Breaking the rules

Cognitive conflict during deliberate rule violations.

Roland
Pfister

Roland Pfister

Breaking the rules

Cognitive conflict during deliberate rule violations

Logos Verlag Berlin

Für André und Sonja

TABLE OF CONTENTS

ZUSAMMENFASSUNG

„Ich werde sie den Ungehorsam lehren, den Widerstand und die Unbeugsamkeit, gegen jeden Befehl aufzubegehren, und nicht zu buckeln vor der Obrigkeit." beschwört der Liedermacher Reinhard Mey und drückt damit aus, dass sich eigene Absichten nicht immer mit bestehenden Regeln vereinbaren lassen. Tatsächlich stellen Konflikte von bestehenden Regeln und eigenen Zielen eine wichtige und alltägliche Herausforderung für jeden einzelnen dar, die es zu bewältigen gilt.

Offensichtlich kann der Verstoß gegen Regeln und soziale Normen bestimmte Konsequenzen nach sich ziehen und diese Konsequenzen lassen sich von verschiedenen Standpunkten aus betrachten, etwa hinsichtlich moralischen Urteilen, ethischen Implikationen oder juristischen Folgen. Im Gegensatz zu den detaillierten Aufarbeitungen die aus den genannten Bereichen vorliegen, sind die kognitiven Prozesse die dem Brechen einer Regel zugrunde liegen bisher nicht systematisch untersucht worden. Als einen ersten Schritt in diese Richtung beschreibe ich drei Experimentalreihen, die untersuchen ob das absichtliche Verstoßen gegen eine Regel kognitiven Konflikt erzeugt. Die Experimente legen hierbei bewusst den Fokus auf eng umschriebene Handlungen und arbiträre Regeln, deren Nicht-Befolgen keinerlei Sanktionen oder andere negative Konsequenzen nach sich zieht.

Tatsächlich führt ein Regelverstoß selbst in diesem umschriebenen Rahmen zu messbarem kognitiven Konflikt, und dies gilt sowohl für einfache Wahlreaktionsaufgaben mit einer bestimmten Zuordnungsregel (Exp. 1-3) als auch für Bewegungstrajektorien und andere Merkmale komplexerer Bewegungen (Exp. 4-6). Diese Experimente legen auch nahe, dass beim Regelverstoß die zu brechende Zuordnungsregel kontinuierlich repräsentiert bleibt und sich daher im Verhalten niederschlägt. Die Regel scheint hingegen deutlich weniger stark repräsentiert zu sein, wenn äquivalente Handlungen nicht als Regelverstoß konzeptualisiert werden, sondern als Befolgung einer alternativen, entgegengesetzten Regel, die von außen vorgegeben wird. Weiterhin zeichnen sich absichtliche Regelverstöße auch durch eine spezifische elektrophysiologische Signatur aus, die mit der Überwindung der ursprünglichen Zuordnungsregel in Zusammenhang zu stehen scheint (Exp. 7-8).

Die erzielten Befunde ermöglichen eine erste, detaillierte Charakterisierung der kognitiven Implikationen absichtlicher Regelverstöße. Hierbei lassen sich verschiedene Mechanismen annehmen, die den beobachteten Effekten zugrunde liegen könnten; plausibel erscheint einerseits eine irrationale Erwartung negativer Konsequenzen oder andererseits eine komplexe Repräsentation von Regelverstößen, etwa als Kombination der Regel selbst sowie ihrer Negation. Die Beiträge dieser Mechanismen zur Entstehung von kognitivem Konflikt bei absichtlichen Regelverstößen sind ein vielversprechendes Feld für zukünftige experimentelle Ansätze. Dasselbe gilt für eine Erweiterung der vorliegenden Untersuchungen auf soziale Situationen und die gezielte Betrachtung von Sanktionen. Hier stellt sich insbesondere die Frage, inwiefern der beschriebene kognitive Konflikt an sich die grundlegende Entscheidung beeinflusst, ob gegen eine unliebsame Regel verstoßen werden soll.

SUMMARY

"I ain't gonna pay no attention to your rules" sings hard-rock legend AC/DC. Violation of rules and social norms is not confined to hard rock musicians, however. Rather, conflict between existing rules and own goals and intentions, is an important daily challenge for human agents.

Rule violations obviously can imply consequences and these consequences can be viewed from different perspectives such as moral reasoning, ethical implications and legal consequences. Whereas these perspectives provide elaborate accounts for rule violation behaviour, the cognitive processes during deliberate rule violations have not been investigated systematically to date. As a first step in this direction, I report three experimental series probing for cognitive conflict while violating a rule. These experiments focus specifically on circumscribed motor actions according to arbitrary experimental rules in a setting that does not enforce any sanctions or otherwise negative consequences upon the agent.

Even in this controlled setting, rule violations cause measurable cognitive conflict that is evident for responses in choice reaction tasks (Exp. 1-3) as well as for movement trajectories and other features of more complex actions (Exp. 4-6). These experiments also suggest the rule representation to remain continuously active while violating a rule, thus leaving a fingerprint on the agent's behaviour. By contrast, the

representation of the original rule seems to be diminished instantly if equivalent actions are not labelled as rule violations but rather as responses according to an alternative rule specifying the opposite mapping. Finally, deliberate rule violations seem to come with a specific electrophysiological signature which likely reflects the need to overcome the active rule representation (Exp. 7-8).

The present results are a first, detailed step towards characterising the cognitive implications of deliberate rule violations. Yet, different mechanisms might account for the observed effects; plausible candidates are continued but irrational expectations of negative consequences or, on a different note, difficulty relating to a complex representation of rule violations, e.g., as a combination of the rule itself and its negation. The differential contribution of both mechanisms to cognitive conflict during rule violations seems to be a promising field for future inquiry. The same is true for a possible extension of the present experimental approach to social situations and the investigation of expected sanctions and their impact on rule violations. Here, it seems especially relevant to address the contribution of the described conflict to decisions on whether or not to violate a rule in the first place.

PART 1: RULES

1 What is a rule? From regularities to moral principles

Rules are an integral part of human life and they govern, shape, and regulate nearly all aspects of modern society. Some of these rules are general and authoritative such as traffic laws or rules governing professional sports, some are less tangible such as laws of etiquette or colloquial agreements.

As a working model for the present series of experiments, I suggest that rules can be defined on three distinct levels (**Fig. 1**). The most basic rules relate to observable regularities in the environment. These regularities may relate equally to both, physical laws and repeating patterns in the social environment. A different class of rules are social rules and norms that include agreements between social partners and institutional regulations. This intermediate level corresponds to the colloquial use of the term "rules" and will be the focus of the following experiments. Finally, some rules have been generalised enough to claim the status of laws or philosophically justified, universal ethical and moral principles.

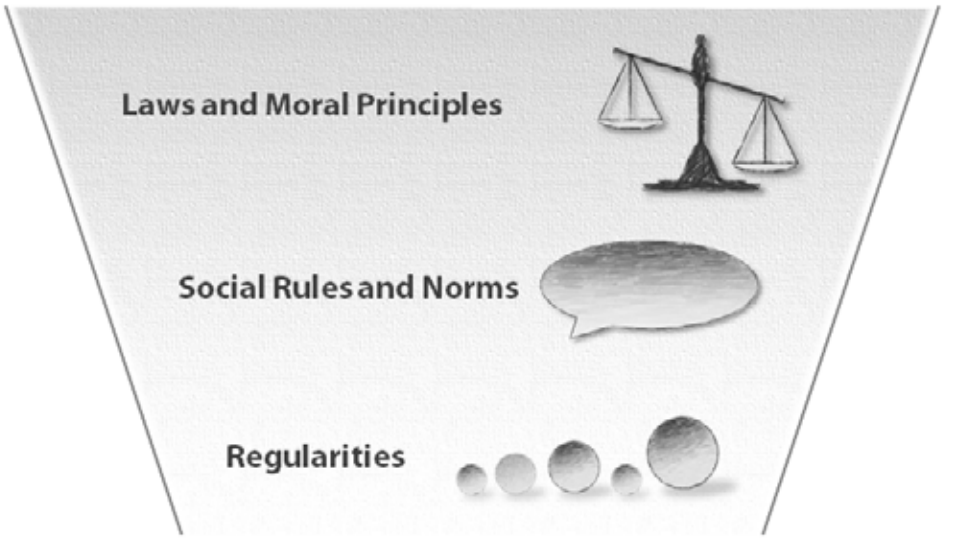

Fig. 1. Three levels to classify different rules. These levels range from regularities in an agent's environment to social rules and norms, such as agreements and regulations, up to laws as well as ethical and moral principles. The current experiments focus on the intermediate level of social rules and norms.

The following sections briefly introduce central aspects of each of the three levels, followed by a more thorough treatment of the intermediate level of social rules and norms. This treatment includes a detailed analysis of how rules affect behaviour in general and how human agents deal with intended or unintended failures to obey a given rule.

1.1 Regularities

The physical and social environment is full of regularities and human agents detect these regularities efficiently and effortlessly. A striking example for the efficiency of such detection processes is the *mismatch negativity* (Näätänen, 1990), a distinct electrophysiological response to stimuli violating a statistical regularity in an ordered sequence. This response seems to be independent of attention (Näätänen & Alho, 1995) and it occurs for different modalities such as with auditory stimuli (e.g.,

Näätänen, Gaillard, & Mäntysalo, 1978), visual stimuli (Czigler, 2007; Pazo-Alvarez, Cadaveira, & Amenedo, 2003) or tactile stimulation (Kekoni et al., 1997). Thus, simple regularities seem to be identified by very basic physiological mechanisms, ensuring that these regularities can be exploited by the agent.

Similarly strong electrophysiological responses arise for stimuli that occur only rarely in a given setting, which can be studied with the oddball paradigm (Sutton, Braren, Zubin, & John, 1965; Squires, Squires, & Hillyard, 1975). In this paradigm, some stimuli occur regularly in a series of trials whereas other stimuli occur only in a few trials and typically demand for a detection response. Just as unpredicted stimuli in case of mismatching stimuli, these rare oddball stimuli trigger a pronounced electrophysiological response over central and parietal electrodes that is often labelled the *P300* component. Naturally, this potential has been related to attention and memory processes (e.g., Donchin, 1981; Polich, 2007) or, alternatively, to processes that mediate between such oddball stimuli and appropriate responses (Verleger, Jaśkowski, & Wascher, 2005).

In addition to these neurophysiological markers of regularity detection, robust behavioural markers can be observed in various *implicit learning paradigms* (Reber, 1989). In these paradigms, human participants typically have the opportunity to identify regularities in their environment, including implicit covariations (Lewicki, 1986), sequential regularities such as finite state grammars (Cleeremans & McClelland, 1991), and differential stimulus probabilities (Hake & Hyman, 1953). Participants readily pick up such relevant information and adapt their behaviour accordingly.

The phenomena sketched above suggest that humans are indeed very sensitive to regularities in their environment. Learning these regularities, in turn, enables efficient behaviour (Reber, 1989) and might even underlie the acquisition of some of the most complex abilities of human agents such as the ability to understand and produce language (e.g., Saffran, 2003;

Seidenberg, 1997). Rules that go beyond such basic regularities usually have to be communicated and explicitly learned. Furthermore, whereas regularities can be seen as mostly descriptive, more explicit rules are clearly normative and classify different kinds of behaviour as either appropriate and rule-conform or inappropriate. These rules – social rules and norms on the one hand, and laws and moral principles on the other hand – are the focus of the following sections.

1.2 Social rules and norms

Many aspects of human life are regulated by explicit rules and these rules can take the form of simple agreements between social partners (e.g., "Lock the door when you leave the apartment.") or regulations in certain institutions (e.g., "Employees have to start their working day at 8 a.m."). During ontogenetic development, rules are first instantiated by a child's parents and internalising these rules is an important challenge during childhood. Even more important than internalising these rules, however, is developing an understanding of what a rule is (Smetana, 1993).

Piaget (1932) distinguished the developing application and the developing understanding of rules. He assumed both processes to occur in parallel even though changes in rule application do not necessarily imply changes in rule understanding, and vice versa. Accordingly, he proposed a separate stage model for each process; these stage models are mostly based on observations of children playing with marbles according to various rules (see **Fig. 2**).[1]

[1] Piaget (1932) uses the term „practice" when referring to the application of rules, and the term "consciousness" when referring to the knowledge about rules in general.

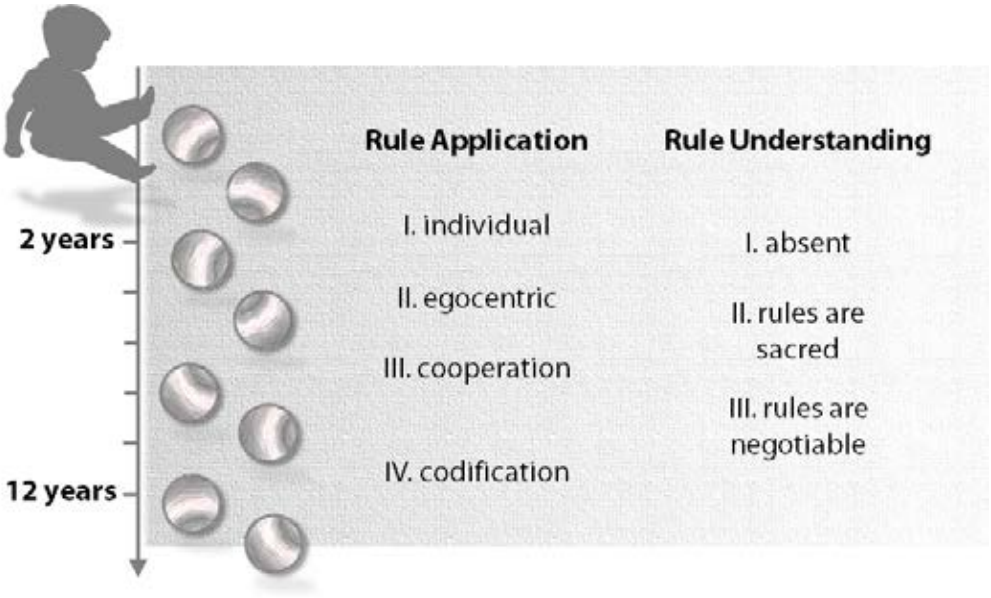

Fig. 2. Piaget's (1932) description of the developing application and understanding of rules as observed in the game of marbles. A thorough understanding of rules as negotiable agreements is only achieved at around ten to eleven years.

Regarding the spontaneous application of rules during a game, Piaget (1932) noticed that children begin with a purely *individual* game that does not involve any social partners (see also Kesselring, 1999). Between the age of two and five years, however, children begin to imitate the rules that are applied by older players. Still, the game is rather *egocentric* and others players are not involved. Children at this developmental stage thus begin to extract regularities by observing others and try to apply the same regularities. Around seven or eight years, children begin to engage in *cooperation* (for general comments on the validity of such age norms, see Brainerd, 1973). Rules are now obeyed and each player observes and controls the others for violations of these rules. Interestingly though, the rules that are applied by each individual player may still vary considerably even across players of a single game. This status changes between eleven and twelve years, when rules are finally *codified* and thus equalized across players.

For the second process – the understanding of rules – Piaget (1932) distinguished three stages (without specific labels as for the application of rules). The first stage extends to the middle of the egocentric stage of rule application and children in this stage do not show any precise understanding of what a rule is. This changes rather dramatically in the second stage (running up to the first half of the stage of cooperation) when rules are seen as sacred, everlasting and as being dictated from a higher authority, i.e., adults. A thorough understanding of what a rule is, however, does not develop until the age of about ten years. Only then do children see rules as explicit agreements between social partners that can, in principle, be debated and altered.

The sketched development indicates that rules are a rather complex concept that takes relatively long to be understood. And even though rules are understood as negotiable from this point onwards, they continue to have a huge impact on human behaviour. This impact is central to the following experiments and I will discuss it in more detail in Chapter 2. Beforehand, however, I briefly turn to the third and most normative class of rules in the following section.

