COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13378

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13378

Following Affirmative and Negated Rules

Robert Wirth, Wilfried Kunde, Roland Pfister

Department of Psychology, Julius-Maximilians-University of Würzburg

Received 30 March 2022; received in revised form 28 June 2023; accepted 2 November 2023

Abstract

Rules are often stated in a negated manner ("no trespassing") rather than in an affirmative manner ("stay in your lane"). Here, we build on classic research on negation processing and, using a finger-tracking design on a touchscreen, we show that following negated rather than affirmative rules is harder as indicated by multiple performance measures. Moreover, our results indicate that practice has a surprisingly limited effect on negated rules, which are implemented more quickly with training, but this effect comes at the expense of reduced efficiency. Only affirmative rules are thus put into action efficiently, highlighting the importance of tailoring how rules are communicated to the peculiarities of the human mind.

Keywords: Rule formulation; Negation; Prescription; Prohibition; Ironic effects; Movement trajectories

1. Introduction

Rules rule our everyday life. They are pivotal for maintaining order in society, and they range from simple prompts of what (not) to do—"do not cross against a red traffic light"—to complex and universal moral maxims—"do no harm."

When confronted with a rule, humans tend to comply directly and immediately. Compliance as the behavioral default is particularly evident for meaningful rules that are commanded by an authority (Milgram, 1963), but it even ensues for arbitrary rules without any apparent plausibility or legitimation (Gozli, 2019; Pfister, Wirth, Schwarz, Steinhauser, & Kunde, 2016, Wirth, Pfister, Foerster, Huestegge, & Kunde, 2016). Possibly, this is for a good reason,

Correspondence should be sent to Robert Wirth, Department of Psychology, Julius-Maximilians-University of Würzburg, Röntgenring 11, 97070 Würzburg, Germany. E-mail: Robert.Wirth@uni-wuerzburg.de

as studies have shown that rule-compliant individuals are viewed as trustworthy and good social partners (Everett, Pizarro, & Crockett, 2016; Gross & De Dreu, 2021), and such side effects of rule-compliance may have rendered rule-based behavior as our evolutionary default (Baum, Richerson, Efferson, & Paciotti, 2004).

Less is known about whether different types of rules can be followed equally easily. Several promising strategies present themselves, however. First, rules should not be implicit ("Do not forget to tip your waiter"), but rather explicit to be communicated easily. Next, they should not be too generic ("Do not disobey traffic lights"), but call for concrete action ("Do not cross the street when the light is red"). They should also not leave too much room for interpretation ("Do not drive drunk"), so they should be unambiguous ("Do not drink and drive"). And finally, rules should not have too many exceptions ("No snacks before dinner").

An observant reader may have noticed one thing at this point: Every single example of a rule given here was a negated rule, stating what not to do. And frankly, it was not hard to come up with these examples. Many rules that we are confronted with are formulated this way. No smoking. No littering. No flash photography. Even the 10 commandments are basically a list of things that thou shalt not do. Unsurprisingly, such formulations are thus recognized as a main type of rule in linguistic classifications of legal text corpora (Biagioli, Francesconi, Passerini, Montemagni, & Soria, 2005; Sandrelli et al., 2018; Waltl, Bonczek, Scepankova, Matthes, 2019).

Findings from psycholinguistics and cognitive psychology document that negations are difficult to process (Jones, 1968; Wason, 1959; see also Beltrán, Orenes, & Santamaría, 2008; Dale & Duran, 2011; Kaup & Dudschig, 2020; Wirth, Kunde, & Pfister, 2019). They have been discussed as a core challenge for language comprehension (Kaup, 2001; MacDonald & Just, 1989), stereotype processing (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008), thought suppression (Wegner, Schneider, Carter, & White, 1987), habit formation (Adriaanse, Van Oosten, De Ridder, De Wit, & Evers, 2011), action control (Dudschig & Kaup, 2020; Wegner, Ansfield, & Pilloff, 1998), and many more (see Kaup & Dudschig, 2020, for a recent overview). Overall, negations are involved in such a wide range of psychological phenomena that they are considered a central cognitive operator. When confronted with a negation, the cognitive system is assumed to represent the non-negated content of the negation in a first step (Hasson & Glucksberg, 2006), only to negate it in a second step (Gilbert, 1991). Crucially, this second step requires cognitive resources, so errors can occur when resources are limited.

