Evidence for an evaluative effect of stimulus co-occurrence may be inflated by evaluative differences between assimilative and contrastive relations

Karoline Corinna Bading[1], Marius Barth[2], & Klaus Rothermund[3]

[1] University of Tübingen

[2] University of Cologne

[3] Friedrich Schiller University Jena

Author Note

Correspondence concerning this article should be addressed to Karoline Corinna Bading, Schleichstraße 4, 72076 Tübingen. E-mail: karoline.bading@uni-tuebingen.de

Abstract

Recent research on relational evaluative conditioning (relational EC) suggests that stimulus co-occurrence can have a direct effect on evaluations over and above the particular relation between the co-occurring stimuli. This research is based on a process dissociation approach where co-occurrence effects are demonstrated via attenuated evaluative learning for co-occurring stimuli that are connected by contrastive in comparison to assimilative relations. Instead of attributing such attenuations to an orthogonal influence of stimulus co-occurrence, we investigated whether (a) contrastive relations tend to produce weaker evaluations than their assimilative counterparts and (b) such evaluative differences can inflate evidence for co-occurrence effects on continuous as well as on categorical evaluation measures. A pilot study ($N = 85$) confirmed notion (a), while a first experiment ($N = 42$) produced preliminary evidence for notion (b) in the context of multinomial processing tree (MPT) modeling. In a second, high-powered experiment ($N = 229$), sub-sample MPT analyses (including only CSs with correct memory for the CS-US proposition) demonstrated that evidence for co-occurrence effects can be inflated by evaluative differences between assimilative vs. contrastive relations. The theoretical and methodological implications of these findings are discussed.

Keywords: evaluative learning, relational evaluative conditioning, mere co-occurrence effects, process dissociation, multinomial modeling

Word count: 14,486

Evidence for an evaluative effect of stimulus co-occurrence may be inflated by evaluative differences between assimilative and contrastive relations

One of the most prominent approaches in contemporary research on evaluative learning is relational evaluative conditioning (hereafter, relational EC). Relational EC procedures are an offshoot of the evaluative conditioning (EC) paradigm in which neutral “conditioned” stimuli (CSs) are repeatedly paired with positive or negative “unconditioned” stimuli (USs). In addition to such CS-US pairings (and the appendant manipulation of US valence), relational EC procedures come with a manipulation of the (stated) relation between the co-occurring stimuli. In a widely used variant (Hu, Gawronski, & Balas, 2017), the relations between CSs (images of fictitious pharmaceutical products) and USs (images of positive vs. negative health states) are manipulated by presenting the conjugated verbs “causes” vs. “prevents” in between the paired images. In another prominent example (Moran & Bar-Anan, 2013), unknown cartoon characters (CSs) either start or stop auditory stimuli (a pleasant melody vs. an unpleasant scream, USs).

The shared feature of these and other manipulations of CS-US relations is their reliance on pairs of antonymic relations (e.g., “cause” vs. “prevent”; “start” vs. “stop”) that imply opposing effects of the valence of the US on the evaluation of the co-occurring CS.[1] In the case of assimilative relations (e.g., “cause” or “start”), conventional reasoning implies a CS evaluation that is in line with the paired US valence (e.g., when a pharmaceutical product causing a negative condition is considered negative). Conversely, for contrastive relations (e.g., “prevent” or “stop”), a CS evaluation of the opposite valence is implied (e.g., when a pharmaceutical product preventing a negative condition is considered positive).

In the vast majority of relational EC studies, the pattern of CS evaluations (assessed

[1] For different manipulations of CS-US relations, see for example Förderer and Unkelbach (2011) and Hughes, Ye, and De Houwer (2019).

after the pairing procedure) supports the previously outlined reasoning. For CSs presented with assimilative relations, such evaluations usually indicate an assimilative effect of US valence (in that CSs paired with positive USs are evaluated more favorably than CSs paired with negative USs); whereas for CSs presented with contrastive relations, a contrastive effect of US valence on CS evaluation is observed (in that CSs paired with positive USs are evaluated *less* favorably than CSs paired with negative USs). This cross-over interaction of US valence and CS-US relation is a crucial finding in evaluative learning research because it indicates that evaluative learning based on stimulus pairings can be mediated by inferential reasoning on propositional statements about the relation between the paired stimuli (hereafter, inferential reasoning on CS-US propositions). It therefore corroborates a core assumption of both single-process propositional and dual-process accounts of evaluative learning (for detailed descriptions, see De Houwer, 2018; Gawronski & Bodenhausen, 2018).

Moreover, the previously mentioned interaction also forms the methodological basis for investigating the intriguing possibility that stimulus co-occurrence has an assimilative effect (on evaluations) despite, and unqualified by, conscious knowledge of a contrastive relation between the co-occurring stimuli. The existence of such unqualified co-occurrence effects would not only have important consequences for everyday life (e.g., in the design of public communication campaigns), but also advance our theoretical knowledge on the mental processes mediating evaluative conditioning and other learning phenomena (e.g., Gawronski, Luke, & Ng, 2021). Based on these practical and theoretical implications, developing reliable tools for inducing and measuring unqualified co-occurrence effects has been a long-standing objective in research on attitude formation and evaluative learning (for an overview, see Bading, 2021).

Using relational EC for investigating unqualified effects of stimulus co-occurrence

Attempts at demonstrating unqualified co-occurrence effects in relational EC procedures have been based on two approaches: a task dissociation approach and a process dissociation approach. In studies based on the task dissociation approach (Hu et al., 2017; Moran & Bar-Anan, 2013), a relational EC procedure is followed by an assessment of CS evaluation on a direct as well as on an indirect measure of evaluation. This approach is based on the assumption that the different processing conditions in the two types of measures should selectively reveal the effects of inferential reasoning on CS-US propositions (on the direct measure) and an additional, unqualified, influence of CS-US associations (on the indirect measure). The task dissociation approach has gained a mixed record in demonstrating unqualified co-occurrence effects: while some studies found the predicted dissociation (e.g., Moran & Bar-Anan, 2013), many others failed to find unqualified EC effects on the indirect evaluation measure (e.g., Hu et al., 2017; Peters & Gawronski, 2011; Zanon, De Houwer, & Gast, 2012). Moreover, successful applications of the task dissociation approach have later been found to suffer from interpretational ambiguity (e.g., Bading, Stahl, & Rothermund, 2020; Hu et al., 2017). Taken together, these issues have led researchers to adopt a process dissociation logic (Jacoby, 1991; Yonelinas & Jacoby, 2012; for a review, see Payne, 2008) to develop a novel approach for investigating mere co-occurrence effects in relational EC.

In this novel approach, a demonstration of unqualified co-occurrence effects is attempted by comparing CS evaluations in the assimilative vs. contrastive conditions within a single measurement procedure. This approach is based on the fact that inferential reasoning (on CS-US propositions) and mere co-occurrence effects should push CS evaluations in the same direction (towards the US valence) in the assimilative conditions and in opposite directions in the contrastive conditions. Taken together, an unqualified influence of stimulus co-occurrence should thus result in overall weaker relational EC effects in the contrastive compared to the assimilative conditions (by subtracting from [adding to] evaluative learning effects based on relational reasoning in the former [latter]). On continuous evaluation measures, such a pattern of asymmetrical evaluative learning in assimilative vs. contrastive conditions has been found in the majority of relational EC studies, pointing to a robust presence of unqualified co-occurrence effects. However, an attenuation of continuous evaluations of CSs presented with contrastive relations can also be explained without such effects. As pointed out by Moran, Bar-Anan, and Nosek (2016), asymmetrical relational EC effects might also be driven by inferential reasoning (on CS-US propositions) assuming that contrastive relations produce weaker continuous evaluations than their assimilative counterparts.

To avoid this interpretational ambiguity (and in keeping with the original process dissociation procedure), the aforementioned comparison (between assimilative vs. contrastive conditions) is usually performed on categorical evaluation measures where CSs are categorized as either positive or negative. Here, it can be argued that inferential reasoning (on CS-US propositions) should lead to propositionally correct classifications regardless of whether the corresponding continuous evaluation is strong or weak (as long as it is located on the "right side" of the positivity/negativity spectrum). Categorical CS evaluations thus allow for more unambiguous data patterns: to the extent that response tendencies towards propositionally correct options are less pronounced in the contrastive (than in the assimilative) conditions, an additional contribution of unqualified co-occurrence effects can be concluded. In addition to neutralizing possibly graded effects of assimilative vs. contrastive relations, categorical CS evaluations also allow for multinomial processing tree (MPT) modeling, a sophisticated technique for quantifying the respective contributions of different processes to a single measurement procedure.

MPT modeling of relational EC data

To disentangle the contributions of inferential reasoning on CS-US propositions and unqualified co-occurrence effects, an MPT model estimating three parameters has been proposed (hereafter, the RCB model, Heycke & Gawronski, 2020; see also Kukken, Hütter, & Holland, 2020). The first parameter, R, quantifies the probability of categorical CS evaluations being driven by inferential reasoning on CS-US propositions (stating the US [and/or its valence] and the specific CS-US relation). The estimation of a single R parameter reflects the ideas that (a) recollection of CS-US propositions is comparable for all four US valence $\times$ relation conditions, and that (b) the reasoning process is equally likely to result in propositionally correct responses in all four conditions of the relational EC procedure (hereafter, the R-invariance assumption). The second parameter, C for co-occurrence, quantifies the conditional probability of categorical CS evaluations being driven by an unqualified effect of US valence (in the absence of inferential reasoning on CS-US propositions). As before, a single C parameter for all four US valence $\times$ relation conditions is estimated (reflecting the idea that unqualified co-occurrence effects have comparable influence across conditions). Finally, the third parameter, B for bias, quantifies a general response tendency whenever categorical CS evaluations are driven neither by inferential reasoning nor by unqualified co-occurrence effects.

The RCB model has been applied to several relational EC procedures (Béna, Mauclet, & Corneille, 2022; for examples, see Heycke & Gawronski, 2020; Kukken et al., 2020; for a similar paradigm, see Gawronski et al., 2021) producing larger than zero estimates of R and C parameters in the vast majority of studies. This consistent pattern has led researchers to conclude that unqualified influences of stimulus co-occurrence have been demonstrated beyond doubt, and to call for intensified research on the mental underpinnings of this intriguing phenomenon (Gawronski, Brannon, & Luke, 2023). The latter call is based on the fact that previous validation studies have been successful in validating the R parameter as an indicator of inferential reasoning (on CS-US propositions), but have so far failed to identify the exact mechanism(s) underlying the mere co-occurrence effects indicated by above zero C parameters (Gawronski, 2022; Heycke & Gawronski, 2020; Kukken et al., 2020).

Importantly, some of these validation studies have also produced findings that are inconsistent with all of the currently discussed mediators of unqualified co-occurrence effects [automatic formation of CS-US associations, partial retrieval accounts and mere co-occurrence propositions; Heycke and Gawronski (2020), Experiment 4], pointing to the possibility that above zero C parameters may (sometimes) reflect other phenomena that are not driven by stimulus co-occurrences. This possibility is corroborated by recent simulation studies showing that a specific violation of the R-invariance assumption (a) produces an above zero C parameter in the complete absence of co-occurrence influences, (b) mimics positive evidence for the C parameter as an indicator of an unqualified co-occurrence effect (i.e., the correlation pattern reported by Kukken et al., 2020) and (c) predicts some of the so far inexplicable results of previous validation studies (for details, see Bading, 2021).

Based on this initial (yet indirect) support for a violation of the R-invariance assumption (as a source of above zero C parameters), it seems worthy to consider its mathematical structure and semantic meaning. Interestingly, the specific violation that produces findings (a) to (c) consists in true R parameters that are larger in the assimilative than in the contrastive conditions of a relational EC procedure. In other words: an above zero C parameter will emerge whenever evaluative categorizations of CSs presented with contrastive relations have a lower probability of being driven by inferential reasoning (even if categorization responses in the absence of such reasoning are based on a general response tendency [B] and not on an unqualified effect of US valence [C]).[2] This particular structure of violated R-invariance ($R_{assimilative} > R_{contrastive}$) is noteworthy because it corresponds to

[2] For a demonstration, see Bading (2021).

the earlier idea that inferential reasoning involving contrastive relations might yield weaker (continuous) evaluations than inferential reasoning based on their assimilative counterparts. This semantic correspondence led us to reconsider the interesting possibility of evaluative asymmetries between antonymic relations and to test whether such asymmetries in continuous evaluations can also produce asymmetrical response patterns in categorical evaluation tasks.

The present research

The present research was motivated by two aims. Firstly, we wanted to establish whether contrastive relations tend to produce weaker continuous evaluations than their assimilative counterparts. Secondly, we wanted to test whether such differences in continuous evaluations lead to a violation of the R-invariance assumption in MPT modeling of relational EC data, thereby inflating (or feigning) evidence for mere co-occurrence effects.

To address our first aim, we conducted a pilot adapting a procedure from a similar study that tested for graded effects of different types of assimilative relations [Hughes, Ye, Van Dessel, and De Houwer (2019); see Appendix A for details]. We presented a student sample ($N = 85$) with statements that each consisted of a generic subject ("X"), a relation (e.g., "causes") and a generic object (either "something positive" or "something negative"). To assess the impact of each relation on the evaluation of the sentence subject (corresponding to the CS in relational EC procedures), we asked participants to rate "X" based on the information contained in a given statement. To express their evaluation, participants were presented with a 21-point scale ranging from -10 (very negative) to +10 (very positive). Across participants, we tested a set of 30 pairs of antonymic relations. The set was selected based on the relations' presumed influence on subject evaluation (assimilative vs. contrastive) and included relation pairs that are commonly used in relational EC studies as well as novel pairs that have not been used before (for details, see Table A1 in Appendix A). In line with Moran et al. (2016), we found that statements containing contrastive relations produced overall weaker evaluations than equivalent statements containing their assimilative counterparts. Mirroring weaker relational EC effects for CSs presented with contrastive relations, the effect of object valence (positive vs. negative) on subject evaluation was significantly smaller for sentences containing contrastive relations ($\bar{d}_{\text{negative}-\text{positive}} = 9.073$, $SD_{\text{d}} = 2.472$) than for sentences containing their assimilative counterparts ($\bar{d}_{\text{positive}-\text{negative}} = 11.150$, $SD_{\text{d}} = 2.200$), $t(29) = 5.53$, $p < .001$.

Moreover, we also found considerable variability across relation pairs: while for many relation pairs, contrastive relations produced weaker subject evaluations, there were also relation pairs where subject evaluations were comparable across relation types (assimilative vs. contrastive) or where contrastive relations produced even stronger evaluations than their assimilative counterparts (see Table A1 in Appendix A).[3] This variability in the evaluative asymmetry across relation pairs is methodologically important for relational EC research in general, and for the present studies in particular. In general, our results show that evaluative differences between relation types are not inevitable and may be avoided (or induced) by careful selection of study materials. In the particular context of the present study, variability in evaluative asymmetry (across relation pairs) allows us to address our second aim (i.e., testing whether such asymmetries lead to inflated C parameters).

To address our second aim, we developed the following study rationale. As an indicator of mere co-occurrence effects, the C parameter is assumed to measure evaluative

[3] Note that such variability can only be produced by actual differences in evaluative asymmetries across relation pairs. By implication, the overall asymmetry (assimilative > contrastive) cannot simply be driven by an assimilative effect of object valence (which would produce identical asymmetries for all relation pairs). If existent, such an effect can be seen as a type of stimulus co-occurrence effect (in that target evaluations are affected by the mere presence of a valenced sentence object). The methods of the pilot study cannot rule out the presence of such an effect in our data (adding to evaluative asymmetries by a constant amount for all relation pairs). The presence of such an effect is, however, inconsequential for interpreting differences in evaluative asymmetry across relation pairs (as explained at the beginning of this footnote).

responding that is driven by US valence and unqualified by CS-US relations (assimilative vs. contrastive). We therefore assumed that a C parameter driven entirely by co-occurrence effects should be unaffected by the specific relation pair that is used in a given relational EC procedure. By contrast, a C parameter driven (at least partly) by violated R-invariance ($R_{assimilative} > R_{contrastive}$) should be increased when the relational EC procedure relies on a relation pair with a large evaluative asymmetry (assimilative > contrastive). Taken together, this implies that a C parameter inflation (based on evaluative asymmetries within relation pairs) can be tested by comparing C parameters across procedures using more vs. less asymmetrical relation pairs.

To implement this rationale, we used the data from the pilot study to select a symmetrical relation pair (where assimilative and contrastive relations produced comparably strong subject evaluations) as well as an asymmetrical relation pair (where the contrastive relation produced weaker subject evaluations). We then designed a relational EC procedure that included all four relations and thus manipulated not only US valence (positive vs. negative) and relation type (assimilative vs. contrastive) but also the difference in continuous evaluations produced by assimilative vs. contrastive relations (relation pair: symmetrical vs. asymmetrical). In this relational EC procedure, we used the same USs for CSs presented with relations from the symmetrical vs. asymmetrical pairs, thereby ensuring that the different relation pairs (specifically, their discrepant differences in continuous evaluations) were the only possible source of variation in the size of the C parameter (between relation pair conditions). We assumed that the pretested differences in continuous evaluations would also affect categorical CS evaluations, thus leading to a violation of the R-invariance assumption ($R_{assimilative} > R_{contrastive}$) in the asymmetrical condition. This logic led us to predict a larger-than-zero C parameter in the asymmetrical condition (prediction $\mathcal{H}_1$: $C_{asymmetrical} > 0$). Moreover, we also expected the C parameter to be larger in the asymmetrical than in the symmetrical condition (prediction $\mathcal{H}_2$: $C_{asymmetrical} > C_{symmetrical}$). Finally, we were also interested in the size of C parameter in the symmetrical condition. Assuming we selected a truly symmetrical pair, a C parameter of negligible size (prediction $\mathcal{H}_3$: $C_{symmetrical} = 0$) would indicate that previous evidence for mere co-occurrence effects was not only inflated, but entirely feigned by evaluative asymmetries between assimilative vs. contrastive relations.[4] In the remainder of this article, we report two studies testing these predictions.

Experiment 1

Methods

All measures, manipulations and data exclusions are reported. Experiment 1 was pre-registered on AsPredicted (#82309). Materials, data and analysis scripts are publicly available on the Open Science Framework: https://osf.io/zfdtb/?view_only=e3d101ec5f474be8be07b7e349376e37.

Participants. We recruited 55 participants through a mailing list for psychology students at Friedrich Schiller University Jena. As compensation, participants received

[4] We acknowledge that the interpretability of a non-significant $C_{symmetrical}$ is complicated by the fact that the presence of an unqualified co-occurrence effect may be concealed if the symmetrical relation pair featured a reversed asymmetry between relation types (with stronger evaluations for the contrastive than for the assimilative relation). If present, such a reversed asymmetry would lead to stronger relational EC effects for contrastive than for assimilative CSs and thereby cancel out (parts of) the co-occurrence effect (resulting in a non-significant $C_{symmetrical}$). Note that, although possible, such an explanation of $C_{symmetrical} = 0$ rests on several equality assumptions. Firstly, to result in $C_{symmetrical} = 0$, the effect of the reversed asymmetry (on CS evaluations) must have the same size as the opposing co-occurrence effect in our relational EC procedure. Secondly, to go undetected in our pilot study (see Appendix B), the reversed asymmetry in the symmetrical relation pair must have been overshadowed by a co-occurrence effect that was present in the pilot study (see previous footnote) and of exactly the same size as the reversed asymmetry between relation types. Since we cannot think of any independent reason to adopt these equality assumptions, we consider a reversed asymmetry in the symmetrical relation pair an unlikely explanation for a non-significant $C_{symmetrical}$ (if obtained).

partial course credit.[5] Based on pre-registered criteria, we excluded 12 participants who failed at least one of two seriousness checks administered at the end of the experiment (see section *Procedure*). In addition, we excluded another participant whose classification data was not available due to an unknown technical error. The final sample consisted of 42 participants (90.48% female; $M_{\text{age}} = 21.69$, $SD_{\text{age}} = 3.41$).

Design. The experiment implemented a 2 (*US valence*: positive vs. negative) × 2 (*relation type*: assimilative vs. contrastive) × 2 (*relation pair*: symmetrical vs. asymmetrical) within-participants design.

Materials. We programmed the experiment with *E-Prime 3.0*.

As CSs, we used eight images of potion bottles of various shapes and colors. Based on a second pilot study (see Appendix B), we selected three positive character traits (courage, patience, self-discipline) as positive USs, and three negative character traits (greed, indifference, cowardice) as negative USs. In the learning procedure, each character trait was represented by three images depicting a scene related to the character trait (to give an example, indifference was represented by three different images of people ignoring a begging person).

Based on the same pilot study, we also selected a symmetrical relation pair (with comparably strong evaluations for assimilative and contrastive relations) and an asymmetrical relation pair (with stronger evaluations induced by the assimilative in comparison to the contrastive relation). The symmetrical relation pair consisted of the German translations of "turn on" and "turn off" (einschalten vs. ausschalten). The asymmetrical relation pair consisted of the German translations of "strengthen" and "weaken" (stärken vs. schwächen; assimilative vs. contrastive relation, respectively).

[5] The study was run as part of a student project. Due to time pressure and lack of monetary funding, we pre-registered a data collection period of approx. 5 weeks. In this period, we collected data from as many participants as possible.

Measures and procedure. The study was run online via *E-Prime Go*. All verbal materials were presented in German.

First, participants were asked to indicate their age and gender. Next, participants were welcomed and thanked for their willingness to participate in the experiment. They were told that the experiment consisted of four parts and would take about 25 minutes. They were also asked to read all instructions carefully.

US pre-rating. Participants were told that they would first be presented with a number of character traits each represented by three images. They were instructed to think about each character trait and its consequences and then to indicate how positive or negative they deemed a given trait. Subsequently, participants were asked to give fine-graded ratings of each character trait on a scale ranging from -10 (very negative) to +10 (very positive). After the instructions, participants worked through all of the six character traits. The order in which the character traits were presented was randomized for each participant anew. Each trial started with the character trait and its three representing images being presented in the upper half of the screen. After 15 seconds, a prompt ("How positive or negative do you deem this character trait?") and a 21-point rating scale appeared in the lower half of the screen. There was no time limit for rating the character traits and participants had to click on a "Continue" button in order to proceed to the next trial.

Conditioning procedure. Participants were told that they would now be presented with image pairs each consisting of a potion and a character trait. It was also announced that for each image pair, they would learn whether the potion turns on, turns off, strengthens or weakens the paired character trait. Participants were instructed to observe carefully the image pairs and the relational information and to think about each potion in terms of its effect on the character traits.

The conditioning procedure consisted of 144 trials separated by a self-paced break after the first 72 trials. On each trial, a CS-relation-US triplet was presented for 5,000 ms. Each triplet consisted of a potion image (left), a relational qualifier (middle) and an image representing a character trait (right). The trials were separated by blank screens presented for 1,000 ms. For each participant, the eight potion images were randomly assigned to the eight US valence × relation type × relation pair conditions. This implies that each potion image was always presented together with one relation ([1] turn of, [2] turn off, [3] strengthen, [4] weaken) that was either assimilative ([1] and [3]) or contrastive ([2] and [4]) in nature and belonged either to the symmetrical ([1] and [2]) or the asymmetrical relation pair ([3] and [4]). Each of the four CSs assigned to the positive (US valence) × relation type × relation pair conditions was paired twice with each of the nine images representing the three positive character traits patience, self-discipline, courage. Conversely, each of the four CSs assigned to the negative (US valence) × relation type × relation pair conditions was paired twice with the nine images representing the three negative character traits (cowardice, greed, indifference). In both blocks of the conditioning procedure, each CS was presented nine times (once with each US image pertaining to its assigned level of US valence). Trial order was randomized for each participant and block.