1.3 Laws and moral principles

Laws and moral principles are strictly authoritative rules and have a much broader scope than simple social agreements. Especially moral principles have a long tradition in philosophy, dating back to Aristotle (cf. Höffe, 2008) and continuing to be an active area in philosophy ever since. Whereas moral philosophy is mainly concerned with the ethical justification of moral principles (ranging from Kant, 1788, to contemporary philosophers such as Habermas, 1983), research on moral psychology has elaborated several intriguing descriptive approaches and taxonomies for moral reasoning and its ontogenetic development.

Based on his observations of rule application and understanding, Piaget (1932) also devised a stage model of moral development that was later adopted and elaborated by Kohlberg (1971; for critical accounts, see Kurtines & Greif, 1974; Modgil & Modgil, 2011). In Kohlberg's model, moral understanding progresses from a pre-conventional level through a conventional level to a post-conventional level. Each level consists of two distinct stages, beginning with an obedience and punishment orientation and changing to a self-interest orientation on the pre-conventional level. In adolescence, moral understanding begins to involve social norms and conformity and, later on, a broader law and order morality. These two stages constitute the conventional level. The final two stages on the post-conventional level (social contract orientation or an orientation at universal ethical principles) are assumed not to be reached by each individual and require rather abstract reasoning about ethics and moral (Colby, Gibbs, Lieberman, & Kohlberg, 1983).

The sketched model is mainly based on observations of how children and adults solve moral dilemmas (Kohlberg, Levine, & Hewer, 1983). This methodology still influences contemporary psychology (Chandler, Greenspan, & Barenboim, 1973; Nichols & Mallon, 2006) and cognitive neuroscience (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004; for a review see Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005).

The use of moral dilemmas thus represents a productive strategy to uncover moral thoughts even though it is not directly equivalent to moral behaviour (e.g., Reynolds & Ceranic, 2007; for critical remarks see Haidt, 2007). Still, there are moderate correlations between moral reasoning on the one hand and moral behaviour on the other hand (Blasi, 1980; Rest, 1986) and these correlations are ascribed to at least four different factors: *Moral sensitivity* to situational and contextual factors, *moral judgment* about possible actions, *moral motivation* and according values, as well as certain personality traits that make a *moral character* such as courage and

perseverance (the four component model; Rest, 1986; see also Rest & Narváez, 1994; Narváez & Rest, 1995).

In sum, moral principles have a measurable impact on human behaviour. The following section takes a step back to social rules and norms that cannot be regarded as moral principles and investigates their influence on human behaviour.

2 Rules and behaviour

This section addresses the impact of explicit rules on human behaviour. I first discuss evidence from social psychology documenting the strength of this impact. Afterwards, I describe the cognitive processes that are influenced by explicit rules.

2.1 Power of rules: The social perspective

The impact of social rules and norms on human behaviour has long been a central theme of social psychology. Most studies in this field naturally investigated rules that were set up by a group, and investigated conformity to such group norms (Cialdini & Goldstein, 2004; Cialdini & Trost, 1998). An elegant paradigm to study this type of conformity is the *Line Judgment Task* in which participants are asked to match the length of comparison lines and that of a standard line (Asch, 1951, 1956). Group norms are set up by several confederates that give their judgments before the actual participant. Critically, in some trials the confederates unanimously gave the same wrong answer. In this situation, participants tend to adjust their answer to that of the group even if this answer does not match the initial instructions.

This basic finding of high conformity to group norms is highly reliable across different settings and tasks (e.g., Deutsch & Gerard, 1955; for recent studies, see Berns et al. 2005; Bond, 2005; Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007; Stallen, De Dreu, Shalvi, Smidts, & Sanfey, 2012; Walther et al., 2002). As such, research on conformity demonstrates that human agents readily follow social rules in their environment (but see Hodges & Geyer, 2006, for an alternative interpretation).

A central feature of these studies is the communication of rules and social norms via the behaviour of social partners. But what happens if these norms are communicated even more explicitly? A rather drastic answer to

this question is the outcome of Milgram's experiments on obedience to authorities (Milgram, 1963, 1974). In these experiments, the simple but explicit instruction to carry on with a straining task – delivering electric shocks to a seemingly suffering confederate – caused a high percentage of participants to deliver shocks with presumably lethal amplitude. Thus, the simple command of an authority seems to create a considerable pressure to obey (see also Blass, 1991, 1999; Elms, 1995; French & Raven, 1959; Cialdini & Goldstein, 2004).

Taken together, rules and norms in the social environment have a great impact on the behaviour of an individual. Notably, the lines of research reviewed above investigated the impact of rules in situations where rules or norms inflicted conflict and pressure on the participants. As such, these studies document the power of rules but they do not address the question of which processes are influenced by rules in non-conflicting situations. These processes were investigated more closely in the domain of cognitive psychology and are discussed in the following.

2.2 Rules and automaticity: The cognitive perspective

A cognitive approach to the processes triggered by new, arbitrary rules is the common choice reaction task: Certain stimuli are mapped onto specific responses by instructions (S-R rules) and participants' performance is assessed in terms of response time (RT) and accuracy. After some practice, these S-R rules become automatically activated when encountering the corresponding stimulus (e.g., Allport, Styles, Hsieh, 1994; Eriksen & Eriksen, 1974; Logan, 1988; Rogers & Monsell, 1995). A particularly interesting property of these instructed S-R rules is that an automatically working association between stimulus and response can be forged by mere intention and without any practice: the *prepared reflex* (Hommel, 2000; Woodworth, 1938).

Evidence for automatic retrieval of merely instructed S-R rules has been reported from several lines of research such as subliminal priming (Kunde, Kiesel, & Hoffmann, 2003), interference by distracting stimuli (Cohen-Kdoshay & Meiran, 2007, 2009; Reisenauer & Dreisbach, 2013), and task-switching (Liefooghe, Wenke, & De Houwer, 2012; Waszak, Wenke, & Brass, 2008; Wenke, Gaschler, & Nattkemper, 2007). These studies provide compelling evidence that a simple instructed S-R rule can be retrieved automatically, thereby activating the associated behaviour. Furthermore, changing instructions can also reduce the impact of learned associations if a newly instructed S-R rule specifies a competing S-R mapping (Waszak, Pfister, & Kiesel, 2013). In other words: Merely instructing an S-R rule can suffice to build up effective S-R associations or it might also counter-act associations that were learned through experience.[2]

Additional evidence for an impact of merely instructed rules comes from research on electrodermal conditioning (see Grings, 1973, for an overview). In typical studies on electrodermal conditioning, certain stimuli were predictably followed by an electrical shock for several times; the association of stimulus and upcoming shock is then evident in an increased skin conductance response. Interestingly, an increase may also occur without explicit conditioning phase when the relation of stimulus and shock is directly instructed by the experimenter (Cook & Harris, 1937), indicating that a physiological response can be triggered automatically by mere knowledge of a certain contingency or rule. Furthermore, the impact of learned associations might also be decreased by explicit instructions stating that the learned contingency is no longer valid (Colgan, 1970; Wilson, 1968).

[2] The above discussion only focuses on the mere existence of such instructed associations. A completely different question relates to their power in comparison to associations that are learned by experience. Merely instructed associations are typically weaker than learned ones and they will not fully counter thoroughly learned associations such as actual habits (but see Gollwitzer, 1999, for a fruitful approach).

The findings reviewed above suggest that merely instructed rules can indeed activate behavioural tendencies and physiological responses. This activation seems to occur automatically via associations between stimuli and appropriate responses. Still, such automatic activations will not always result in behaviour that conforms to the corresponding rule. Implications of such failures to act according to a rule are discussed in the following section.

3 Mismatches of rule and behaviour

As described above, human agents readily tend to obey social rules and norms. But not all rules can be obeyed at all times – and failures to obey a rule can occur intendedly or unintendedly. These two types of failures are typically labelled *violations* and *errors*, respectively (Reason, 1990; Reason, Manstead, Stradling, Baxter, & Campbell, 1990), and both types can be dissected into various different behaviours, according to their context and the underlying psychological processes (cf. **Fig. 3**). The following section gives an overview of the processes involved in both, errors and violations, from a cognitive perspective. This overview also sketches the central questions of the following experiments.

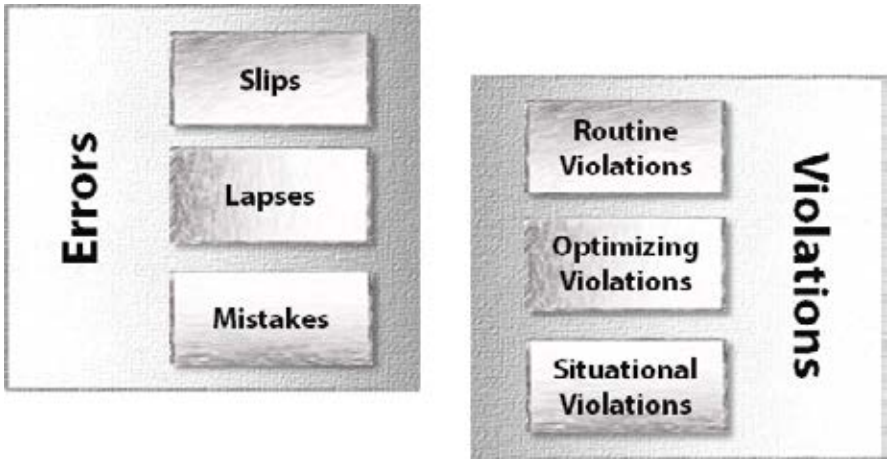

Fig. 3. Behaviour that counteracts a rule is an *error* if the failure to obey the rule is unintended whereas the behaviour is labelled *violation* if the failure to obey the rule is intended. Both, errors and violations, can be broken down to more specific behaviours according to context and underlying processes (according to Reason, 1990, 1995).

3.1 Errors

Errors, i.e., unintended failures to obey a rule, can arise for two distinct reasons: Either because a rule is misunderstood in the first place (mistakes) or because of failures during memory retrieval (lapses) or action execution (slips; Reason, 1990). In the following I will use the term error exclusively with the connotation of slips and lapses.

Research on error genesis and error processing has uncovered several behavioural and physiological correlates of such slips and lapses. For instance, errors leave a fingerprint on subsequent behaviour; this fingerprint is mainly visible in terms of decreased speed to initiate following actions and it is typically labelled *Post-Error-Slowing* (*PES*; Laming, 1968; Rabbitt, 1966). Different theoretical accounts have been proposed to explain PES, including suppression of automatic error-correction responses (Rabbitt & Rodgers, 1977), increased response caution (Laming, 1979; Dutilh et al., 2010), attentional orienting (Notebaert et al., 2009), motor inhibition (Marco-Pallarés, Camara, Münte, & Rodrígues-Fornells, 2008), and continued preoccupation with error monitoring (Jentzsch & Dudschig, 2009). Each of these accounts seems to cover distinct aspects of PES and may be better suited for one situation or another – in any case, however, PES is an established behavioural fingerprint of error processing.

A similarly established neurophysiological marker is the *Error-Related Negativity* (*ERN*) or *Error Negativity* (N_E), a negative-going deflection of the event-related potential (ERP) that peaks within 100 ms after erroneous responses (Falkenstein, Hohnsbein, & Hoormann, 1990; Gehring, Goss, Coles, Meyer, & Donchin, 1993; Renault, Ragot, & Lesevre, 1980; for reviews, see Falkenstein, Hoormann, Christ, & Hohnsbein, 2000; Gehring, Liu, Orr, & Carp, 2012). In contrast to PES, the ERN has a more widely accepted interpretation and is typically assumed to reflect a detection mechanism signalling either conflicting response tendencies (Botvinick,

Braver, Barch, Carter, & Cohen, 2001) or, alternatively, signalling that the outcome of a given action is not the intended one (Holroyd & Coles, 2002; Holroyd, Yeung, Coles, & Cohen, 2005). Both mechanisms highlight the role of the ERN for triggering behavioural adjustments after errors and, consequently, the ERN has been linked to brain structures such as the anterior cingulate cortex (Dehaene, Posner, & Tucker, 1994; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004).

In sum, errors seem to have a unique behavioural and electrophysiological signature setting them apart from normal behaviour. Furthermore, understanding the processes triggered by committing an error does not only inform about how human agents cope with such unintended failures to obey a rule but they also inform about more general mechanisms that enable rapid adjustments to unexpected events.

3.2 Violations

Research on violations has typically adopted a quite different perspective than research on errors. Most notably, research on violations has highlighted situational and organizational determinants of rule violations instead of focusing on the acting individual (Reason, 1990, 1995).

Still, the agent's perspective plays an important role for rule violations. *Routine violations*, for instance, are based on shortcuts to simplify a given task in light of suitable situational opportunities. Similarly, *optimising violations* (or "violations for kicks") are committed to render unloved tasks more exciting. Only *necessary violations*, by contrast, are mainly determined by situations that do not allow for rule-based behaviour (Reason, 1995). Previous research on the implications of these violations for the agent himself is summarized on the following page.

This page is deliberately left blank.

In other words, previous work assumed violations to depend mainly on critical "violation producing conditions" that need to be countered in order to foster rule-based performance (Johnson, Charlton, Oxley, & Newstead, 2013; Perneger, 2005; Phipps et al., 2008; Reason, 2000; Vincent et al., 2000). These conditions include factors such as inadequate training, time pressure, poor supervision and checking, lacking organizational safety culture, and conflict between management and staff (cf. Reason, 1995, for a detailed overview). In this framework, personal (i.e., individual) factors are only considered relevant as far as they allow for predicting whether a violation will occur or not. Consequently, research on such personal factors has highlighted predictors such as poor morale or beliefs that violations will not lead to negative outcomes (Reason, 1995). While certainly being useful on its own right, this approach needs to be complemented by research on the processes that occur in an individual's mind during violation behaviour. The present experiments are a first step towards understanding these processes.

In the following sections, I will present three different experimental approaches to studying intended rule violations. A first set of experiments uses a choice reaction task to study intended rule violations and probes for after-effects of such violations, similar to the PES commonly elicited by errors (Exp. 1-3). A second set of studies specifically targets cognitive conflict while committing violations by using movement trajectory analyses (Exp. 4-6). Finally, I explore the electrophysiological signature of rule violations in comparison to unintended errors (Exp. 7 and 8). Taken together, the corresponding findings indicate that rule violations are indeed separable from normal, rule-based behaviour on the one hand, and from unintended errors on the other hand. The results point towards several possible mechanisms underlying the observed effects which I will address in the General Discussion (Chapter 13).

PART 2: A BASIC EXPERIMENTAL APPROACH TO RULE VIOLATIONS

The following three experiments explore the effects and after-effects of intended rule violations in a simple choice reaction task. A specific focus of these experiments is thus (a) to compare rule violations to normal, rule-based behaviour, and (b) to investigate potential after-effects of such rule violations.

Regarding the first comparison – *rule violations vs. rule-based behaviour* – two contradicting predictions can be derived from the literature. On the one hand, research documenting automatic retrieval of S-R rules suggests that rule violations should be more complex than rule-based actions because the S-R rule needs to be actively suppressed in the former case (e.g., Hommel, 2000; Logan, 1988; see also Section 2.2). Accordingly, competition between the original S-R rule and the intention to violate this rule should result in cognitive conflict which, in turn, should increase RTs for rule violations. On the other hand, several studies indicated that learned S-R associations can be countered by merely re-instructing a different mapping rule (Colgan, 1970; Waszak, et al., 2013; Wilson, 1968).

Particularly relevant for the present experiment are two previous studies that have explicitly addressed *instructed* rule-reversals in choice reaction tasks (Schroder, Moran, Moser, & Altmann, 2012; Waszak et al., 2013). Waszak and colleagues (2013) asked their participants to respond to coloured shapes with either a left or a right keypress response in a task-

switching setup, randomly alternating between colour and shape classification. The stimulus set comprised six colours and six shapes, three of each being mapped to left responses and the remaining three being mapped to right responses. Participants practiced this mapping to reinforce automatic S-R translations that were measured in terms of congruency effects of the currently irrelevant dimension of the target (i.e., shape when participants were to respond to colour and colour when participants were to respond to shape). After a training block, the instructed mapping was switched for two shapes that, from this point onward, only appeared as distractors in the colour task but no longer as targets in the shape task. Still, congruency effects elicited by these shapes were reduced or even inverted following this re-instruction. This study clearly suggests that instructing changes of an S-R rule can be sufficient to block automatic retrieval.

