|\text{ProSmaller}) = (1-C_2) \times (1-N_{pro}) \times S$$

$$P(\text{action}|\text{PreGreater}) = C_3 + (1-C_3) \times N_{pre} + (1-C_3) \times (1-N_{pre}) \times (1-S) \times (1-I)$$

$$P(\text{inaction}|\text{PreGreater}) = (1-C_3) \times (1-N_{pre}) \times (1-S) \times I$$

$$P(\text{skip}|\text{PreGreater}) = (1-C_3) \times (1-N_{pre}) \times S$$

$$P(\text{action}|\text{PreSmaller}) = (1-C_4) \times N_{pre} + (1-C_4) \times (1-N_{pre}) \times (1-S) \times (1-I)$$

$$P(\text{inaction}|\text{PreSmaller}) = C_4 + (1-C_4) \times (1-N_{pre}) \times (1-S) \times I$$

$$P(\text{skip}|\text{PreSmaller}) = (1-C_4) \times (1-N_{pre}) \times S$$

For the *i*th participant, a data vector, $\boldsymbol{y}_i$, consisting of counts of each of these three response categories (action, inaction, skip) across the four CNI conditions (ProGreater, ProSmaller, PreGreater, PreSmaller) is formed. Via the CNIS model equations, these counts are modeled through a vector of 8 theta parameters for each participant, $\boldsymbol{\theta}_i$. In the four-parameter version, the invariance assumption is made, whereby $N_{pre} = N_{pro} = N$ and $C_1 = C_2 = C_3 = C_4 = C$, resulting in a vector of 4 theta parameters for each participant, $\boldsymbol{\theta}_i$.

In the standard CNI model, the inaction bias corresponding to the I parameter governs responses when neither moral cue (norms or consequences) compels a response. Similarly, in the extended CNIS model, the skip option comes into play, if participants have no guidance as to their response from norms and consequences and thus, in the $(1\text{-}C_j) \times (1\text{-}N_k)$ cases. This dovetails with the instruction to be permitted to skip in case participants are undecided about whether the described action is morally acceptable or inacceptable. In the original model, participants have the choice between action and inaction in this state of uncertainty (cases with (1-C) × (1-N)) with preferences governed by parameter I. One consequence is that although the skip parameter S is constant, the actual frequency of the use of the skip option can differ between the four types of dilemmas to the extent that C and N differ between them.

In the extended CNIS model, we offer participants three choices instead of only two in the case of reaching the uncertainty state with probability $(1\text{-}C_j) \times (1\text{-}N_k)$: They can then skip, choose action, or choose inaction with probabilities S, (1-S) × (1-I) and (1-S) × I. Both the S and I parameters can also vary between persons, and in the model with random effects by scenario as a function of scenario (see Appendix B). Yet, both the S and I parameters remain invariant across the four CNI conditions within every scenario and person.