CS evaluation phase. All participants performed a speeded classification task (SCT) followed by a continuous evaluation task (CET). For the SCT, participants were told that they would now be presented with single potion images. Participants were instructed to place their index fingers on the “K” and “F” keys and to indicate as fast as possible whether they would drink a given potion (by pressing the “K” and “F” keys for “yes” and “no”, respectively).[6] The SCT consisted of 72 trials (nine presentations per CS) and trial order was randomized for each participant anew. Each trial started with an

[6] The present SCT asked for hypothetical decisions (about drinking the potions) rather than CS evaluations (as positive vs. negative), and might therefore measure other, non-evaluative constructs (such as memory for the CS-US propositions or curiosity about drinking the potions). To check whether our SCT was indeed an evaluation measure, we calculated the Pearson’s product moment correlation between the share of propositionally correct CS classifications and the mean CS rating. To create a common metric for

empty screen (1,000 ms) followed by a centrally positioned fixation cross. After 900 ms the fixation cross was replaced by a potion image which remained on screen until participants gave a response. If participants did not respond within 2,000 ms, the potion image was replaced by a verbal message (“Faster!”) printed in red.

For the CET, participants were again presented with single potion images (presentation order was randomized for each participant anew). Each potion image (displayed in the upper half of the screen) was accompanied by a prompt (“How positive or negative do you deem this potion?”) and a 21-point rating scale ranging from -10 (very negative) to +10 (very positive). There was no time limit for rating the potions and participants had to click on a “Continue” button in order to proceed to the next trial. Trials were separated by blank screens presented for 2,000 ms.

Seriousness checks. After the CET, participants were told that the main part of the experiment was now finished and that they would now be asked to answer a few questions about themselves.

For the first seriousness check, a text consisting of four long sentences was displayed. The first three sentences referred to attitude research and through length and writing style

CSs with positive vs. negative meanings, evaluations of CSs with a negative meaning (from the “assimilative” × “negative”, “contrastive” × “positive” conditions) were multiplied with -1. Aggregated across conditions, we found a strongly positive correlation, $r = .66$, 95% CI [.45, .80], $t(40) = 5.59$, $p < .001$. We repeated this analysis for the data from Experiment 2 and found an even stronger association, $r = .91$, 95% CI [.88, .93], $t(216) = 32.11$, $p < .001$. Since Experiment 2 included a test of memory for the CS-US propositions, we also calculated the Pearson’s product moment correlations between the share of correctly remembered CS-US propositions and the share of propositionally correct CS classifications, $r = .38$, 95% CI [.26, .49], $t(216) = 6.09$, $p < .001$, as well as between the share of correctly remembered CS-US propositions and the mean CS rating, $r = .45$, 95% CI [.33, .55], $t(216) = 7.34$, $p < .001$. Both correlations were strongly positive, but noticeably weaker than the positive correlation between the share of propositionally correct CS classifications and the mean CS rating. Based on these findings, we concluded that the present SCT tapped into the same evaluative contents as the CET (rather than into other, non-evaluative constructs).

were meant to discourage participants from reading the whole text. In the very last sentence, participants were instructed to ignore the upcoming question about their exercise habits (in order to demonstrate that they had read the entire passage). On the next screen, participants were presented with a list of seven physical activities and were asked to indicate which of these activities they performed regularly (by ticking a small box next to the respective activity). After having ticked all relevant boxes (or none at all), participants proceeded to the next screen by clicking on the “Continue” button at the bottom of the screen.

For the second seriousness check participants were simply asked to indicate whether they had participated conscientiously and with full attention. Finally, participants were thanked and debriefed and received a participation code (for claiming their partial course credit).

Data analysis. The data analyses were performed in R (R Core Team, 2023). The aggregated response frequencies from the SCT (see Table 1) were analyzed with the RCB model (see below) using the R package *MPTinR* (Singmann & Kellen, 2013).[7]

Based on the pre-registration, we initially fitted the RCB model with separate sets of parameters for CSs presented with relations from the symmetrical vs. asymmetrical relation pair ($R_{symmetrical}$, $C_{symmetrical}$, $B_{symmetrical}$ and $R_{asymmetrical}$, $C_{asymmetrical}$, $B_{asymmetrical}$, respectively). Because the original RCB model did not produce adequate fit, we ran several less restrictive model extensions (see Appendix D for details). Based on absolute fit and model selection criteria (AIC), an extended RCB model estimating separate R parameters for positively vs. negatively paired CSs provided the best account of the data (hereafter, RCB4a model). The RCB4a model is based on the plausible idea that propositional learning effects might differ between US valence conditions and has precedence in the literature (see Kukken et al., 2020, Experiment 2). Furthermore, the crucial parameter

[7] Traditional ANOVA-based analyses of SCT and CET data are reported in Appendix C.

estimates ($C_{symmetrical}$ and $C_{asymmetrical}$) based on the RCB4a were practically identical to those produced by the original RCB model (see Appendix D). Moreover, additional analyses using hierarchical model extensions showed that individual parameter estimates from the RCB4a model predicted continuous CS evaluations in a comparable manner to individual parameter estimates from the original (three-parameter) RCB model [c.f., Kukken et al. (2020); see Appendix E]. Taken together, we therefore deemed the RCB4a model as an appropriate baseline model for testing predictions $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$.

The three predictions were tested via formal model comparisons based on the G^2 statistic. For each prediction, we fitted an additional RCB4a model implementing the parameter restriction implied by the respective prediction ($\mathcal{H}_1$: $C_{asymmetrical} = 0$, $\mathcal{H}_2$: $C_{asymmetrical} = C_{symmetrical}$, $\mathcal{H}_3$: $C_{symmetrical} = 0$). We then compared the model fit of the restricted models with that of the unrestricted RCB4a model. To do so, we calculated the ΔG^2 value and its associated p value for each restricted model (in comparison to the unrestricted RCB4a model). Based on convention, p values $< .05$ will be taken to indicate that a given parameter restriction cannot be implemented without a significant decrease in model fit.

Results

RCB model. The RCB4a model estimating separate sets of parameters for CSs presented with symmetrical vs. asymmetrical relations fit the data well, $G^2 = 0$.[8] Parameter estimates and 95% confidence intervals are reported in Table 2.

In line with our first prediction ($\mathcal{H}_1$), the C parameter in the asymmetrical condition was significantly larger than zero, $\Delta G^2(1) = 6.82$, $p = .009$. Moreover, and in line with our

[8] Note that the four-parameter variants of the RCB model are "saturated" in the sense that they have as many freely estimated parameters as independent category counts. Parameters of MPT models are, however, bound to the unit interval, which is why saturated models in the above sense may still produce substantial misfit (then indicated by $G^2 > 0$).

Table 1

Experiment 1: Aggregated frequencies of 'yes' responses in the speeded classification task as a function of US valence, relation type and relation pair.

		Asymmetrical		Symmetrical	
US valence	Relation type	*N*	%	*N*	%
positive	assimilative	300	80.43	289	77.48
	contrastive	87	23.14	48	12.83
negative	assimilative	71	19.03	61	16.31
	contrastive	273	73.19	262	70.05

Table 2

Experiment 1: Parameter estimates (with 95% confidence intervals) based on the unrestricted RCB4a model.

	Asymmetrical		Symmetrical	
Parameter	$\hat{\theta}$	95% CI	$\hat{\theta}$	95% CI
R_{positive}	.573	[.514, .632]	.646	[.592, .701]
R_{negative}	.542	[.481, .602]	.537	[.478, .597]
C	.127	[.032, .221]	.010	[-.088, .109]
B	.475	[.421, .529]	.356	[.307, .406]

second prediction ($\mathcal{H}_2$), the C parameter was descriptively larger in the asymmetrical than in the symmetrical condition. However, the decrease in model fit produced by a restricted RCB4a model with $C_{asymmetrical} = C_{symmetrical}$ failed to reach significance, $\Delta G^2(1) = 2.78$, $p = .096$. Finally, and in line with our third prediction ($\mathcal{H}_2$), the C parameter in the symmetrical condition was close to zero and did not differ from it significantly, $\Delta G^2(1) = 0.04$, $p = .835$.

Discussion

Experiment 1 produced first evidence that differences in continuous evaluations between assimilative vs. contrastive relations can result in a violation of the R-invariance assumption in MPT modeling of relational EC, thereby inflating the C parameter as an indicator of mere co-occurrence effects. In line with predictions, we found a descriptively larger C parameter for CSs presented with relations from an asymmetrical relation pair than for CSs presented with relations from a symmetrical relation pair. The larger C parameter in the asymmetrical condition is well explained by a violation of the R-invariance assumption (due to substantially stronger continuous evaluations produced by the assimilative in comparison to the contrastive relation) and cannot be based on a stronger influence of stimulus co-occurrence (since USs and number of CS-US pairings were constant across conditions). Moreover, we also found that the C parameter in the symmetrical condition was negligible in size, indicating that evidence for mere co-occurrence effects is absent whenever R-invariance ($R_{assimilative} = R_{contrastive}$) is ensured through use of a symmetrical relation pair. Taken together, these findings suggest that previous studies may have overestimated evidence for mere co-occurrence effects (by drawing on asymmetrical relation pairs) and that such evidence would have been absent if symmetrical relation pairs had been used.

Though highly relevant for theoretical debates on evaluative learning and attitude formation, the conclusiveness of the present findings is undermined by two issues related to the small sample size realized in Experiment 1. Firstly, though substantial in size, the difference in C parameters across relation pair conditions failed to reach conventional levels of significance. As indicated by a post hoc power analysis, this lack of statistical significance is likely due to insufficient power: based on the current sample size and effect size estimates, formal model comparison (via G^2 statistics) was found to achieve only (approx.) 38.5 % power to indicate a significant difference between $C_{asymmetrical}$ and $C_{symmetrical}$. Secondly, the interpretability of the non-significant C parameter in the symmetrical condition (in terms of an absence of genuine mere co-occurrence effects) is also compromised by the small sample size of Experiment 1. To assess the severity of the problem, we conducted a post hoc power analysis using the size of the smallest yet significant C parameter in a comparable study ($C = .06$, see Heycke & Gawronski, 2020, Experiment 4) as an estimate of $C_{symmetrical}$: given the current sample size and estimates of the remaining MPT parameter (all but $C_{symmetrical}$), formal model comparison achieved only (approx.) 22.5 % power to indicate a significant difference between $C_{symmetrical}$ and zero.

In order to gather more conclusive evidence for inflated C parameters (when using asymmetrical relation pairs) and an absence of genuine mere co-occurrence effects (when using symmetrical relation pairs), we decided to replicate Experiment 1 with a sample size sufficiently large to ensure adequate power for testing all three predictions.

Experiment 2

Methods

All measures, manipulations and data exclusions are reported. Experiment 2 was pre-registered on the OSF: https://osf.io/3b28c. The data from Experiment 2 are publicly available on the Open Science Framework:
https://osf.io/zfdtb/?view_only=e3d101ec5f474be8be07b7e349376e37.

Participants. Participants were recruited through Prolific and received monetary compensation for their participation. The sampling pool was restricted to Prolific users aged between 18 and 35 years, with German as their first language and a Prolific approval rate of 100 percent. We collected data until 229 non-excluded participants were reached.

To determine sample size, we conducted power analyses in *multiTree* (Moshagen, 2010) focusing on prediction $\mathcal{H}_2$ ($C_{asymmetrical} > C_{symmetrical}$) as the critical test. We ran the same power analysis twice: once with the parameter estimates from the unrestricted RCB4a model (see Table 2) and a second time using the parameter estimates from the original RCB model implemented without restrictions (see Table D1 in Appendix D). Since the second analysis revealed a higher number of required observations, we decided to determine sample size based on the estimates from the unrestricted RCB model (to ensure a sufficiently powered test based on either model). Supposing true parameter values of $R_{asymmetrical} = .557$, $R_{symmetrical} = .592$, $C_{asymmetrical} = .127$, $C_{symmetrical} = .020$, $B_{asymmetrical} = .475$, $B_{symmetrical} = .355$, and $\alpha = .05$, the power analysis revealed that 16,219 observations were necessary to achieve 95% power for detecting a significant difference between $C_{asymmetrical}$ and $C_{symmetrical}$ (via model comparison based on the G^2 statistic).

We then checked whether the targeted number of observations also ensured adequate power for testing predictions $\mathcal{H}_1$ and $\mathcal{H}_3$. We used the previously listed estimates for true parameter values and found that 16,219 observations allowed for 99.99 % power to detect a significant difference between $C_{asymmetrical}$ and zero ($\alpha = .05$). To determine power for testing prediction $\mathcal{H}_3$, we again used the size of the smallest yet significant C parameter in a comparable study as an estimate of the C parameter in the symmetrical condition. Thus supposing true parameter values of $R_{asymmetrical} = .557$, $R_{symmetrical} = .592$, $C_{asymmetrical} = .127$, $C_{symmetrical} = .060$, $B_{asymmetrical} = .475$, $B_{symmetrical} = .355$, and $\alpha = .05$, we found that the targeted number of observations allowed for 80.61% power to detect a significant difference between $C_{symmetrical}$ and zero ($\alpha = .05$). Taken together, these calculations led us to conclude that 16,219 observations were enough to ensure adequate power in testing all three predictions.

To determine the required number of participants, we divided the required number of observations by the expected number of valid observations provided by each participant. Based on Experiment 1, we expected participants to respond in time on 98.8 percent of trials in the SCT. Given a total number of 72 trials, we thus expected 71 valid responses from each participant (on average). Taken together, these calculations implied a required sample size of 229 participants ($16,219/71 \approx 229$). To avoid a final sample size smaller than $N = 229$, we checked exclusion criteria after collecting 229 participants. We then re-sampled participants (again checking exclusion criteria) until 229 non-excluded participants were reached.

Design. The experiment followed a 2 (*US valence*: positive vs. negative) × 2 (*relation type*: assimilative vs. contrastive) × 2 (*relation pair*: symmetrical vs. asymmetrical) within-participants design.

Materials. The experiment was programmed in *lab.js* (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2021). Other than that, we used the same materials as in Experiment 1.

Measures and procedure. We used JATOS (Lange, Kühn, & Filevich, 2015) to run the study online. Instructions, measures and procedure were identical to Experiment 1 with a few exceptions.

Firstly, participants were given a 11-point scale (ranging from -5 [very negative] to +5 [very positive]) to express their evaluation of the six character traits at the beginning of the study (instead of the original 21-point scale, see *Procedure* section of Experiment 1). Secondly, we tested whether participants remembered the character trait represented by each image set (right after the rating of the character traits and before the start of learning procedure). To this aim, participants were presented with each image set (one after the other) and a list of the six traits (identical on all trials). For each image set, they were asked to select the represented trait and to click on a "Continue" button in order to proceed to the next trial. As a third change, the SCT was performed by pressing the "K" (yes) and "A" (no) keys (instead of "K" and "F" as in Experiment 1). Fourthly, participants were given a 11-point scale (ranging from -5 [very negative] to +5 [very positive]) to express their continuous CS evaluation at the end of the study (instead of a 21-point scale as in Experiment 1). Finally, we included a memory measure (for the CS-US propositions) at the end of the experiment (after the CS evaluation phase and before administering the seriousness checks).[9] This measure was included to control for memory differences between experimental condition (if existent) allowing for a better test of our focal predictions. In the memory measure, participants were again presented with single potion images (presentation order was randomized for each participant anew). Each potion image (displayed in the upper half of the screen) was accompanied by a list of nine response options: (1) turns on positive character traits,(2) turns on negative character traits,(3) turns off positive character traits,(4) turns off negative character traits,(5) strengthens positive character traits, (6) strengthens negative character traits, (7) weakens positive character traits, (8) weakens negative character traits, (9) I don't remember. There was no time limit for selecting a response option and participants had to click on a "Continue" button in order to proceed to the next trial.

Apart from these changes, Experiment 2 was identical to Experiment 1. Importantly, we implemented the same seriousness checks and used them to exclude inattentive participants.

Inclusion and exclusion criteria. We planned to exclude participants with no data in at least one of the cells of the *US valence* × *relation type* × *relation pair* design.

[9] We disclose that this measure was added to the study protocol only after the registered report had received stage-1 acceptance. The change was approved by the action editor before Experiment 2 was pre-registered on the OSF and before we began data collection.

However, this criterion did not apply to any participant (for SCT and CET data). As pre-registered, data from participants who failed at least one seriousness check (by checking at least one physical activity and/or responding "no" to the question about conscientious and attentive participation) were excluded. This criterion applied to 34 participants. For specific measures/analyses, we also excluded participants who gave the same response on all trials of the measurement task.[10] For the SCT (CET), this criterion applied to two (five) participants.

Data analysis. All data analyses were performed in R (R Core Team, 2023). We analyzed the aggregated response frequencies from the SCT with the RCB model using the R package *MPTinR* (Singmann & Kellen, 2013).

We followed the pre-registered data analysis protocol. The pre-registered protocol included complementary RCB model analyses on (a) the whole sample (including all CSs without memory exclusions) and (b) on the sub-sample of CSs with correct memory for the CS-US proposition (i.e., the paired US valence together with the specific CS-US relation). We disclose that the sub-sample analyses ([b]) were added to the data analysis protocol only after the registered report had received stage-1 acceptance. However, the logic and inclusion of these analyses was approved by the action editor before Experiment 2 was pre-registered on the OSF and before we began data collection. As mentioned earlier, the sub-sample RCB analyses were included to control for memory differences between experimental conditions (if present). As explained by Kukken et al. (2020), better memory for assimilative than for contrastive propositions produces inflated C parameter estimates (in comparison to the actual size of the co-occurrence effect). In the present context, discrepant C parameter estimates in the two relation pair conditions might therefore be explained by differences in evaluative asymmetry (as intended) or, alternatively, by differences in memory asymmetry (see above). Moreover, previous simulation studies

[10] We disclose that this exclusion criterion was not included in the pre-registration.

showed that evaluative asymmetries (between relation type conditions) result in strongly inflated C parameters when overall memory levels are high, but produce only weakly inflated C parameter estimates when overall memory levels are low (Bading, 2021). In the present context, C parameter differences between relation pair conditions may therefore be blurred by low levels of memory for CS-US propositions. In a simulation study, we determined that sub-sample RCB analyses (including only CSs with correct memory for the CS-US proposition) can alleviate both of these issues and thus allow for more informative tests of our focal predictions (when memory accuracy is low and/or when memory differences are present).[11]

For both samples ([a] and [b]), we planned to fit the original RCB model with separate sets of parameters for CSs presented with relations from the symmetrical vs. asymmetrical relation pairs. Given adequate model fit based on the G^2 statistic, we planned to test predictions $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ via formal model comparison. To do so, we planned to fit additional RCB models implementing the parameter restriction implied by the respective prediction ($\mathcal{H}_1$: $C_{asymmetrical} = 0$, $\mathcal{H}_2$: $C_{asymmetrical} = C_{symmetrical}$, $\mathcal{H}_3$: $C_{symmetrical} = 0$). To compare model fit of the restricted models with that of the unrestricted RCB model, we planned to calculate the ΔG^2 value and its associated p value for each restricted model (in comparison to the unrestricted RCB model). Based on convention, we planned that p values $< .05$ would be taken to indicate that a given parameter restriction cannot be implemented without a significant decrease in model fit.

If the original RCB model did not fit the data, we planned to run the less restrictive RCB4a model (estimating separate R parameters for positively vs. negatively paired CSs) as well as two other model extensions (model RCB4b [RCB4c] estimating separate C [B] parameters for positively vs. negatively paired CSs, see Appendix D). Based on absolute fit and model selection criteria (AIC), we planned to select the best fitting model (among all

[11] For details, see “Simulation study 1: effects of memory differences on C parameter estimation with asymmetrical relation pairs” in the OSF repository.

four variants of the RCB model) and to use it as a baseline model for testing our three predictions (see above). Moreover, we also planned to perform additional analyses using hierarchical extensions of the selected model (results are reported in see Appendix E). In particular, we planned to check whether individual parameter estimates from the selected model predicted continuous CS evaluations in a comparable manner to individual parameter estimates based on the original RCB model (c.f., Kukken et al., 2020).

Finally, we also performed parallel tests of predictions $\mathcal{H}_1$ to $\mathcal{H}_3$ on CS ratings from the CET (again on the whole sample [a] and the sub-sample [b]). We disclose that these analyses were not included in the pre-registered data analysis protocol, but mirrored CET data analyses performed in Experiment 1 (see Appendix C for details). The results of the parallel tests and other analyses (on CS classifications, US ratings and memory data) are reported in Appendix C.

Results

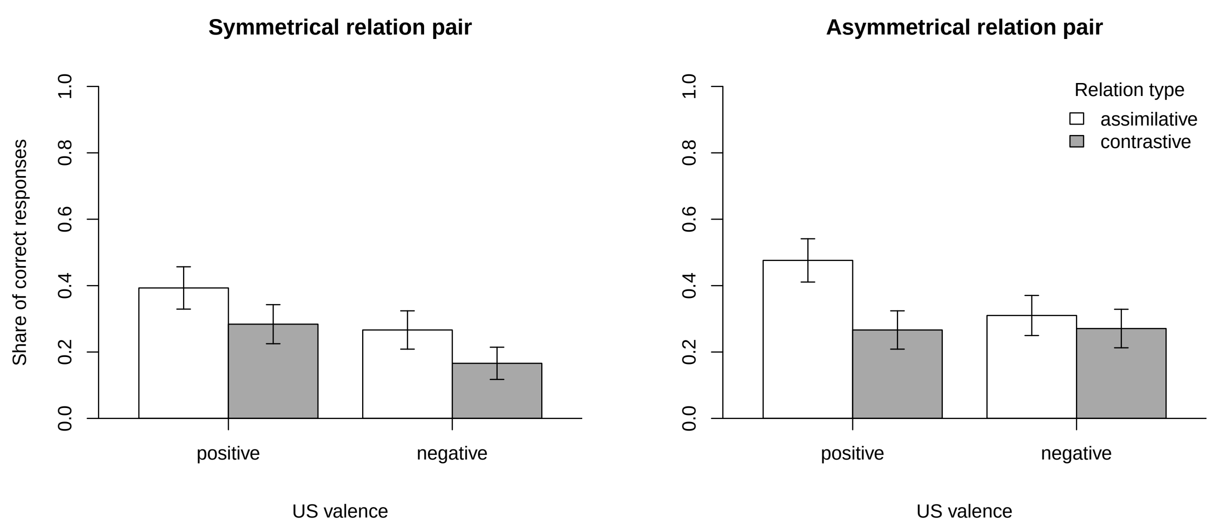

Figure 1. Experiment 2, memory test: memory accuracy as a function of US valence, relation type and relation pair.

Memory for CS-US propositions. Figure 1 shows memory accuracy as a function of US valence, relation type and relation pair.

To test for memory differences between experimental conditions, we calculated a 2 (US valence) × 2 (relation type) × 2 (relation pair) logistic regression on the shares of correctly remembered CS-US propositions. The logistic regression revealed significant main effects of US valence, $b = 0.24$, 95% CI $[0.14, 0.35]$, $z = 4.65$, $p < .001$, of relation type, $b = 0.28$, 95% CI $[0.17, 0.38]$, $z = 5.24$, $p < .001$, and of relation pair, $b = -0.14$, 95% CI $[-0.24, -0.03]$, $z = -2.59$, $p = .010$. Moreover, the three-way interaction was also significant, $b = -0.10$, 95% CI $[-0.21, 0.00]$, $z = -1.99$, $p = .046$. All other interactions did not reach significance, all ps $\geq .143$.