A different picture emerges for the study of Schroder et al. (2012) who employed a flanker task with varying letters being assigned to either a left or a right response button. Each block of trials only used two letters and the stimulus set was changed every two blocks. Thus, participants performed two blocks with the first pair of letters, then two blocks with another pair of letters etc. Crucially, the letter-key mapping switched between the two blocks in each pair. Participants thus practiced a mapping in the first block of each pair and had to apply the reversed mapping in every other block. Responses in these "switch blocks" were consistently slower than responses in non-switch blocks, indicating that the initial mapping could not be suppressed completely.

The two studies sketched above yield different predictions regarding the impact of rule violations on RTs in a choice reaction task. Yet, the design of Schroder et al. (2012) seems to mirror the situation of rule violations more closely because this study addressed the application of reversed rules directly (in contrast of the indirect approach via target congruency effects by Waszak et al., 2013). I thus hypothesized violations to yield costs in terms of prolonged RTs and this prediction is tested in

Experiment 1 and 2. It should be noted, however, that this prediction is based on studies that investigated instructed rule reversals rather than explicit rule violations. This important difference will be addressed in Experiment 3.

The second effect of interest – *after-effects of rule violations* – draws on research that documented PES as a robust consequence of unintended failures to obey a rule (cf. Section 3.1). Accordingly, I chose to focus exclusively on PES and did not investigate other potential post-error adjustments such as post-error reduction of interference or post-error adjustments of accuracy, because these effects are less firmly established in the literature and tend to be less reliable (Danielmeier & Ullsperger, 2011).

Yet, one has to be aware of at least one methodological pitfall regarding the comparison of after-effects of rule violations and PES. Because PES is typically addressed in terms of the RT in correct trials following an error compared to the RT in correct trials following another correct trial, it seems plausible to construct a similar measure for rule violations. Accordingly *post-violation slowing* (*PVS*) can be computed as the RT for correct, rule-based responses following rule violations minus RT for correct, rule-based responses following another correct response. This measure is most likely to show pronounced effects for a trivial reason, however: Each transition from rule violation to rule-based behaviour necessarily engenders a task switch (or response switch in the terminology of Schroder et al., 2012) whereas two subsequent rule-based responses are task repetition trials. PVS is thus artificially inflated by switch costs (Allport et al., 1994; Rogers & Monsell, 1995) and a proper assessment of PVS can only be achieved by a smart experimental design, a first attempt of which is the following Experiment 1.

4 Experiment 1: Effects and after-effects of violating a rule

To allow for an unbiased assessment of both, effects and after-effects of rule violations, I employed a paradigm that involved three different tasks. Each of the three tasks used its own, distinct set of two target stimuli (**Fig. 4**), whereas the response set was the same across tasks and participants always responded with the left or right index finger. One stimulus of each task was mapped to a left response key and a second stimulus was mapped to a right response key; the task sequence was randomized across trials.

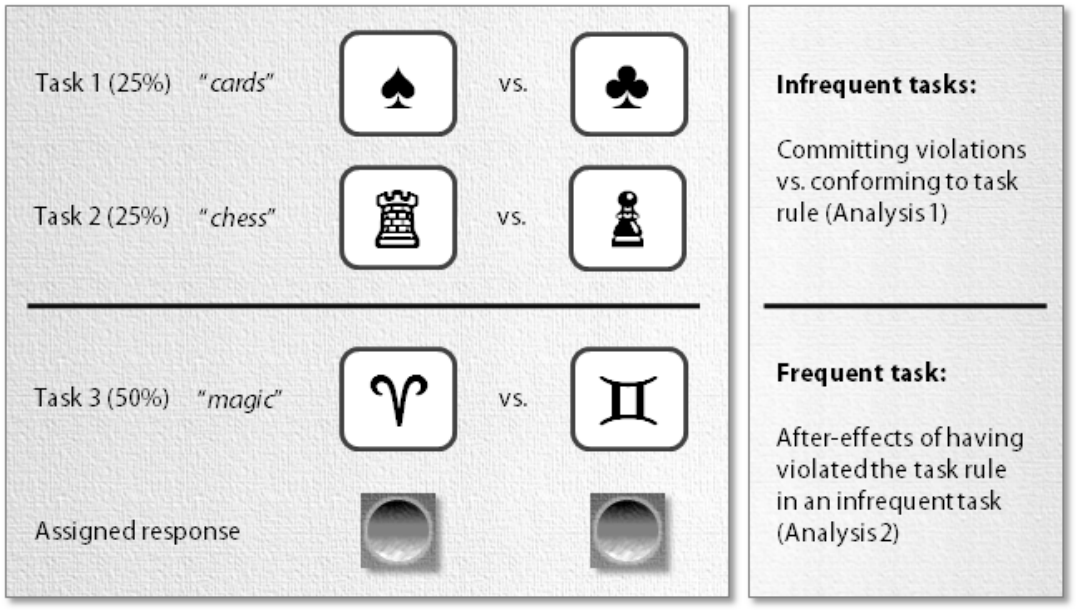

Fig. 4. Stimulus-set and exemplar mapping for the three tasks in Experiment 1. In some blocks, participants were instructed to violate the mapping rule of either Task 1 or Task 2 whenever this task came up. This setup allows studying (a) possible differences between rule violations and normal, rule-based behaviour when actually performing a response and (b) the after-effects of having violated a rule.

One of the three tasks appeared in 50% of the trials (the "magic" task in **Fig. 4**), whereas the other two tasks appeared in 25% each. Crucially, participants were instructed to violate the original mapping rule of either Task 1 or Task 2 in separate blocks of trials. This setup allows studying (a) possible differences between rule violations and normal, rule-based behaviour when actually performing a response and (b) the after-effects of having violated a rule on unrelated behaviour.

The chosen operationalization of rule violations obviously maximizes internal validity: It allows to clearly identify rule violations and to separate these intended failures to follow the mapping rule from unintended errors. This focus on internal validity, however, comes at the cost of decreased external validity: Outside the laboratory, violations are rarely committed because one is told to violate a rule; rather, they are driven by more intrinsic factors. As a consequence, the behaviour studied here does not comply fully with Reason's (1990, 1995) taxonomy of rule violations, a conceptual issue that I will address in Experiment 6. Still, the present setup allows for clear conclusions about the effects of labelling a very simple behaviour as rule violation.

4.1 Method

4.1.1 Participants, apparatus, and stimuli

Twenty-four participants were recruited and received either course credit or monetary compensation (mean age: 24.8 years, 20 females, 2 left-handers). All participants were naive concerning the hypothesis underlying the experiment.

Stimuli appeared on a 17'' monitor and participants responded on a left and a right response key. Stimuli were two card hands ("cards task"; spades vs. clubs), two chess pieces ("chess task"; rook vs. pawn), and two astrological symbols ("magic task"; Aries vs. Gemini), displayed centrally in

60 pt MS Gothic font, corresponding to a stimulus size of approximately 1 cm x 2 cm. Two stimulus pairs were designated as infrequent tasks (Task 1 and 2; each appearing in 25% of the trials) whereas the remaining stimulus pair was designated as frequent task (Task 3; appearing in 50% of the trials).

4.1.2 Procedure

Trials started with a fixation dot in the center of the screen (500 ms), followed by the target stimulus. The target stayed on screen until a response was given or until 800 ms had elapsed (whichever occurred first). The next trial started after an inter-trial interval (ITI) of 500 ms. Failures to apply the correct mapping rule, i.e., errors and violations, immediately triggered a feedback message ("Fehler", German for "error", displayed for 500 ms).

The experiment consisted of 12 blocks of 80 trials each (960 trials in total). One task appeared in 50% of the trials throughout the experiments whereas the other two tasks appeared in 25% each. The critical manipulation was implemented in terms of different instructions across blocks. The first block always was a training block in which participants simply practiced the mapping rules and responded according to these rules in all three tasks. This block was followed by two error blocks: Participants were now instructed to violate the mapping rule for one of the infrequent tasks by written instructions such as "*The card task and the chess task require correct responses. When the magic task comes up, you have to commit an error by intention!*".[3]

[3] As a direct consequence of this design, rule violations comprise two aspects: First, the labelling of the response as rule violation and, secondly, the reversed mapping of stimuli to responses. Thus, any potential effects of rule violations might simply be a function of applying a reversed mapping rule that differs from what was learned previously. This important alternative explanation is addressed in Experiment 3.

The violation instruction referred to each of the two infrequent tasks in two subsequent blocks (Block 2 and 3), followed by a training block (Block 4) and, again, by two blocks with violation instructions (Blocks 5 and 6), etc. This procedure ensured that both infrequent tasks appeared with standard instructions and violation instructions for each participant.

4.1.3 Data treatment

I ran two different analyses to probe for the effects of violating a rule: The first one targeted the effects for actually committing a violation and the second one targeted the after-effects of having violated a rule. Both analyses were run only on the data of those blocks in which participants violated a rule (thus omitting the data of the training blocks). For all analyses, I corrected for outliers by removing trials with RTs that deviated more than 2.5 standard deviations from the corresponding cell mean (1.5% of the trials across all analyses).

The analysis of *committing violations* used only data of trials with an infrequent task in the current trial (trial n) as well as in the preceding trial (trial n-1). These data consist of four possible sequences: $Correct_{Infrequent}$ ➤ $Correct_{Infrequent}$, Violation ➤ Violation, Violation ➤ $Correct_{Infrequent}$, and $Correct_{Infrequent}$ ➤ Violation. Accordingly, I analysed mean RTs with a 2 x 2 repeated-measures Analysis of Variance (ANOVA) with the factors rule compliance (rule-based responses vs. violations) and task sequence (task switch vs. task repetition). Increased conflict when violating a rule should be mirrored in a significant main effect of rule compliance. Most importantly, a putative difference between rule-based responses and violations should be present also for the direct comparison of task repetition trials (i.e., violations following another violation vs. correct responses following another correct response). This contrast was addressed by a paired-samples *t*-test.

The logic behind this pairwise comparison of repeated violations and repeated correct responses in the infrequent tasks draws on two arguments. First, assuming that normal, correct responses represent a

default mode for participants, switching towards rule violations might incur costs simply due to this task switch (e.g., Allport et al., 1994; Rogers & Monsell, 1995). Performance for repeated rule violations, however, is less likely to be affected by such switch costs and represents a purer measure of the impact of rule violations. Secondly, a similar argument applies to switches from violations to correct responding in the other infrequent task. Here, RT might be increased due to potential slowing after violations. Thus, to evaluate the effects of actually violating a rule it seems imperative to compare trials in which the same task has been performed repeatedly.

The second analysis, targeting *after-effects of having violated* a rule, focused on the data of the frequent task in trial n. RTs in this task were analysed as a function of the preceding trial type and, because the omnibus ANOVA results are rather uninformative in this case, I will report four specific contrasts, each of them addressing a specific comparison to the RT of correct responses following another correct response of the frequent task. More precisely, these contrasts address the following four effects.

- Post-error slowing (PES):

$$RT_{Correct(Frequent)\rightarrow Correct(Frequent)} - RT_{Error(Frequent)\rightarrow Correct(Frequent)}$$

- Switch costs due to *task switches (TS)*:

$$RT_{Correct(Frequent)\rightarrow Correct(Frequent)} - RT_{Correct(Infrequent)\rightarrow Correct(Frequent)}$$

- The combined effects of task switch and errors (TS+PES):

$$RT_{Correct(Frequent)\rightarrow Correct(Frequent)} - RT_{Error(Infrequent)\rightarrow Correct(Frequent)}$$

- Post-violation slowing (PVS),:

$$RT_{Correct(Frequent)\rightarrow Correct(Frequent)} - RT_{Violation(Infrequent)\rightarrow Correct(Frequent)}$$

4.2 Results

On average, participants committed errors in 13.6% of the trials of the frequent task and failed to respond in another 2.1% of the trials. When prompted to perform correctly in the infrequent tasks, participants committed errors in 7.0% of the trials and failed to respond in another 4.6% of the trials. Both types of errors (i.e., errors of commission and errors of omission) were used in the analyses to avoid missing data for some participants (especially for errors in the infrequent task).

When prompted to violate the mapping rule in an infrequent task, participants mistakenly followed the original rule in overall 15.5% of the trials. This relatively high rate seems to suggest that the rule-based response is at times retrieved automatically against the agent's intention. Even though this interpretation would fit nicely to the results of the following RT analysis, this pattern of results can also be accounted for by a number of alternative explanations. For instance, again assuming rule-based responses to be a behavioural default, participants might also fail simply to retrieve the instructions to violate the mapping rule in several trials. Inspection of the raw data also suggested this effect to be mainly driven by individual participants who failed to violate the mapping rule in the majority of trials of a single block – presumably because they did not attend to the instructions between blocks carefully enough.

For these reasons, any data regarding frequencies of violations and mistakenly correct responses are to be taken with caution. I will therefore not elaborate on these data in any of the following experiments but rather focus on the results of the RT analyses that can be interpreted more clearly. The complete descriptive statistics for these RT analyses are listed in **Appendix A**.

4.2.1 Analysis 1: Committing violations

Means for violations and correct responses in the infrequent task are plotted in **Figure 5**. Descriptively, RTs for rule violations were slower than for correct responses for both, task repetition trials and task switch trials.

These observations were qualified by a 2 x 2 ANOVA with the factors rule compliance and task sequence, which clearly yielded a significant main effect of rule compliance, $F(1, 23) = 20.52$, $p < .001$, $\eta_p^2 = .47$. Additionally, responses were faster in repetition trials than in switch trials, $F(1, 23) = 124.53$, $p < .001$, $\eta_p^2 = .84$, whereas the interaction of rule compliance and task sequence did not approach significance, $F(1, 23) = 1.00$, $p = .328$, $\eta_p^2 = .04$.

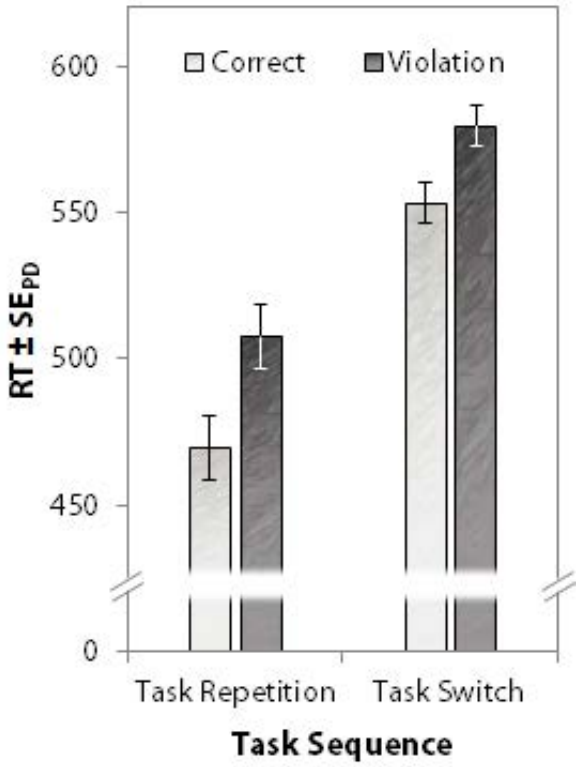

Fig. 5. The effects of committing violations versus performing rule-based responses in Experiment 1. Error bars show the standard error of paired differences, computed separately for task repetitions and task switches (Pfister & Janczyk, 2013).

Moreover, the direct comparison of violations and rule-based responses for repetition trials only was significant as well (cf. the left-hand side of **Fig. 5**), $t(23) = 3.43$, $p = .002$, d = 0.70, indicating that also repeated violations took longer to initiate than repeated rule-based responses.