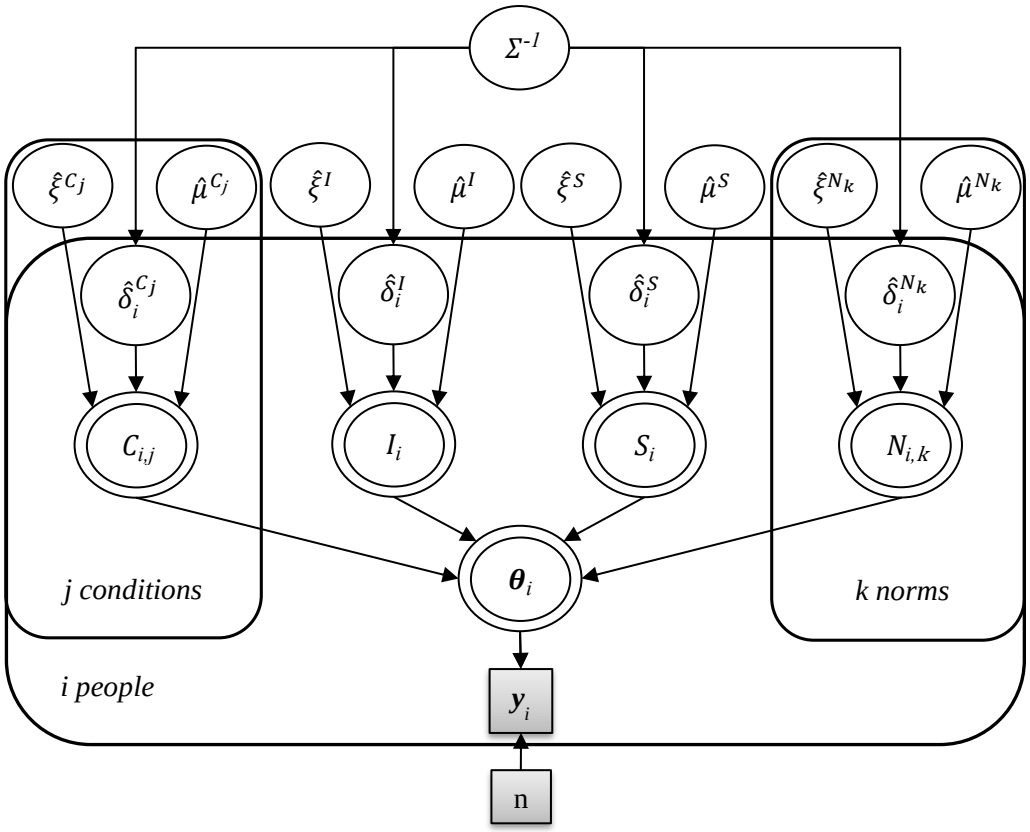
Bayesian Hierarchical Implementation

To estimate the MPT parameters of the CNIS model for each participant separately, we here follow the hierarchical latent trait model of Klauer (2010), which has also been implemented in the `TreeBUGS` R-package by Heck et al. (2018).

In this approach, a probit link function is used to transform MPT parameters (representing probabilities between 0 and 1) to the real line, $\Phi^{-1}(\theta)$. The transformed parameters are then modeled via a multivariate normal distribution while estimating mean, μ, and covariance matrix, Σ, from the data. The advantage of this approach is that heterogeneity in parameter estimates across participants and correlations among MPT parameters can be accommodated while allowing for partial aggregation of statistical information across participants in the posterior parameters of the multivariate normal distribution (Klauer, 2010). Accordingly, for each participant, *i*, the probit-transformed parameters are additively decomposed into a group mean, μ, and a random effect, $\Phi^{-1}(\theta) = \mu + \delta_i$.

We contrasted two hierarchical multinomial models following this approach with different numbers of MPT parameters (4 vs. 8). Table A1 illustrates $CNIS_8$, whereby a distinct C parameter is estimated for each of the $j = 1, \ldots, 4$ CNI conditions, and a distinct N parameter is estimated for each of the $k = 1, 2$ types of norms. For $CNIS_4$, one shared C parameter is estimated ($j = 1$) together with one shared N ($k = 1$) parameter.

Table A1. Hierarchical Latent Trait MPT Model



$$\hat{\mu}^{C_j}, \hat{\mu}^{N_k}, \hat{\mu}^{I}, \hat{\mu}^{S} \sim \text{Gaussian}(0,1)$$
$$\hat{\xi}^{C_j}, \hat{\xi}^{N_k}, \hat{\xi}^{I}, \hat{\xi}^{S} \sim \text{Uniform}(0,10)$$
$$\Sigma^{-1} \sim \text{Wishart}(\mathbf{I}, df)$$
$$\left(\hat{\delta}_i^{C_j}, \hat{\delta}_i^{N_k}, \hat{\delta}_i^{I}, \hat{\delta}_i^{S}\right) \sim \text{MvGaussian}(\mathbf{0}, \Sigma^{-1})$$
$$C_{i,j} \leftarrow \Phi\left(\hat{\mu}^{C_j} + \hat{\xi}^{C_j}\hat{\delta}_i^{C_j}\right)$$
$$N_{i,k} \leftarrow \Phi\left(\hat{\mu}^{N_k} + \hat{\xi}^{N_k}\hat{\delta}_i^{N_k}\right)$$
$$I_i \leftarrow \Phi\left(\hat{\mu}^{I} + \hat{\xi}^{I}\hat{\delta}_i^{I}\right)$$
$$S_i \leftarrow \Phi\left(\hat{\mu}^{S} + \hat{\xi}^{S}\hat{\delta}_i^{S}\right)$$
$$\boldsymbol{y}_i \sim \text{Multinomial}(\boldsymbol{\theta}_i, n)$$

Note. There are four CNI conditions with three categorical responses (action, inaction, skip). Via the CNIS model equations displayed above, the outcome probabilities of the responses in the data vector, $\boldsymbol{y}_i$, are represented by 8 theta parameters. For each participant, a vector of 8 theta parameters, $\boldsymbol{\theta}_i$, is estimated. The inverse Wishart distribution has 8+1 degrees of freedom, *df*, and a 8×8 identity matrix, **I**, as scale matrix.