The main effect of US valence reflected relatively higher memory accuracy for positively paired CSs than for negatively paired CSs. The main effect of relation type reflected relatively higher memory accuracy for assimilative than for contrastive CSs. The main effect of relation pair reflected relatively higher memory accuracy in the asymmetrical than in the symmetrical condition.

To disentangle the three-way interaction, we calculated 2 (US valence) × 2 (relation type) logistic regressions (one per relation pair condition). In the symmetrical condition, the logistic regression revealed significant main effects of US valence, $b = 0.32$, 95% CI $[0.17, 0.47]$, $z = 4.14$, $p < .001$, and of relation type, $b = 0.27$, 95% CI $[0.12, 0.42]$, $z = 3.57$, $p < .001$. The two-way interaction between US valence and relation type was non-significant, $b = -0.03$, 95% CI $[-0.18, 0.12]$, $z = -0.36$, $p = .717$.

In the asymmetrical condition, the logistic regression revealed significant main effects of US valence, $b = 0.17$, 95% CI $[0.03, 0.31]$, $z = 2.37$, $p = .018$, and of relation type, $b = 0.28$, 95% CI $[0.14, 0.42]$, $z = 3.86$, $p < .001$. The two-way interaction between US valence and relation type was also significant, $b = 0.18$, 95% CI $[0.04, 0.32]$, $z = 2.53$, $p = .011$. This two-way interaction reflected relatively higher memory accuracy in the

Table 3

Experiment 2, whole sample without memory exclusions: Parameter estimates (with 95% confidence intervals) based on the unrestricted RCB4a model.

	Asymmetrical		Symmetrical	
Parameter	$\hat{\theta}$	95% CI	$\hat{\theta}$	95% CI
R_{positive}	.458	[.431, .485]	.572	[.547, .598]
R_{negative}	.234	[.205, .264]	.276	[.247, .305]
C	.070	[.038, .102]	.085	[.050, .121]
B	.438	[.421, .454]	.435	[.416, .454]

assimilative condition than in the contrastive condition only for positively paired CSs, $b = 0.46$, 95% CI $[0.26, 0.66]$, $z = 4.59$, $p < .001$, but not for negatively paired CSs, $b = 0.10$, 95% CI $[-0.11, 0.30]$, $z = 0.93$, $p = .355$.

RCB model.

Whole sample (without memory exclusions). For the whole sample (including all CSs without memory exclusions), the RCB model did not provide adequate fit, $G^2(2) = 338.30$, $p < .001$. As in Experiment 1, the RCB4a model (estimating separate R parameters for positively vs. negatively paired CSs) provided the best account of the data (according to the AIC and statistics of absolute fit, see Appendix D for details). The absolute fit of the RCB4a model was adequate, $G^2 = 0$.[12] Parameter estimates and 95%

[12] Note that four-parameter variants of the RCB model are "saturated" in the sense that they have as many freely estimated parameters as independent category counts. Parameters of MPT models are, however, bound to the unit interval, which is why saturated models in the above sense may still produce substantial misfit (then indicated by $G^2 > 0$).

Table 4

Experiment 2, sub-sample with memory exclusions: Parameter estimates (with 95% confidence intervals) based on the unrestricted RCB4a model.

	Asymmetrical		Symmetrical	
Parameter	$\hat{\theta}$	95% CI	$\hat{\theta}$	95% CI
R_{positive}	.739	[.702, .775]	.787	[.754, .820]
R_{negative}	.627	[.583, .672]	.711	[.662, .759]
C	.288	[.203, .372]	.127	[.017, .237]
B	.510	[.451, .569]	.449	[.386, .513]

confidence intervals are reported in Table 3.

The C parameter in the asymmetrical condition was significantly larger than zero, $\Delta G^2(1) = 18.50$, $p < .001$, as was the C parameter in the symmetrical condition, $\Delta G^2(1) = 21.93$, $p < .001$. Contrary to expectations, $C_{symmetrical}$ was descriptively larger than $C_{asymmetrical}$. However, the difference between the two C parameters was non-significant, $\Delta G^2(1) = 0.40$, $p = .526$.[13].

[13] We also tested R and B parameters against 0 and .5, respectively. In the symmetrical condition, all three parameters differed significantly from their reference value [R_{pos} vs. 0: $\Delta G^2(1) = 1,412.13$, $p < .001$; R_{neg} vs. 0: $\Delta G^2(1) = 229.46$, $p < .001$; B vs. .5: $\Delta G^2(1) = 42.70$, $p < .001$] Moreover, we also found that, in the symmetrical condition, R_{pos} was significantly larger than R_{neg}, $\Delta G^2(1) = 221.39$, $p < .001$. In the asymmetrical condition, we found the same pattern [R_{pos} vs. 0: $\Delta G^2(1) = 883.80$, $p < .001$; R_{neg} vs. 0: $\Delta G^2(1) = 320.35$, $p < .001$; B vs. .5: $\Delta G^2(1) = 51.20$, $p < .001$; R_{pos} vs. R_{neg}: $\Delta G^2(1) = 116.91$, $p < .001$]. Furthermore, we found that both R_{pos} and significantly larger in the symmetrical than in the asymmetrical condition, $\Delta G^2(1) = 36.12$, $p < .001$. The same was true for R_{neg}, $\Delta G^2(1) = 3.84$, $p = .050$. Finally, the difference between $B_{symmetrical}$ and $B_{asymmetrical}$ was non-significant, $\Delta G^2(1) = 0.04$, $p = .835$.

Sub-sample (with memory exclusions). To control for the previously described memory differences, the following analyses included only CSs with correct memory for the CS-US proposition (30.5% of all SCT trials).[14]

As before, the original RCB model did not provide adequate fit, $G^2(2) = 21.30$, $p < .001$. According to the AIC and absolute fit statistics, the RCB4a model was again the best fitting model (see Appendix D for details). The absolute fit of the RCB4a model was adequate, $G^2 = 0$. Parameter estimates and 95% confidence intervals are reported in Table 4.

In line with our first prediction, the C parameter in the asymmetrical condition was significantly larger than zero, $\Delta G^2(1) = 39.83$, $p < .001$. In line with our second prediction, $C_{asymmetrical}$ was significantly larger than $C_{symmetrical}$, $\Delta G^2(1) = 5.19$, $p = .023$. Finally, and contrary to our third prediction, the C parameter in the symmetrical condition was also significantly larger than zero, $\Delta G^2(1) = 4.97$, $p = .026$.[15]

---

[14] Note that inclusion in these analyses required correct recollection of the specific relation, not just of relation type. To give an example, selecting option "turns off positive character traits" instead of the correct option "weakens positive character traits" would have led to exclusion from the sub-sample analyses.

[15] We also tested R and B parameters against 0 and .5, respectively. In the symmetrical condition, both R parameters differed significantly from 0 [R_{pos}: $\Delta G^2(1) = 956.30$, $p < .001$; R_{neg}: $\Delta G^2(1) = 511.60$, $p < .001$], while the difference between B and .5 was non-significant ($\Delta G^2(1) = 2.48$, $p = .115$). Moreover, we also found that, in the symmetrical condition, R_{pos} was significantly larger than R_{neg}, $\Delta G^2(1) = 6.80$, $p = .009$. In the asymmetrical condition, we found an equivalent pattern [R_{pos} vs. 0: $\Delta G^2(1) = 873.70$, $p < .001$; R_{neg} vs. 0: $\Delta G^2(1) = 477.27$, $p < .001$; B vs. .5: $\Delta G^2(1) = 0.11$, $p = .740$; R_{pos} vs. R_{neg}: $\Delta G^2(1) = 14.51$, $p < .001$]. Furthermore, we found that R_{neg} was significantly larger in the symmetrical than in the asymmetrical condition, R_{neg}: $\Delta G^2(1) = 6.02$, $p = .014$). For R_{pos}, the difference between relation pair conditions was marginally significant, $\Delta G^2(1) = 3.69$, $p = .055$. Finally, the difference between $B_{symmetrical}$ and $B_{asymmetrical}$ was non-significant, $\Delta G^2(1) = 1.90$, $p = .168$.

Discussion

In Experiment 2, we tested predictions $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ in RCB analyses on the whole sample of all CSs (without memory exclusions) and on the sub-sample of CSs with correct memory for the CS-US proposition. Departing from Experiment 1, RCB analyses on the whole sample revealed above-zero C parameters estimates (for symmetrical and asymmetrical conditions) that were similar in size and did not differ from each other significantly. Results from this sample type thus confirmed prediction $\mathcal{H}_1$, disconfirmed prediction $\mathcal{H}_3$, and failed to provide evidence for C parameter inflation due to evaluative differences between assimilative vs. contrastive relations (prediction $\mathcal{H}_2$).

However, we also found substantial differences in memory accuracy (for CS-US propositions) across conditions, limiting the interpretability of the previous results. In the symmetrical condition, memory accuracy was substantially higher for assimilative than for contrastive CSs for both levels of US valence. As explained by Kukken et al. (2020), such memory differences lead to above-zero C parameter estimates in the absence of genuine co-occurrence effects (or evaluative differences between assimilative vs. contrastive relations). In the asymmetrical condition, a main effect of relation type (assimilative > contrastive) on memory accuracy was present only among positively paired CSs (suggesting that $C_{asymmetrical}$ should be less affected by estimation bias from memory differences). Importantly, however, memory accuracy was also rather low in both relation pair conditions, implying that, if present, C parameter inflation in the asymmetrical condition (due to evaluative asymmetries) will be underestimated.[16] Taken together, the reported memory levels and differences (in the two relation pair conditions) create a bias against our central prediction ($C_{asymmetrical} > C_{symmetrical}$) and therefore render results from whole sample RCB analyses uninformative with respect to C parameter inflation due to

[16] For an illustration, see "Simulation study 1: effects of memory differences on C parameter estimation with asymmetrical relation pairs" in the OSF repository.

evaluative differences between assimilative vs. contrastive relations.

To counteract estimation biases from memory differences and low memory levels, we performed pre-registered RCB analyses on the sub-sample of CSs with correct memory for the CS-US proposition. In these sub-sample analyses, $C_{asymmetrical}$ was significantly larger than zero (confirming prediction $\mathcal{H}_1$), as was $C_{symmetrical}$ (disconfirming prediction $\mathcal{H}_3$). Most importantly, prediction $\mathcal{H}_2$ was confirmed by a significantly larger C parameter estimate in the asymmetrical than in the symmetrical condition. When controlling for differences in memory accuracy, we therefore obtained evidence for C parameter inflation due to evaluative differences between assimilative vs. contrastive relations. Before addressing the theoretical and methodological implications of the present findings, we will discuss (and dispel) an alternative explanation that focuses on the use of sub-sample RCB analyses and explains the crucial finding ($C_{asymmetrical} > C_{symmetrical}$) as an artifact of this analytical approach.

As mentioned earlier, we decided to incorporate sub-sample RCB analyses based on the findings of a first simulation study.[17] In the context of C parameter estimation with asymmetrical relation pairs, this simulation study showed (a) that memory differences (assimilative > contrastive) and low memory levels produce estimation biases in whole-sample C parameters, and (b) that these biases are eliminated (entirely or in part) in sub-sample RCB analyses. Moreover, the degree of success (in eliminating these estimation biases) was shown to depend on the chance level of the memory test[18]: the

[17] For details, see "Simulation study 1: effects of memory differences on C parameter estimation with asymmetrical relation pairs" in the OSF repository.

[18] By chance level of the memory test, we refer to the probability for guessing the correct CS-US proposition when actual memory is absent/inaccessible. When the entire CS-US proposition is inaccessible, this probability should correspond to the inverse of the number of response options in the memory test. Note, however, that the chance level of the memory test is effectively higher (than the inverse) when parts of the CS-US proposition can be recollected.

lower the chance level, the weaker the left-over estimation bias (from memory differences [assimilative > contrastive] and/or low memory levels) in sub-sample C parameters.[19] When assuming that $C_{asymmetrical}$ was indeed partly driven by an evaluative asymmetry between relation type conditions, sub-sample RCB analyses are thus methodologically unproblematic and therefore superior to their whole-sample counterparts.

However, since we cannot simply assume what the present studies are supposed to demonstrate (C parameter inflation due to evaluative asymmetries), we deemed it indispensable to check for methodological problems of sub-sample RCB analyses also in alternative scenarios (where $C_{asymmetrical}$ is not driven by evaluative asymmetries). Most importantly, we considered the possibility that both relation pairs (turn on vs. turn off and strengthen vs. weaken) performed symmetrically in the CS classification task[20], and asked ourselves which other factors may then have produced $C_{asymmetrical} > C_{symmetrical}$ in sub-sample RCB analyses. To identify these factors (if present), we conducted a second simulation study exploring the effects of memory levels and differences on C parameter estimation with symmetrical relation pairs (i.e., relation pairs that perform symmetrically in the CS classification task).[21] In this simulation study, we again found that estimation bias from memory differences (assimilative > contrastive) is eliminated in sub-sample RCB analyses when the memory test has a chance level of zero. With above-zero chance levels,

---

[19] Note that with a chance level of zero, estimation bias from memory differences is entirely eliminated (in sub-sample RCB analyses).

[20] In our usage of terms, a relation pair that performs symmetrically in the CS classification task is a relation pair with comparable inference probabilities for assimilative vs. contrastive CSs. Note that in our evaluative asymmetry account of results from Experiment 2, we assume that turn on vs. turn off (the symmetrical relation pair) performed symmetrically in the CS classification task, while strengthen vs. weaken (the asymmetrical relation pair) performed somewhat asymmetrically (with a larger inference probability for assimilative than for contrastive CSs).

[21] For details, see "Simulation study 2: effects of memory differences on C parameter estimation with symmetrical relation pairs" in the OSF repository.

however, we found that memory differences (assimilative > contrastive) continue to produce estimation bias in sub-sample C parameters and can lead to pronounced overestimation of the actual co-occurrence effect. Importantly, the degree of left-over C parameter inflation (produced by a given memory difference) in sub-sample RCB analyses was found to reflect a complex interaction between various factors: the overall memory level, the inference probability of the CS-US propositions[22], the size of the actual co-occurrence effects and, again, the chance level of the memory test.

Based on findings from the second simulation study, we concluded that results from Experiment 2 may be compatible with an alternative account where the two relation pair conditions featured comparable co-occurrence effects and performed equally symmetrically in the CS classification task. In this alternative account, the C parameter difference in sub-sample RCB analyses ($C_{asymmetrical} > C_{symmetrical}$) may then be explained by differences, across relation pair conditions, in the particular value combination for the previously listed factors (such that in the asymmetrical [symmetrical] condition, the particular combination of inference probability, overall memory level and chance level results in more [less] left-over estimation bias from a given memory difference).[23]

While demonstrating that, in general, $C_{asymmetrical} > C_{symmetrical}$ (in sub-sample RCB analyses) may be explained by such an alternative account, the second simulation study did not show that the present results (including whole- vs. sub-sample R and C parameters as well as memory test data) can actually be explained by any specific combinations of the relevant factors. We therefore conducted a third simulation study implementing a

[22] By inference probability, we refer to the probability of drawing an evaluative inference from a given CS-US proposition (via inferential reasoning). In symmetrical relation pairs, inference probabilities are comparable for assimilative vs. contrastive CS-US propositions. In asymmetrical relation pairs, inference probabilities are higher for assimilative than for contrastive CS-US propositions.

[23] For an illustration, see “Simulation study 3 - Reproduction of results from Experiment 2 with a symmetrical model” in the OSF repository.

systematic 7-stage procedure in which specific value combinations were constructed and tested against well-defined criteria (reflecting actual result patterns in Experiment 2).[24] Despite using a wide-ranging and fine-grained set of value combinations, the 7-stage procedure was unsuccessful at reproducing findings from the asymmetrical condition (whereas, for the symmetrical condition, viable value combinations were easily obtained). In our view, this lack of success in the asymmetrical condition discredits the proposed alternative account, and suggests that results from this condition cannot be explained without assuming that "strengthening" vs. "weakening" CSs performed somewhat asymmetrically in the CS classification task.

To corroborate our own evaluative asymmetry account for the present results, we conducted a fourth and final simulation study.[25] In this simulation study, we implemented the same 7-stage procedure as in Simulation study 3, but included value combinations where inference probabilities for assimilative CSs were higher than inference probabilities for contrastive CSs (in addition to value combinations where inference probabilities were identical across levels of relation type). In line with our evaluative asymmetry account, we found that findings from the asymmetrical condition were reliably reproduced by value combinations with relatively larger inference probability differences (assimilative > contrastive), while findings from the symmetrical condition were reliably reproduced by value combinations in which such differences were absent or relatively small. Against the backdrop of Simulation study 3 and 4, the reported difference in sub-sample C parameters ($C_{asymmetrical} > C_{symmetrical}$) therefore provide positive evidence for C parameter inflation due to evaluative asymmetries between assimilative vs. contrastive relations.

[24] For details, see "Simulation study 3 - Reproduction of results from Experiment 2 with a symmetrical model" in the OSF repository.

[25] For details, see "Simulation study 4 - Reproduction of results from Experiment 2 with an asymmetrical model" in the OSF repository.

General discussion

In the present research, we tested whether evidence for co-occurrence effects can be overestimated due to evaluative differences between relation types. As a first step, we conducted a pilot study estimating evaluative differences in a set of 30 antonymic relation pairs. This pilot study showed (a) that contrastive relations produced weaker continuous evaluations (than their assimilative counterparts) in the majority of these relation pairs, and (b) that such evaluative differences (assimilative > contrastive) were absent or reversed in the remaining relation pairs.

In a second step, we used these findings to design an experimental procedure testing whether evaluative differences between relation types (assimilative > contrastive) lead to C parameter inflation in RCB analyses of categorical relational EC data. In this experimental procedure, we compared C parameter estimates from an asymmetrical relation pair (with a pre-tested evaluative difference between relation types [assimilative > contrastive]) to C parameter estimates from a symmetrical relation pair (where the pre-test did not indicate such an evaluative difference). Across two experiments, we tested a total of three predictions: ($\mathcal{H}_1$) $C_{asymmetrical} > 0$, ($\mathcal{H}_2$) $C_{asymmetrical} > C_{symmetrical}$, and ($\mathcal{H}_3$) $C_{symmetrical} = 0$. In Experiment 1, these predictions were tested in whole-sample RCB analyses (including all CSs presented in the conditioning procedure). In Experiment 2, we added a memory test to the experimental procedure. This allowed us to conduct complementary tests on the sub-sample of CSs for which the CS-US propositions had been correctly remembered.

In the following sections, we will summarize results from Experiments 1 and 2 (separately for each prediction), and explain their differences. Subsequently, we will present theoretical and methodological implications of the present findings, and discuss a potential confound in our experimental procedure (between evaluative asymmetry and applicability/judged usability). Finally, we will address a number of limitations and open questions and make suggestions for future research.

Summary of results from Experiments 1 and 2

Prediction $\mathcal{H}_1$. Our first prediction ($C_{asymmetrical} > 0$) received reliable support from whole-sample RCB analyses in Experiment 1, and whole- as well as sub-sample RCB analyses in Experiment 2. Taken together, these findings therefore show that, in the asymmetrical condition, the regular EC effect for assimilative CSs was reliably stronger than the reversed EC effect for contrastive CSs.

A noteworthy difference between the two experiments concerns the size of $C_{asymmetrical}$ in whole-sample RCB analyses (which was slightly larger in Experiment 1 than in Experiment 2). As explained earlier, an evaluative difference between relation types (assimilative > contrastive) will result in stronger C parameter inflation when memory levels are high rather than low (for a demonstration, see Bading, 2021). The difference in whole-sample C parameter estimates from asymmetrical conditions (across experiments) may therefore be explained by higher memory levels in Experiment 1 than in Experiment 2. Note that, since Experiment 1 did not include a test of memory for CS-US propositions, this explanation cannot be tested.

Prediction $\mathcal{H}_2$. Our second prediction ($C_{asymmetrical} > C_{symmetrical}$) received descriptive support from whole-sample RCB analyses in Experiment 1 and statistically reliable support from sub-sample RCB analyses in Experiment 2 (whereas whole-sample RCB analyses in Experiment 2 did not support $\mathcal{H}_2$).

A noteworthy inconsistency between the two experiments concerns the C parameter difference ($C_{asymmetrical} - C_{symmetrical}$) in whole-sample RCB analyses (which was clearly positive in Experiment 1, but close to zero in Experiment 2). This discrepancy (in the numerical differences between whole-sample C parameters) was driven by opposing shifts in the two components of the C parameter difference (across experiments): while the size of $C_{symmetrical}$ increased from Experiment 1 to Experiment 2, the size of $C_{asymmetrical}$ decreased (from the first to the second experiment). As explained in the previous section, the relatively smaller estimate of whole-sample $C_{asymmetrical}$ in Experiment 2 may be explained by lower memory levels (see above). Similarly, the relatively larger estimate of whole-sample $C_{symmetrical}$ in Experiment 2 may be explained by a relatively stronger difference in memory accuracy for assimilative vs. contrastive CSs (in the symmetrical condition of Experiment 2 in comparison to the symmetrical condition of Experiment 1). Note that, since memory for CS-US propositions was not assessed in Experiment 1, we cannot verify this explanation (by testing whether the symmetrical condition in Experiment 2 featured a stronger memory difference [assimilative > contrastive] than its counterpart in Experiment 1).

Another noteworthy inconsistency concerns the difference in result patterns from whole- vs. sub-sample RCB analyses in Experiment 2. This discrepancy is produced by a sharp increase in $C_{asymmetrical}$ from whole- to sub-sample RCB analyses (which is unmatched by a small increase in $C_{symmetrical}$). As stated above, this increase in $C_{asymmetrical}$ is explained by the fact that sub-sample RCB analyses eliminate estimation bias from low memory levels (thereby revealing the true magnitude of C parameter inflation due to evaluative differences between assimilative vs. contrastive relations), and could not be reproduced by an alternative model including only memory differences and co-occurrence effects but no evaluative differences between relation types (for details, see Simulation 3 in the OSF repository).

In summary, the apparent inconsistencies (between tests of $\mathcal{H}_2$) can be explained by a common logic (variability in memory accuracy across conditions and experiments) that is perfectly compatible with an evaluative asymmetry account of $C_{asymmetrical} > C_{symmetrical}$ in sub-sample RCB analyses from Experiment 2 (and in whole-sample RCB analyses from Experiment 1). Taken together, the present findings therefore demonstrate that evaluative differences between assimilative vs. contrastive relations can lead to inflated C parameter estimates (and, by implication, to overstated evidence for co-occurrence effects).

Prediction $\mathcal{H}_3$. Our third prediction ($C_{symmetrical} = 0$) received support from Experiment 1 (where $C_{symmetrical}$ was practically zero), but not from Experiment 2 (where $C_{symmetrical}$ was significantly larger than zero in whole- as well as sub-sample RCB analyses). Before we discuss these findings, we will address a number of challenges connected to their interpretation.