4.2.2 Analysis 2: After-effects of having violated a rule

The four effects of interest are shown in **Figure 6** and, as can be seen from the figure, all four effects were highly significant – PES: $t(23) = 3.268$, $p = .003$, d = 0.67; TS: $t(23) = 11.19$, $p < .001$, d = 2.28; TS+PES: $t(23) = 9.81$, $p < .001$, d = 2.00; PVS: $t(23) = 12.61$, $p < .001$, d = 2.57. Most relevant for the present experiment, however, TS and PVS did not differ significantly from one another, $t(23) = 1.47$, $p = .156$, d = 0.30.

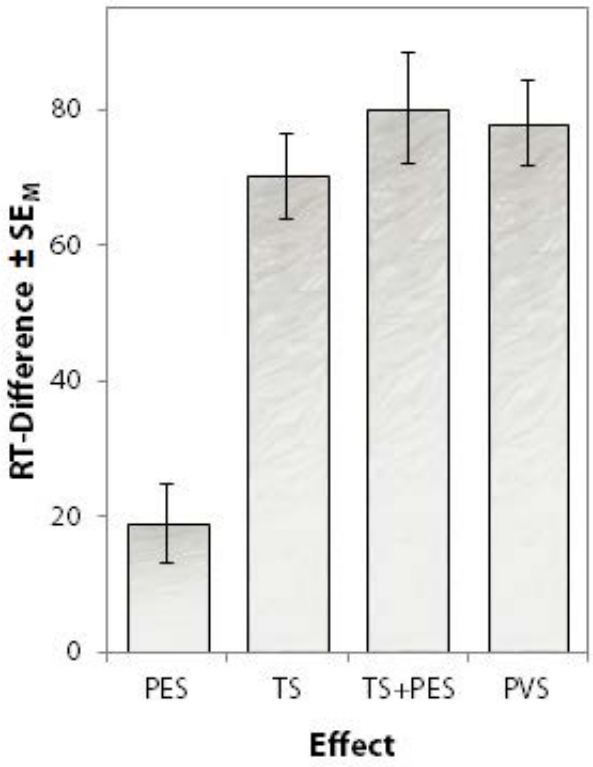

Fig. 6. Sequential effects on correct responses in the frequent task of Experiment 1. Error bars indicate standard errors of the mean differences (Pfister & Janczyk, 2013).

4.3 Discussion

Experiment 1 investigated effects and after-effects of rule violations in a simple choice reaction task. The data clearly attest considerable difficulty of violating a rule in terms of increased RTs when committing a violation as compared to normal responses in a similarly infrequent task. Most notably, this was even true when only considering task repetition trials, i.e., when any types of switch costs cannot explain potential differences in RTs. Thus, Experiment 1 demonstrates a clear performance difference between rule violations and normal, correct responses. The after-effects of having violated a rule, however do not seem to exceed the costs associated with a normal task switch (Allport et al., 1994; Rogers & Monsell, 1995) as suggested by the non-significant comparison of switch costs due to task switching (TS) and PVS.

It should be noted, however, that the design of Experiment 1 allows for a critical alternative explanation for the effects observed in the first analysis because rule violations triggered error messages just as unintended errors did. This procedure was implemented to reinforce the concept of violating a rule but might have had other side effects. For instance, merely anticipating the negative feedback or one's emotional reaction to it might have caused the observed conflict (e.g., Baumeister, Vohs, DeWall, & Zhang, 2007). This alternative explanation is addressed in Experiment 2.

5 Experiment 2: Controlling for feedback

In Experiment 2, I did not give any feedback for either rule violations or unintended errors. If the violation-specific effects observed in Experiment 1 were mainly driven by anticipated or observed feedback, rule violations should behave as normal, rule-based responses for all analyses. By contrast, similar effects as in Experiment 1 should emerge if the observed difficulty is indeed caused by certain peculiarities of violating a rule.

5.1 Method

A new sample of twenty-four participants was recruited for either course credit or monetary compensation (mean age: 24.3 years, 17 females, 1 left-hander). All participants were naive concerning the hypothesis underlying the experiment.

Stimuli, apparatus, and procedure were identical to Experiment 1 with the only exception that no feedback was given throughout the experiment (neither for unintended errors nor for violations). All analyses were conducted as described for Experiment 1 and I excluded 1.6% of the trials as outliers due to the same criterion.

5.2 Results

Commission errors occurred in 8.9% of the trials of the frequent task and response omissions occurred in another 3.5% of the trials. When instructed to respond correctly in an infrequent task, participants committed errors in 6.0% of the trials and failed to respond in another 6.5% of the trials. Again, both types of errors were used in the analyses to avoid missing data. By contrast, when prompted to violate the mapping

rule in an infrequent task, participants mistakenly applied the original rule in 12.9% of the trials. As for Experiment 1, the complete descriptive statistics for all following RT analyses are listed in **Appendix A**.

5.2.1 Analysis 1: Committing violations

Replicating the analysis of Experiment 1, the 2 x 2 ANOVA on the data of the infrequent tasks yielded a significant main effect of rule compliance with violations again being slower than rule-based responses, $F(1, 23) = 33.56$, $p < .001$, $\eta_p^2 = .59$ (**Fig. 7**). Responses in repetition trials were faster than in switch trials, $F(1, 23) = 111.75$, $p < .001$, $\eta_p^2 = .83$, whereas the interaction of rule compliance and task sequence was not significant ($F < 1$). Similarly, the direct comparison of violations and rule-based responses for repetition trials only was significant, $t(23) = 3.92$, $p < .001$, $d = 0.80$.

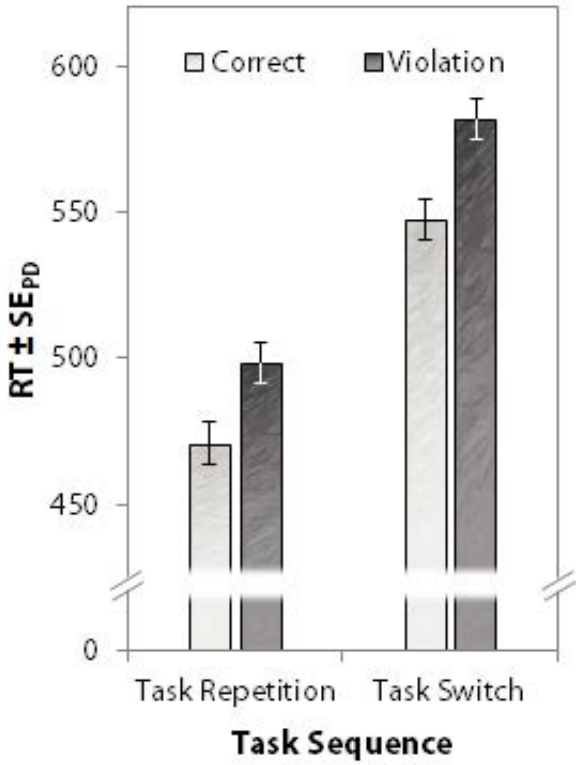

Fig. 7. The effects of committing violations versus performing rule-based responses in Experiment 2. The design was similar to Experiment 1 but neither errors nor violations triggered any feedback. Error bars show the standard error of paired differences, computed separately for task repetitions and task switches.

5.2.2 Analysis 2: After-effects of having violated a rule

The four effects of interest are shown in **Figure 8**. Unexpectedly, PES did not differ significantly from zero, $t(23) = 1.68$, $p = .107$, d = 0.34, whereas the remaining effects were clearly significant – TS: $t(23) = 11.57$, $p < .001$, d = 2.36; TS+PES: $t(23) = 7.35$, $p < .001$, d = 1.50; PVS: $t(23) = 11.97$, $p < .001$, d = 2.44. Replicating the findings of Experiment 1, TS and PVS did not differ significantly from one another, $t(23) = 1.71$, $p = .100$, d = 0.35.

To follow up on the unexpected result of absent PES, I also computed the same measure for the training blocks where PES was clearly significant (27 ms), $t(23) = 2.88$, $p = .008$, d = 1.70. Interestingly though, a 2 x 2 repeated-measures ANOVA with the factors block type (training vs. violation) and preceding trial type (correct vs. error in the frequent task), did not yield a significant interaction, $F(1, 23) = 1.65$, $p = .212$, $\eta_p^2 = .07$.

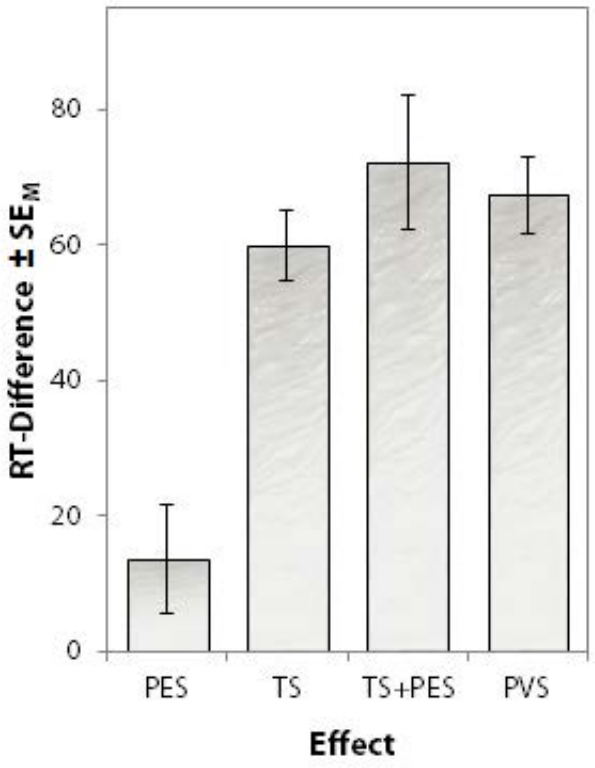

Fig. 8. Sequential effects on correct responses in the frequent task of Experiment 2. Error bars indicate standard errors of the mean differences.

Instead, the main effect of preceding trial type was significant, $F(1, 23) = 8.62$, $p = .007$, $\eta_p^2 = .27$. In addition, responses were descriptively faster in training blocks than in violation blocks (446 ms vs. 458 ms), even though the main effect of block type was not significant, $F(1, 23) = 2.28$, $p = .145$, $\eta_p^2 = .09$.

5.3 Discussion

Experiment 2 replicated the central findings of Experiment 1. Again, violating a rule took considerably longer than performing a rule-based response whereas I did not observe any violation-specific after-effects on behaviour. Before drawing any strong conclusions from this finding, however, Experiment 3 addresses a final alternative explanation by asking whether violating a rule is not only distinct from normal-rule based behaviour but also from adhering to *changing* rules. As mentioned above, rule violations as operationalized in Experiment 1 and 2 differ not only with regard to labelling ("rule violation" vs. "correct responses") but rule violations also imply working against a previously learned mapping. Even though this property clearly belongs to any kind of rule violation, similar effects on performance might occur when rules actually change. Experiment 3 tests whether this speculation holds. Afterwards, I will compare the obtained effects across the first three experiments to pinpoint potential effects that are indeed specific for rule violations as compared to rule-based behaviour.

6 Experiment 3: Changing rules

As described above, two previous studies have examined the impact of instructed rule-reversals on behaviour (Schroder et al., 2012; Waszak et al., 2013). These studies employed different choice reaction tasks and came to opposing conclusions. Most importantly for the present question, Schroder et al. (2012) found instructed rule-reversals to deteriorate performance. Arguably, any effects of violating a rule found in Experiment 1 and 2 could thus be driven entirely by difficulty to overcome the previously learned mappings, independently of whether the behaviour is actually to be seen as a rule violation or as performance according to changed task rules.

To address this potential alternative explanation, I replicated Experiment 2 with changed instructions. Participants were no longer asked to violate the mapping rule of the infrequent tasks. Rather, they were informed that the mapping rule for one of the infrequent tasks was reversed for some of the blocks. This procedure ensured that participants carried out the very same actions as in Experiment 2 with the only difference that these actions were no longer labelled as rule violations but simply as performance according to a changed mapping rule.

6.1 Method

A new sample of twenty-four participants was recruited for either course credit or monetary compensation (mean age: 28.9 years, 14 females, 2 left-handers). All participants were naive concerning the hypothesis underlying the experiment.

Stimuli, apparatus, and procedure were identical to Experiment 2 with the only exception that the instructions did no longer refer to rule violations. Rather, participants were now told that the mapping for one of the tasks would reverse for the following block. Accordingly, the current

correct mapping for each of the three tasks was displayed at the beginning of each block. This procedure ensured that the participants performed exactly the same task as in Experiment 2 but with a different labelling of the responses. All analyses were again conducted as described for Experiment 1 and I excluded 1.7% of the trials as outliers due to the same criterion. When describing the results, I will use the term post-reversal slowing (PRS) instead of the previously used PVS.

6.2 Results

Commission errors occurred in 6.6% of the trials of the frequent task and response omissions occurred in another 2.7% of the trials. When instructed to respond correctly in an infrequent task, participants committed errors in 5.0% of the trials and failed to respond in another 6.1% of the trials. Again, both types of errors were used in the analyses to avoid missing data. By contrast, when prompted to violate the mapping rule in an infrequent task, participants mistakenly applied the original rule in 10.4% of the trials. As for the previous experiments, the complete descriptive statistics for all following RT analyses are listed in **Appendix A**.

6.2.1 Analysis 1: Applying a reversed mapping

The 2 x 2 ANOVA on the data of the infrequent tasks yielded a significant main effect of rule compliance with responses using the reversed mapping being slower than normal responses, $F(1, 23) = 4.74$, $p = .040$, $\eta_p^2 = .17$ (**Fig. 9**). Responses were faster in repetition trials than in switch trials, $F(1, 23) = 138.32$, $p < .001$, $\eta_p^2 = .86$, whereas the interaction of rule compliance and task sequence was not significant, $F(1, 23) = 2.31$, $p = .143$, $\eta_p^2 = .09$. Considered separately, the direct comparison of responses using the reversed and the normal mapping was significant for task switches, $t(23) = 2.25$, $p = .035$, $d = 0.46$, but not for task repetitions, $t(23) = 0.90$, $p = .378$, $d = 0.18$.

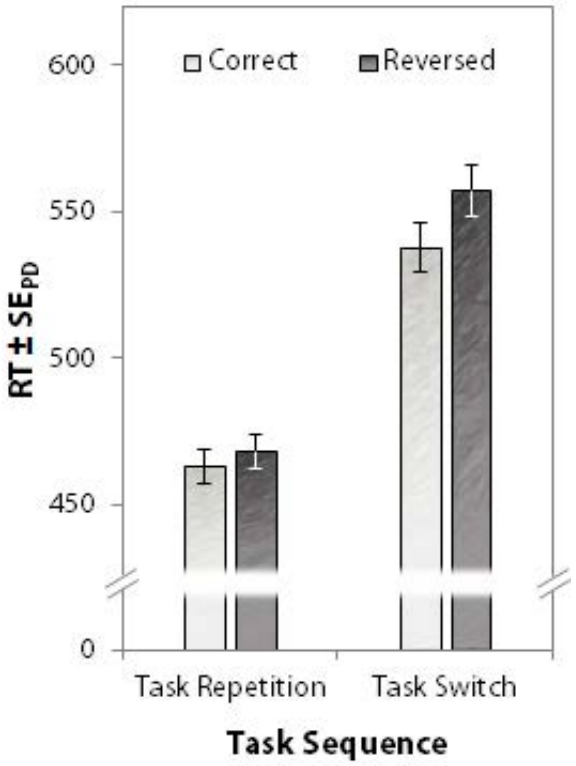

Fig. 9. The effects of acting according to a reversed mapping rule as compared to normal rule-based responses in Experiment 3. Error bars show the standard error of paired differences, computed separately for task repetitions and task switches.