Thereby, negations often produce so-called ironic effects when the second step of the negation process is hindered, producing the exact mental representation of what you were not to think of (Wegner, 2009). A "no smoking" prompt should trigger a mental image of clean breathable air, but what comes to mind for most people is an image of a lit and smoking cigarette. Affirmative rules, on the contrary, should not come with such ironic effects. Despite conceptual differences between affirmative and negated rules (commands vs. prohibitions, von Wright, 1963) and the prevalence of negated rules in everyday life, research on rules has only rarely addressed differences between different rule types (Malle et al., 2021). Instead, cognitive accounts of rule representation and retrieval tend to generalize across different rule types on both theoretical (e.g., Sellars, 1949) and empirical grounds (e.g., Brass, Liefooghe, Braem, & De Houwer, 2017; Meiran, Liefooghe, & De Houwer, 2017).

Based on the profound challenges attached to negation processing, it is high time to compare negated and affirmative rules. In the context of rules, another inherent property of negations is that they only inform about what is not true. In linguistics, the phrase "the box is not filled with cookies" does not inform about what the box actually contains (if anything at all). Similarly, in action control, negations do not support agents a lot in selecting appropriate actions. Not crossing a road at a red traffic light is compatible with many behavioral alternatives of a pedestrian, like checking the smartphone, talking to a nearby fellow, or doing cartwheels. Consequently, the action selection system would have to engage in a search process of what to do instead. Affirmative rules (like "stand still at the stop line") conceivably support action selection more, as they highlight behavioral options that are obviously rule-compliant. There are exceptions to the rule of negation difficulty, however. In linguistics, there are combinations that are considered *pragmatically licensed* (Nieuwland & Kuperberg, 2008). Sentences like "the reviews weren't too bad" are easily and readily understood. However, preconditions for such pragmatic licensing are highly constrained (Nieuwland, 2016), and this has mostly been shown for linguistic negations. Here, we took the perspective of action control on negation processing, that is, how a corresponding behavior is enacted when confronted with affirmative and negative rules. For such cases, pragmatic licensing has barely been shown (but see Dale & Duran, 2011).

An early attempt to assess performance for negated rules is research on negative instructions (Geissler, 1912; Langfeld, 1910, 1913; for a review, see Proctor & Xiong, 2017). Participants were presented with photographs of objects and were asked to say the first thing that came to mind, but not to name the object. This task posed considerable difficulty, suggesting that with negative instructions, participants usually activate the name of the object, which then must be inhibited to select an alternative response (see also Beltrán, Morera, García-Marco, & Vega, 2019). What previous research in this area did not agree on is whether practice could improve performance, with some evidence for decreasing reaction times with repeated picture presentation in some studies (Langfeld, 1910) but no evidence for repetition benefit in others (Geissler, 1912).

A century later, it is still an open question whether negation costs can diminish with practice. One recent study found that when participants perform a visual search task that allows for ignoring a certain feature throughout the task ("ignore the red targets"), this instruction produces initial costs that turn into overall search benefits with practice (Cunningham & Egeth, 2016). Note, however, that this instruction does in fact not include a negation (see also Jones, 1966). One study found no overall practice-related negation skill improvements, but occasional performance improvements that could be linked to memory traces (Deutsch, Gawronski, & Strack, 2006). Similarly, a more recent study found that true negation effects can indeed diminish, but only if two specific criteria are met: There must be a massive frequency of negations and a negation must have been processed very recently (Wirth et al., 2019).

With the current experiment, we wanted to test how rule formulation affects how efficiently rules are put into action. Our first aim was to directly compare how people behave when confronted with an affirmative rule ("do this") compared to a negated rule ("do not do this"), expecting superior performance with affirmative rules. Second, we explored if and how

performing compliant behavior may change with practice. Therefore, we confronted participants with one of two rules, either affirmative or negated, for about half an hour. If participants improve when practicing negated rules, performance differences between rule types should diminish with increasing trials.

To measure the efficiency of translating different rule types into action, we used an innovative task setup that required movements as responses. For example, Dale and Duran (2011) used mouse movements to the upper left and right corners of the screen using negated or non-negated instructions. They show more complex movements with negation instructions. Our setup is similar to established mouse tracking setups (Freeman & Ambady, 2010; Kieslich, Henninger, Wulff, Haslbeck, Schulte-Mecklenbeck, 2019; Spivey, Grosjean, & Knoblich, 2005), with the difference that we measure finger movements on a touchscreen, so there is no movement transformation between the mouse and the cursor movement (see Wirth, Foerster, Kunde, & Pfister, 2020 for an overview). This let us analyze both temporal parameters of how long it took to execute compliance responses, and spatial parameters that let us estimate the assumed ironic effects via the movement trajectories (moving toward something while being asked to not do that).