The models were fitted in a Bayesian framework through a Gibbs sampler, which estimates the posterior distributions of model parameters by means of Monte Carlo-Markov chains.

Appendix B: Item Effects

Baron and Goodwin (2020) suggest that both item and participants effects should be estimated for the CNI parameters. Above, we have already estimated the CNI parameters for each participant to test individual variation. In an exploratory analysis, we also estimated item effects in a model with crossed random effects for participants and scenarios (Matzke et al., 2013) to test whether the invariance assumption would be violated within the individual scenarios used in the experiment. For an analysis with less uncertainty in the estimates, a larger sample size would be required. But the exploratory analysis displayed in Table B1 below already suggests violations of the invariance assumption almost in every scenario investigated.

Table B1. Item Effects and the Invariance Assumption

Scenario	$C_{ProGreater}$	$C_{ProSmaller}$	$C_{PreGreater}$	$C_{PreSmaller}$	I	N_{pre}	N_{pro}	S
Assisted-suicide	$\tilde{\Delta}$ = .34 [.23, .44]	$\tilde{\Delta}$ = .01 [$3.0\cdot10^{-9}$, 0.05]	$\tilde{\Delta}$ = .05 [.004, 0.12]	$\tilde{\Delta}$ = .05 [.001, .11]	$\tilde{\Delta}$ = .58 [.52, .64]	$\tilde{\Delta}$ = .08 [.03, .15]	$\tilde{\Delta}$ =.18 [.05, .33]	$\tilde{\Delta}$ = .12 [.09, .15]
Bishop	$\tilde{\Delta}$ = .0003 [$6.7\cdot10^{-11}$, .002]	$\tilde{\Delta}$ = .49 [.26, .70]	$\tilde{\Delta}$ = .25 [.07, .43]	$\tilde{\Delta}$ = .0002 [$5.0\cdot10^{-16}$, .002]	$\tilde{\Delta}$ = .62 [.54, .69]	$\tilde{\Delta}$ = .31 [.19, .42]	$\tilde{\Delta}$ = .16 [$3.8\cdot10^{-16}$, 0.63]	$\tilde{\Delta}$ = .15 [.04, .36]
Construc-tionsite	$\tilde{\Delta}$ = .009 [.0006, .02]	$\tilde{\Delta}$ = .14 [.004, .33]	$\tilde{\Delta}$ = .02 [$1.4\cdot10^{-7}$, .09]	$\tilde{\Delta}$ = .006 [$7.7\cdot10^{-5}$, .02]	$\tilde{\Delta}$ = .59 [.52, .66]	$\tilde{\Delta}$ = .43 [.31, .54]	$\tilde{\Delta}$ = .37 [.20, .53]	$\tilde{\Delta}$ = .04 [.02, .06]
Dialysis	$\tilde{\Delta}$ = .12 [.06, .18]	$\tilde{\Delta}$ = .31 [.15, .48]	$\tilde{\Delta}$ = .43 [.29, .57]	$\tilde{\Delta}$ = .03 [$1.2\cdot10^{-5}$, .08]	$\tilde{\Delta}$ = .59 [.53, .66]	$\tilde{\Delta}$ = .03 [.005, .07]	$\tilde{\Delta}$ = $5.77\cdot10^{-5}$ [$5.5\cdot10^{-18}$, .003]	$\tilde{\Delta}$ = .08 [.06, .11]
Immune-deficiency	$\tilde{\Delta}$ = .0002 [$1.5\cdot10^{-14}$, .003]	$\tilde{\Delta}$ = .30 [.09, .52]	$\tilde{\Delta}$ = .17 [.02, .35]	$\tilde{\Delta}$ = $4.3\cdot10^{-5}$ [$9.3\cdot10^{-13}$, .0007]	$\tilde{\Delta}$ = .66 [.58, .73]	$\tilde{\Delta}$ = .62 [.51, .72]	$\tilde{\Delta}$ = .26 [.08, .45]	$\tilde{\Delta}$ = .06 [.04, .08]
Mother	$\tilde{\Delta}$ = .10 [.05, .16]	$\tilde{\Delta}$ = .45 [.21, .69]	$\tilde{\Delta}$ = .29 [.12, .46]	$\tilde{\Delta}$ = .01 [$4.6\cdot10^{-5}$, .03]	$\tilde{\Delta}$ = .61 [.55, .68]	$\tilde{\Delta}$ = .46 [.34, .57]	$\tilde{\Delta}$ = .34 [.08, .41]	$\tilde{\Delta}$ = .13 [.09, .17]
Peanuts	$\tilde{\Delta}$ = .08 [.03, .13]	$\tilde{\Delta}$ = .03 [$9.7\cdot10^{-10}$, 0.16]	$\tilde{\Delta}$ = .004 [$1.1\cdot10^{-7}$, .03]	$\tilde{\Delta}$ = .009 [.0001, .03]	$\tilde{\Delta}$ = .42 [.13, .74]	$\tilde{\Delta}$ = .30 [.20, .39]	$\tilde{\Delta}$ = .23 [.08, .40]	$\tilde{\Delta}$ = .05 [.03, .06]
Torture	$\tilde{\Delta}$ = .38 [.27, .48]	$\tilde{\Delta}$ = .56 [.34, .76]	$\tilde{\Delta}$ = .26 [.11, .42]	$\tilde{\Delta}$ = .09 [.02, .17]	$\tilde{\Delta}$ = .59 [.52, .67]	$\tilde{\Delta}$ = .21 [.10, .34]	$\tilde{\Delta}$ = .03 [$2.2\cdot10^{-6}$, .14]	$\tilde{\Delta}$ = .06 [.04, .09]
Transplant	$\tilde{\Delta}$ = .002 [$1.5\cdot10^{-17}$, .01]	$\tilde{\Delta}$ = .0006 [$2.4\cdot10^{-21}$, 0.01]	$\tilde{\Delta}$ = .0002 [$1.3\cdot10^{-21}$, .006]	$\tilde{\Delta}$ = $2.6\cdot10^{-5}$ [$8.4\cdot10^{-20}$, .002]	$\tilde{\Delta}$ = .55 [.48, .61]	$\tilde{\Delta}$ = .11 [.05, .17]	$\tilde{\Delta}$ = .14 [.04, .26]	$\tilde{\Delta}$ = .05 [.03, .06]
Vaccine	$\tilde{\Delta}$ = $5.5\cdot10^{-5}$ [$1.3\cdot10^{-23}$, .002]	$\tilde{\Delta}$ = .24 [.05, .44]	$\tilde{\Delta}$ = .009 [$2.7\cdot10^{-7}$, .04]	$\tilde{\Delta}$ = $1.9\cdot10^{-6}$ [$2.1\cdot10^{-18}$, .0001]	$\tilde{\Delta}$ = .60 [.52 .66]	$\tilde{\Delta}$ = .27 [.18, .36]	$\tilde{\Delta}$ = .15 [.03, .30]	$\tilde{\Delta}$ = .06 [.04, .08]