6.2.2 Analysis 2: After-effects of reversed mapping rules

The four effects of interest are shown in **Figure 10**. As in Experiment 2, PES did not differ significantly from zero, $t(23) = 0.89$, $p = .384$, d = 0.18, whereas the remaining effects were again significant – TS: $t(23) = 11.59$, $p < .001$, d = 2.37; TS+PES: $t(23) = 9.23$, $p < .001$, d = 1.88; PRS: $t(23) = 12.52$, $p < .001$, d = 2.56. Interestingly, TS and PRS did differ significantly from one another for the data of Experiment 3, $t(23) = 2.79$, $p = .010$, d = 0.57 with PRS being slightly *larger* than TS.

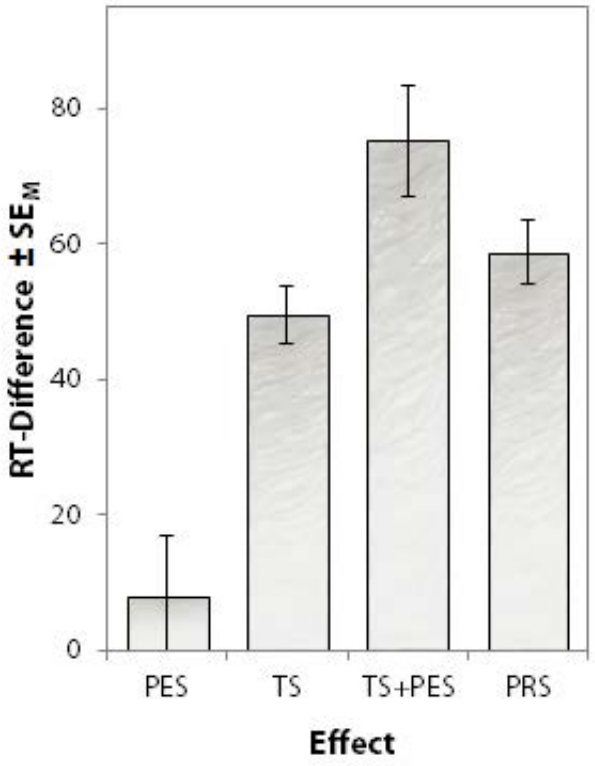

Fig. 10. Sequential effects on correct responses in the frequent task of Experiment 3. Error bars indicate standard errors of the mean differences.

As for Experiment 2, I also computed a separate estimate of PES for the training blocks where the effect was significant again (25 ms), $t(22) = 3.54$, $p = .001$, d = 1.35.[4] Similarly, a 2 x 2 repeated-measures ANOVA with the factors block type (training vs. reversal) and preceding trial type (correct vs. error in the frequent task), did not yield a significant interaction, $F(1, 22) = 1.78$, $p = .196$, $\eta_p^2 = .07$, whereas the main effect of preceding trial type was significant again, $F(1, 22) = 8.38$, $p = .008$, $\eta_p^2 = .28$. In addition, responses were descriptively faster in training blocks than in reversal blocks (444 ms vs. 456 ms), giving rise to a marginally significant main effect of block type, $F(1, 22) = 4.17$, $p = .053$, $\eta_p^2 = .16$.

[4] One participant did not commit a single error in the training blocks. Consequently, PES was computed for the remaining 23 participants only.

6.3 Discussion

In Experiment 3, I instructed participants to perform according to a reversed mapping rule (as compared to violating the rule in question as in Experiment 1 and 2). Most notably, when performing two reversed-rule responses in a row, performance with the second response did not differ from performance with normal, rule-based responses. Response costs were visible, however, when switching from a normal, rule-based response to a response according to a reversed mapping rule. It thus seems as if labelling the reversed responses as correct behaviour did indeed reduce the impact on behaviour (as compared to rule violations) even though some residual effects remained.

As speculated in Section 4.1.3, switching from the default mode of normal, rule-based behaviour seems to be an effortful process. Implementing a changed and less thoroughly practiced mapping rule during this process might add additional difficulty which might, in turn, drive the effects observed for task switches. Once the representation of the changed rule is activated, however, it does not seem to affect performance negatively – in contrast to the enduring effects of rule violations observed in Experiment 1 and 2.

Unexpectedly, I also observed a significant difference between post-reversal slowing and task-switching, indicating additional after-effects of having performed according to the reversed mapping. This effect was not significant in Experiment 1 and 2 even though similar tendencies emerged for these experiments. I therefore decided to run two additional between-experiment analyses to clarify what can be safely concluded from the present experiments. This analysis is presented in the following section.

7 Preliminary conclusions

The two most informative measures of Experiments 1-3 were (a) the difference between repeatedly performing a rule violation (or reversed rule response) as compared to normal, rule-based behaviour and (b) the after-effects in terms of PVS/PRS as compared to normal switch costs. **Figure 11** summarizes these measures by showing the difference between repeated rule violations (or reversed rule responses) and repeated responses according to the initial mapping rule (left panel) as well as the effects of task-switching and PVS/PRS (right panel).

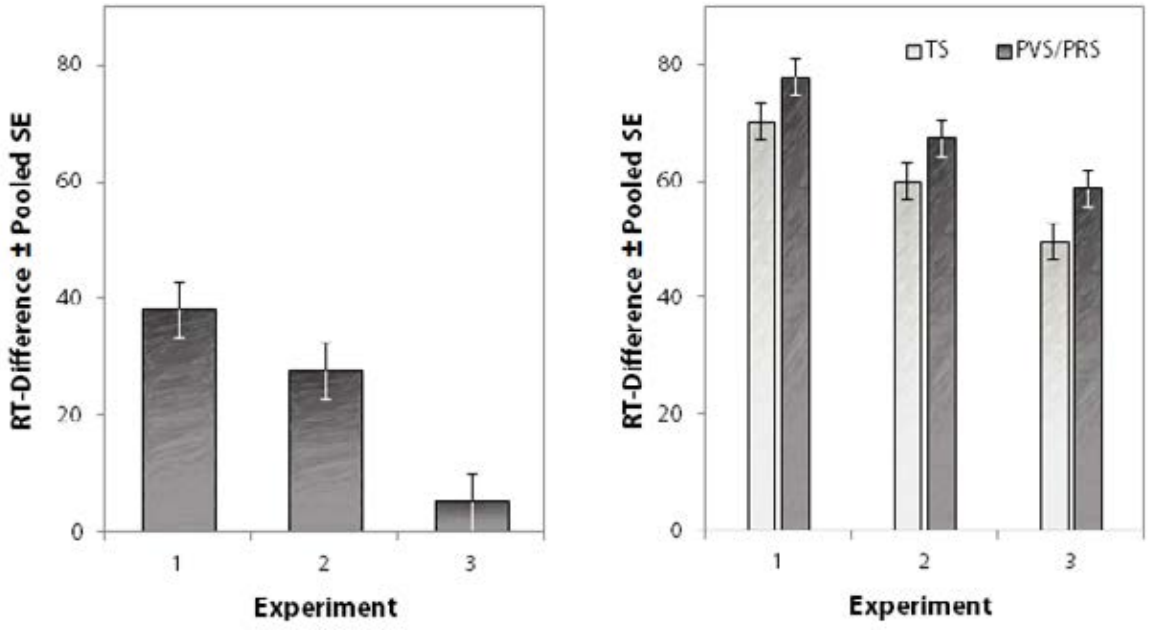

Fig. 11. Comparison of the two main measures of interest across Experiments 1-3. The left panel shows the effects of repeatedly violating a rule (Exp. 1 and 2) or repeatedly applying a reversed mapping rule (Exp. 3) as compared to normal, rule-based behaviour in infrequent tasks. Error bars indicate the pooled between-subjects standard error as derived from the error term of a one-way ANOVA. The right panel shows the after-effects of switching from an infrequent task to the frequent task separately for normal task switches (light gray bars) and task swtiches after violations (dark gray bars; Exp. 1 and 2) or reversed rule responses (Exp. 3). Error bars represent pooled standard errors as derived from separate one-way ANOVAs for task switches and rule violations / reversed rule responses, respectively (cf. Estes, 1997).

The effects of repeatedly committing a violation (or reversed rule response) clearly differed between experiments as attested by a one-way ANOVA on the difference scores, $F(2, 69) = 4.13$, $p = .020$, $\eta^2 = .11$. Separate *t*-tests for independent samples suggested the effect of violations to be comparable for Experiment 1 and 2, $t(46) = 0.80$, $p = .428$, $d = 0.23$, whereas the effects of Experiment 2 and 3 differed significantly, $t(46) = 2.47$, $p = .017$, $d = 0.71$. These results confirm the visual impression of **Figure 11** (left panel): Labelling an action as violation clearly affected behaviour even for repeated rule violations whereas labelling the same action as adhering to a reversed mapping rule caused less difficulty in this situation.

A second between-experiment analysis targeted the after-effects of rule violations and reversed rule responses (**Figure 11**, right panel) by means of a 2 x 3 split-plot ANOVA with the within-subjects factor preceding trial (normal task switch vs. violation / reversed rule response) and the between-subjects factor experiment. Most importantly, this analysis revealed a main effect of preceding trial, $F(1, 69) = 10.30$, $p = .002$, $\eta_p^2 = .13$, but the interaction of preceding trial and experiment did not approach significance, $F(2, 69) = 0.51$, $p = .951$, $\eta_p^2 < .01$. Additionally, the main effect of experiment indicated a nearly linear decrease of effects from Experiment 1 to Experiment 2 and 3, $F(2, 69) = 4.01$, $p = .021$, $\eta_p^2 = .11$. These results call for a different interpretation than the pattern of significances for each individual experiment. In fact, the increased power of the pooled analysis suggests a small additional after-effect of rule violations and reversed rule responses alike as compared to typical task switches. Most likely, this additional after-effect indicates difficulty in overcoming the previously learned mapping rule and activating a representation of the currently intended behaviour – a process that is obviously shared by rule violations and responses according to a changed mapping rule.

The emerging pattern of results for after-effects of rule violations is thus rather mixed and does not allow for any clear-cut interpretations; a definite answer to this question does certainly require a different experimental approach. A promising avenue for these experiments might be the comparison of different ITIs, which has been shown to be an important factor for post-error adjustments of behaviour (Jentzsch & Dudschig, 2009; Danielmeier & Ullsperger, 2011). More precisely, PES is much stronger at short ITIs (e.g., 50-100 ms in the case of Jentzsch & Dudschig, 2009) as compared to longer intervals (e.g., 1000 ms), with the present ITI being situated in between those values (500 ms). Furthermore, in the case of PES, different intervals seem to index distinct processes: Response-monitoring at short intervals and strategic adaptations of response criteria at long intervals. Even though these processes do certainly not map directly on possible processes that follow rule violations, investigating the time course of such after-effects might still be an informative endeavour.

Furthermore, the rather short response deadline of the present experiments might have worked against potential after-effects of rule violations (see Steinhauser & Hübner, 2006, for a related discussion for PES). More precisely, if after-effects of rule violations primarily affected strategic parameters such as speed-accuracy trade-offs, the present response deadline might have limited possible variations on the speed-accuracy dimension, thus working against possible post-violation adjustments.

A much clearer picture emerges for the effects of committing a rule violation right at the moment it takes place. Here, the results of Experiment 1-3 suggest a considerable degree of conflict that seems to be rather specific for rule violations (as compared to reversed rule responses). This perspective on the individual agent goes beyond previous accounts for rule violations (Perneger, 2005; Reason, 2000; Vincent et al., 2000) but it does not speak to the processes underlying this conflict. This question is addressed in Experiment 4 and 5.

PART 3: VIOLATING A RULE CHANGES THE WAY WE MOVE

The results of Experiment 1-3 clearly suggest that intended rule violations create some form of cognitive conflict for the agent who violates the rule. Most importantly, this was true when participants had violated a mapping rule not only once but repeatedly, i.e., when they did not need to switch from the assumed default mode of responding according to the rules. Yet, the results do not inform about the mechanisms underlying this effect. As noted above, a likely source of such cognitive conflict could be difficulty to suppress the original mapping rule (e.g., Logan, 1988; Schroder et al., 2012). A direct test of this hypothesis, however, calls for a refined experimental approach.

In Experiments 4 and 5, I thus used a mouse-tracking setup in which participants pointed to one or another target stimulus. The corresponding trajectory data provide detailed information about cognitive conflict during task performance (Freeman, Dale, & Farmer, 2011; Song & Nakayama, 2009). Following these analyses, Experiment 6 examines the external validity of the reported findings in a task that did not enforce rule violations directly and allows thus for analysing spontaneous violations.

8 Experiment 4: How violations are performed

Analyses of trajectory data have become increasingly popular in behavioural research because they provide a window on the online competition of different processes. For instance, when participants have to select a target picture following a verbal instruction, their movement trajectories are strongly biased towards phonologically similar distractors on the screen as compared to phonologically unrelated distractors, providing an index of on-going conflict between the stimuli (Spivey, Grosjean, & Knoblich, 2005).

Similar approaches have been reported in a variety of settings, covering such diverse fields as perceptual decision making (Resulaj, Kiani, Wolpert, & Shadlen, 2009; Song & Nakayama, 2008), word and sentence processing (Dale & Duran, 2011; Dale, Kehoe, & Spivey, 2007; McKinstry, Dale, & Spivey, 2008), and person perception (Freeman & Ambady, 2009, 2011; Freeman, Ambady, Rule, & Johnson, 2008)

Experiment 4 used a similar mouse-tracking setup to study cognitive conflict during rule violations in which participants performed a simple pointing task with the computer mouse. Departing from a home area in the bottom center of the screen, they moved towards a target area to the upper left or right (based on a target stimulus). Crucially, at the beginning of each trial, participants indicated whether to act according to the mapping rule or whether to violate the rule and commit an error by intention (**Fig. 12**).[5] Based on the findings of Experiment 1 and 2, I expected movement trajectories during rule violations to be attracted towards the rule-based response.

[5] The compliance prompt obviously allowed participants to prepare for the upcoming action. This feature of the design of Experiment 4 serves the same purpose as the analysis of task repetition trials for Experiments 1-3 (cf. Section 4.1.3 for a description of the underlying logic).

Moreover, I compared the performance in this *violation group* to a second group that received different instructions. Similarly to Experiment 3, participants in this group chose whether to perform a standard task or a task that employed the exact opposite stimulus-response mapping (*reversed rule group*). Accordingly, similar actions were labelled as violations for the violation group and as different rule-based responses for the reversed rule group. Just as for the comparison of Experiment 2 and 3, I expected stronger effects for the violation group than for the reversed rule group – most importantly concerning the corresponding movement trajectories.

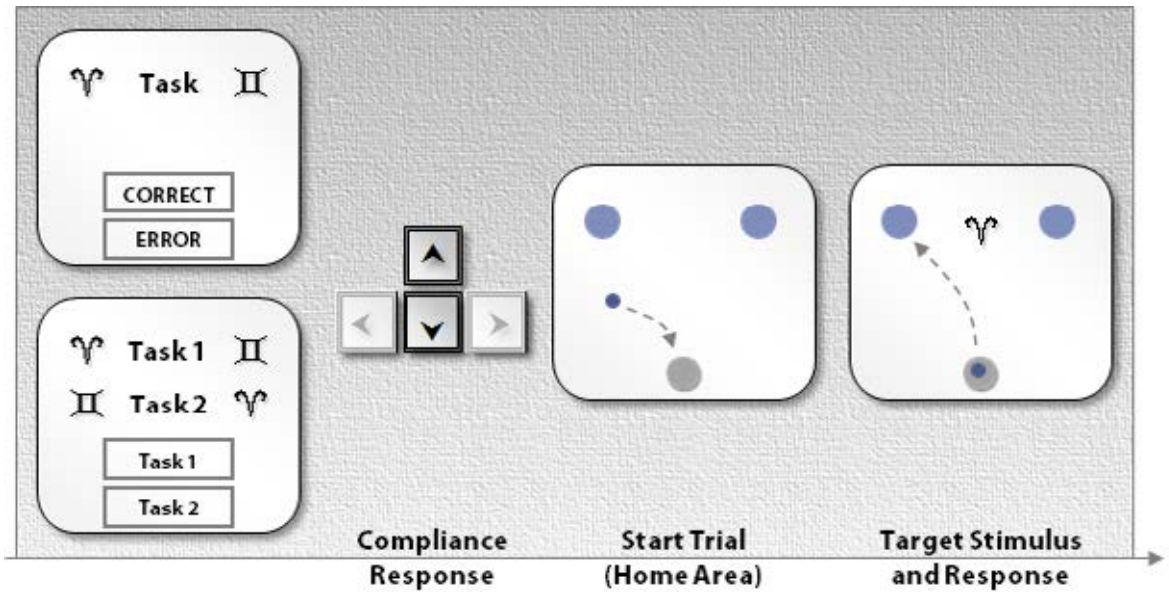

Fig. 12. Trial procedure of Experiment 4. Crucially, at the beginning of each trial, participants indicated their intention for the upcoming response. That is, participants in the *violation group* indicated whether they would respond according to the mapping rule or whether they would violate this rule and commit an error by intention. By contrast, participants in the *reversed rule group* indicated whether they would perform either a Task 1 or a Task 2, whereas the two tasks employed the exact opposite stimulus-response mapping. Accordingly, similar actions were labelled as violations for the violation group and as different but rule-based responses for the reversed rule group. The target stimulus set on after the participant had given the compliance response and had moved the mouse cursor to the home area in the bottom center of the screen.