To remind the reader, prediction $\mathcal{H}_3$ is concerned with the size of the C parameter after controlling for evaluative differences between relation types (assimilative > contrastive). By selecting a symmetrical relation pair, we sought to ensure that $C_{symmetrical}$ would be unbiased by evaluative differences and therefore reveal the true magnitude of the unqualified co-occurrence effect. Assuming we selected a truly symmetrical pair, $C_{symmetrical} \approx 0$ would then indicate that evidence for unqualified co-occurrence effects is absent once evaluative differences between relation types are eliminated (suggesting that previous evidence was artifactual). Conversely, instances of $C_{symmetrical} > 0$ would show that evidence for co-occurrence effects is still present even after controlling for such differences (suggesting that previous evidence may have been inflated but not entirely feigned by a lack of control for evaluative differences between relation types). Unfortunately, the interpretability of both result patterns is compromised in the present studies.

As explained in the introduction, the interpretation of $C_{symmetrical} \approx 0$ (in terms of an absent co-occurrence effect) is generally complicated by the fact that the presence of an unqualified co-occurrence effect may be concealed if the symmetrical relation pair featured a reversed asymmetry between relation types (with stronger evaluations for the contrastive than for the assimilative relation) that went undetected in our second pilot study (based on which we selected character traits and relation pairs; see Appendix B). As mentioned earlier, such an explanation of $C_{symmetrical} = 0$ rests on a number of auxiliary assumptions which we deemed unlikely in the present study.

The second challenge refers to the interpretability of $C_{symmetrical} > 0$. In the context of Experiment 2, the interpretation of this result pattern is complicated by the fact that memory for the CS-US propositions was better for assimilative than for contrastive CSs (in the symmetrical condition). As explained earlier, such memory differences (assimilative > contrastive) are likely to produce inflated C parameter estimates (Kukken et al., 2020). In the context of Experiment 2, whole-sample estimates of $C_{symmetrical}$ (without control for memory differences between assimilative vs. contrastive conditions) are therefore unlikely to reveal the true magnitude of the unqualified co-occurrence effect (if present). By contrast, sub-sample estimates of $C_{symmetrical}$ may be unbiased by memory differences if the sub-sample does not include false positives (i.e., CSs for which the correct CS-US proposition was guessed rather than recollected). Such a "false-positive-free" sub-sample of CS classifications would require that all participants refrain from guessing in the memory test, which, unfortunately, seems unlikely in the context of Experiment 2.[26] As demonstrated in Simulation 2 (see OSF repository), C parameter inflation from memory differences remains present when the sub-sample of CS classifications contains false positives (with the exact degree of left-over estimation bias being determined by a host of factors). Taken together, these points imply that the interpretation of sub-sample $C_{symmetrical} > 0$ (in Experiment 2) is complicated by the (likely) presence of left-over estimation bias from memory differences between assimilative vs. contrastive conditions (and that the exact degree of left-over C parameter inflation cannot be determined).

Based on the previous considerations, it seems obvious that the present findings do not allow for strong conclusions about the true magnitude of the unqualified co-occurrence effect in the present studies. Due to the presence of memory differences (between

[26] This assessment is based on the fact that a total refrain from guessing (in the memory test) should result in a response distribution including only correct and "I don't know" responses (and no incorrect responses). By contrast, the memory data from Experiment 2 showed 30.4% correct responses, 19.9% "I don't know" responses and 50.3% incorrect responses.

assimilative vs. contrastive conditions), the above-zero estimate of whole-sample $C_{symmetrical}$ in Experiment 2 cannot be taken to show that evidence for co-occurrence effects is still present even after controlling for evaluative differences between relation types. Similarly, the above-zero estimate of sub-sample $C_{symmetrical}$ (Experiment 2) is likely to contain left-over estimation bias (from the memory differences) the exact size of which cannot be determined. This implies that any conclusion about the true magnitude of unqualified co-occurrence effects (from the size of sub-sample $C_{symmetrical}$) must rest on (currently) untestable assumptions about the size of the left-over estimation bias. To avoid such assumptions (as much as possible), we will refrain from drawing strong conclusions about the presence (or absence) of genuine co-occurrence effects in our studies. Instead, we will conclude this section with a remark on the possible range of the unqualified co-occurrence effect in the present studies (which is based on a simulation study that does not commit to any specific assumption about the size of the left-over estimation bias).

As mentioned earlier, we conducted Simulation study 3 to explore whether findings from Experiment 2 could be reproduced with a simple data-generating model that includes memory differences and co-occurrence effects (but excludes evaluative differences between relation types). Importantly, Simulation study 3 included parallel sets of data simulations with co-occurrence effects of 0, 0.05 and .1, and found that findings from the symmetrical condition of Experiment 2 could be reliably reproduced when co-occurrence effects were set to 0 and .05, but not when co-occurrence effects were set to .1. Since we did not implement any data simulations with co-occurrence effects between .05 and .1, we cannot give a specific upper limit for the possible range of the unqualified co-occurrence effect in the symmetrical condition of Experiment 2. Nevertheless, results from Simulation study 3 clearly show that the reported instances of $C_{symmetrical} > 0$ are perfectly compatible with absent or small co-occurrence effects (between 0 and .05).

Theoretical implications

In the previous sections, we explained that the present findings provide direct support for C parameter inflation due to evaluative differences ($\mathcal{H}_2$). Since previous process dissociation studies (Heycke & Gawronski, 2020; Kukken et al., 2020) did not control for evaluative differences between relation types (or equivalent differences in memory accuracy), our findings therefore imply that earlier evidence for co-occurrence effects (from these studies) may have been inflated (by uncontrolled differences between relation types).

A related implication refers to the interpretation of previous research on the moderators of the C parameter (e.g., Heycke and Gawronski (2020)). In this context, the present findings imply that any significant moderation of the C parameter (by a given experimental manipulation) may be driven by the portion of the parameter estimate that is due to an evaluative difference between relation types (instead of reflecting the functional properties of the underlying co-occurrence effect). In previous research (Bading, 2021), we showed that C parameter inflation from evaluative differences between relation types will be more pronounced in experimental conditions that give rise to better memory retrieval (for the CS-US propositions). Against this backdrop, the present findings may therefore shed new light on a previously inexplicable moderation of the C parameter by the duration of the response window in the CS classification task (and on its theoretical implications). Specifically, the larger C parameter in a condition with a longer response window in Heycke and Gawronski's (2020) Experiment 4 can now be explained by an evaluative asymmetry (between the assimilative vs. contrastive relations used in Heycke and Gawronski [2020]) that produces stronger C parameter inflation when memory retrieval is improved by a longer response window in the CS classification task. In summary, the present findings therefore show that (past and future) research on the moderators of the C parameters can only gain valid theoretical insights (about the nature of unqualified co-occurrence effects) if evaluative differences between relation types are controlled for.

A third implication of the present findings refers to the overall state of evidence for the existence of unqualified co-occurrence effects (and its consequences for the long-standing debate over single- vs. dual-process models of evaluative learning). As discussed by Hütter (2022), unqualified co-occurrence effects can be explained by all major accounts of evaluative conditioning (De Houwer, 2018; Gawronski & Bodenhausen, 2018; Stahl & Aust, 2018), so that simply demonstrating their existence does not advance this theoretical debate (Hütter, 2022). This fact, however, does not imply that a lack of reliable demonstrations of such effects cannot advance this debate either. On the contrary, it is easy to see that ongoing failures in producing reliable evidence for unqualified co-occurrence effects pose a grave challenge to dual-process models (by rendering associative processes empirically unnecessary). At the same time, such empirical failures do not seem to challenge propositional models (which assume that all instances of EC are driven by inferential reasoning on a propositional statement about the relation between CS and US). In these models, an unqualified effect of stimulus co-occurrence is explained by inferential reasoning on a simple proposition stating that CS and US have previously co-occurred in time and space (while qualified EC effects are explained by inferential reasoning on propositional statements including other, more informative relations).[27] For propositional models, consistent failures in demonstrating unqualified co-occurrence effects therefore have limited theoretical implications (in showing that simple propositions about stimulus co-occurrence do not suffice for the emergence of EC), but do not question their overarching principle (since successful demonstrations of EC can still be attributed to

[27] Note that an assimilative effect of stimulus co-occurrence (despite conscious knowledge of a contrastive CS-US relation) seems at odds with the idea that evaluative learning is always driven by intentional reasoning processes. This view is based on the fact that co-occurrence effects produce CS evaluations that are in clear violation of a simple and logically sound inferential rule and are thus unlikely to be based on intentional processes. From this perspective, one may argue that explaining unqualified co-occurrence effects by inferential reasoning on co-occurrence propositions amounts to postulating that evaluative learning is (sometimes) mediated by unintentional reasoning processes.

inferential reasoning on other, more complex relations).

As previously mentioned, the present findings do not allow for strong conclusions about the true magnitude of unqualified co-occurrence effects (in the present studies and/or in general). However, the present findings are in line with previous research showing that evidence for unqualified co-occurrence effects from the relational EC paradigm is (at the very least) interpretationally ambiguous (for an overview, see Bading, 2021). In combination with an absence of evidence for non-propositional co-occurrence effects from other paradigms (Högden, Hütter, & Unkelbach, 2018; Stahl & Bading, 2020; Stahl, Haaf, & Corneille, 2016), the present findings therefore add to a growing challenge for dual-process models (by questioning the empirical basis for postulating a contribution of associative processes to evaluative learning). At the same time, however, we also believe that future research (e.g., on the size of $C_{symmetrical}$ without bias from memory differences) is indispensable for drawing any strong conclusions on the presence or absence of unqualified co-occurrence effects. Until such research has been conducted, we recommend to refrain from strong claims on the existence of such effects (e.g., Gawronski et al., 2023).

Finally, while the present findings do not challenge the core principle of the propositional approach, they do highlight a certain lack of theoretical elaboration in propositional models of EC. Specifically, one cannot help but wonder whether the RCB model (with its built-in requirement of evaluative symmetry between relation types) would have been developed at all if current propositional theorizing were more specific on the factors that determine the strength of evaluative learning, and not just its direction (for a notable exception, see Hughes, Ye, & De Houwer, 2019). From this perspective, the present findings may thus be seen as a reminder that, though empirically successful, propositional models do have theoretical blind spots that need to be filled in by future research.

Methodological implications

The present findings have a number of important implications for studying co-occurrence effects with the RCB model (and with the process dissociation approach in general).

Most obviously, future studies with the RCB model need to incorporate relation pairs that do not possess an evaluative asymmetry between their assimilative and contrastive relations. Using such symmetrical relation pairs is important not only to achieve convincing demonstrations of unqualified co-occurrence effects (via above-zero C parameter estimates), but also to ensure that statistically significant moderations of the C parameter reflect the functional properties of actual co-occurrence effects (and not those of an evaluative asymmetry between relation types). To ensure that a given relation pair performs symmetrically in the context of a particular relational EC procedure (and not just in a generic test as implemented in Pilot study 1), future applications of the RCB model should therefore include reliable demonstrations of evaluative symmetry for the specific combinations of CSs, USs and CS-US relations that will be used in the relational EC procedure (for an example, see Pilot study 2 in Appendix B).

The present findings also imply that future process dissociation studies should always include a test of memory for the CS-US propositions. As illustrated by memory data from Experiment 2, memory accuracy may be higher for assimilative than for contrastive CSs (leading to C parameter inflation due to memory differences instead of evaluative asymmetries). Future process dissociation studies should therefore test for memory differences between experimental conditions. If memory differences (assimilative > contrastive) are shown to be present, whole-sample C parameters should not be interpreted (at least not without having estimated the degree of C parameter inflation that can be expected from the memory difference). In these situations, one may also try to obtain unbiased C parameter estimates from sub-sample RCB analyses that include only CSs for which the CS-US proposition was correctly remembered (as was done in Experiment 2 of the present study). Note, however, that C parameter inflation may remain present in these analyses when the chance level of the memory test is larger than zero (i.e., when the correct CS-US proposition can be guessed with an above-zero probability).

Finally, the present research also suggests that, when memory differences are present, simulation studies may be helpful (and sometimes indispensable) for a valid interpretation of results (e.g., by estimating the expected degrees of whole-sample C parameter inflation from a given memory difference, or by testing for left-over estimation bias in sub-sample RCB analyses).

Applicability and judged usability as alternative explanations

In the following section, we discuss an alternative explanation for $C_{asymmetrical} > C_{symmetrical}$ in terms of (unintended) differences in applicability and/or judged usability (Higgins, 1996) across relation pairs. If found to be plausible, this alternative explanation would question whether the reported instances of $C_{asymmetrical} > C_{symmetrical}$ were indeed driven by (intended) differences in evaluative asymmetry between relation pairs.

According to (Higgins, 1996), applicability refers to "the goodness of fit between some stored knowledge and the attended features of a stimulus" and is assumed to affect the likelihood of knowledge activation (the greater the fit, the greater the likelihood). In the same publication, judged usability is defined as the "judged appropriateness or relevance of applying stored knowledge to a stimulus" and is assumed to affect the likelihood of knowledge use (i.e., whether activated knowledge will actually be used in a given task and/or when assigning meaning to a stimulus).

To explain the present results (from both experiments), one may assume that our asymmetrical relation pair (strengthen vs. weaken) had lower applicability and/or judged usability with respect to the other learning materials (potions [CSs] and character traits [USs]) than our symmetrical relation pair (such that differences in evaluative asymmetry and differences in applicability/judged usability are completely confounded across relation pairs). If true, this assumption implies that relations from the asymmetrical pair had a lower probability (a) of being activated by later CS presentations (e.g., during the CS classification task) and/or (b) of being used to evaluate a CS (once activated). In the absence of an activated or usable CS-US relation, participants may then have based their CS classification on the US valence (if activated), leading to a stronger co-occurrence effect (and, by implication, a larger C parameter) in the asymmetrical condition.

To judge the plausibility of applicability differences between the two relation pairs (in the present experiments), we considered the memory data from Experiment 2 (as a proxy for the [knowledge] activation of the CS-US relations by later CS presentations). As previously mentioned, our analyses on the memory data from Experiment 2 revealed overall higher memory accuracy (for the CS-US propositions) in the asymmetrical than in the symmetrical condition. In our view, this pattern is at odds with the idea that the asymmetrical relation pair was less applicable to the CSs (and their paired USs) than the symmetrical relation pair. This view is based on the fact that lower applicability is assumed to result in a decreased likelihood of knowledge activation and should therefore have led to lower (instead of higher) memory accuracy in the asymmetrical condition. Based on this logic, we deemed a confound between evaluative asymmetry and applicability unlikely in the present experiments.

To assess whether the present relation pairs differed in terms of judged usability, we considered the evaluation data from our second pilot study (based on which we selected character traits and relation pairs). Looking at these data was based on the assumption that judged usability affects the likelihood of knowledge use (so that differences in judged usability between relations should show up in differences in the absolute strength of the evaluative responses produced by these relations). As reported in Appendix B, Pilot 2 showed that “turn on” vs. “strengthen” produced comparably strong evaluations, while “turn off” produced stronger evaluations than “weaken”. This pattern suggests that, if anything, differences in judged usability must have been restricted to the contrastive relations (so that a lower judged usability for “weaken” might have led to a lower likelihood of knowledge use and, consequently, weaker evaluative responses) and did not extend to the assimilative relations that were used in the present experiments. From this perspective, usability differences between the contrastive relations may thus be viewed as a possible reason for the (intended) differences in evaluative asymmetry across relation pairs. Importantly, this perspective therefore implies that judged usability and evaluative asymmetry do not pose separate (and competing) explanations for $C_{asymmetrical} > C_{symmetrical}$; instead, evaluative asymmetry is viewed as mediating the effect of judged usability on the C parameter. By contrast, a genuine alternative account (where $C_{asymmetrical} > C_{symmetrical}$ is driven by overall lower judged usability of the asymmetrical relation pair) seems to be dispelled by Pilot 2 showing that evaluative responses were comparably strong for the two assimilative relations (which is at odds with the idea that the asymmetrical relation pair possessed generally lower judged usability than the symmetrical relation pair).

In summary, a confound between evaluative asymmetry and applicability or judged usability (across relation pairs) seems unlikely in the present experiments. We therefore conclude that the reported instances of $C_{asymmetrical} > C_{symmetrical}$ can be attributed to the (intended) differences in evaluative asymmetry across relation pairs.

Limitations, open questions and future research

The present findings raise a number of interesting questions that may be addressed in future research.

A first question concerns the true size of $C_{symmetrical}$ (without estimation bias from memory differences) and is perhaps the hardest to address. The difficulty in obtaining an unbiased estimate of $C_{symmetrical}$ is based on the fact that such an estimate is only possible when memory differences are absent (or sufficiently small) or when the memory test has a chance level of zero. Unfortunately, experimenters do not have perfect control over either condition: memory differences may emerge despite optimized learning conditions and performance incentives, and some participants may still try to guess the correct CS-US proposition (in the memory test) despite being instructed not to do so. Despite these limits in experimental control, future research must still strive to achieve both conditions (and check for their attainment). If one or both conditions are found to be violated in a given study, interpretations of (whole- or sub-sample) $C_{symmetrical}$ as evidence for co-occurrence effects should be backed up by data simulations showing that the obtained estimate of $C_{symmetrical}$ cannot be explained by memory differences alone (for a possible procedure, see Simulation studies 3 and 4 in the OSF repository).

A second question concerns the effects of evaluative differences (between relation types) in other approaches for studying co-occurrence effects. Most importantly, Béna et al. (2022) reported higher attitudinal ambivalence for contrastive than for assimilative CSs and interpreted this pattern in terms of a conflict between opposing evaluations from inferential reasoning (on CS-US propositions), on the one hand, and unqualified co-occurrence effects, on the other hand. Based on the present findings, one cannot help but wonder whether higher attitudinal ambivalence (for contrastive CSs) may be driven by weaker CS evaluations produced by contrastive relations. To test this possibility, future research may combine the present relational EC procedure (including symmetrical vs. asymmetrical relation pairs) with the ambivalence measures used by Béna et al. (2022) and test whether differences in attitudinal ambivalence (contrastive > assimilative) are more pronounced in the asymmetrical than in the symmetrical condition.

A third and final question concerns the origins of evaluative asymmetries between relation types. In demonstrating substantial variability across relation pairs, Pilot study 1 suggests that evaluative differences between antonymic relations (assimilative vs. contrastive) do not have a single, hard-wired cause (such as greater processing ease for assimilative relations) but may be present or absent depending on certain properties of these relations. One such property of CS-US relations may be their diagnosticity with respect to the (propositionally correct) valence of the CS. Specifically, it may be the case that contrastive relations are seen as less diagnostic (for drawing an inference about the nature of the CS) and therefore result in weaker CS evaluations.[28] Another important property of CS-US relations may be their context-dependency. Specifically, certain contrastive relations may be more context-dependent and therefore require specific conditions to result in pronounced CS evaluations.[29] In our view, studying diagnosticity, context-dependency and other properties of assimilative and contrastive relations will not only bring theoretical progress (by enriching propositional models of evaluative learning) but may also help to improve research methods for studying evaluative effects of stimulus co-occurrence.

Conclusions

The present research demonstrates that C parameter estimates from the RCB model are influenced by various factors that are unrelated to the co-occurrence of CS and US, and thus cannot be taken as valid indicators for unqualified effects of stimulus co-occurrence

[28] As an illustration, consider the diagnosticity of being equal vs. unequal to something positive. Logically, being equal to something positive implies positivity, while being unequal to something positive implies negativity or neutrality. By implying a smaller range of possible values, being equal to something positive is therefore more diagnostic than being unequal to something positive (which implies a wider range of possible values), and may thus lead to stronger evaluative responses.

[29] As an example, consider the context-dependency of causing vs. preventing undesirable health states: while causing such health states should be seen as negative regardless of context, preventing these health states may be seen as positive in a context where their occurrence is probable (but as neutral in a context where these health states are improbable).

(on CS evaluation). Most importantly, evaluative asymmetries between the assimilative and contrastive relations of a given relation pair will lead to an overestimation of the C parameter (in comparison to actual co-occurrence effects). Likewise, memory asymmetries between assimilative and contrastive conditions can also produce inflated C parameter estimates. We recommend to carefully select relation pairs which are symmetrical in evaluative strength, and to carefully control for memory asymmetries when using the RCB model to estimate unqualified co-occurrence effects in evaluative learning.

Data availability statement

The data supporting our findings are publicly available on the Open Science Framework: https://osf.io/zfdtb/?view_only=e3d101ec5f474be8be07b7e349376e37.

References

Bading, K. (2021). *Does association formation contribute to evaluative conditioning? A review of current methodologies and their findings.* Retrieved from https://kups.ub.uni-koeln.de/54059/

Bading, K., Stahl, C., & Rothermund, K. (2020). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion*, *34*(1), 128–143. https://doi.org/10.1080/02699931.2019.1604322

Béna, J., Mauclet, A., & Corneille, O. (2022). Does co-occurrence information influence evaluations beyond relational meaning? An investigation using self-reported and mouse-tracking measures of attitudinal ambivalence. *Journal of Experimental Psychology: General*.

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3, 3), e28046. https://doi.org/10.5964/spb.v13i3.28046

Förderer, S., & Unkelbach, C. (2011). Hating the cute kitten or loving the aggressive pit-bull: EC effects depend on CS-US relations. *Cognition & Emotion*, *26*(3), 534–540. https://doi.org/10.1080/02699931.2011.588687

Gawronski, B. (2022). Attitudinal effects of stimulus co-occurrence and stimulus relations: Paradoxical effects of cognitive load. *Personality and Social Psychology Bulletin*, *48*(10), 1438–1450.

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin*, *13*(3, 3), 1–33. https://doi.org/10.5964/spb.v13i3.28024

Gawronski, B., Brannon, S. M., & Luke, D. M. (2023). Unintentional influences in intentional impression formation. In *The handbook of impression formation: A social psychological approach* (pp. 199–219). New York: Routledge.

Gawronski, B., Luke, D. M., & Ng, N. L. (2021). Is less of an unhealthy ingredient healthy or unhealthy? Effects of mere co-occurrence and quantitative relations on attribute judgments. *Journal of Experimental Social Psychology*, *97*.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). Lab. Js: A free, open, online study builder. *Behavior Research Methods*, 1–18.

Heycke, T., & Gawronski, B. (2020). Co-occurrence and relational information in evaluative learning: A multinomial modeling approach. *Journal of Experimental Psychology: General*, *149*(1), 104–124. https://doi.org/10.1037/xge0000620

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). The Guilford Press.

Högden, F., Hütter, M., & Unkelbach, C. (2018). Does evaluative conditioning depend on awareness? Evidence from a continuous flash suppression paradigm. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *44*(10), 1641–1657. https://doi.org/10.1037/xlm0000533

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *43*(1), 17–32.

Hughes, S., Ye, Y., & De Houwer, J. (2019). Evaluative conditioning effects are modulated by the nature of contextual pairings. *Cognition and Emotion*, *33*(5), 871–884.

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019). When people co-occur with good or bad events: Graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin*, *45*(2), 196–208. https://doi.org/10.1177/0146167218781340

Hütter, M. (2022). An integrative review of dual-and single-process accounts of evaluative conditioning. *Nature Reviews Psychology*, *1*(11), 640–653.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Klauer, K. C. (2010). Hierarchical Multinomial Processing Tree Models: A Latent-Trait Approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Kukken, N., Hütter, M., & Holland, R. W. (2020). Are there two independent evaluative conditioning effects in relational paradigms? Dissociating the effects of CS-US pairings and their meaning. *Cognition and Emotion*, *34*(1), 170–187. https://doi.org/10.1080/02699931.2019.1617112

Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies”(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6).