2. Methods

2.1. Participants

Ninety-six participants were recruited (mean age = 26.0 years, SD = 4.8) and received either course credit or monetary compensation.[1] Assuming a sizeable effect of negation processing on movement trajectories (e.g., $d_z = 1.30$ for Exp. 3 in Wirth et al., 2016), this sample size should provide high power ($1\text{-}\beta > .99$) to find the expected effect and allow for counterbalancing. All participants gave informed consent, were naïve to the purpose of the experiment, and were debriefed after the session.

2.2. Apparatus and stimuli

The experiment was run on an iPad (12.9-inch screen diagonal, resolution of 2048×2732 px) in portrait mode with a viewing distance of about 50 cm. Participants used the index finger of their dominant hand to operate the touchscreen, which sampled the finger movements at 120 Hz.

We used six different symbols of two sets (astrology symbols: ♈ / ♊ / ♉; card symbols: ♠ / ♣ / ♢). Different sets of symbols for both rule conditions avoided possible carryover effects. For each participant, one symbol of each set was randomly chosen as the symbol they had to reach (affirmative rule) or not reach (negated rule). The mapping of symbol set and rule condition was counterbalanced between participants. In every trial, two symbols of one set would appear in the upper left and right corners of the screen, prompting movements to the left or right. At the bottom center of the screen was the starting position (a circle of 0.6 cm in diameter) from which all movements had to be initiated (see Fig. 1). In between trials, a

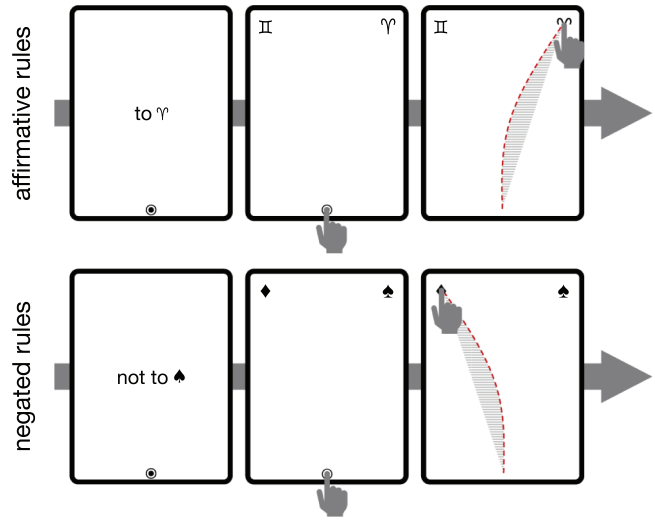

Fig. 1. Procedure of the experiment. Between trials, participants were confronted with the currently relevant rule that, depending on the current rule condition, told them where to go (affirmative rules, upper half) or where not to go (negated rules, lower half). When participants touched the starting area, the two possible targets appeared and prompted movements to the left or right upper corners.

reminder of the current rule (e.g., "to ♈" for affirmative rules; "not to ♠" for negated rules) was displayed.

2.3. Procedure

Participants were confronted with two rule types, each with their individual symbols, and rules were varied in a blocked manner. Half of the participants started with the affirmative rule and switched to the negated rule halfway through the experiment, the other half of the participants experienced the reverse order for counterbalancing. Before each trial, the currently relevant rule was displayed as a reminder in the center of the screen (e.g., "to ♈"; "not to ♠"). This rule stayed constant for each participant.

When touching the starting area, this written reminder disappeared, and the two symbols in the upper corner appeared, prompting a movement to the upper left or right corner. The currently relevant symbol was always paired with another symbol of the same set. Each set comprised three symbols to avoid easy recoding strategies that would be possible with two symbols (for which not reaching one symbol is always equivalent to reaching the alternative). A trial ended when the finger was lifted from the touchscreen, and the next trial started immediately by displaying the rule reminder. Error feedback was displayed if participants reached the wrong target or failed to hit any target at all.

Instructions stressed that responses had to be delivered quickly and accurately. Participants completed 20 blocks of 40 trials each, lasting about 1 h in total.