8.1 Method

8.1.1 Participants, Apparatus, and Stimuli

Twenty participants were assigned to the violation group (mean age = 20.5 years, 14 females, 2 left-handers) and another 20 participants were assigned to the reversed rule group (mean age = 21.4 years, 17 females, 2 left-handers).

Participants operated a standard computer mouse with their right hand and placed their left hand on the arrow keys of the keyboard. Stimuli appeared on a 17'' computer monitor at a viewing distance of about 60 cm. Target stimuli were two astrological symbols (Aries vs. Gemini, displayed in 60 pt. MS Gothic font), mapped to a left and a right response, respectively (counterbalanced across participants).

8.1.2 Procedure

Each trial started with a compliance prompt (left-most screen in **Fig. 12**). For the violation group, the bottom of the screen showed two boxes that contained the German words "KORREKT" (correct) and "FEHLER" (error). The locations of the correct box and the error box were counterbalanced across participants but constant for each individual. Participants responded whether they would comply to the mapping rule by pressing either the up- or the down-key on the computer keyboard with their left hand. For the reversed rule group, the boxes of the compliance prompt contained the labels "Task 1" and "Task 2" instead of correct and error and both possible mapping rules were presented in the upper half of the screen. As in the violation group, participants in the reversed rule group pressed the up- or the down-key of the keyboard to indicate which of the two tasks they would like to perform in the upcoming trial. Thus, choosing Task 2 was identical to choosing to violate the mapping rule (i.e., committing an error by intention) except for the different framing of the response.

Terminating the compliance cue made three areas appear: The home area in the bottom center and the two target areas to the upper right and upper left of the screen. From this point onward, the mouse cursor was displayed as a small circle (0.5 cm in diameter) and the program waited for the participants to move inside the home area. Each area measured 1.6 cm in diameter and the inter-center distance between home area and each target area was 14 cm, whereas the two target areas were separated by an inter-center distance of 15.2 cm.

The target stimulus appeared in the upper center of the screen after the cursor had spent a dwell time of 500 ms in the home area. Participants were to move towards one of the target areas as quickly as possible (according to the target stimulus and the preceding compliance response). From this point on, the program sampled the x- and y-coordinates of the mouse cursor at 100 Hz.[6] Response time was defined as the time from target onset until the cursor had left the home area. Movement time (MT) was recorded once the cursor hit one of the target areas. Accordingly, the trial ended and the cursor shrank and disappeared from the screen. The screen was cleared 500 ms later and the next trial began after an additional ITI of 1000 ms. I did not display any error feedback but participants were encouraged to respond more quickly when they did not start their movement within 500 ms after target onset.

Participants completed 9 blocks of 50 trials each (25 trials with Aries and 25 trials with Gemini as target stimulus) and blocks were separated by short breaks. The first block was considered practice and excluded from all analyses.

[6] It should be noted that earlier mouse-tracking studies often displayed the target stimuli after movement initiation (e.g., Spivey et al., 2005). I departed from this procedure to arrive at a clearer separation of effects on decision time (as measured in terms of RT) and effects on the movement trajectories. This procedure, however, also calls for a refined approach to trajectory analysis as outlined in **Appendix B**.

8.1.3 Data treatment

Trajectory data were preprocessed using custom MATLAB scripts (The MathWorks, Inc.) to determine measures of maximum absolute distance (MAD) and area under the curve (AUC) for each trial. Movements to the left were mirrored at the vertical midline. For both measures, I used a straight line from the movement's start point to its final point for reference as described in **Appendix B**. I then stripped off all dwell time data that was recorded until the cursor had left the target area (i.e., until RT was measured) and time-normalized the remaining data to 101 points by linear interpolation. MAD was then computed as the signed maximum Euclidean distance from each of these points to the reference line (in px), with positive values indicating deviation in direction of the opposite target. Similarly, AUC was computed as the signed area between the interpolated points and the reference line (in px^2).

For statistical analyses I considered only trials in which participants acted according to the compliance prompt (95.2%). Furthermore, trials were discarded as outliers if any measure (RT, MT, MAD, or AUC) deviated by more than 2.5 standard deviations from the respective cell mean (6.3%). Confidence intervals around effect size estimates (**Fig. 13**) were computed with the R package *MBESS* (Kelley, 2007).

8.2 Results

Mean trajectories are plotted in **Figure 13** (cf. **Fig. 14** for descriptive statistics). Normal, correct responses were initiated more quickly than violations or responses using the reversed mapping rule, as indicated by a significant effect of rule compliance on RTs, $F(1, 38) = 26.30$, $p < .001$, $\eta_p^2 = .41$. Unexpectedly, these effects did not differ between the instruction groups and the overall RT level was also comparable across groups ($ps > .238$, $\eta_p^2 < .04$).

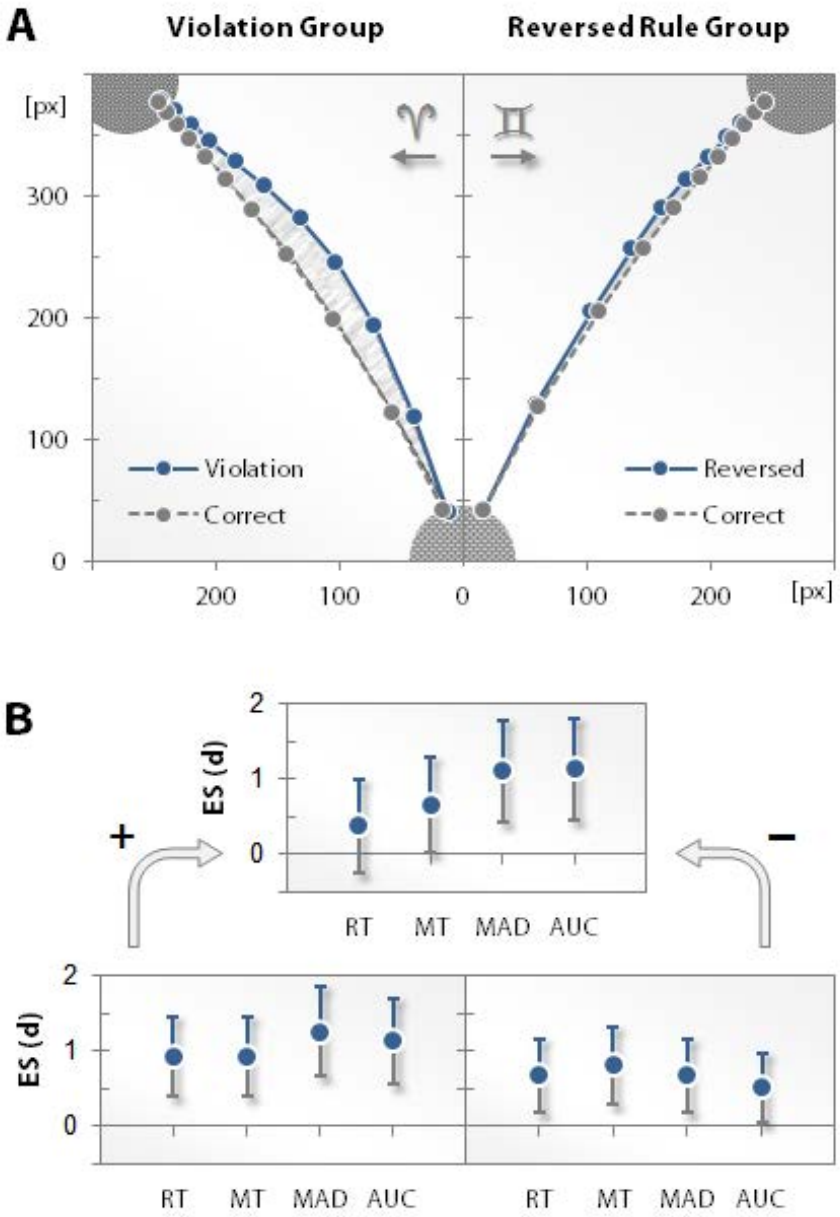

Fig. 13. Design and results for both instruction groups in Experiment 4. **(A)** When violating the mapping rule, trajectories were strongly biased towards the target that was indicated by the mapping rule. This impact was also present but clearly reduced for the reversed rule group. **(B)** Key results of the trajectory analysis. Dots indicate effect size estimates and error bars indicate the 95% confidence interval for these effect sizes. The lower panels show the effect sizes of each pairwise, within-group difference whereas the upper panel compares these differences across groups (i.e., corresponding to the interaction of group and rule compliance). RT = response time, MT = movement time, MAD = maximum absolute distance, AUC = area under the curve.

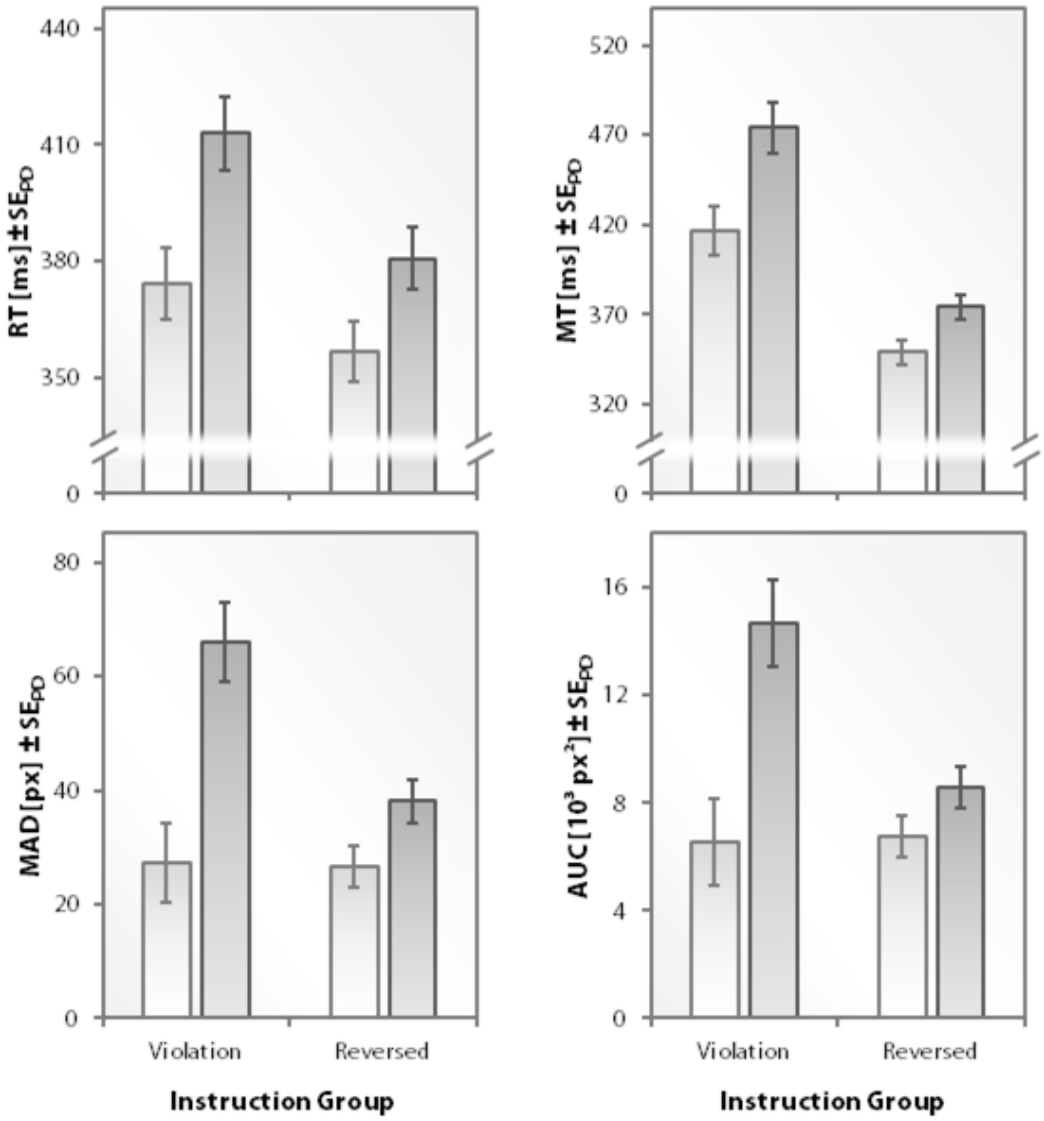

Fig. 14. Mean response time (RT), movement time (MT), maximum absolute distance (MAD), and area under the curve (AUC) for both instruction groups in Experiment 4. Light gray bars indicate normal, rule-based responses whereas the dark gray bars indicate violation responses for the violation group, and responses according the reversed mapping for the reversed rule group. Error bars indicate standard errors of paired difference scores, computed separately for each group.

A similar main effect of rule compliance emerged for MTs, $F(1, 38) = 28.22$, $p < .001$, $\eta_p^2 = .43$. This analysis, however, also yielded a significant interaction of rule compliance and instruction group, $F(1, 38) = 4.35$, $p = .044$, $\eta_p^2 = .10$, with a stronger impact of rule compliance for the violation group. This group also showed generally longer MTs, $F(1, 38) = 12.22$, $p = .001$, $\eta_p^2 = .24$.

The most important data for the present question are those relating to MAD and AUC, and both measures yielded converging results. For MADs, a significant main effect of rule compliance indicated stronger deviations to the alternative target area for both, violations and responses applying the reversed mapping rule, $F(1, 38) = 41.21$, $p < .001$, $\eta_p^2 = .52$, and this effect was clearly more pronounced for the violation group than for the reversed rule group, $F(1, 38) = 13.32$, $p = .001$, $\eta_p^2 = .24$. Overall, MADs did not differ between the two groups, $F(1, 38) = 2.78$, $p = .104$, $\eta_p^2 = .07$. For AUCs, a similar main effect of rule compliance, $F(1, 38) = 31.31$, $p < .001$, $\eta_p^2 = .45$, was moderated by a significant interaction, $F(1, 38) = 12.81$, $p < .001$, $\eta_p^2 = .25$, again indicating a stronger effect for the violation group than for the reversed rule group. Overall, AUCs did not differ between groups, $F(1, 38) = 1.91$, $p = .175$, $\eta_p^2 = .05$.

Follow-up correlation analyses across trials also targeted the intercorrelations of the four measures (**Tab. 1**). Most importantly, RT was not correlated significantly with any other measure, indicating that participants did not trade-off processing time between response initiation and the following movement phase.

Moreover, the different labelling of the responses had a profound impact on choice frequencies: Participants in the violation group chose to violate in 39.5% of the trials whereas participants in the reversed rule group chose to use the reversed mapping rule in 47.6% of the trials. The mean number of violations differed from chance for the violation group, $t(19) = -3.52$, $p = .002$, $d = -0.79$ but not for the reversed rule group $t(19) = -1.38$, $p = .183$, $d = -0.31$, and both percentages differed between groups, $t(38) = 2.35$, $p = .026$, $d = 0.23$, corrected for unequal variances.