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743–752. https://doi.org/10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition*, *34*(5), 435–461. https://doi.org/10.1521/soco.2016.34.5.435

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass*, *2*(2), 1073–1092.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557–569. https://doi.org/10.1177/0146167211400423

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*(3), 523–539. https://doi.org/10.1093/biostatistics/kxm049

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575. https://doi.org/10.3758/s13428-012-0259-0

Stahl, C., & Aust, F. (2018). Evaluative Conditioning as Memory-Based Judgment. *Social Psychological Bulletin*, *13(3)*, e28589. https://doi.org/10.5964/spb.v13i3.28589

Stahl, C., & Bading, K. (2020). Evaluative conditioning of pattern-masked nonwords requires perceptual awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 822–850. https://doi.org/10.1037/xlm0000757

Stahl, C., Haaf, J., & Corneille, O. (2016). Subliminal evaluative conditioning? Above-chance CS identification may be necessary and insufficient for attitude learning. *Journal of Experimental Psychology: General*, *145*(9), 1107–1131. https://doi.org/10.1037/xge0000191

van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Modeling*, *17*(3), 443–463.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*, 3571–3594. Retrieved from http://www.jmlr.org/papers/v11/watanabe10a.html

Yonelinas, A. P., & Jacoby, L. L. (2012). The process-dissociation approach two decades later: Convergence, boundary conditions, and new directions. *Memory & Cognition*, *40*(5), 663–680. https://doi.org/10.3758/s13421-012-0205-5

Zanon, R., De Houwer, J., & Gast, A. (2012). Context effects in evaluative conditioning of implicit evaluations. *Learning and Motivation*, *43*(3), 155–165.

Appendix A

Pilot study 1 – Evaluative differences between assimilative vs. contrastive relations in antonymic relation pairs

Methods

All measures, manipulations and data exclusions are reported. Data and analysis script are publicly available on the Open Science Framework: https://osf.io/zfdtb/?view_only=e3d101ec5f474be8be07b7e349376e37.

Participants. We recruited 87 participants through a mailing list for psychology students at Friedrich Schiller University Jena. As compensation, participants received partial course credit. We had to exclude two participants whose data sets were incomplete due to an unknown technical error. The final sample consisted of 85 participants (68.24% female; $M_{age} = 27.64$, $SD_{age} = 12.94$).

Materials, measures and procedure. Based on previous relational EC studies and brainstorming, a set of 30 antonymic relation pairs was assembled. We then created two separate lists (A and B) each containing 15 relation pairs.

The experiment was programmed in *E-Prime 3.0* and was run online via *E-Prime Go*. Each participant was randomly assigned to list A vs. B and saw a total of 60 statements (15 relation pairs × 2 relation types × 2 sentence objects). Each statement consisted of a generic sentence subject ("X", hereafter referred to as target), a relation (e.g., "causes") as the sentence predicate and a generic sentence object (either "something positive" or "something negative"). For each statement, participants were asked to evaluate the target based solely on the information provided in a given statement (i.e., X's relation with the sentence object). To express their evaluation, participants were presented with a 21-point scale ranging from -10 (very negative) to +10 (very positive). The order in which statements were presented was randomized for each participant anew.

Data processing. We assumed that, if understood correctly, statements containing assimilative (contrastive) relations should produce target evaluations in line with (opposite to) the valence of the sentence object (positive vs. negative). We therefore excluded all trials in which target evaluations did not comply with the aforementioned logic (20.47 %, neutral evaluations of zero were not excluded).

We were primarily interested in whether statements containing contrastive relations produced weaker evaluations of the target than statements containing their assimilative counterparts. To assess whether this was case, we calculated a double difference score for each relation pair. In a first step, we subtracted the mean target evaluation based on the statement containing the assimilative relation and the negative sentence object from the mean target evaluation based on the statement containing the assimilative relation and the positive sentence object ($d_a = x_{positive|assimilative} - x_{negative|assimilative}$). In a next step, we subtracted the mean target evaluation based on the statement containing the contrastive relation and the positive sentence object from the mean target evaluation based on the statement containing the contrastive relation and the negative sentence object ($d_c = x_{negative|contrastive} - x_{positive|contrastive}$). Finally, the double difference score d_{a-c} was calculated by subtracting d_c from d_a.

Results and discussion

Target evaluations and (double) difference scores for all 30 relation pairs are reported in Table A1. As expected, the mean double difference d_{a-c} across relation pairs was significantly larger than zero, $M = 2.08$, 95% CI $[1.31, 2.84]$, $t(29) = 5.53$, $p < .001$. As displayed in Table A1, the mean double difference for individual relation pairs was significantly larger than zero in the majority of cases. However, there was also considerable variance in the size of the double difference across relation pairs. This "across-pair" variance in d_{a-c} has important implications (a) for the interpretation of the present results, and (b) for future studies working with antonymic relation pairs. With regard to result interpretation, the across-pair variance in d_{a-c} shows that the reported asymmetries between assimilative vs. contrastive relations cannot simply be driven by assimilative

Table A1

Pilot study: Target evaluations as a function of object valence and relation type for all thirty relation pairs. Test statistics refer to the double difference score d_{a-c} calculated for each relation pair.

		assimilative relation			contrastive relation						
Relation pair (assimilative - contrastive)	List	$\bar{x}_{positive}$	$\bar{x}_{negative}$	d_a	$\bar{x}_{positive}$	$\bar{x}_{negative}$	d_c	d_{a-c}	t	n	p
speed up - slow down	A	6.24	-6.10	12.34	-3.00	3.05	6.05	6.29	10.40	40.00	0.00
maximize - minimize	A	7.18	-7.67	14.85	-3.59	5.05	8.64	6.21	9.64	38.00	0.00
enlarge - make smaller	A	6.61	-6.37	12.98	-3.37	4.30	7.67	5.30	9.14	45.00	0.00
strengthen - weaken	B	6.34	-6.03	12.38	-3.47	3.69	7.16	5.22	8.77	31.00	0.00
give - take	B	4.29	-3.14	7.43	-1.43	1.14	2.57	4.86	1.68	6.00	0.14
promote - undermine	B	6.84	-7.08	13.92	-5.60	4.40	10.00	3.92	5.32	24.00	0.00
increase - decrease	A	5.85	-5.94	11.79	-3.85	4.32	8.17	3.62	6.65	46.00	0.00
endorse - reject	B	5.74	-6.29	12.03	-3.97	5.10	9.06	2.97	4.37	30.00	0.00
affirm - negate	B	5.78	-4.63	10.41	-3.74	3.74	7.48	2.93	3.96	26.00	0.00
praise - rebuke	B	5.84	-5.69	11.53	-4.69	4.19	8.88	2.66	3.53	31.00	0.00
love - hate	A	5.85	-5.05	10.90	-4.54	3.95	8.49	2.41	4.12	38.00	0.00
be similar - be dissimilar	A	4.74	-4.02	8.77	-3.05	3.67	6.72	2.05	3.03	42.00	0.00
deal with - ignore	A	4.36	-3.43	7.79	-3.29	2.57	5.86	1.93	1.73	13.00	0.11
habituate - dishabituate	A	5.79	-4.76	10.55	-3.57	5.07	8.64	1.90	3.32	41.00	0.00
allow - suppress	B	5.00	-5.23	10.23	-4.83	3.63	8.47	1.77	1.57	29.00	0.13
integrate - exclude	B	5.93	-4.59	10.52	-4.79	4.00	8.79	1.72	2.93	28.00	0.01
deem good - deem bad	A	4.57	-4.60	9.17	-3.55	3.91	7.47	1.70	3.32	46.00	0.00
turn to - turn away	B	5.63	-4.53	10.17	-3.70	4.77	8.47	1.70	2.23	29.00	0.03
create - destroy	A	7.62	-7.53	15.16	-7.47	6.04	13.51	1.64	2.72	44.00	0.01
cause - prevent	B	6.77	-7.00	13.77	-5.77	6.97	12.73	1.03	1.35	29.00	0.19
favor - disadvantage	B	4.23	-4.00	8.23	-4.19	3.04	7.23	1.00	1.77	25.00	0.09
start - stop	A	6.70	-6.11	12.81	-5.43	6.45	11.87	0.94	1.46	46.00	0.15
enable - throttle	B	6.42	-5.10	11.52	-5.42	5.32	10.74	0.77	1.40	30.00	0.17
turn on - turn off	A	5.35	-5.93	11.28	-4.88	5.67	10.56	0.72	1.56	42.00	0.13
institute - abolish	A	6.58	-6.29	12.88	-6.21	6.17	12.38	0.50	1.01	47.00	0.32
reward - punish	B	5.77	-6.03	11.80	-7.07	4.43	11.50	0.30	0.50	29.00	0.62
be the same - be the opposite	B	6.91	-6.59	13.50	-6.88	6.47	13.35	0.15	0.25	33.00	0.80
continue - terminate	A	5.59	-4.91	10.50	-5.24	5.91	11.15	-0.65	-0.97	45.00	0.34
acknowledge - deny	A	4.00	-2.00	6.00	-3.20	4.20	7.40	-1.40	-0.32	4.00	0.76
permit - forbid	B	4.89	-4.43	9.32	-6.18	5.00	11.18	-1.86	-2.30	27.00	0.03

Note. Original relation pairs in German (in order of appearance): beschleunigen - verlangsamen, maximieren - minimieren, vergrößern - verkleinern, stärken - schwächen, geben - nehmen, fördern - untergraben, erhöhen - vermindern, befürworten - ablehnen, bejahen - verneinen, loben - tadeln, lieben - hassen, ähnlich sein - unähnlich sein, sich mit etwas beschäftigen - ignorieren, angewöhnen - abgewöhnen, zulassen - unterdrücken, integrieren - ausschließen, gut finden - schlecht finden, zuwenden - abwenden, erschaffen - zerstören, verursachen - verhindern, bevorzugen - benachteiligen, starten - stoppen, ermöglichen - unterbinden, einschalten - ausschalten, einführen - abschaffen, belohnen - bestrafen, das Gleiche sein - das Gegenteil sein, fortsetzen - beenden, anerkennen - leugnen, erlauben - verbieten effects of the sentence object's valence (something positive vs. something negative). If such an effect is present in our data (which might be the case), it will add to the size of d_{a-c} by the same amount for all relation pairs. By implication, the reported variance in d_{a-c} across relation pairs can only be driven by across-pair variance in evaluative asymmetry between relation types (assimilative vs. contrastive). With regard to future studies (working with antonymic relation pairs), the size differences in d_{a-c} show that evaluative differences between relation types are not inevitable and may be avoided by careful selection of study materials.

Appendix B

Pilot study 2 - potion evaluations based on statements containing different relations and character traits

In the present pilot study, we sought to identify two relation pairs and character traits (positive and negative in equal numbers) that could be combined to create credible effects of potions (the CS category in the main study). Aside from creating credible potion effects, the selected materials were supposed to meet three requirements. When combined with the selected character traits (as USs), one relation pair was supposed to produce equally strong CS evaluations in the assimilative vs. contrastive conditions (and therefore act as a symmetrical relation pair; hereafter, requirement [1]), whereas the other relation pair was supposed to produce stronger CS evaluations in the assimilative than in the contrastive conditions (and therefore act as an asymmetrical relation pair; hereafter, requirement [2]). To avoid overall differences (in evaluative strength) between the two relation pairs (aside from the difference in symmetry), we sought to identify two relation pairs where differences in evaluative strength were restricted to one relation type condition (hereafter, requirement [3]).

Methods

All measures, manipulations and data exclusions are reported. Data and analysis script are publicly available on the Open Science Framework: https://osf.io/zfdtb/?view_only=e3d101ec5f474be8be07b7e349376e37.

Participants. We recruited 37 participants through a mailing list for psychology students at Friedrich Schiller University Jena. As compensation, participants received partial course credit. We had to exclude four participants who failed the seriousness check at the end of the experiment. Moreover, two participants were excluded because their data sets were incomplete due to an unknown technical error. The final sample consisted of 31 participants (87.09% female; $M_{age} = 21.19$, $SD_{age} = 5.19$).

Materials, measures and procedure. We assembled six antonymic relation pairs with conceptual fit to the CS category (potions) and the US category (character traits). Based on the previous pilot study, three of these relation pairs were expected to “act” symmetrically (turn on vs. turn off, induce vs. throttle, cause vs. prevent), whereas the other three relation pairs were expected to “act” somewhat asymmetrically (strengthen vs. weaken, increase vs. decrease, promote vs. hinder). Via brainstorming, we assembled two lists each of which contained five positive and five negative character traits (list A: indifference, envy, stubbornness, greed, cowardice, courage, diligence, patience, helpfulness, faithfulness; list B: laziness, ruthlessness, irascibility, impatience, egoism, warmth, tolerance, discipline, self-confidence, empathy).

The experiment was programmed in *E-Prime 3.0* and was run online via *E-Prime Go*. Participants were randomly assigned to list A vs. B. At the beginning of the experiment, participants were told that they would learn about 120 potions and their effects on various character traits. They were instructed to vividly imagine each potion’s effect and then to evaluate each potion based on its stated effect. Next, participant worked through a total of 120 statements (6 relation pairs × 2 relation types × 10 character traits from list A vs. list B). The statements were presented in 4 blocks of 30 statements, separated by self-paced breaks. For each statement, participants were asked to evaluate the potion based on its stated effect (i.e., the potion’s relation with the character trait). To express their evaluation, participants were presented with a 21-point scale ranging from -10 (very negative) to +10 (very positive). The order in which statements were presented was randomized for each participant anew.

Data processing. We assumed that, if understood correctly, statements containing assimilative (contrastive) relations should produce potion evaluations in line with (opposite to) the valence of the character trait. We therefore excluded all trials in which potion evaluations did not comply with the aforementioned logic (4.47 %, neutral evaluations of zero were not excluded). To compare the absolute strength of potion evaluations across conditions, negative evaluations (from the "assimilative" × "negative" and "contrastive" × "positive" conditions) were multiplied with −1. Next, we plotted these absolute values as a function of relation type (assimilative vs. contrastive) separately for each character trait and relation pair. Through visual inspection, we identified two relation pairs (turn on vs. turn off, strengthen vs. weaken) and 6 character traits (positive: patience, courage, self-discipline; negative: cowardice, indifference, greed) that seemed to comply with requirements (1), (2) and (3). To test whether this was indeed the case, we implemented a linear mixed-effects model using the "lmer" function of the lme4 packages (Bates, Maechler, Bolker, & Walker, 2014) in R. For each participant, we calculated the mean of absolute potion evaluation in all trait valence × relation type × relation pair conditions for which data was available for a given participant. In the linear mixed-effects model, these means (including missing values) were regressed on trait valence (positive vs. negative), relation type (assimilative vs. contrastive) and relation pair (turn on/turn off vs. strengthen/weaken). The model included fixed effects for the three factors as well as for all possible (two- and three-way) interaction terms. Effects coding was used for all factors (trait valence: positive = 1, negative = −1; relation type: assimilative = 1, contrastive = −1; relation pair: symmetrical = 1, asymmetrical = −1). To clarify significant interactions, we also implemented separate linear mixed-effects models for the two relation pairs and for the two relation types. As before, we used effects coding for all factors (see above) and included fixed effects for all involved factors and interaction terms (see below).

Results and discussion

Absolute target evaluations as a function of trait valence, relation type and relation pair are reported in Figure B1.

The linear mixed-effects model revealed a significant main effect of relation pair, $F(1, 143.29) = 10.05$, $p = .002$, and a significant two-way interaction between relation pair and relation type, $F(1, 143.73) = 9.50$, $p = .002$. All other main effects and interactions did

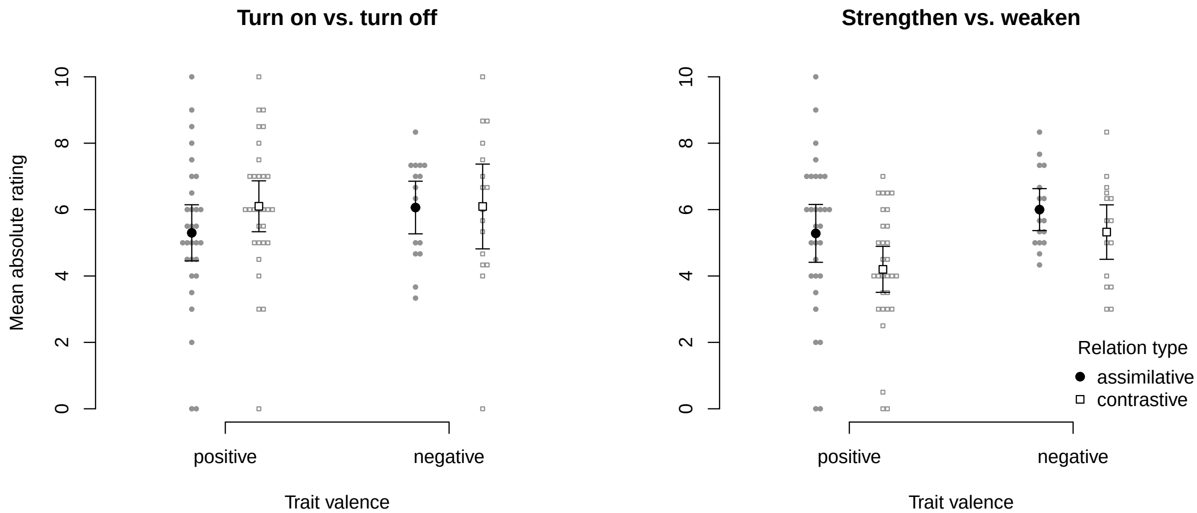

Figure B1. Absolute values of target evaluations as a function of trait valence, relation type and relation pair.

not reach significance, all ps $\geq .140$.

The main effect of relation pair reflected the fact that target evaluations were overall stronger when character traits were turned on or turned off ($M = 5.83$, $SD = 2.10$) in comparison to when character traits were strengthened or weakened ($M = 5.06$, $SD = 1.97$). To clarify the two-way interaction, we implemented separate linear mixed-effects models for the two relation pairs. In both of these follow-up models, the main effects of trait valence and relation type as well as the two-way interaction of trait valence and relation type were included as predictors.

The "strengthen vs. weaken" model revealed a significant main effect of relation type, $F(1, 53.90) = 10.38$, $p = .002$. In line with the idea of an asymmetrical relation pair, the main effect of relation type reflected overall stronger target evaluations when targets strengthened a character trait ($M = 5.53$, $SD = 2.02$) in comparison to when targets weakened a character trait ($M = 4.59$, $SD = 1.82$). The main effect of trait valence and the two-way interaction between trait valence and relation type did not reach significance, all ps $\geq .150$.

In the "turn on vs. turn off" model, the main effect of relation type did not reach significance, $F(1, 54.74) = 1.64$, $p = .205$. In line with the idea of a symmetrical relation pair, the absence of this main effect suggested that "turn on" and "turn off" sentences produced target evaluations that were comparably strong ($M_{turnon} = 5.57$, $SD_{turnon} = 2.04$; $M_{turnoff} = 6.10$, $SD_{turnoff} = 2.15$). As before, the main effect of trait valence and the two-way interaction between trait valence and relation type did not reach significance, all ps $\geq .239$.

We also implemented separate linear mixed-effects models for the two relation types (assimilative vs. contrastive). In both of these follow-up models, the main effects of trait valence and relation pair as well as the two-way interaction of trait valence and relation pair were included as predictors. Replicating the (overall) main effect of relation pair, the "contrastive" model revealed significantly stronger target evaluations when character traits ($M = 6.09$, $SD = 2.15$) were turned off in comparison to when character traits were weakened ($M = 4.59$, $SD = 1.82$), $F(1, 55.33) = 13.74$, $p < .001$. The main effect of trait valence and the two-way interaction between trait valence and relation pair did not reach significance, all ps $\geq .117$. In the "assimilative" model, the main effect of relation pair did not reach significance, $F(1, 54.44) = 0.00$, $p = .986$. The absence of this main effect suggested that "turn on" and "strengthen" sentences produced target evaluations that were comparably strong ($M_{turnon} = 5.57$, $SD_{turnon} = 2.04$; $M_{strengthen} = 5.53$, $SD_{strengthen} = 2.02$). As before, the main effect of trait valence and the two-way interaction between trait valence and relation pair did not reach significance, all ps $\geq .431$.

Taken together, the previous results complied with requirements (1), (2) and (3). In line with requirement (1), "turn on" vs. "turn off" seemed to constitute a symmetrical relation pair inducing comparably strong target evaluations (as indicated by the non-significant main effect of relation type in the "turn on vs. turn off" model). Furthermore, and in line with requirement (2), "strengthen" vs. "weaken" seemed to constitute an asymmetrical relation pair inducing stronger target evaluations in the assimilative than in the contrastive condition (as reflected by the main effect of relation type in the “strengthen vs. weaken” model). Finally, and in line with requirement (3), there was no overall difference in evaluative strength between the two relation pairs; instead, differences in evaluative strength were restricted to one relation type condition (as indicated by the significant [non-significant] main effect of relation pair in the “contrastive” [“assimilative”] conditions)[30].

[30] As previously mentioned, we excluded all trials in which the sign of the potion evaluation did not match the evaluative implication of the US valence × relation type combination assigned to the potion. To ensure that our conclusions were not distorted by these data exclusions, we repeated the previous analyses on the whole data set without exclusions (evaluations from the “assimilative” × “negative” and “contrastive” × “positive” conditions were again multiplied with -1). In these complementary analyses, we again found a significant two-way interaction between relation type and relation pair (which followed the exact same pattern as in the original analyses). Based on these findings, our conclusions regarding requirements (1), (2) and (3) can be seen as robust across data exclusion strategies.

Appendix C

Frequentist analyses on categorical and continuous CS evaluations and other measures

Experiment 1

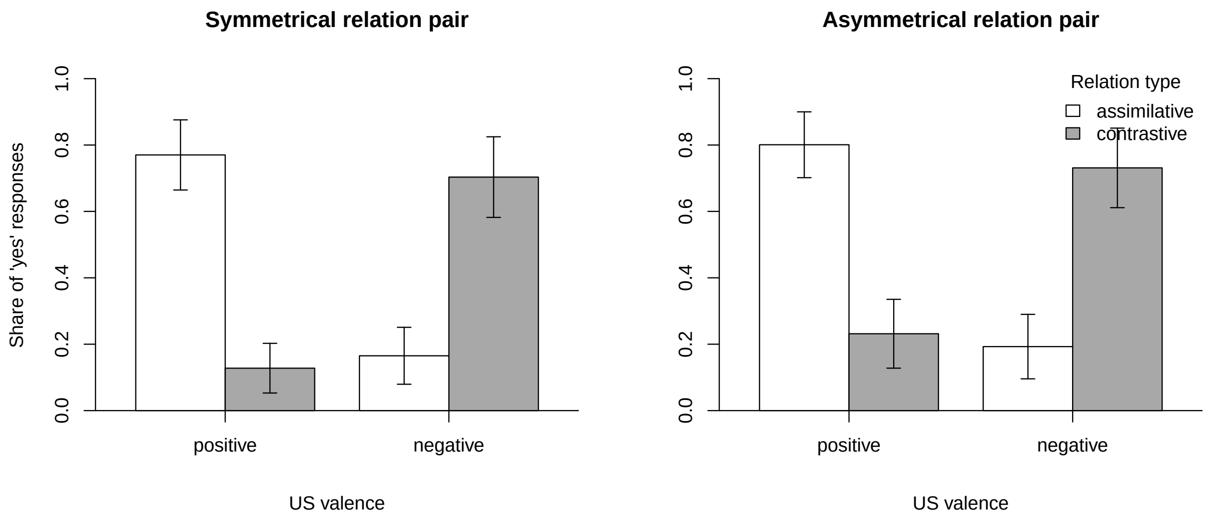

Figure C1. Experiment 1: Shares of 'yes' responses in the speeded classification task as a function of US valence, relation type and relation pair. Error bars represent 95% within-subjects confidence intervals.