Note. The square brackets indicate 95% highest density intervals (HDI).

Appendix C: Extended SEM Analysis

The aim of this appendix is to reanalyze the data in Experiment 1 with 1) SEM models that include the S parameter, and 2) an integrative model that combines $CNIS_8$ with the best fitting SEM model to exploit the propagation of uncertainty from the estimated MPT parameters to the structural equation analysis.

First, four SEM models were fitted to the data from Experiment 1 that included the S parameter to determine the optimal SEM model with the S parameter.

M1: SEM model with direct effects of Gender and mediation analysis + S parameter with Time and Primary Psychopathy as predictors of the C, N, I, S parameters (wherein the ‚C' and ‚N' parameters are the estimates based on $CNIS_8$).

M2: Like M2 but without the direct effects of Gender and the mediation analysis.

M3: Same as M1 but without Time and Primary Psychopathy as predictors of S.

M4: Same as M2 but without Time and Primary Psychopathy as predictors of S.

Table C1. Model Comparison

	LOOIC	Δelpd	SE	WAIC	Weight
M1	14396.4	-8527.9	146.6	11442.5	.000
M2	-1207.3	-726.0	16.6	-3721.9	.000
M3	21961.0	-12310.2	198.2	25829.1	.001
M4	-2659.3	0	--	-5487.7	.999

Note. 'elpd' = expected log predictive density. elpd is a measure of out-of-sample predictive adequacy. LOOIC = -2*elpd. The weights are stacking weights based on LOOIC. Note that information criteria can take both positive and negative values and that the lowest value on the real line still indicates best fit.