As for Experiment 1-3, failures to act according to the compliance prompt were also more prevalent for rule violations (11.0%) than for normal, rule-based responses (2.7%) in the violation group. This trend was

Tab. 1. Intercorrelations of the four measures of Experiment 4 across trials for both groups. The upper diagonal shows mean correlation coefficients that were computed by Z-transforming each correlation for each participant, averaging the resulting Z-scores and re-transforming the results to correlation coefficients. The lower diagonals show the *p*-values when testing mean Z-scores against 0.

	Violation Group				**Reversed Rule Group**			
	RT	MT	MAD	AUC	RT	MT	MAD	AUC
RT		0.00	-0.10	-0.10		-0.04	-0.10	-0.10
MT	.991		0.25	0.23	.875		0.27	0.24
MAD	.684	.279		0.94	.684	.256		0.94
AUC	.674	.330	<.001		.677	.304	<.001	

RT = response time, MT = movement time, MAD = maximum absolute distance, AUC = area under curve.

also visible for the reversed rule group (4.8% vs. 1.9%). These numbers, however, should again be taken with caution due to the alternative interpretations discussed in Section 4.2.

A final analysis targeted whether the frequency of rule violations committed by each participant was correlated to the degree of conflict caused by these rule violations. To this end, I re-analysed the data of the violation group only and correlated the frequencies of rule violations with the effect of these violations across participants. This analysis showed the violation effect for RT ($RT_{Violations}$–$RT_{Correct\ Responses}$) not to be correlated with the frequency of violations r = -0.09, $t(18) = -0.38$, $p = .705$. However, a different picture emerged for the remaining three variables, where larger effects were significantly related to fewer rule violations for MT: r = -0.47, $t(18) = -2.27$, $p = .036$. Similar significant or marginally significant correlations emerged for the trajectory measures; MAD: r = -0.41, $t(18) = -1.91$, $p = .072$; AUC: r = -0.46, $t(18) = -2.21$, $p = .040$.

8.3 Discussion

Experiment 4 provided compelling evidence for the hypothesis that cognitive conflict during rule violations is – at least partly – driven by a failure to suppress the original mapping rule. When committing a violation, movement trajectories were strongly biased towards the rule-conform target; when performing the same behaviour as a response according to a reversed mapping rule, however, movement trajectories were not attracted towards the target implied by the original mapping rule.

These results extend previous cognitive studies on the automatic implementation of new mapping rules (e.g., Cohen-Kdoshay & Meiran, 2009; Kunde, Kiesel, & Hoffmann, 2003; Reisenauer & Dreisbach, 2013; Wenke, Gaschler, & Nattkemper, 2007) in showing that a rule representation cannot be suppressed easily when an agent deliberately violates this rule. This conflict was again present even though violating the rule did not cause any negative consequences. By contrast, instructing equivalent actions in terms of a new, reversed mapping rule did not yield comparable conflict (cf. also Colgan, 1970; Waszak et al., 2013; Wilson, 1968).

Unexpectedly, however, the effects of rule compliance did not differ significantly in the RT analysis. This finding stands in contrast to the comparison of Experiment 2 and 3 which showed stronger effects for rule violations than for reversed rule responses even though RT was the main dependent variable. Possible reasons for this difference might relate to the design of Experiment 4, such as the use of continuous instead of discrete responses, the possibility to prepare for rule violations and the free choice of whether to violate or not. On the other hand, the non-significant difference might also reflect a Type II error for at least two reasons. First, the confidence interval around the effect size of the between-groups comparison (**Fig. 13**) clearly includes the effect size of d = 0.71 that had

resulted for the comparison of Experiment 2 and 3. Secondly, the following Experiment 5 used essentially the same setup as Experiment 4 but also found the effects of rule compliance to differ between groups. Given that Experiment 4 tested multiple dependent variables (RT, MT, MAD, AUC, and choice frequencies) with each variable increasing the probability of at least one Type I or Type II error, I thus prefer not to interpret the differing results regard the RT data.

As a final noteworthy result, Experiment 4 suggests that labelling responses in terms of rule violations does affect response choices as is evident in the lower frequency of violation choices in the violation group as compared to Task 2 choices in the reversed rule group. Further, it seems as if the degree of cognitive conflict is also related to the individual frequency of rule violations. This latter, correlational finding obviously allows for at least two interpretations. For one, agents who commit many violations might not be affected as much by these violations as are agents who are less prone to violating rules. Speculatively, however, the observed correlation might also indicate the reverse process: People who have difficulty with suppressing a rule might also refrain from violating it (possibly in anticipation of the increased effort). The relevance of this speculation is further enhanced by Experiment 6 which indicates that reliable conflict also arises for purely self-chosen rule violations (see Section 10.3, for an extended discussion). Yet, it should also be noted that participants in the violation group of Experiment 4 were still instructed to violate the mapping rule at times. Experiment 6 presents a different approach to study spontaneous rule violations without enforcing this kind of behaviour onto participants.

Before focusing on such spontaneous response choices, however, a possible confound in the design of Experiment 4 deserves to be addressed, and this confound relates to the difference in choice frequencies between both instruction groups. Because rule violation responses in the violation group were less frequent than Task 2 responses in the reversed rule group,

the observed differential effects of rule compliance might be driven partly or even entirely by the frequency of the responses. Experiment 5 thus tested whether the observed effects still hold when controlling for the frequencies of violation responses and Task 2 responses in a forced choice setting.

9 Experiment 5: Only a matter of choice?

Experiment 5 aimed at replicating the effects of deliberate rule violations on movement trajectories as observed in Experiment 4 while controlling for alternative explanations in terms of different choice frequencies between both instruction groups. To this end, I replaced the compliance prompt of Experiment 4 with a cue that instructed participants whether to follow a rule or not (violation group) or whether to respond according to Task 1 or Task 2 (reversed rule group; cf. **Fig. 15**).

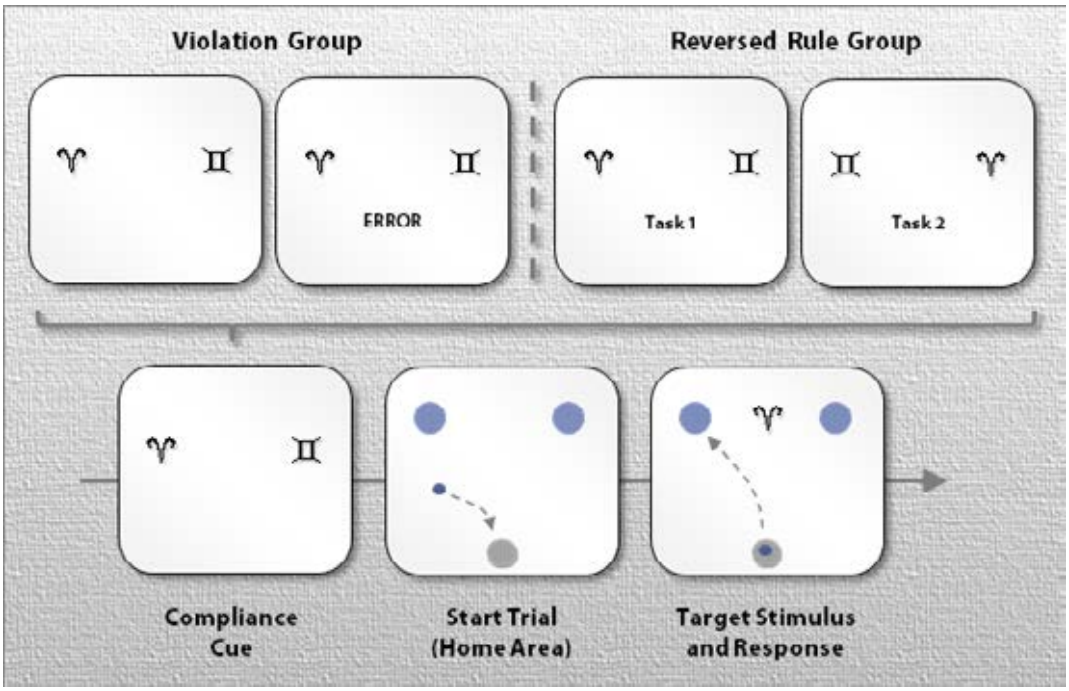

Fig. 15. Task cue and trial procedure of Experiment 5. Participants in the violation group were first informed whether to respond according to the mapping rule or to violate this rule and commit an error by intention. By contrast, participants in the reversed rule group were informed to perform either a Task 1 or a Task 2, whereas the two tasks again employed the exact opposite stimulus-response mapping. Accordingly, similar actions were labelled as violations for the violation group and as different rule-based responses for the reversed rule group. Rule violations and Task 2 responses were cued equally often (in 25% of the trials).

9.1 Method

9.1.1 Participants, Apparatus, and Stimuli

Twenty new participants were assigned to the violation group (mean age = 26.1 years, 16 females, 1 left-hander) and another 20 participants were assigned to the reversed rule group (mean age = 27.4 years, 14 females, 3 left-handers, 1 ambidextrous). Stimuli and apparatus were as in Experiment 4. The number of trials per block was 48 with 36 trials being normal, rule-based responses or Task 1 responses and 12 trials being rule violations or Task 2 responses for the violation group and the reversed rule group, respectively.

9.1.2 Procedure

In contrast to Experiment 4, each trial started with a compliance cue (upper row in **Fig. 15**), informing participants whether to follow the rule or not (violation group) or whether to perform Task 1 or Task 2 (reversed rule group).

For the violation group, this display featured simply the correct task mapping in the upper half of the screen in most trials (75%) which called for a normal response in the upcoming trial. In some trials (25%), the compliance cue also featured the word "FEHLER" (error) in the center of the screen, calling for a rule violation instead. For the reversed rule group, the compliance cue instructed the participants to perform either Task 1 (75% of the trials) or Task 2 that featured the reversed mapping (25% of the trials). Participants were instructed to carefully attend to the compliance cue in each trial and press the spacebar whenever they felt ready to perform the task. The remaining trial procedure was as in Experiment 4.

9.2 Results

Mean trajectories are plotted in **Figure 16** and summary statistics are plotted in **Figure 17**. Rule-based responses were again initiated more quickly than violations or responses that used the reversed mapping rule, as indicated by a significant main effect of rule compliance in the RT analysis, $F(1, 38) = 55.78$, $p < .001$, $\eta_p^2 = .59$. In contrast to Experiment 4, the effects of rule compliance differed between the instruction groups, $F(1, 38) = 20.10$, $p < .001$, $\eta_p^2 = .35$, with a stronger impact of rule compliance for the violation group. The overall RT level was comparable across groups, $F(1, 38) = 1.88$, $p = .179$, $\eta_p^2 = .05$. Similarly, a main effect of rule compliance emerged for MTs, $F(1, 38) = 14.51$, $p < .001$, $\eta_p^2 = .28$, and the effect of rule compliance on MTs differed between the instruction groups, $F(1, 38) = 5.33$, $p = .027$, $\eta_p^2 = .12$. The main effect of instruction group did not approach significance, $F(1, 38) = 0.99$, $p = .326$, $\eta_p^2 = .03$.

Again, the most important performance data for the present question are those relating to MAD and AUC. For MADs, a significant main effect of rule compliance indicated stronger deviations to the alternative target area for both, violations and responses applying the reversed mapping rule, $F(1, 38) = 24.37$, $p < .001$, $\eta_p^2 = .39$, and this effect was again more pronounced for the violation group than for the reversed rule group, $F(1, 38) = 16.99$, $p < .001$, $\eta_p^2 = .31$. Furthermore, the violation group also showed overall larger MADs, $F(1, 38) = 7.05$, $p = .012$, $\eta_p^2 = .16$. For AUCs, a similar main effect of rule compliance, $F(1, 38) = 21.79$, $p < .001$, $\eta_p^2 = .36$, was moderated by a significant interaction, $F(1, 38) = 14.68$, $p < .001$, $\eta_p^2 = .28$, again indicating a stronger effect for the violation group than for the reversed rule group. Overall AUCs also differed between groups, $F(1, 38) = 4.54$, $p = .040$, $\eta_p^2 = .11$.

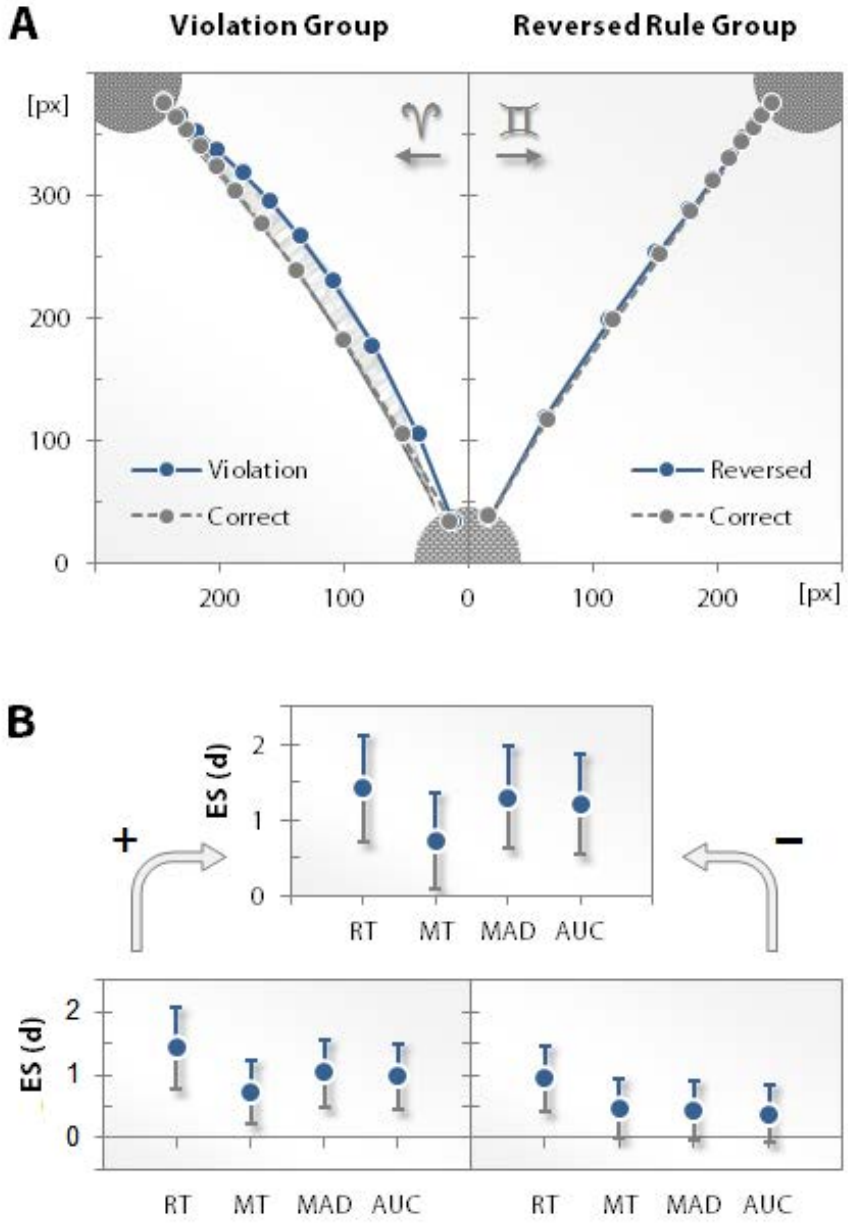

Fig. 16. Design and results for both instruction groups in Experiment 5. **(A)** Time-normalized movement trajectories. As in Experiment 4, trajectories of rule violations were strongly biased towards the target that was indicated by the mapping rule. This impact was again reduced for the reversed rule group. **(B)** Key results of the trajectory analysis. Dots indicate effect size estimates and error bars indicate the 95% confidence interval for these effect sizes. The lower panels show the effect sizes of each pairwise, within-group difference whereas the upper panel compares these differences across groups (i.e., corresponding to the interaction of group and rule compliance). RT = response time, MT = movement time, MAD = maximum absolute distance, AUC = area under the curve.