Categorical CS evaluations (SCT). Figure C1 shows shares of "yes" responses as a function of US valence, relation type and relation pair.

We analyzed the shares of "yes" responses using a 2 (*US valence*: positive vs. negative) × 2 (*relation type*: assimilative vs. contrastive) × 2 (*relation pair*: symmetrical vs. asymmetrical) repeated-measures ANOVA. We found a significant two-way interaction between *US valence* and *relation type*, $F(1, 41) = 112.25$, $p < .001$, $\hat{\eta}^2_G = .438$, 90% CI $[.249, .582]$, indicating that CS classifications were determined by the integration of US valence and CS-US relation, all other ps $\geq .182$. In Appendix A, we already introduced the double-difference score d_{a-c} to quantify how much stronger evaluations from assimilative relations are compared with evaluations from contrastive relations. We also calculated this double difference for both relation pairs. In line with expectations, the double difference was descriptively larger for the asymmetrical relation pair. However, there was no significant difference between the two double differences, $d_{a-c;asymmetrical} - d_{a-c;symmetrical} = 0.08$, 95% CI $[-0.13, \infty]$, $t(41) = 0.62$, $p = .268$. Moreover, neither of the two double differences differed from zero (symmetrical pair: $d_{a-c} = 0.03$, 95% CI $[-0.12, \infty]$, $t(41) = 0.32$, $p = .374$; asymmetrical pair: $d_{a-c} = 0.11$, 95% CI $[-0.05, \infty]$, $t(41) = 1.17$, $p = .124$).

CS ratings (CET). Figure C2 shows evaluative ratings as a function of US valence, relation type and relation pair.

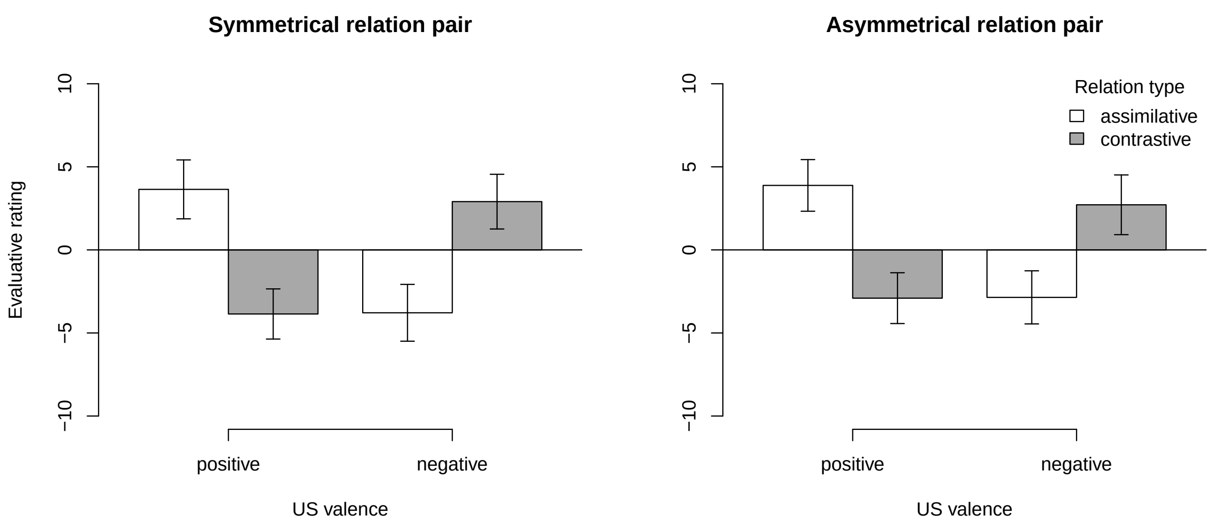

Figure C2. Experiment 1: Evaluative ratings as a function of US valence, relation type and relation pair. Error bars represent 95% within-subjects confidence intervals.

We analyzed evaluative ratings using a 2 (*US valence*: positive vs. negative) $\times$ 2 (*relation type*: assimilative vs. contrastive) $\times$ 2 (*relation pair*: symmetrical vs. asymmetrical) repeated-measures ANOVA. Similar to the previous analysis on data from the speeded classification task, we found a significant two-way interaction between *US valence* and *relation type*, $F(1, 41) = 41.66$, $p < .001$, $\hat{\eta}^2_G = .302$, 90% CI $[.121, .467]$, indicating that ratings were determined by the integration of US valence and CS-US relation, all other ps $\geq .243$. We also calculated double differences d_{a-c} for both relation pairs. In line with expectations, the double difference was descriptively larger for the asymmetrical relation pair. However, there was no significant difference between the two double differences, $d_{a-c;asymmetrical} - d_{a-c;symmetrical} = 0.45$, 95% CI $[-2.94, \infty]$, $t(41) = 0.22$, $p = .412$. Moreover, neither of the two double differences differed from zero (symmetrical pair: $d_{a-c} = 0.67$, 95% CI $[-1.48, \infty]$, $t(41) = 0.52$, $p = .302$; asymmetrical pair: $d_{a-c} = 1.12$, 95% CI $[-1.21, \infty]$, $t(41) = 0.81$, $p = .212$).

Experiment 2

Categorical CS evaluations (SCT). We excluded two participants who gave the same response on all trials of the task. We therefore analyzed 227 participants (all of which had passed the two seriousness checks).

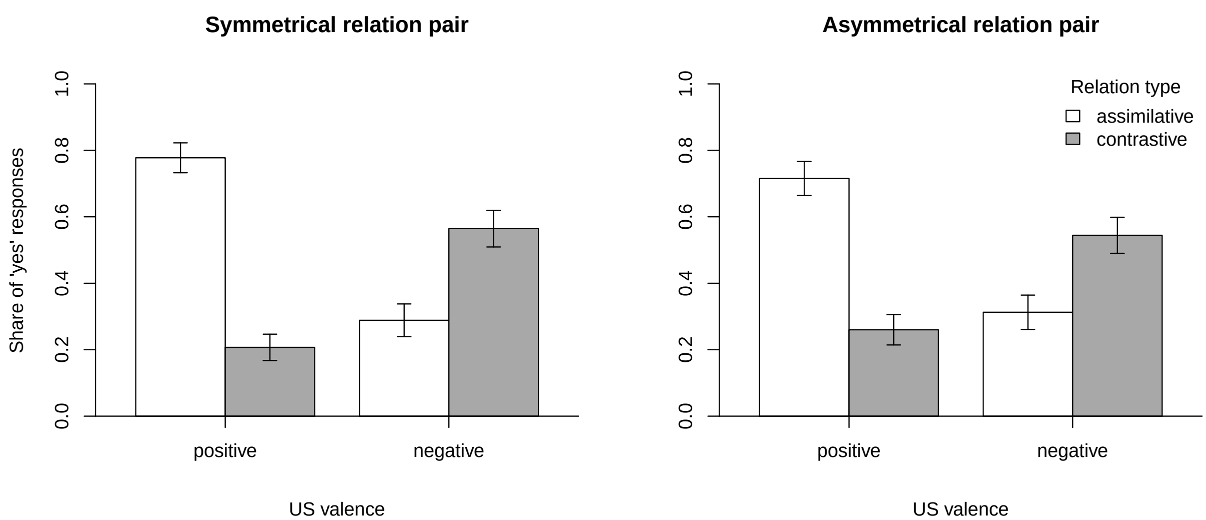

Figure C3. Experiment 2, whole sample (without memory exclusions): Shares of 'yes' responses in the speeded classification task as a function of US valence, relation type and relation pair. Error bars represent 95% within-subjects confidence intervals.

Whole sample (without memory exclusions). Figure C3 shows shares of "yes" responses as a function of US valence, relation type and relation pair.

We analyzed the shares of "yes" responses using a hierarchical linear model that included a random intercept and the three factors (US valence, relation type and relation pair) as effect-coded fixed effects. We found significant main effects of US valence (positive > negative), $F(1, 1808) = 12.62$, $p < .001$, and of relation type (assimilative > contrastive), $F(1, 1808) = 54.39$, $p < .001$. The two-way interaction between US valence and relation type was also significant, $F(1, 1808) = 475.83$, $p < .001$. This interaction reflected a regular EC effect for "assimilative" CSs and a reversed EC effect for "contrastive" CSs. Moreover, we found a significant three-way interaction between US valence, relation type and relation pair, $F(1, 1808) = 5.13$, $p = .024$. All other main effects and interactions did not reach significance, all ps $\geq .313$.

To disentangle the three-way interaction, we calculated separate hierarchical linear models for the two relation pair conditions. Both models included random intercepts and the two factors (US valence, relation type) as effect-coded fixed effects. For the symmetrical condition, we found significant main effects of US valence, $F(1, 904) = 7.43$, $p = .007$, and of relation type, $F(1, 904) = 37.18$, $p < .001$, and a significant two-way interaction, $F(1, 904) = 306.65$, $p < .001$. For the asymmetrical condition, the two main effects and the two-way interaction were significant as well (US valence: $F(1, 904) = 5.34$, $p = .021$; relation type: $F(1, 904) = 19.21$, $p < .001$; US valence $\times$ relation type: $F(1, 904) = 181.16$, $p < .001$). The difference in effect structure between the two relation pair conditions consisted in the fact the all three effects were stronger in the symmetrical than in the asymmetrical condition.

In Appendix A, we already introduced the double-difference score d_{a-c} to quantify how much stronger evaluations from assimilative relations are compared with evaluations from contrastive relations. We also calculated this double difference for both relation pair conditions. Both double differences were significantly larger than zero (symmetrical: $d_{a-c} = 0.13$, 95% CI $[0.05, \infty]$, $t(1582) = 2.65$, $p = .004$; asymmetrical: $d_{a-c} = 0.12$, 95% CI $[0.04, \infty]$, $t(1582) = 2.37$, $p = .009$), but did not differ from each other,

$d_{a-c;asymmetrical} - d_{a-c;symmetrical} = -0.01$, 95% CI $[-0.13, \infty]$, $t(1582) = -0.20$, $p = .578$.

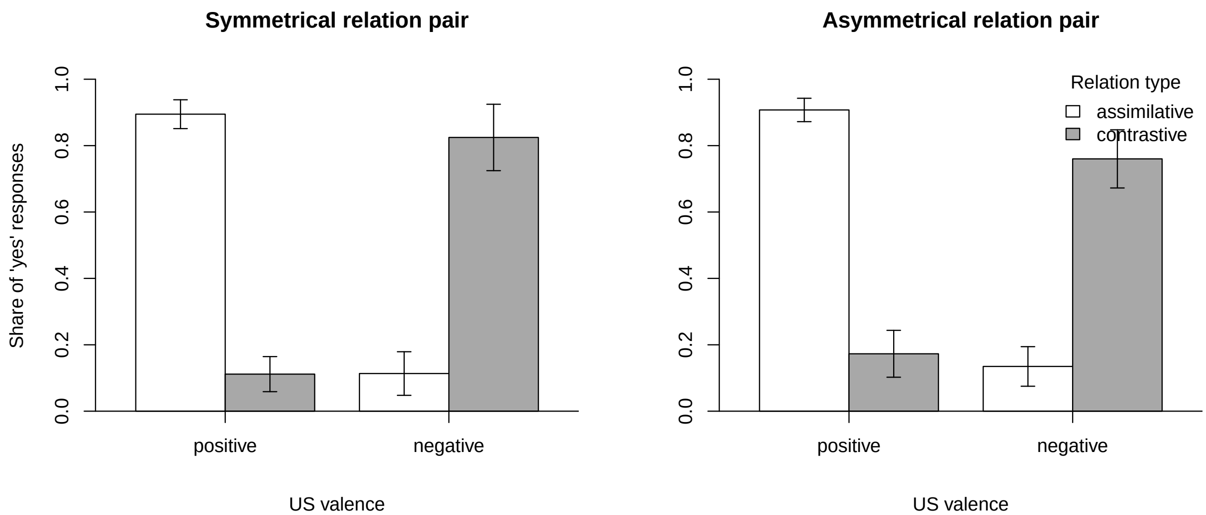

Figure C4. Experiment 2, sub-sample (with memory exclusions): Shares of 'yes' responses in the speeded classification task as a function of US valence, relation type and relation pair. Error bars represent 95% confidence intervals.

Sub-sample (with memory exclusions). To control for memory differences between US valence × relation type × relation pair condition, the following analyses included only CSs with correct recollection of both US valence and CS-US relation (30.5% of all SCT trials). Figure C4 shows shares of "yes" responses as a function of US valence, relation type and relation pair.

As before, we analyzed the shares of "yes" responses using a hierarchical linear model that included a random intercept and the three factors (US valence, relation type an relation pair) as effect-coded fixed effects. We found significant main effects of US valence (positive > negative), $F(1, 526.14) = 8.89$, $p = .003$, and of relation type (assimilative > contrastive), $F(1, 519.91) = 4.11$, $p = .043$. The two-way interaction between US valence and relation type was also significant, $F(1, 489.46) = 1,080.10$, $p < .001$. This interaction reflected a regular EC effect for "assimilative" CSs and a reversed EC effect for "contrastive" CSs. All other main effects and interactions did not reach significance, all

ps $\geq .127$.

We again calculated double differences for both relation pair conditions. In the asymmetrical condition, the double difference was significantly larger than zero, $d_{a-c} = 0.18$, 95% CI $[0.09, \infty]$, $t(508.24) = 3.17$, $p < .001$. In the symmetrical condition, the double difference was also larger than zero, but did not differ from it significantly, $d_{a-c} = 0.08$, 95% CI $[-0.03, \infty]$, $t(513.48) = 1.17$, $p = .121$ (one-tailed test). As expected, the double difference was larger in the asymmetrical than in the symmetrical condition. However, the difference between the two conditions failed to reach significance on a one-tailed test, $d_{a-c;asymmetrical} - d_{a-c;symmetrical} = 0.11$, 95% CI $[-0.03, \infty]$, $t(490.64) = 1.25$, $p = .106$.

CS ratings (CET). We excluded five participants who gave the same response on all trials of the task. We therefore analyzed 224 participants (all of which had passed the two seriousness checks).

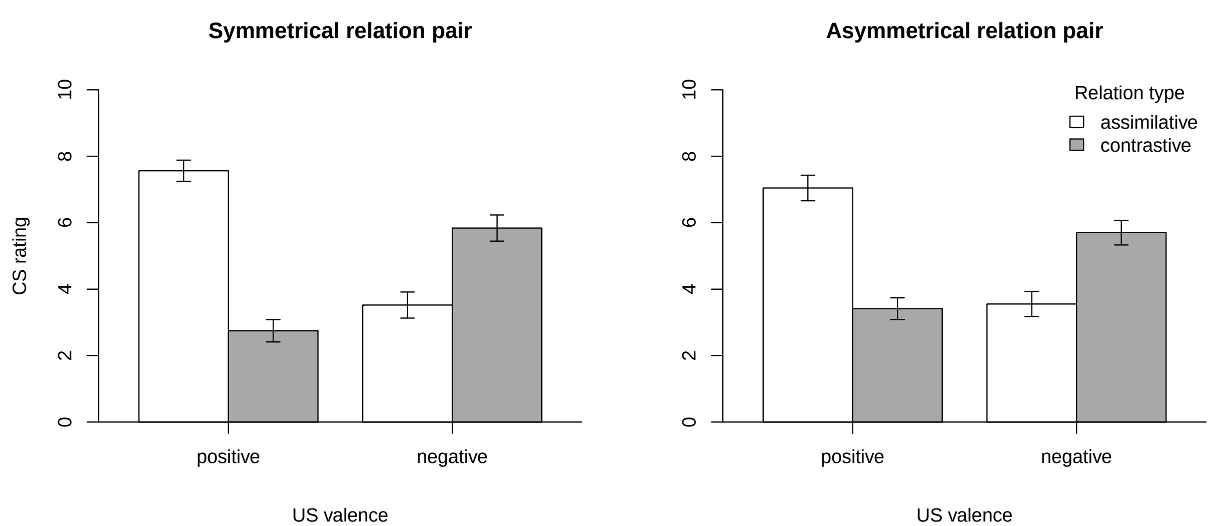

Figure C5. Experiment 2, whole sample (without memory exclusions): Evaluative ratings as a function of US valence, relation type and relation pair. Error bars represent 95% within-subjects confidence intervals.

Whole sample (without memory exclusions). Figure C5 shows evaluative ratings as a function of US valence, relation type and relation pair.

CS ratings were analyzed with a hierarchical linear model that included a random intercept and the three factors (US valence, relation type an relation pair) as effect-coded fixed effects. We found significant main effects of US valence (positive > negative), $F(1, 1784) = 18.12$, $p < .001$, and of relation type (assimilative > contrastive), $F(1, 1784) = 62.44$, $p < .001$. The two-way interaction between US valence and relation type was also significant, $F(1, 1784) = 655.36$, $p < .001$. This interaction reflected a regular EC effect for “assimilative” CSs and a reversed EC effect for “contrastive” CSs. Moreover, we also found a significant two-way interaction between relation type and relation pair, $F(1, 1784) = 4.03$, $p = .045$. This interaction reflected a stronger simple main effect of relation type (assimilative > contrastive) in the symmetrical condition than in the asymmetrical condition. Finally, the three-way interaction between US valence, relation type and relation pair was also significant, $F(1, 1784) = 7.19$, $p = .007$. All other main effects and interactions did not reach significance, all ps $\geq .614$.

To disentangle the three-way interaction, we calculated separate hierarchical linear models for the two relation pair conditions. Both models included random intercepts and the two factors (US valence, relation type) as effect-coded fixed effects. For the symmetrical condition, we found significant main effects of US valence, $F(1, 892) = 7.00$, $p = .008$, and of relation type, $F(1, 892) = 48.86$, $p < .001$, and a significant two-way interaction, $F(1, 892) = 397.86$, $p < .001$. For the asymmetrical condition, the two main effects and the two-way interaction were significant as well (US valence: $F(1, 892) = 11.39$, $p < .001$; relation type: $F(1, 892) = 17.46$, $p < .001$; US valence × relation type: $F(1, 892) = 264.00$, $p < .001$). The difference in effect structure between the two relation pair conditions consisted in the fact the two-way interaction between US valence and relation type was stronger in the symmetrical than in the asymmetrical condition.

Parallel hypothesis tests. We also performed parallel hypothesis tests on CET data. Based on the hierarchical linear model, we calculated linear contrasts directly testing the three hypotheses (in one-tailed tests). Degrees of freedom were derived via the Kenward-Roger method. In two contrasts (mirroring hypotheses 1 and 3), we compared the strength of the regular EC effect in the assimilative condition to the strength of the reversed EC in the contrastive condition, separately for both relation pair conditions[31]. In a third contrast (mirroring hypothesis 2), we compared the degree of asymmetry in the strength of relational EC effects (assimilative vs. contrastive) between relation pair conditions[32].]

For both relational pair conditions, the regular EC effect in the assimilative condition was stronger than the reversed EC effect in the contrastive condition (symmetrical: $d_{a-c} = 0.95$, 95% CI $[0.36, \infty]$, $t(1561) = 2.65$, $p = .004$; asymmetrical: $d_{a-c} = 1.20$, 95% CI $[0.61, \infty]$, $t(1561) = 3.37$, $p < .001$). The asymmetry in the strength of relational EC effects (assimilative > contrastive) did not differ between relation pair conditions, $d_{a-c;asymmetrical} - d_{a-c;symmetrical} = 0.25$, 95% CI $[-0.58, \infty]$, $t(1561) = 0.50$, $p = .307$.

Sub-sample (with memory exclusions). Again controlling for memory differences between experimental conditions, the following analyses included only CSs with correct recollection of both US valence and CS-US relation (30.5% of all CET trials). Figure C6 shows CS ratings as a function of US valence, relation type and relation pair.

As before, CS ratings were analyzed with a hierarchical linear model that included a

---

[31] These contrasts were based on the following double differences: $d_{a-c;symmetrical} = (PA_{symmetrical} - NA_{symmetrical}) - (NC_{symmetrical} - PC_{symmetrical})$ and $d_{a-c;asymmetrical} = (PA_{asymmetrical} - NA_{asymmetrical}) - (NC_{asymmetrical} - PC_{asymmetrical})$. In these formulas, pairs of capital letters represent mean CS rating in the four US valence (positive [P] vs. negative [N]) × relation type (assimilative [A] vs. contrastive [C]) conditions.

[32] This contrast was based on the difference between the two double differences from the two relation pair conditions: $d_{a-c;asymmetrical} - d_{a-c;symmetrical} = (PA_{asymmetrical} - NA_{asymmetrical}) - (NC_{asymmetrical} - PC_{asymmetrical}) - ((PA_{symmetrical} - NA_{symmetrical}) - (NC_{symmetrical} - PC_{symmetrical}))$.

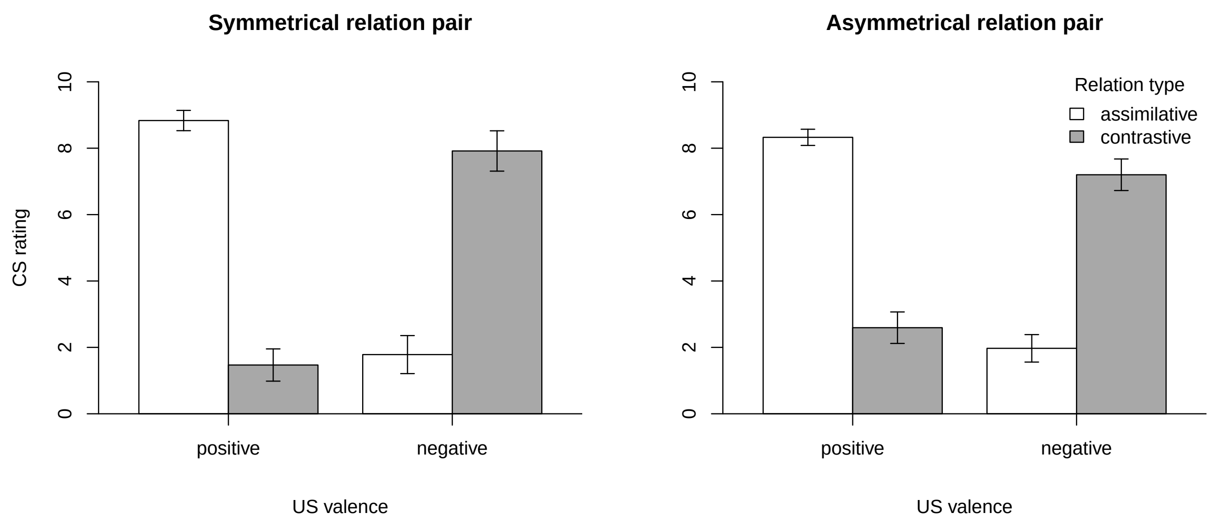

Figure C6. Experiment 2, sub-sample (with memory exclusions): Evaluative ratings as a function of US valence, relation type and relation pair. Error bars represent confidence intervals.

random intercept and the three factors (US valence, relation type and relation pair) as effect-coded fixed effects. We found significant main effects of US valence (positive > negative), $F(1, 1784) = 18.12$, $p < .001$, and of relation type (assimilative > contrastive), $F(1, 1784) = 62.44$, $p < .001$. The two-way interaction between US valence and relation type was also significant, $F(1, 1784) = 655.36$, $p < .001$. This interaction reflected a regular EC effect for "assimilative" CSs and a reversed EC effect for "contrastive" CSs. Moreover, we also found a significant two-way interaction between relation type and relation pair, $F(1, 1784) = 4.03$, $p = .045$. This interaction reflected a stronger simple main effect of relation type (assimilative > contrastive) in the symmetrical condition than in the asymmetrical condition. Finally, the three-way interaction between US valence, relation type and relation pair was also significant, $F(1, 1784) = 7.19$, $p = .007$. All other main effects and interactions did not reach significance, all ps $\geq .614$.