As Table C1 shows, a SEM model (M4) without direct effects of gender is preferred after adding the S parameter to the model, thus replicating the mediation analysis reported in the paper. Furthermore, the model comparison shows a model without Time and Primary Psychopathy as predictors of the S parameter is preferred.

Next, $CNIS_8$ was combined with M4 in one integrative model ($CNIS_{8\_SEM}$) to permit the propagation of uncertainty from the posterior draws of the MPT parameters to the

structural equation analysis. In contrast, $CNIS_8$ was fitted independently of the SEM model in the paper and the SEM models compared did not include the S parameter. It was found that this integrative model showed a similar performance on the posterior checks proposed in Klauer (2010) as the $CNIS_8$ model reported in the paper ($p_{T1} < .02$, $p_{T2} < .01$).

Figure C1 plots the posterior distributions of the MPT parameters of $CNIS_{8\_SEM}$.

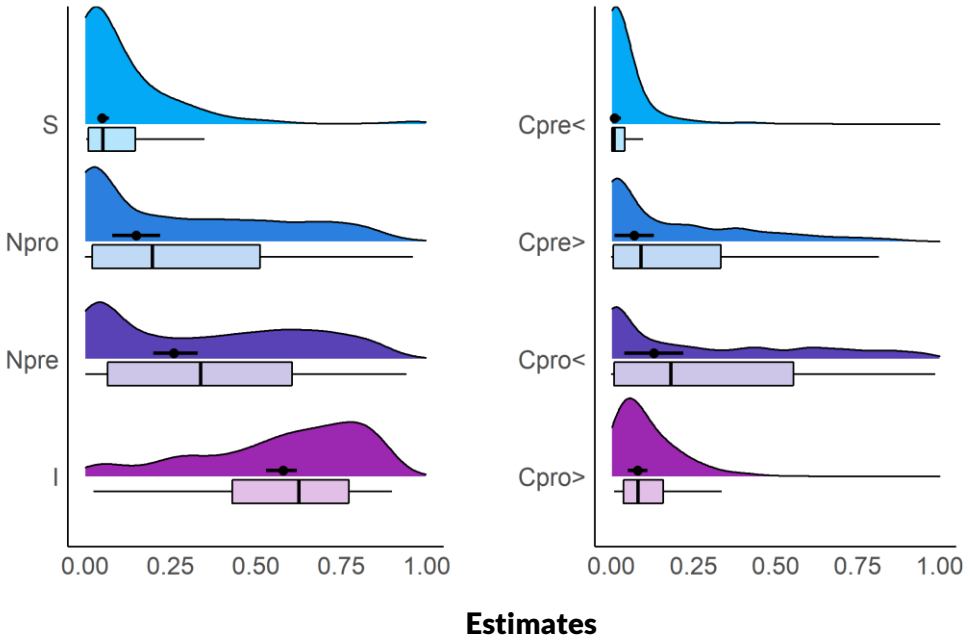


Figure C1. Distributions of the CNIS parameters estimated for each participant with boxplots indicating the quartiles of the individual estimates. The black points and lines indicate the group-level posterior medians and their 95% HDI. 'Cpro>' = $C_{ProGreater}$, 'Cpro<' = $C_{ProSmaller}$, 'Cpre>' = $C_{PreGreater}$, 'Cpre<' = $C_{PreSmaller}$.

These posterior parameter distributions were found to be similar to those that we report based on $CNIS_8$ in the paper. To test possible violations of the invariance assumption that Npro = Npre and that $C_{proGreater} = C_{proSmaller} = C_{preGreater} = C_{preSmaller}$, we analyzed contrasts of pairs of parameters on the probability scale by identifying whether the 95% HDI intervals for the difference between the two parameters included zero. Credible differences were found for Npre > Npro ($\Delta^{N}_{pro\text{-}pre}$ = - 0.11, 95% HDI [-0.21, -0.03]), $C_{proGreater} > C_{preSmaller}$ ($\Delta^{C}_{proGreater\text{-}preSmaller}$ = 0.07, 95% HDI [0.04, 0.10]), and $C_{proSmaller} > C_{preSmaller}$ ($\Delta^{C}_{proSmaller\text{-}preSmaller}$ = 0.12, 95% HDI [0.03, 0.22]). In contrast, the 95% HDI interval for the $C_{preGreater} > C_{preSmaller}$ comparison just included zero

($\Delta^{C}_{preGreater\text{-}preSmaller}$ = 0.06, 95% HDI [0.00, 0.12]). It was thus found that the invariance violations of $CNIS_8$ reported in the paper were replicated with $CNIS_{8\_SEM}$. The only exception was the $\Delta^{C}_{preGreater\text{-}preSmaller}$ contrast.