3. Results

3.1. Preprocessing

We measured several variables of each movement (see Supplementary Material) and, for brevity, decided to report two main measures: First, the time to initiate a movement, defined as the time between touching the starting area and leaving it (initiation time; IT), indexing the time to complete the purely cognitive operations of processing affirmative versus negated rules (response planning), and second, the area between the actual movement trajectory and a straight line from start to endpoint (area under the curve; AUC; shaded area in Fig. 1) as a spatial measure for the ironic effect during response execution. AUC was computed from the time-normalized coordinate data of each trial by using custom MATLAB scripts. Movements to the left were mirrored at the vertical midline for all analyses. AUC was computed as a signed area so that positive values indicate attraction toward the opposite target area (in case of negations, larger AUCs stand for stronger ironic effects), and negative values would indicate attraction toward the nearest edge of the display. All data, analysis scripts, and results of additional measures (see Wirth et al., 2020) are available at osf.io/96t3z.

3.2. Data selection and analyses

We omitted trials in which participants failed to enact the instruction and landed on the wrong target area (0.3%, with no difference between rule conditions, $|t(95)|<1$), and trials in which participants failed to hit any of the target areas at all (2.1%, with no difference between rule conditions, $|t(95)|<1$). Trials were discarded as outliers if any measures (IT, AUC) deviated more than 2.5 SDs from a participant's individual cell mean (5.0%). Data for each measure were then aggregated separately for each participant and each combination of rule condition (affirmative vs. negated) and block within each condition (1–10).

Mean ITs and AUCs were analyzed in a 2×10 analysis of variance with rule condition and block as within-subject factors. To avoid violations of sphericity, we used the multivariate approach for all analyses. Planned post-hoc tests probed for the effect of rule condition separately in each block via *t*-tests. In post-hoc analyses, we computed $d_z = \frac{t}{\sqrt{n}}$ and $\Delta = DV_{negated} - DV_{affirmative}$.

3.3. ITs

Fig. 2 shows ITs as a function of rule condition and block. A significant effect of rule condition, $F(1,95) = 67.53$, $p<.001$, $\eta_p^2 = .42$, indicated faster response initiations for affirmative rules (323 ms) than for negated rules (374 ms). A significant contribution of block, $F(9,87) = 14.87$, $p<.001$, $\eta_p^2 = .61$, showed faster response initiation of later blocks (Block 1: 382 ms; Block 10: 326 ms), producing a significant linear trend, $F(1,95) = 95.13$, $p<.001$, $\eta_p^2 = .50$. There was no interaction between both factors, $F(9,87) = 1.47$, $p = .171$, $\eta_p^2 = .13$. Planned post-hoc analyses demonstrated significant negation effects for all blocks, ts>5.53, ps<.001, d_zs>0.56, Δs>39 ms.

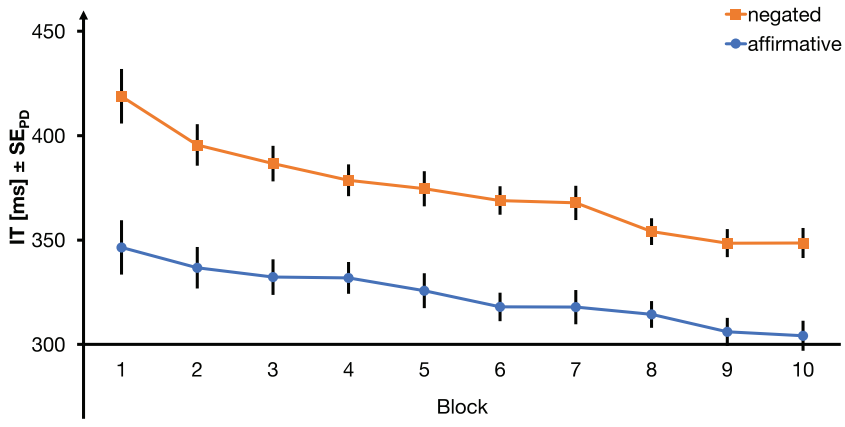

Fig. 2. Initiation time (IT) results. Mean ITs are plotted as a function of block (abscissa) and rule condition (blue circles for affirmative rules, orange squares for negated rules). Error bars represent standard errors of paired differences, computed separately for each block (Pfister & Janczyk, 2013).