To disentangle the three-way interaction, we calculated separate hierarchical linear models for the two relation pair conditions. Both models included random intercepts and the two factors (US valence, relation type) as effect-coded fixed effects. For the symmetrical condition, we found a significant main effect of relation type, $F(1, 246) = 6.32$, $p = .013$, and a significant two-way interaction between US valence and relation type, $F(1, 246) = 759.20$, $p < .001$. The main effect of US valence was non-significant, $F(1, 246) = 1.51$, $p = .220$. For the asymmetrical condition, we found a significant main effect of US valence, $F(1, 292) = 20.23$, $p < .001$, and a significant two-way interaction between US valence and relation type, $F(1, 292) = 794.70$, $p < .001$. The main effect of relation type was non-significant, $F(1, 292) = 1.69$, $p = .195$.

Parallel hypothesis tests. We used the same approach as for the whole sample (see above). In the asymmetrical condition, the regular EC effect in the assimilative condition was stronger than the reversed EC effect in the contrastive condition, $d_{a-c} = 1.75$, 95% CI $[1.07, \infty]$, $t(510.85) = 4.23$, $p < .001$. In the symmetrical condition, the regular EC effect in the assimilative condition was descriptively but non-significantly larger than the reversed EC effect in the contrastive condition, $d_{a-c} = 0.60$, 95% CI $[-0.16, \infty]$, $t(518.17) = 1.31$, $p = .096$. Finally, and in line with expectations, the asymmetry in the strength of relational EC effects (assimilative > contrastive) was significantly larger in the asymmetrical than in the symmetrical condition, $d_{a-c;asymmetrical} - d_{a-c;symmetrical} = 1.15$, 95% CI $[0.13, \infty]$, $t(496.96) = 1.85$, $p = .032$ (in a one-tailed test).

US ratings. There were zero participants who gave the same response on all trials of the US rating task. The following analyses thus included all 229 participants who passed the two seriousness checks. Figure C7 shows US ratings as a function of trait.

In a first step, the US ratings were submitted to a within-subjects ANOVA with trait valence (positive vs. negative) as the only factor. The main effect of trait valence was significant, $F(1, 228) = 4,172.61$, $p < .001$, $\hat{\eta}^2_G = .916$, 90% CI $[.901, .927]$, and reflected on average more positive ratings for positive traits (self-discipline, patience and courage) than for negative traits (greed, indifference and cowardice).

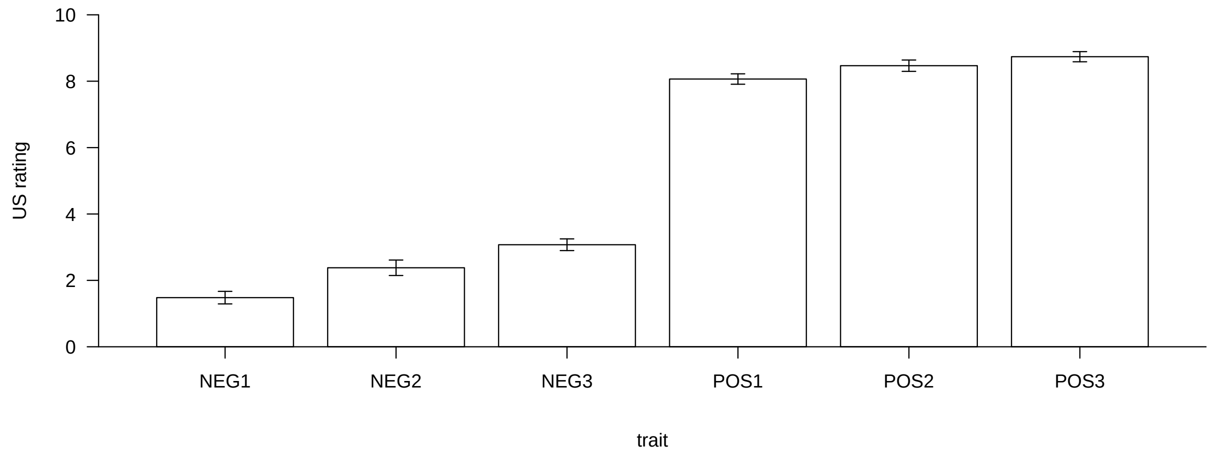

Figure C7. US ratings as a function of character trait (NEG1 = greed, NEG2 = indifference, NEG3 = cowardice, POS1 = courage, POS2 = self-discipline, POS3 = patience).

In a second step, we tested for differences between mean US ratings within the set of negative traits. US ratings for cowardice were significantly higher than US ratings for greed, $M_D = 1.59$, 95% CI $[1.34, 1.84]$, $t(228) = 12.56$, $p < .001$, and US ratings for indifference, $M_D = 0.69$, 95% CI $[0.42, 0.97]$, $t(228) = 4.97$, $p < .001$. Moreover, US ratings for indifference were significantly higher than US ratings for greed, $M_D = 0.90$, 95% CI $[0.62, 1.18]$, $t(228) = 6.40$, $p < .001$.

In a third step, we also tested for differences between mean US ratings within the set of positive traits. US ratings for patience were significantly higher than US ratings for courage, $M_D = 0.67$, 95% CI $[0.48, 0.87]$, $t(228) = 6.77$, $p < .001$, and US ratings for self-discipline, $M_D = 0.27$, 95% CI $[0.08, 0.46]$, $t(228) = 2.75$, $p = .006$. Moreover, US ratings for self-discipline were significantly higher than US ratings for courage, $M_D = 0.40$, 95% CI $[0.20, 0.60]$, $t(228) = 3.99$, $p < .001$.

Finally, we also tested whether US ratings for positive vs. negative values differed in absolute value. To this aim, we calculated the absolute difference between each US rating and 5 (the mid-point of the rating scale). A paired t-test on these (absolute) differences revealed that US ratings for positive traits were stronger (in terms of absolute value) than US ratings for negative traits, $M_D = 0.63$, 95% CI $[0.49, 0.76]$, $t(228) = 9.11$, $p < .001$.

Memory for image-trait relationships. There were zero participants who gave the same response on all trials of the task. We therefore analyzed all 229 participants who passed the two seriousness checks. Due to the extremely high share of correct responses (see below), we refrained from performing any statistical tests.

224 participants had perfect memory performance; i.e., they selected the correct trait on all 12 trials of the task (2 trials for each of the 6 image sets). One participant gave two incorrect responses (both for the images representing patience). Another four participant each gave one incorrect response (two for the images representing patience, one for the images representing cowardice, one for the images representing indifference).

Appendix D

Additional MPT analyses on the aggregated response frequencies

Experiment 1

Model selection. The RCB model estimating separate sets of parameters for CSs presented with symmetrical vs. asymmetrical relations did not fit the data, $G^2(2) = 7.54$, $p = .023$. We ran several less restrictive model extensions (each estimating separate sets of parameters for CSs presented with symmetrical vs. asymmetrical relations). In addition to $C_{symmetrical}$, $C_{asymmetrical}$, $B_{symmetrical}$ and $B_{asymmetrical}$, the RCB4a model estimates two R parameters in each relation pair condition: one for CSs paired with positive USs and one for CSs paired with negative USs. Similarly, the RCB4b model estimates separate C parameters for positively vs. negatively paired CSs in each relation pair condition (in addition to $R_{symmetrical}$, $R_{asymmetrical}$, $B_{symmetrical}$ and $B_{asymmetrical}$). Finally, the RCB4c model also estimates four parameters per relation pair condition: parameters R and C as well as two B parameters (again, one for positively and one for negatively paired CSs). As reported in the main text, the RCB4a model fit the data well, $G^2 = 0$. Both the RCB4b and the RCB4c model attained G^2 values greater than zero. We used a parametric double bootstrap (c.f., van de Schoot, Hoijtink, & Deković, 2010) with 1,000 first-level samples to estimate the empirical distribution of G^2 for these two models, and 1,000,000 second-level samples to estimate the empirical distribution of p values to determine adjusted α levels α^*. We found that the misfit was substantial for both the RCB4b and the RCB4c model ($G^2 = 7.54$, $p = .029$, $\alpha^* = .057$ and $G^2 = 7.54$, $p = .025$, $\alpha^* = .056$, respectively). We also compared the models using the AIC: Model RCB4a attained the lowest AIC score and outperformed all other models ($\text{AIC}_{RCB4a} = 16.00$, $\text{AIC}_{RCB3} = 19.54$, $\text{AIC}_{RCB4b} = \text{AIC}_{RCB4c} = 23.54$).

Hypothesis tests based on the RCB model. Parameter estimates and 95% confidence intervals are reported in Table D1. The C parameter in the asymmetrical condition was significantly larger than zero, $\Delta G^2(1) = 6.87$, $p = .009$. Moreover, the C parameter was noticeably larger in the asymmetrical than in the symmetrical condition. However, the decrease in model fit produced by a restricted RCB model with $C_{asymmetrical} = C_{symmetrical}$ failed to reach significance, $\Delta G^2(1) = 2.40$, $p = .121$. Finally, the C parameter in the symmetrical condition was close to zero and did not differ from it significantly, $\Delta G^2(1) = 0.15$, $p = .694$.

Table D1

Experiment 1: Parameter estimates (with 95% confidence intervals) based on an MPT analysis of aggregated response frequencies with the unrestricted RCB model.

	Asymmetrical		Symmetrical	
Parameter	$\hat{\theta}$	95% CI	$\hat{\theta}$	95% CI
R	.557	[.515, .599]	.592	[.551, .632]
C	.127	[.032, .221]	.020	[-.078, .117]
B	.475	[.421, .529]	.355	[.305, .405]

Experiment 2

Model selection.

Whole sample. The RCB model estimating separate sets of parameters for CSs presented with symmetrical vs. asymmetrical relations did not fit the data, $G^2(2) = 338.30$, $p < .001$. As pre-registered, we ran the less restrictive RCB4a, RCB4b and RCB4c models (see above). As reported in the main text, the RCB4a model fit the data well, $G^2 = 0$. Both the RCB4b and the RCB4c model attained G^2 values greater than zero. We again used a parametric double bootstrap with 1,000 first-level samples to estimate the empirical distribution of G^2 for these two models, and 1,000,000 second-level samples to estimate the empirical distribution of p values to determine adjusted α levels α^*. We found that the misfit was substantial for both the RCB4b and the RCB4c model ($G^2 = 338.30$, $p < .001$, $\alpha^* = .057$ and $G^2 = 338.30$, $p < .001$, $\alpha^* = .049$, respectively). We also compared the models using the AIC: Model RCB4a attained the lowest AIC score and outperformed all other models ($\text{AIC}_{RCB4a} = 16.00$, $\text{AIC}_{RCB3} = 350.30$, $\text{AIC}_{RCB4b} = \text{AIC}_{RCB4c} = 354.30$).

Sub-sample. The RCB model estimating separate sets of parameters for CSs presented with symmetrical vs. asymmetrical relations did not fit the data, $G^2(2) = 21.30$, $p < .001$. As pre-registered, we ran the less restrictive RCB4a, RCB4b and RCB4c models (see above). As reported in the main text, the RCB4a model fit the data well, $G^2 = 0$. Both the RCB4b and the RCB4c model attained G^2 values greater than zero. We again used a parametric double bootstrap with 1,000 first-level samples to estimate the empirical distribution of G^2 for these two models, and 1,000,000 second-level samples to estimate the empirical distribution of p values to determine adjusted α levels α^*. We found that the misfit was substantial for both the RCB4b and the RCB4c model ($G^2 = 21.30$, $p < .001$, $\alpha^* = .045$ and $G^2 = 21.30$, $p < .001$, $\alpha^* = .043$, respectively). We also compared the models using the AIC: Model RCB4a attained the lowest AIC score and outperformed all other models ($\text{AIC}_{RCB4a} = 16.00$, $\text{AIC}_{RCB3} = 33.30$, $\text{AIC}_{RCB4b} = \text{AIC}_{RCB4c} = 37.30$).

Hypothesis tests based on the RCB model.

Whole sample. Parameter estimates and 95% confidence intervals are reported in Table D2. The C parameter in the asymmetrical condition was significantly larger than zero, $\Delta G^2(1) = 24.71$, $p < .001$, as was the C parameter in the symmetrical condition, $\Delta G^2(1) = 30.10$, $p < .001$. Contrary to expectations, $C_{symmetrical}$ was descriptively larger than $C_{asymmetrical}$. However, the difference between the two C parameters was non-significant, $\Delta G^2(1) = 0.53$, $p = .468$.

Table D2

Experiment 2 (whole sample): Parameter estimates (with 95% confidence intervals) based on an MPT analysis of aggregated response frequencies with the unrestricted RCB model.

	Asymmetrical		Symmetrical	
Parameter	$\hat{\theta}$	95% CI	$\hat{\theta}$	95% CI
R	.345	[.325, .365]	.423	[.403, .442]
C	.079	[.048, .110]	.096	[.062, .130]
B	.434	[.418, .451]	.430	[.412, .449]

Sub-sample. Parameter estimates and 95% confidence intervals are reported in Table D3. The C parameter in the asymmetrical condition was significantly larger than zero, $\Delta G^2(1) = 46.49$, $p < .001$, as was the C parameter in the symmetrical condition, $\Delta G^2(1) = 4.50$, $p = .034$. In line with hypothesis 1, $C_{asymmetrical}$ was significantly larger than $C_{symmetrical}$, $\Delta G^2(1) = 6.80$, $p = .009$.

Table D3

Experiment 2 (sub-sample): Parameter estimates (with 95% confidence intervals) based on an MPT analysis of aggregated response frequencies with the unrestricted RCB model.

	Asymmetrical		Symmetrical	
Parameter	$\hat{\theta}$	95% CI	$\hat{\theta}$	95% CI
R	.686	[.657, .714]	.758	[.731, .786]
C	.305	[.223, .388]	.122	[.011, .233]
B	.525	[.466, .584]	.464	[.401, .527]

Appendix E

Additional MPT analyses on the individual response frequencies

Experiment 1

The following MPT analyses were carried out using the *TreeBUGS* package (Heck, Arnold, & Arnold, 2018). Individual response frequencies from the SCT were analyzed with the RCB model (see below) using the latent-trait approach which allows for heterogeneity in parameter values between participants (Klauer, 2010). Parameters were estimated by running four Markov chains with 200,000 iterations each (100,000 were discarded as burn-in iterations). We used 20,000 adaptation iterations and a thinning rate of 40. Convergence was monitored by means of the Gelman-Rubin statistic (Gelman & Rubin, 1992) using a criterion of $\hat{R} < 1.02$ for all parameters. Model fit was assessed by means of posterior predictive model checks T_1 and T_2 as proposed by (Klauer, 2010).

Model selection. As a first modelling step, we fitted the three-parameter RCB model (RCB3) with separate sets of parameters for CSs presented with relations from the symmetrical vs. asymmetrical relation pairs ($R_{symmetrical}$, $C_{symmetrical}$, $B_{symmetrical}$ and $R_{asymmetrical}$, $C_{asymmetrical}$, $B_{asymmetrical}$, respectively), and assessed the model's adequacy by calculating posterior-predictive model checks as proposed by Klauer (2010). Model checks showed substantial deviations between the model's predictions and the data, $T_1^{\text{observed}} = 0.44$, $T_1^{\text{expected}} = 0.12$, $p = .004$, $T_2^{\text{observed}} = 130.26$, $T_2^{\text{expected}} = 14.53$, $p < .001$.

It might be possible that a small sub-set of participants caused such misfit. We therefore calculated fit statistic T_1 for each participant at the individual level: using a criterion of $p \leq .05$, the RCB3 model fit the data of only 17 of 42 participants. We therefore concluded that misfit of the RCB3 model is arguably not caused by a small sub-set of participants.

In a second step, we fitted the previously mentioned four-parameter variants of the RCB model (RCB4a, RCB4b and RCB4c model) and compared their ability to predict the

Table E1

Experiment 1: Absolute fit and WAIC for the hierarchical extensions of the unrestricted RCB, RCB4a, RCB4b and RCB4c models.

	RCB3	RCB4a	RCB4b	RCB4c
Goodness of fit: Means				
T_1^{observed}	0.44	0.31	0.49	0.35
T_1^{expected}	0.12	0.10	0.12	0.12
p	.004	.022	.001	.009
Goodness of fit: Covariances				
T_2^{observed}	130.26	33.48	130.26	88.02
T_2^{expected}	14.53	10.81	14.59	14.23
p	$< .001$	$< .001$	$< .001$	$< .001$
Relative predictive accuracy				
WAIC	1,486.29	809.79	1,476.16	1,428.08
SE	89.40	49.09	88.19	82.65

data. We used the widely applicable information criterion (WAIC, Watanabe, 2010), a model selection criterion similar to AIC and DIC (Plummer, 2008) that is asymptotically equal to leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2017). For each of these models, we again calculated model fit measures T_1 and T_2 to assess absolute fit between each model's predictions and the observed data.

Table E1 shows the results for all models. With respect to WAIC, model RCB4a clearly outperforms all other models (it attains the lowest WAIC score). Regarding fit statistics T_1 and T_2, none of the considered models seemed to provide a satisfactory account of the data. However, it might still be possible that a sub-set of participants caused misfit, even for a model that is adequate for the vast majority of participants.

Table E2

Experiment 1: Parameter estimates (with 95% credible intervals) based on a hierarchical extension of the unrestricted RCB4a model.

	Asymmetrical		Symmetrical	
Parameter	*M*	95% CI	*M*	95% CI
R_{positive}	.713	[.437, .916]	.825	[.550, .974]
R_{negative}	.721	[.326, .963]	.574	[.214, .876]
C	.174	[.009, .539]	.043	[.000, .234]
B	.321	[.034, .768]	.271	[.057, .641]

Individual-level fit statistics indicated that models RCB3, RCB4a, RCB4b and RCB4c accounted well for the data of 25, 38, 24, 28 participants, respectively. Taken together, we concluded that the RCB4a model represents the best account of the data, and should therefore be used as a starting point for conducting hypothesis tests.

We also re-estimated model RCB4a with only the data of the 38 participants that were well-accounted for by the model. For this final model, both T_1 and T_2 indicated satisfactory model fit, $T_1^{\text{observed}} = 0.17$, $T_1^{\text{expected}} = 0.10$, $p = .196$, $T_2^{\text{observed}} = 17.27$, $T_2^{\text{expected}} = 10.84$, $p = .103$. Parameter estimates and 95% credible intervals are reported in Table E2.

Hypothesis tests based on the RCB4a model. Hypotheses 1 and 3 ($C_{asymmetrical} > 0$ and $C_{symmetrical} = 0$, respectively) were tested via formal model comparisons using the WAIC. For each hypothesis, we fitted an additional RCB4a model where the respective C parameter was set to zero. We then compared model adequacy (WAIC) of the restricted model with that of the unrestricted RCB4a model. Based on convention, a WAIC difference > 10 counts as strong evidence in favor of the model with the smaller WAIC. To test hypothesis 2 ($C_{asymmetrical} > C_{symmetrical}$), the posterior difference $C_{asymmetrical} - C_{symmetrical}$ was calculated. In case of a positive difference and a posterior distribution excluding zero, it can be concluded that the C parameter is larger for CSs presented with asymmetrical relations than for CSs presented with symmetrical relations.

In line with our first prediction ($\mathcal{H}_1 : C_{asymmetrical} > 0$), the C parameter in the asymmetrical condition was noticeably larger than zero. Moreover, a restricted RCB4a model with $C_{asymmetrical}$ set to zero led to a substantial increase in the WAIC, $\Delta\text{WAIC} = 66.45, SE = 19.62$, indicating that such a restriction cannot be made without significant loss in model adequacy. We therefore concluded that $C_{asymmetrical}$ is larger than zero.

In line with our third prediction ($\mathcal{H}_3 : C_{symmetrical} = 0$), the C parameter in the symmetrical condition was close to zero. However, the hypothesis test based on formal model comparison was inconclusive: a restricted RCB4a model with $C_{symmetrical}$ set to zero produced hardly any change in the WAIC in comparison to the unrestricted RCB4a model, $\Delta\text{WAIC} = 0.91, SE = 3.38$. We could therefore not conclude that $C_{symmetrical}$ is equal to zero.

In line with our second prediction, $C_{asymmetrical}$ was noticeably larger than $C_{symmetrical}$, resulting in a posterior difference of $C_a - C_s = 0.13$. However, the posterior distribution did not exclude zero, 95% CI $[-0.09, 0.49]$, Bayesian $p = .126$. We could therefore not conclude that the C parameter is larger in the asymmetrical than in the symmetrical condition.

MPT parameters and evaluative ratings. To validate the RCB4a model as an alternative to the original RCB model, we tested whether MPT parameters based on the RCB4a model predicted continuous CS evaluations in a comparable manner to MPT

Table E3

Experiment 1: Evaluative ratings as a function of US valence, relation type and relation pair.

		Asymmetrical		Symmetrical	
US valence	Relation type	*M*	*SD*	*M*	*SD*
positive	assimilative	3.88	4.75	3.64	4.75
	contrastive	-2.90	4.76	-3.86	4.76
negative	assimilative	-2.86	4.93	-3.79	4.93
	contrastive	2.71	5.61	2.90	5.61

parameters based on the original RCB model (Kukken et al., 2020). We also wanted to see if MPT parameters in the symmetrical vs. asymmetrical condition predicted continuous CS evaluations in a similar way. To this aim, we implemented a linear mixed model approach using the "lmer" function of the lme4 packages (Bates, Maechler, Bolker, & Walker, 2014) in R. The model included a fixed intercept, fixed effects for the R- and C-predictors, and fixed effects for the factors US valence, relation type and relation pair. In addition, fixed effects for all interaction terms between US valence, relation type, relation pair and the R parameter, and US valence, relation type, relation pair and the C parameter were included. The random-effects part of the model included parameter R. MPT parameter estimates were centered and effects coding was used for all factors (US valence: positive = 1, negative = -1; relation type: assimilative = 1, contrastive = -1; relation pair: symmetrical = 1, asymmetrical = -1). Descriptive statistics of the dependent variable (evaluative ratings) can be found in Table E3.

Most importantly, we found a significant two-way interaction between US valence and the C parameter that did not enter into any three- or four-way interactions involving relation type and/or relation pair (see Table E4). Replicating Kukken et al. (2020), the C

Table E4

Experiment 1: Modeling results on the relationship between MPT parameters and continuous CS evaluations.