Finally, Figure C2 plots the path diagram for the SEM model included in $CNIS_{8\_SEM}$.

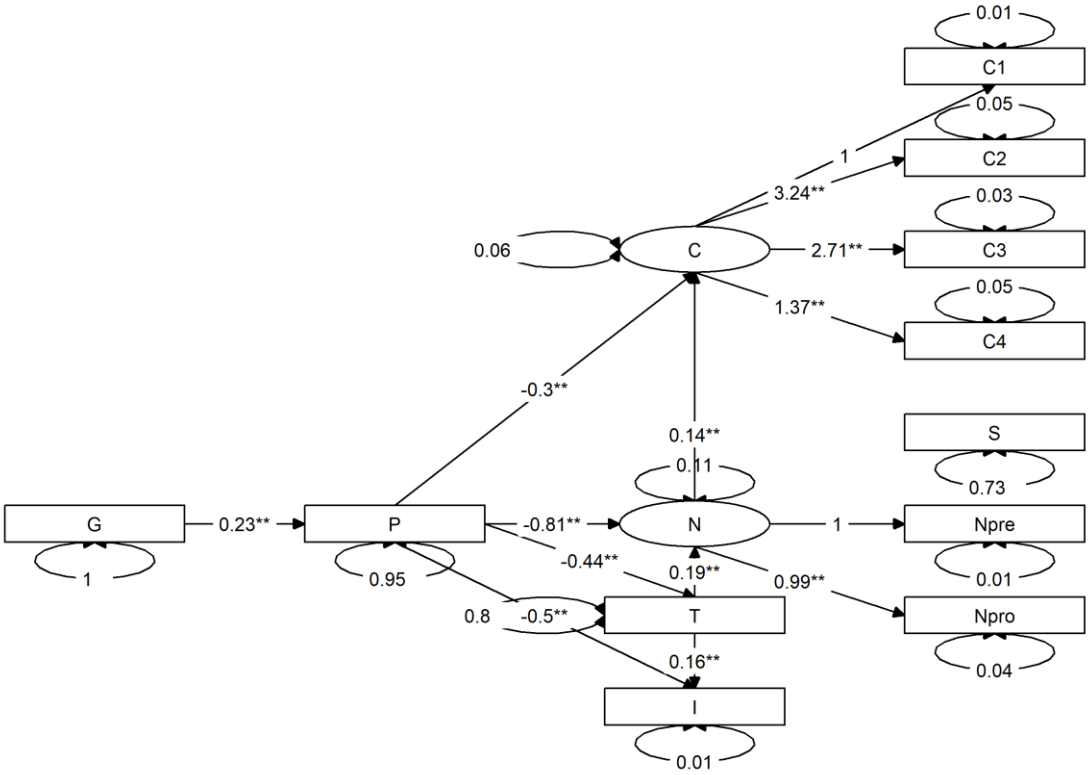


Figure C2. M4 SEM model. Path coefficients indicate posterior medians and are marked with '**', if the 95% HDI interval does not include 0. 'G' = self-reported gender with 'male' = 1 and 'female' = 0 (excluding five participants, who preferred not to respond), 'P' = Primary psychopathy, 'T' = total response time for all the items. Both 'P' and 'T' were scaled to take values between 0 and 1 before fitting the model to prevent large differences in the variances of the different parameters. Direct and indirect effects are encoded via arrows. The loops indicate variances. The covariances have been left out to simplify the graph. Latent variables are marked with circles. The scales of the latent variables were fixed by setting the first path coefficient equal to 1.0. Note that whereas all the variables were z-transformed in Figure 4, only the G, P, and T variables were z-transformed in Figure C2; the other variables were left on the probability scale. 'C1' = $C_{ProGreater}$, 'C2' = $C_{ProSmaller}$, 'C3' = $C_{PreGreater}$, 'C4' = $C_{PreSmaller}$.

Again, the path coefficients of the SEM model included in $CNIS_{8\_SEM}$ are similar to those that we report based on $CNIS_8$ in the paper. Since the winning model of Table A1 did not include direct paths from primary psychopathy and time to the S parameter, such paths are absent from Figure C2.