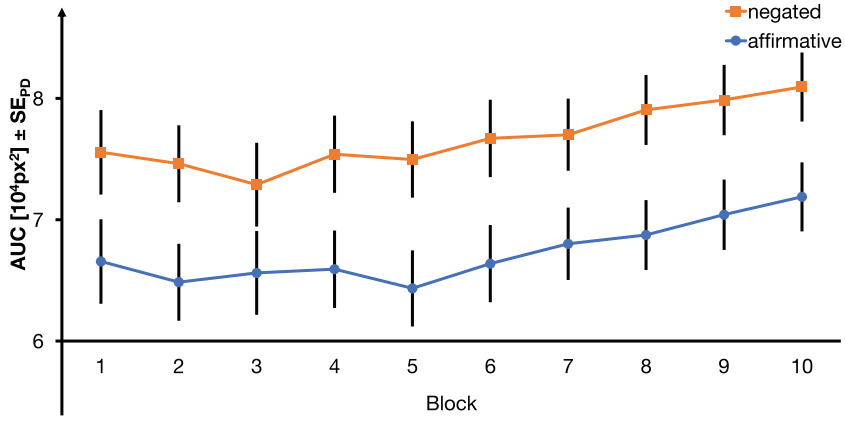

Fig. 3. Area under the curve (AUC) results. Mean AUCs are plotted as a function of block (abscissa) and rule condition (blue circles for affirmative rules, orange squares for negated rules). Error bars represent standard errors of paired differences, computed separately for each block (Pfister & Janczyk, 2013).

3.4. AUCs

Fig. 3 shows AUCs as a function of rule condition and block. A significant effect of rule condition, $F(1,95) = 24.40$, $p<.001$, $\eta_p^2 = .20$, indicated more direct responses for affirmative rules (67771px^2) than for negated rules (77734px^2). A significant contribution of block, $F(9,87) = 2.02$, $p = .046$, $\eta_p^2 = .17$, showed less direct response execution of later blocks (Block 1: 71901px^2; Block 10: 77088px^2), producing a significant linear trend, $F(1,95) = 8.60$, $p = .004$, $\eta_p^2 = .08$. There was no interaction between both factors, $F<1$. Planned post-hoc analyses demonstrated significant negation effects for all blocks, ts>2.26, ps$<.013$, d_zs>0.23, Δs>7941px^2.

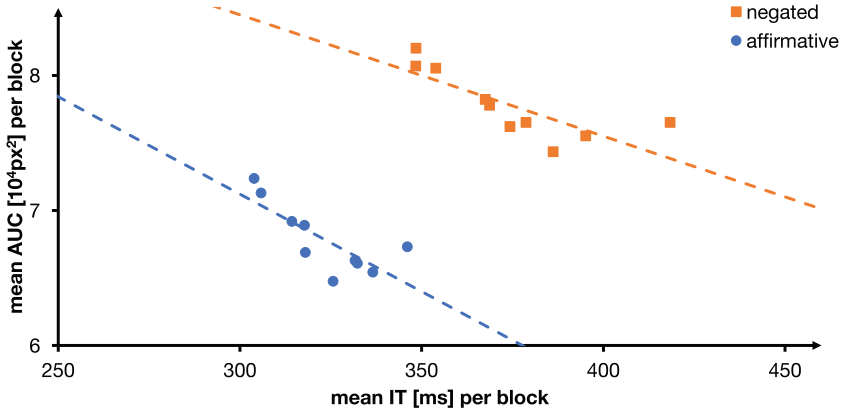

Fig. 4. Correlation and regression between IT and AUC data. Mean ITs (abscissa) are plotted against mean AUCs (ordinate), separately for each block (1–10, each data point represents one block: blue dots for affirmative rules, orange squares for negated rules).

3.5. Spatio-temporal tradeoff

The spatial efficiency of movement execution surprisingly declined with increasing practice. To highlight the spatio-temporal dynamics during practice, we analyzed both measures via correlations between the mean ITs and AUCs for each block (1–10) separate for each rule condition (Fig. 4). We found a strong negative correlation for both the affirmative, $r(8) = -.788$, $p = .007$, and the negated rules, $r(8) = -.765$, $p = .010$, suggesting a proportional impairment in the spatial domain with every improvement in the temporal domain.

3.6. Partialling out IT influences in AUC

To arrive at a measure of practice effects over time that is adjusted for this spatio-temporal tradeoff, we computed individual regressions for AUCs based on ITs for each participant and each rule condition over the 10 blocks. Based on these regressions, we first computed $AUC_{predicted}$ based on their observed mean ITs for each participant and each block of each rule condition, and then the residuals as $AUC_{residual} = AUC_{observed} - AUC_{predicted}$. This way, we could partial the systematic decline in ITs out of the systematic increase in AUCs. Consequently, if the increase in AUCs could fully be explained by the decrease in ITs, then $AUC_{residual}$ should no longer show increasing values with increasing blocks. As expected, individual regression slopes on $AUC_{residual}$, separately for each participant and each rule condition over the 10 blocks, did not differ from zero for neither affirmative nor negated rules, $ts < 1$. This suggests that the increase in AUCs can fully be attributed to the spatio-temporal tradeoff.