Term	$\hat{\beta}$	95% CI	t	df	p
Intercept	-0.12	[-0.76, 0.52]	-0.36	218.42	.717
US valence	0.19	[-0.43, 0.81]	0.60	276.31	.550
Relation type	0.04	[-0.55, 0.64]	0.14	249.92	.887
Relation pair	-0.45	[-1.07, 0.16]	-1.45	279.08	.147
R	0.40	[-1.42, 2.22]	0.43	46.42	.668
C	-0.80	[-4.19, 2.59]	-0.46	256.67	.643
US valence × Relation type	3.61	[3.01, 4.20]	11.85	249.92	$< .001$
US valence × Relation pair	0.44	[-0.20, 1.07]	1.35	258.71	.178
Relation type × Relation pair	-0.29	[-0.89, 0.30]	-0.96	249.92	.339
US valence × R	-1.53	[-3.04, -0.02]	-1.99	277.27	.048
US valence × C	4.46	[1.11, 7.81]	2.61	275.96	.010
Relation type × R	0.32	[-1.15, 1.80]	0.43	249.92	.668
Relation type × C	-0.75	[-3.95, 2.46]	-0.46	249.92	.648
Relation pair × R	0.76	[-0.75, 2.26]	0.99	274.24	.324
Relation pair × C	-0.81	[-4.07, 2.45]	-0.49	272.77	.628
US valence × Relation type × Relation pair	0.14	[-0.45, 0.74]	0.47	249.92	.637
US valence × Relation type × R	4.35	[2.87, 5.82]	5.78	249.92	$< .001$
US valence × Relation type × C	0.38	[-2.82, 3.59]	0.23	249.92	.815
US valence × Relation pair × R	0.55	[-0.98, 2.08]	0.71	280.00	.480
US valence × Relation pair × C	0.03	[-3.35, 3.41]	0.02	267.06	.986
Relation type × Relation pair × R	0.61	[-0.86, 2.09]	0.81	249.92	.417
Relation type × Relation pair × C	-0.84	[-4.04, 2.37]	-0.51	249.92	.609
US valence × Relation type × Relation pair × R	0.88	[-0.59, 2.36]	1.17	249.92	.241
US valence × Relation type × Relation pair × C	0.05	[-3.15, 3.25]	0.03	249.92	.976

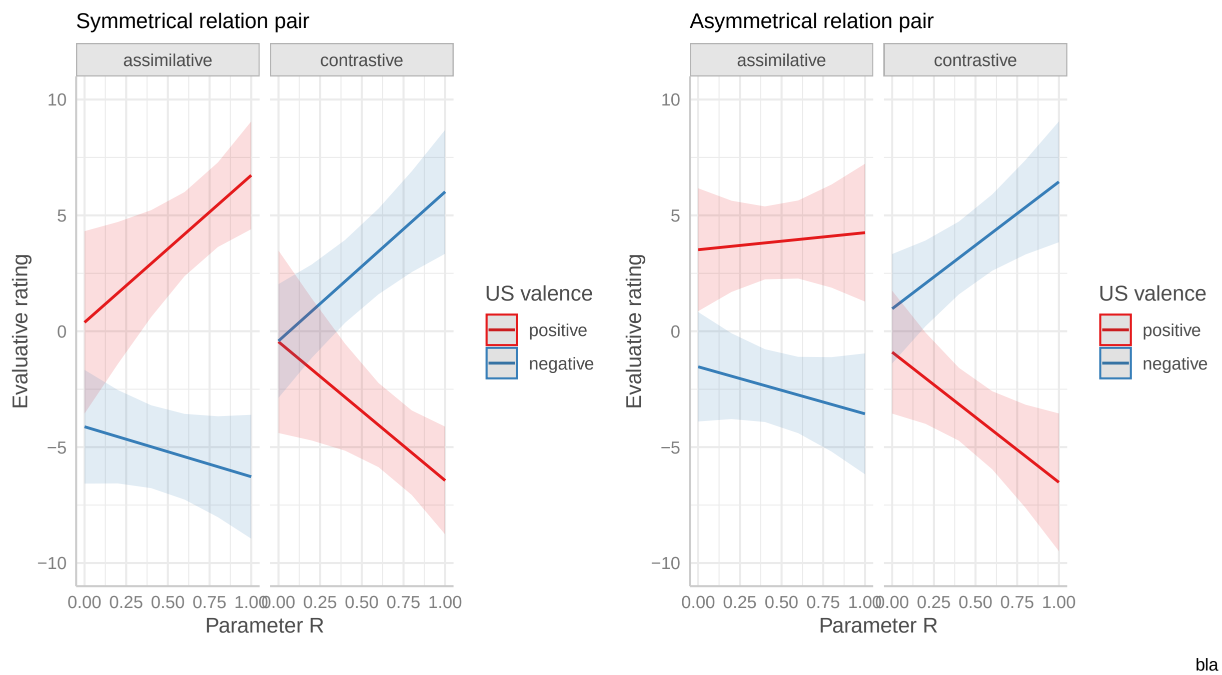

Figure E1. Experiment 1: Relationship between the R parameter and continuous CS evaluations.

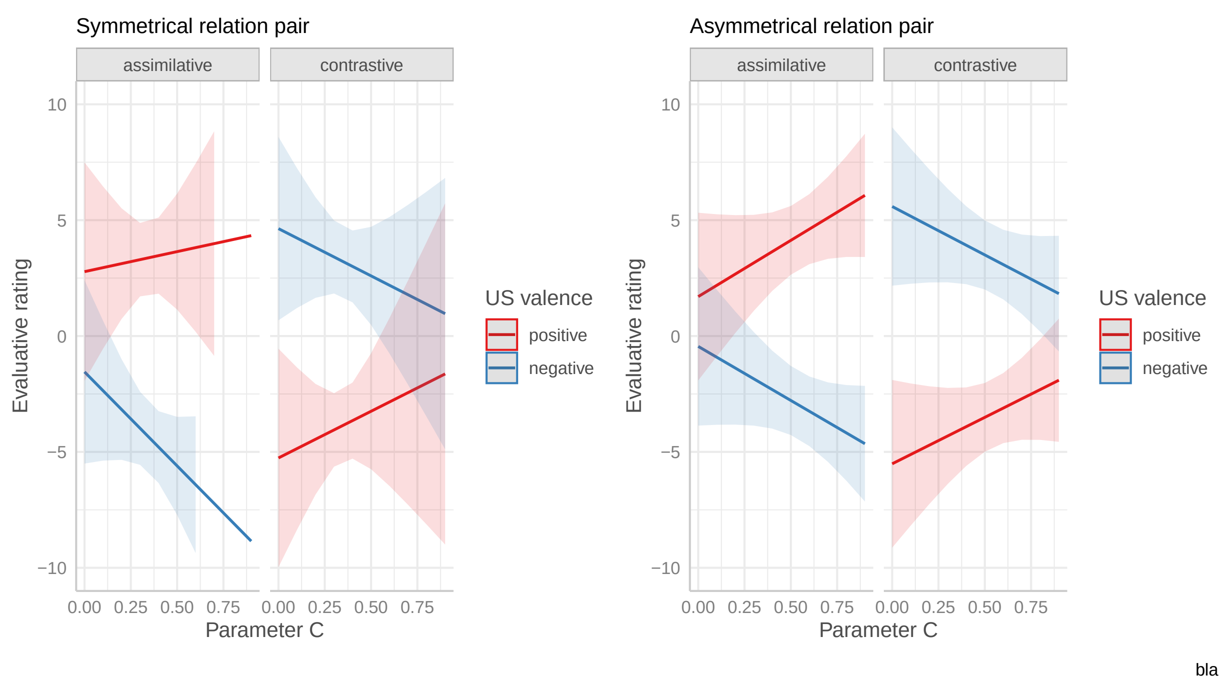

Figure E2. Experiment 1: Relationship between the C parameter and continuous CS evaluations.

parameter was related to assimilative effects of US valence: while the C parameter predicted positive CS evaluations for positively paired CSs, it predicted negative CS evaluations for negatively paired CSs (see Figure E2). As indicated by the lack of significant three- and four-way interactions, this pattern was comparable across levels of relation type and relation pair. We also found a significant three-way interaction between US valence, relation type and the R parameter. Again replicating Kukken et al. (2020), the R parameter was related to assimilative effects of US valence for CSs presented with assimilative relations, but related to contrastive effects of US valence for CSs presented with contrastive relations (see Figure E1). As indicated by a non-significant four-way interaction between US valence, relation type, relation pair and the R parameter, this pattern was comparable for CSs presented with symmetrical vs. asymmetrical relations.

Experiment 2

The following analyses were conducted for the whole sample of SCT data, but not for the sub-sample of SCT data from CSs with correct memory for the CS-US proposition (due to empty cells for individual participants). We used the same approach, settings and criteria as in Experiment 1 (see above).

Model selection. As a first modelling step, we fitted the three-parameter RCB model (RCB3) with separate sets of parameters for CSs presented with relations from the symmetrical vs. asymmetrical relation pairs, and assessed the model's adequacy by calculating posterior-predictive model checks as proposed by Klauer (2010). Model checks showed substantial deviations between the model's predictions and the data, $T_1^{\text{observed}} = 1.87$, $T_1^{\text{expected}} = 0.02$, $p < .001$, $T_2^{\text{observed}} = 196.78$, $T_2^{\text{expected}} = 2.50$, $p < .001$.

In a second step, we fitted the previously mentioned four-parameter variants (RCB4a, RCB4b and RCB4c model) and compared their ability to predict the data. We again used the widely applicable information criterion (WAIC, Watanabe, 2010). For each of these

Table E5

Experiment 2 (whole sample): Absolute fit and WAIC for the hierarchical extensions of the unrestricted RCB, RCB4a, RCB4b and RCB4c models.

	RCB3	RCB4a	RCB4b	RCB4c
Goodness of fit: Means				
T_1^{observed}	1.87	1.08	1.87	1.55
T_1^{expected}	0.02	0.02	0.02	0.02
p	$< .001$	$< .001$	$< .001$	$< .001$
Goodness of fit: Covariances				
T_2^{observed}	196.78	110.67	196.77	152.50
T_2^{expected}	2.50	2.04	2.51	2.46
p	$< .001$	$< .001$	$< .001$	$< .001$
Relative predictive accuracy				
WAIC	10,251.00	6,353.52	10,100.36	9,827.14
SE	233.08	182.11	227.66	218.28

models, we again calculated model fit measures T_1 and T_2 to assess absolute fit between each model's predictions and the observed data.

Table E5 shows the results for all models. With respect to WAIC, model RCB4a clearly outperforms all other models (it attains the lowest WAIC score). Regarding fit statistics T_1 and T_2, none of the considered models seemed to provide a satisfactory account of the data. However, it might still be possible that a sub-set of participants caused misfit, even for a model that is adequate for the vast majority of participants. Individual-level fit statistics indicated that models RCB3, RCB4a, RCB4b and RCB4c accounted well for the data of 113, 163, 109, 126 participants, respectively. As in Experiment 1, we concluded that the RCB4a model represents the best account of the

Table E6

Experiment 2 (whole sample): Parameter estimates (with 95% credible intervals) based on a hierarchical extension of the unrestricted RCB4a model.

	Asymmetrical		Symmetrical	
Parameter	*M*	95% CI	*M*	95% CI
R_{positive}	.503	[.322, .680]	.715	[.594, .824]
R_{negative}	.091	[.016, .223]	.133	[.029, .298]
C	.049	[.005, .143]	.031	[.002, .102]
B	.491	[.373, .614]	.515	[.379, .657]

data, and should therefore be used as the baseline model for conducting hypothesis tests.

Hypothesis tests based on the RCB4 model. Parameter estimates and 95% credible intervals from the unrestricted RCB4a model are reported in Table E6.

Hypotheses 1 and 3 ($C_{asymmetrical} > 0$ and $C_{symmetrical} = 0$, respectively) were tested via formal model comparisons using the WAIC. For each hypothesis, we fitted an additional RCB4a model where the respective C parameter was set to zero. We then compared model adequacy (WAIC) of the restricted model with that of the unrestricted RCB4a model. Based on convention, a WAIC difference > 10 counts as strong evidence in favor of the model with the smaller WAIC. To test hypothesis 2 ($C_{asymmetrical} > C_{symmetrical}$), the posterior difference $C_{asymmetrical} - C_{symmetrical}$ was calculated. In case of a positive difference and a posterior distribution excluding zero, it can be concluded that the C parameter is larger for CSs presented with asymmetrical relations than for CSs presented with symmetrical relations.

In line with our first prediction ($\mathcal{H}_1 : C_{asymmetrical} > 0$), the C parameter in the asymmetrical condition was larger than zero. Moreover, a restricted RCB4a model with $C_{asymmetrical}$ set to zero led to a substantial increase in the WAIC, ΔWAIC = 498.05, $SE = 70.13$, indicating that such a restriction cannot be made without significant loss in model adequacy. We therefore concluded that $C_{asymmetrical}$ is larger than zero.

Contrary to our third prediction ($\mathcal{H}_3 : C_{symmetrical} = 0$), the C parameter in the symmetrical condition was larger than zero. Moreover, a restricted RCB4a model with $C_{symmetrical}$ set to zero led to a substantial increase in the WAIC, ΔWAIC = 397.86, $SE = 63.38$, indicating that such a restriction cannot be made without significant loss in model adequacy. We therefore concluded that $C_{symmetrical}$ is larger than zero.

In line with our second prediction, $C_{asymmetrical}$ was larger than $C_{symmetrical}$, resulting in a posterior difference of $C_a - C_s = 0.02$. Moreover, a restricted RCB4a model with $C_{asymmetrical}$ set equal to $C_{symmetrical}$ led to a substantial increase in the WAIC, ΔWAIC = 313.42, $SE = 54.96$, indicating that such a restriction cannot be made without significant loss in model adequacy. We therefore concluded that $C_{asymmetrical}$ is larger than $C_{symmetrical}$.

MPT parameters and evaluative ratings. We implemented a linear mixed model approach using the “lmer” function of the lme4 packages (Bates, Maechler, Bolker, & Walker, 2014) in R. The model included a fixed intercept, fixed effects for the R- and C-predictors, and fixed effects for the factors US valence, relation type and relation pair. In addition, fixed effects for all interaction terms between US valence, relation type, relation pair and the R parameter, and US valence, relation type, relation pair and the C parameter were included. The random-effects part of the model included parameter R. MPT parameter estimates were centered and effects coding was used for all factors (US valence: positive = 1, negative = -1; relation type: assimilative = 1, contrastive = -1; relation pair: symmetrical = 1, asymmetrical = -1). Descriptive statistics of the dependent variable (evaluative ratings) can be found in Table E7. The modeling results are reported in Table E8.

Table E7

Experiment 2 (whole sample): Evaluative ratings as a function of US valence, relation type and relation pair.

		Asymmetrical		Symmetrical	
US valence	Relation type	*M*	*SD*	*M*	*SD*
positive	assimilative	7.04	2.71	7.56	2.71
	contrastive	3.41	2.40	2.75	2.40
negative	assimilative	3.55	2.80	3.52	2.80
	contrastive	5.70	2.72	5.84	2.72

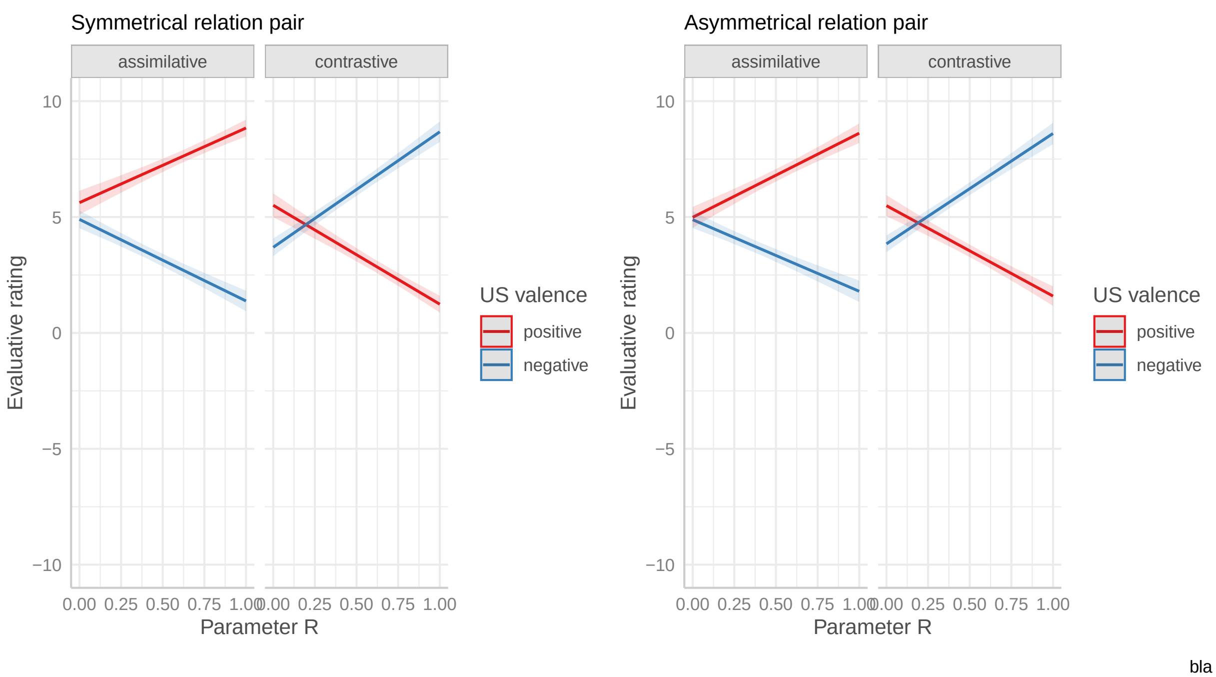

Figure E3. Experiment 2 (whole sample): Relationship between the *R* parameter and continuous CS evaluations.

Table E8

Experiment 2 (whole sample): Modeling results on the relationship between MPT parameters and continuous CS evaluations.

Term	$\hat{\beta}$	95% CI	t	df	p
Intercept	4.98	[4.88, 5.08]	101.31	693.26	$< .001$
Us valence	0.26	[0.16, 0.35]	5.28	1718.72	$< .001$
Relation type	0.15	[0.05, 0.24]	3.07	1537.39	.002
Relation pair	0.00	[-0.09, 0.10]	0.06	1765.19	.956
R	0.23	[-0.01, 0.47]	1.86	239.36	.064
C	-0.06	[-0.42, 0.31]	-0.30	1579.10	.762
Us valence × Relation type	1.63	[1.54, 1.73]	33.69	1537.39	$< .001$
Us valence × Relation pair	0.06	[-0.03, 0.16]	1.29	1767.97	.196
Relation type × Relation pair	0.05	[-0.04, 0.15]	1.11	1537.39	.267
Us valence × R	-0.56	[-0.79, -0.32]	-4.66	1710.51	$< .001$
Us valence × C	2.37	[2.01, 2.74]	12.74	1712.50	$< .001$
Relation type × R	-0.17	[-0.40, 0.07]	-1.40	1537.39	.161
Relation type × C	-0.23	[-0.60, 0.13]	-1.26	1537.39	.207
Relation pair × R	-0.12	[-0.36, 0.11]	-1.03	1657.11	.302
Relation pair × C	0.18	[-0.18, 0.55]	0.99	1689.26	.322
Us valence × Relation type × Relation pair	0.10	[0.00, 0.19]	1.96	1537.39	.050
Us valence × Relation type × R	3.92	[3.69, 4.16]	32.88	1537.39	$< .001$
Us valence × Relation type × C	0.52	[0.16, 0.89]	2.83	1537.39	.005
Us valence × Relation pair × R	-0.07	[-0.31, 0.16]	-0.59	1685.65	.552
Us valence × Relation pair × C	0.16	[-0.20, 0.53]	0.87	1575.64	.386
Relation type × Relation pair × R	-0.09	[-0.32, 0.15]	-0.73	1537.39	.468
Relation type × Relation pair × C	-0.23	[-0.59, 0.13]	-1.23	1537.39	.218
Us valence × Relation type × Relation pair × R	0.08	[-0.16, 0.31]	0.64	1537.39	.522
Us valence × Relation type × Relation pair × C	0.09	[-0.27, 0.46]	0.51	1537.39	.610

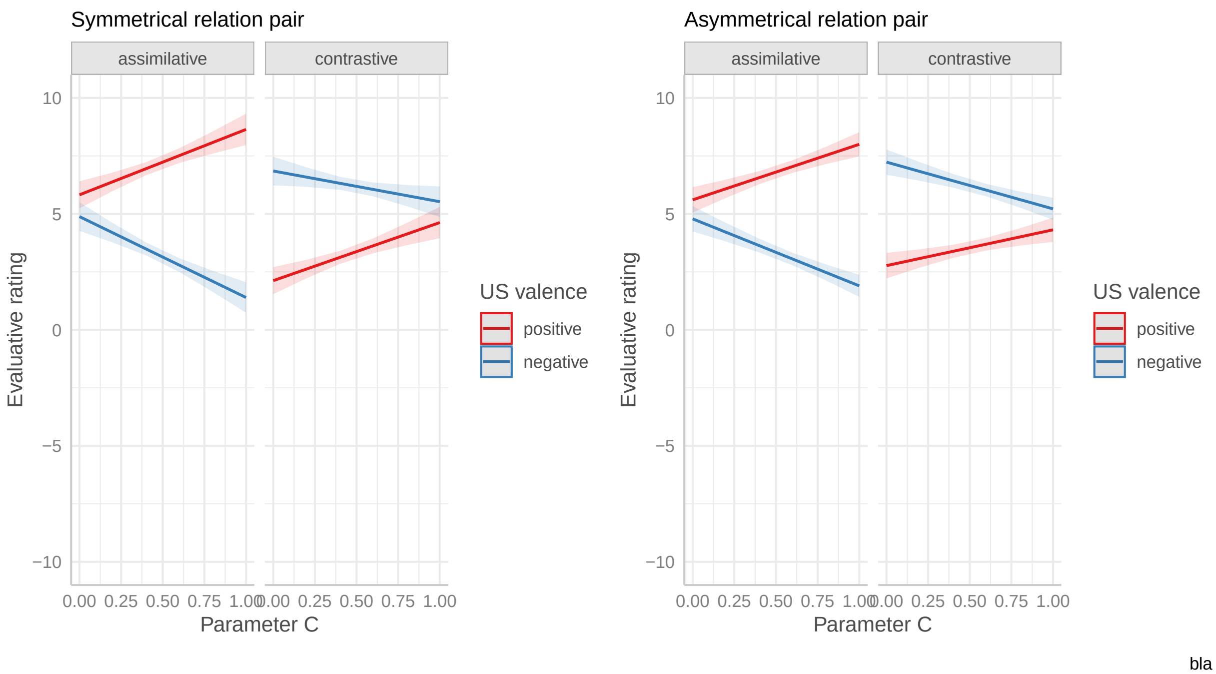

Figure E4. Experiment 2 (whole sample): Relationship between the C parameter and continuous CS evaluations.

As in Experiment 1, we found a significant three-way interaction between US valence, relation type and the R parameter. Again replicating Kukken et al. (2020), the R parameter was related to assimilative effects of US valence for CSs presented with assimilative relations, but related to contrastive effects of US valence for CSs presented with contrastive relations (see Figure E3). As indicated by a non-significant four-way interaction between US valence, relation type, relation pair and the R parameter, this pattern was comparable for CSs presented with symmetrical vs. asymmetrical relations.

In line with results from Experiment 1, we also found a significant two-way interaction between US valence and the C parameter (see Table E8). Again replicating Kukken et al. (2020), the C parameter was related to assimilative effects of US valence: while the C parameter predicted positive CS evaluations for positively paired CSs, it predicted negative CS evaluations for negatively paired CSs (see Figure E4). As indicated by a non-significant three-way interaction between US valence, relation pair and the C parameter, this pattern was comparable for CSs presented with symmetrical vs. asymmetrical relations.