3.7. Partialling out AUC influences in IT

To partial the increase in AUCs out of ITs, we conducted an analog analysis as reported above. Note that since the final analysis is solely based on the residuals, these analyses are

not commutative. We again computed individual regressions for ITs based on AUCs, $IT_{residual}$ as the difference between $IT_{observed}$ and $IT_{predicted}$, and then independent regressions of the residual ITs as a function of the experimental block. Interestingly, even after partialling out the influence of AUC, there was a significant decrease in ITs for later blocks for both affirmative (at a slope of −2.47 ms per block), $t(95) = 6.10$, $p<.001$, $d_z = 0.62$, and negated rules (−5.29 ms per block), $t(95) = 6.72$, $p<.001$, $d_z = 0.69$. Even after accounting for the spatio-temporal tradeoff, response initiation seems to show an index of performance improvements. Specifically, when directly comparing the slopes after partialling out the AUC influences, we find that performance for negated rules benefits more from practice than for affirmative rules, $t(95) = 3.16$, $p = .002$, $d_z = 0.32$.

Further, by computing ratios of the remaining slope of the regression after partialling out the spatio-temporal tradeoff and the slope of the regressions based on the observed data, we can estimate the relative contribution of practice effect to overall performance. These ratios suggest that $\frac{-2.47}{-4.45} \approx 55\%$ of improvements with affirmative rules and $\frac{-5.29}{-7.04} \approx 75\%$ of improvements with negated rules can be attributed to practice effects.

3.8. Control experiment

Based on the feedback of two anonymous reviewers, we conducted a control experiment to address several potential confounds, namely (a) possible recoding strategies due to the constant mapping per participant, (b) the separate symbol sets for both rule types, (c) the blocked manipulation of rule type, and (d) potential visual priming effects. The control experiment can be found in the Supplementary Material. To sum up the results, even when controlling for these influences, the control experiment replicates the results of the main experiment.

4. Discussion

In the current study, we probed rule-following behavior for affirmative and negated rules, and we tested how practice effects would moderate a potential performance difference between rule types. We employed a finger-tracking design on a touchscreen that let us access not only the temporal characteristics of responding to these rules, but also provided a spatial signature mirroring possible ironic negation effects (Freeman, Dale, & Farmer, 2011; Dale & Duran, 2011; Wirth et al., 2016, 2019). While we found an overall decrease of response initiation times for both affirmative and negated rules, even with massive practice there were still stable response costs for negated compared to affirmative rules (see Dudschig & Kaup, 2018, for converging evidence from analyses of response times).

Interestingly, while the temporal markers showed an overall improvement with later blocks, the spatial deviation became more prominent, showing that movements were less direct in later blocks. This suggests that with every apparent increase in temporal performance goes a proportional decrease in spatial performance. At first glance, it appears participants do not improve (neither with affirmative nor with negated rules), but rather they seem to gradually shift their strategy during the course of the experiment: While at first, they take their time to

plan their response, producing higher ITs and lower AUCs, with some practice they start their responses more liberally, producing lower ITs at the cost of higher AUCs.

Negations under practice seem to produce a spatio-temporal tradeoff in the present design. Interestingly, a small index of practice effects seems to survive when partialling out this tradeoff, suggesting that practice does nominally improve performance with negations, and negations benefit from practice to a larger degree than affirmative responses, but its influence is negligible in the broader view of action control. This result could only be obtained by extending classical chronometric methods with movement responses and spatial markers. This may also explain why some authors found practice benefits (Langfeld, 1910; Cunningham & Egeth, 2016), while others have not (Geissler, 1912), as apparent temporal performance improvements may go with the detriment of other aspects of performance that had yet to be studied. Thereby, the current results may help answering the century-old question of whether performance under negative instructions improves as a function of training—it may so, but only if you look at single performance measures.

As a supplementary explanation, we may consider possible recoding strategies. Obviously, participants have trouble enacting responses based on negated rules. Possibly, participants avoid dealing with negations by cognitively recoding them: Instead of "not to ♠," they may represent "to either ♣ or ♢." This highlights another inherent difficulty of negations, while they instruct you what not to do, they often leave you with a spectrum of alternatives. This uncertainty may add to the observed results, but even when negations are designed in a way that they are as concrete as affirmative instructions, they come with performance decrements (e.g., Dudschig & Kaup, 2018, 2020; Wirth et al., 2019).

This leads to the question of whether compliance rates differ between affirmative and negative rules if participants could choose whether they want to follow a rule or not. Previous studies have shown that with affirmative rules, violations are only chosen infrequently (Pfister et al., 2016, 2019), and that enacting these rule violations comes with significant performance costs (Wirth et al., 2016, 2018a, 2018b; see also Imhof & Rüsseler, 2019). We reason that violating affirmative rules requires a negation, producing both the choice bias and performance costs. If that were the case, we might find more violations and superior violation performance with negated rules, as violating negated rules should reduce the ironic effects (Rück et al., 2021). Future research should, therefore, explore this intriguing intersection of nonconformity and rule formulation.

There are situations in which it is not possible to opt for an affirmative formulation, and at times negated rules may satisfy the requirements of being explicit and efficient more than any alternative. For example, it may be easier to communicate to not pass the red light of a busy street to a child than to list all the behavioral options that are compatible with the child's survival. It also highlights the crucial information here (crossing the red light is dangerous), so the implied dangers may be better understood when explicitly calling out the unwanted behavior, even in its negated form.

Overall, however, we can show that negated rules are difficult to enact, even after extensive practice. While we find markers for performance improvements with practice, these benefits are negligible in the scope of the overall costs of negation processing. If rules can equally be formulated affirmatively and in a negated manner, affirmative formulations are clearly

preferable. Instead of asking people to "not walk on the grass," we should, therefore, ask them to "stay on the paths" to make following these rules maximally easy, and we should do so even after having told them to "not walk on the grass" hundreds of times before.

Acknowledgments

Open access funding enabled and organized by Projekt DEAL.

Note

1 We originally conducted an experiment with 48 participants and a fixed mapping of symbols to rule types. Following the suggestion of two anonymous reviewers, we ran a replication study with the reversed mapping to ensure that the different symbols did not exert any confounding influences on our results. Both experiments yielded highly similar results so that we decided to report both samples in a pooled analysis here. The data of both individual experiments are available online (osf.io/96t3z).

References

Adriaanse, M. A., Van Oosten, J. M., De Ridder, D. T., De Wit, J. B., & Evers, C. (2011). Planning what not to eat: Ironic effects of implementation intentions negating unhealthy habits. *Personality and Social Psychology Bulletin*, *37*(1), 69–81.

Baum, W. M., Richerson, P. J., Efferson, C. M., & Paciotti, B. M. (2004). Cultural evolution in laboratory microsocieties including traditions of rule giving and rule following. *Evolution and Human Behavior*, *25*(5), 305–326.

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., & Soria, C. (2005). Automatic semantics extraction in law documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law* (pp. 133–140).

Beltrán, D., Orenes, I., & Santamaría, C. (2008). Context effects on the spontaneous production of negation. *Intercultural Pragmatics*, *5*, 409–419.

Beltrán, D., Morera, Y., García-Marco, E., & Vega, M. D. (2019). Brain inhibitory mechanisms are involved in the processing of sentential negation, regardless of its content. Evidence from EEG theta and beta rhythms. *Frontiers in Psychology*, *10*, 1782. https://doi.org/10.3389/fpsyg.2019.01782

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience & Biobehavioral Reviews*, *81*, 16–28.

Cunningham, C. A., & Egeth, H. E. (2016). Taming the white bear. Initial costs and eventual benefits of distractor inhibition. *Psychological Science*, *27*(4), 476–485.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, *35*(5), 983–996.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology*, *91*(3), 385–405.

Dudschig, C., & Kaup, B. (2018). How does "not left" become "right"? Electrophysiological evidence for a dynamic conflict-bound negation processing account. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 716–728.

Dudschig, C., & Kaup, B. (2020). Negation as conflict: Conflict adaptation following negating vertical spatial words. *Brain and Language*, *210*, 104842.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, *44*(2), 370–377.

Geissler, L. R. (1912). Analysis of consciousness under negative instruction. *American Journal of Psychology*, *23*, 183–213.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107–119.

Gozli, D. (2019). *Experimental psychology and human agency*. Cham: Springer.

Gross, J., & De Dreu, C. K. (2021). Rule following mitigates collaborative cheating and facilitates the spreading of honesty within groups. *Personality and Social Psychology Bulletin*, *47*(3), 395–409.

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation?: An examination of negated metaphors. *Journal of Pragmatics*, *38*(7), 1015–1032.

Imhof, M. F., & Rüsseler, J. (2019). Performance monitoring and correct response significance in conscientious individuals. *Frontiers in Human Neuroscience*, *13*, 239.

Jones, S. (1966). The effect of a negative qualifier in an instruction. *Journal of Verbal Learning and Verbal Behavior*, *5*(5), 497–501.

Jones, S. (1968). Instructions, self-instructions and performance. *Quarterly Journal of Experimental Psychology*, *20*, 74–78.

Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory & Cognition*, *29*(7), 960–967.

Kaup, B., & Dudschig, C. (2020). Understanding negation: Issues in the processing of negation. In V. Déprez & M. T. Espinal (Eds.), *The Oxford handbook of negation* (pp. 634–655). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198830528.013.33

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M., & Schulte-Mecklenbeck, M. (2019). Mouse-Tracking. *A Handbook of Process Tracing Methods*; Routledge: Abingdon, UK, 111–130.

Langfeld, H. S. (1910). Suppression with negative instruction. *Psychological Bulletin*, *7*, 200–208.

Langfeld, H. S. (1913). Voluntary movement under positive and negative instruction. *Psychological Review*, *20*, 459–478.

MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 633–642.

Malle, B. F., Austerweil, J. L., Chi, V. B., Kenett, Y., Beck, E. D., Thapa, S., & Allaham, M. (2021). Cognitive properties of norm representations. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Meiran, N., Liefooghe, B., & De Houwer, J. (2017). Powerful instructions: Automaticity without practice. *Current Directions in Psychological Science*, *26*(6), 509–514.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: Calculation, interpretation, and a few simple rules. *Advances in Cognitive Psychology*, *9*(2), 74–80.

Pfister, R., Wirth, R., Schwarz, K. A., Steinhauser, M., & Kunde, W. (2016). Burdens of non-conformity: Motor execution reveals cognitive conflict during deliberate rule violations. *Cognition*, *147*, 93–99.

Pfister, R., Wirth, R., Weller, L., Foerster, A., & Schwarz, K. A. (2019). Taking shortcuts: Cognitive conflict during motivated rule-breaking. *Journal of Economic Psychology*, *71*, 138–147.

Proctor, R. W., & Xiong, A. (2017). The method of negative instruction: Herbert S. Langfeld's and Ludwig R. Geissler's 1910–1913 insightful studies. *American Journal of Psychology*, *130*(1), 11–21.

Rück, F., Dudschig, C., Mackenzie, I. G., Vogt, A., Leuthold, H., & Kaup, B. (2021). The role of predictability during negation processing in truth-value judgment tasks. *Journal of Psycholinguistic Research*, *50*(6), 1437–1459.

Sandrelli, A. (2018). Observing Eurolects: Corpus analysis of linguistic variation in EU law.

Sellars, W. (1949). Language, rules and behavior. In S. Hook (Ed.), *John Dewey: Philosopher of science and freedom*.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.

von Wright, G. H.(1963). *Norm and action: A logical enquiry*. Humanities Press. https://doi.org/10.2307/2218217

Waltl, B., Bonczek, G., Scepankova, E., & Matthes, F. (2019). Semantic types of legal norms in German laws: Classification and analysis using local linear explanations. *Artificial Intelligence and Law*, *27*(1), 43–71.

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *11*(2), 92–107.

Wegner, D. M. (2009). How to think, say, or do precisely the worst thing for any occasion. *Science*, *325*(5936), 48–50.

Wegner, D. M., Ansfield, M., & Pilloff, D. (1998). The putt and the pendulum: Ironic effects of the mental control of action. *Psychological Science*, *9*(3), 196–199.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*(1), 5–13.

Wirth, R., Foerster, A., Herbort, O., Kunde, W., & Pfister, R. (2018a). This is how to be a rule breaker. *Advances in Cognitive Psychology*, *14*(1), 21–37.

Wirth, R., Foerster, A., Kunde, W., & Pfister, R. (2020). Design choices: Empirical recommendations for designing two-dimensional finger tracking experiments. *Behavior Research Methods*, *52*, 2394–2416.

Wirth, R., Foerster, A., Rendel, H., Kunde, W., & Pfister, R. (2018b). Rule-violations sensitise towards negative and authority-related stimuli. *Cognition & Emotion*, *32*(3), 480–493.

Wirth, R., Kunde, W., & Pfister, R. (2019). How not to fall for the white bear: Combined frequency and recency manipulations diminish negation effects on overt behavior. *Journal of Cognition*, *2*(1), 1–18.

Wirth, R., Pfister, R., Foerster, A., Huestegge, L., & Kunde, W. (2016). Pushing the rules: Effects and aftereffects of deliberate rule violations. *Psychological Research*, *80*(5), 838–852.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Results for the Main Experiment
Control Experiment