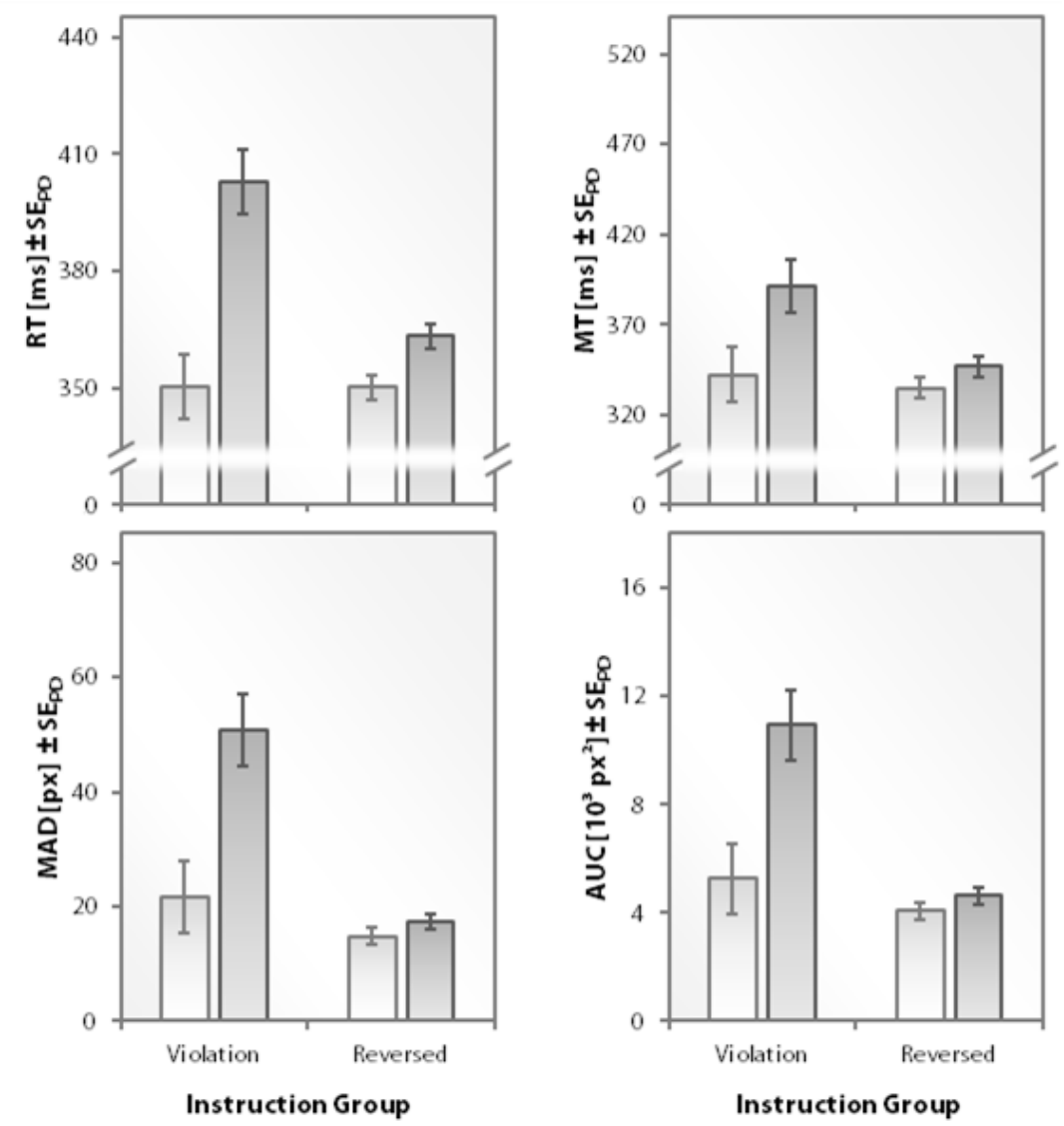

Fig. 17. Mean response time (RT), movement time (MT), maximum absolute distance (MAD), and area under the curve (AUC) for both instruction groups in Experiment 5. Light gray bars indicate normal, rule-based responses, whereas the dark gray bars indicate violation responses for the violation group, and responses according the reversed mapping in the reversed rule group. Error bars indicate standard errors of paired difference scores, computed separately for each group.

As in Experiment 4, failures to act according to the intention indicated by the compliance cue (4.8% in total) were more prevalent for rule violations (6.2%) than for normal, rule-based responses (2.8%) for the violation group. This trend was equally present for the reversed rule group

Tab. 2. Intercorrelations of the four measures of Experiment 5 across trials for both groups. The upper diagonal shows mean correlation coefficients that were computed by Z-transforming each correlation for each participant, averaging the resulting Z-scores and re-transforming the results to correlation coefficients. The lower diagonals show the *p*-values when testing mean Z-scores against 0.

	Violation Group				**Reversed Rule Group**			
	RT	MT	MAD	AUC	RT	MT	MAD	AUC
RT		-0.04	-0.08	-0.09		0.02	-0.06	-0.06
MT	.883		0.25	0.23	.920		0.13	0.13
MAD	.725	.294		0.95	.802	.591		0.94
AUC	.699	.337	<.001		.787	.594	<.001	

RT = response time, MT = movement time, MAD = maximum absolute distance, AUC = area under curve.

where failures to perform Task 2 (8.2%) occurred more frequently than failures to perform Task 1 (2.8%). Finally, follow-up correlation analyses across trials ensured that the effects for MAD and AUC were not correlated with those for RT (see **Tab. 2**).

9.3 Discussion

Experiment 5 replicated the central findings of Experiment 4 by showing that rule representations cannot be suppressed easily when an agent decides to violate a rule. This was true even though different frequencies of rule violations as compared to Task 2 responses cannot account for the observed effects. It thus seems as if the original mapping rule would indeed remain actively represented during rule violations, biasing behaviour towards rule-based responding. Possible mechanisms that might account for these effects will be discussed in detail in Chapter 13. Beforehand, however, Experiment 6 explores a different approach to study spontaneous rule violations without enforcing this kind of behaviour onto participants.

10 Experiment 6: Hot delivery

Experiment 4 provided first evidence for an impact of rule violations even when an agent decides to violate a rule. The setup of Experiment 4, however, still prompted participants to commit violations by explicitly asking whether participants would conform to a mapping rule or whether they would violate it. This setup is thus rather far from investigating spontaneous, freely chosen rule violations (Schüür & Haggard, 2011). Experiment 6 aimed at filling this gap by using a task in which participants were not explicitly prompted to violate rules; rather the task was designed to encourage self-chosen violations (see **Fig. 18**).

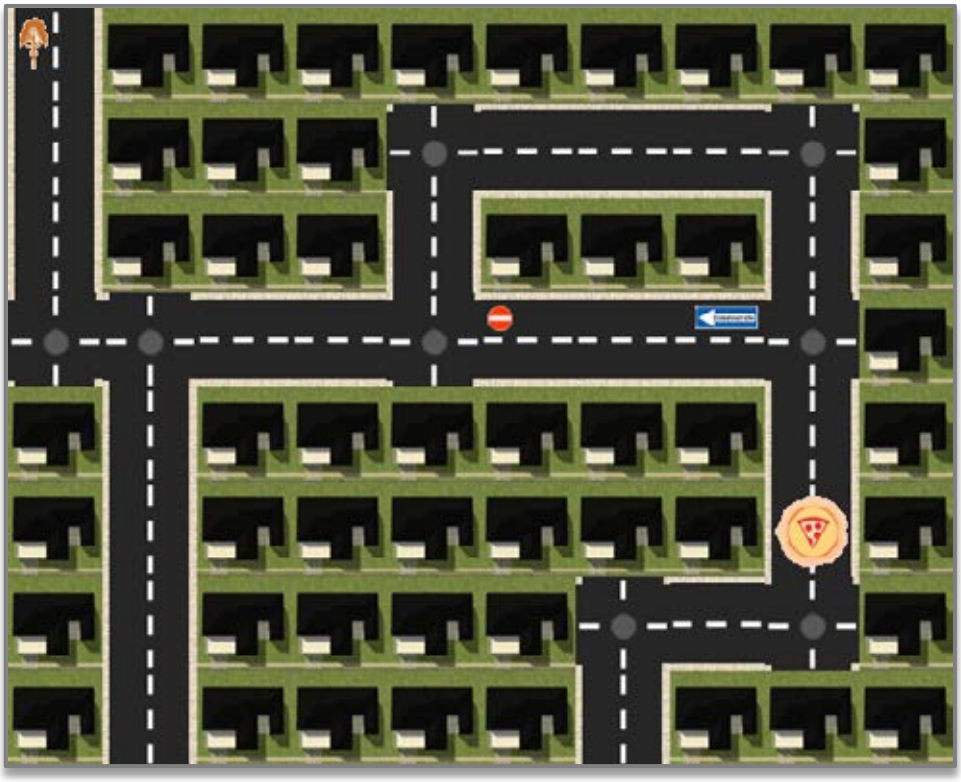

Fig. 18. Exemplar trial of the pizza task of Experiment 6. Participants navigated a bicycle courier (here: top-left corner) to deliver a virtual pizza (goal location; bottom-right corner) and each keypress moved the courier for one tile (10 x 8 tiles in total). The map consisted of roads, non-passable houses, and a goal location. Some roads could be designated as one-way.

Participants were asked to take control of a virtual bicycle courier who was to deliver a pizza in a two-dimensional city map and the only instruction was to deliver the pizza as quickly as possible. Crucially, I implemented one-way roads in some of the maps and violating these one-way roads could speed up the delivery at times. Accordingly, I expected participants to use these shortcuts, (i.e., to perform optimising violations sensu Reason, 1990, 1995). Still, if the findings of the previous tasks generalize to such freely chosen actions, cognitive conflict should again be evident in the participants' behaviour.

In addition to investigating cognitive conflict during spontaneous, freely chosen rule violations, Experiment 6 also targeted the relation of this behaviour to other cognitive shortcuts that human agents tend to apply in order to facilitate various tasks. Such shortcuts range from explicit, strategic shortcuts in mental arithmetic (Haider & Frensch, 1996, 1999a, 1999b) and heuristics in decision making (Tversky & Kahnemann, 1974) to implicit categorization shortcuts (Pashler & Baylis, 1991).

Experiment 6 focused on the latter type of shortcuts. These are typically observed in choice reaction tasks in which participants respond to target stimuli in a succession of trials. If the current target stimulus matches the stimulus that was encountered in the preceding trial (*stimulus repetitions*), responses are typically much faster than with changing stimuli. This finding has been taken to indicate that the time-consuming categorization of the imperative stimulus is skipped when a stimulus is repeated (Bertelson, 1963; Pashler & Baylis, 1991; see also Tan & Dixon, 2011). Categorization shortcuts can thus be construed as a tendency to not select an appropriate response according to a specific mapping rule but rather bypass this rule by relying on previous experiences.

These categorization shortcuts arguably are quite different from the intended rule violations that are the focus of this experiment: Categorization shortcuts take place on a scale of a few hundred

milliseconds and they are rarely employed deliberately (Pashler & Baylis, 1991), whereas the type of rule violations studied here takes place on a larger timescale and – assuming that participants are aware of the one-way signs – are based on a deliberate decision to violate this rule.

Despite these differences, categorization shortcuts and optimising violations have in common that the agent reaches a desired end – a correct categorization or successful performance, respectively – by other means than suggested by the task at hand. I thus speculated that both behaviours could tap on similar processes and decided to correlate the tendency to violate rules (in the *pizza task*) with the tendency to exploit stimulus repetitions in an unrelated cognitive task (the *celebrity task*). Because I expected generally small effect sizes for the correlation analysis as well as for the analysis of the violation behaviour proper, I used a larger sample size than in the previous experiments.

10.1 Method

10.1.1 Participants

Seventy-two undergraduate students participated for course credit (mean age: 20.5 years, 61 females, 8 left-handers). One participant partly guessed the purpose of the experiment and was replaced.

10.1.2 Celebrity task: Measuring categorization shortcuts

For the celebrity task, participants responded with the keys *J*, *K*, and *L* of a standard computer keyboard, operated by the index, middle, and ring finger of the right hand. The keys were marked with coloured patches (orange, green, and white) and instructions always referred to these colours. Target stimuli were grayscale portraits of six celebrities (3.5 cm x 3.5 cm) that appeared on a 17'' monitor. All six celebrities are likely to be well-known among German university students: Angela Merkel (German chancellor), Queen Elizabeth II., Angelina Jolie (actress), Günther Jauch

(German quizmaster), Johnny Depp (actor), and Dirk Nowitzki (Würzburgian basketball player). Two portraits, one male and one female, were mapped to each response key and the S-R mapping was counterbalanced across participants.

Each trial simply featured a target stimulus and participants were to respond as quickly and accurately as possible with the assigned key. The stimulus remained on screen until a response was given and wrong responses triggered error feedback for 1500 ms. The next trial started after 500 ms; responses during the ITI produced an error message.

Participants worked through eleven blocks of 54 trials each (i.e., each stimulus was displayed nine times per block in a random order) and feedback after each block informed the participants about their mean response time and the number of errors to ensure a high motivation. The first block was considered practice and did not enter the analyses.

10.1.3 Pizza task: Measuring optimising violations

For the pizza task, participants responded with the four arrow keys of the keyboard to navigate a bicycle courier through city-like 2D-mazes (see **Fig. 18**). These mazes consisted of 10 x 8 tiles (1.5 cm x 1.5 cm) and each map contained roads, non-passable houses, and a goal location that was signalled by a pizza icon. Some maps additionally contained one or more designated one-way roads.

Pressing a key moved the courier forward one tile and the bicycle movement was always coded relative to the map, i.e., pressing the left arrow moved the bicycle to the left, irrespective of the bicycles' orientation. The program logged inter-keystroke-intervals (IKIs), responses, and corresponding bicycle locations throughout the trial and the trial ended as soon as the bicycle reached the goal location. The final map stayed on screen for 500 ms and the next trial started after additional 1000 ms.

The experiment started with a training block of five maps that did not contain any one-way roads and participants were not informed about these upcoming stimuli. Then, the experimenter left the room and the participant worked through two blocks of 60 trials each. The two blocks used the same maps in a fixed sequence. Overall, the participants thus completed 120 trials, 30 of which did not contain any one-way roads, 20 contained one-way roads that did not help to cut short to the goal location if used in the wrong direction, and 70 contained one-way roads that helped to cut short to the goal location by violating the indicated direction.

10.1.4 Ad-hoc questionnaire

After the experiment, I administered a short ad-hoc questionnaire (in German language) that featured three questions on a visual analogue scale (length: 7.1 cm) with verbal anchors at both ends. The questionnaire probed for the participant's attitude towards rule violations. The first question translated to "If you violate a rule, how guilty do you feel?" (*'feeling guilty'*; *not very guilty* to *very guilty*) whereas the second question targeted directly how prone participants were to committing violations: "How often do you violate rules?" ('violation frequency'; *very rarely* to *very frequently*). The final question translates to "How strongly would you condemn others for breaking a rule on purpose?" ('condemn others', *not very much* to *very strongly*).

10.2 Results

10.2.1 Celebrity task: Inter-trial repetitions

For analysis of the celebrity task, I considered only RTs of correct trials that were not preceded by errors (which occurred in 5.3% of all trials). RTs that deviated by more than 2.5 standard deviations from their cell mean were discarded as outliers (3.0%).

The remaining RTs were aggregated to separate means for the three conditions of interest: stimulus repetitions (444 ms), response repetitions (614 ms), and complete alternations (602 ms). These means differed significantly, as indicated by a repeated-measures ANOVA, $F(2, 142) = 331.84$, $p < .001$, $\eta_p^2 = .82$. The critical effect for the current study, however, was not the omnibus ANOVA but rather the pairwise comparison of complete alternations and stimulus repetitions. Considered separately, this repetition benefit ($RT_{Complete\ Alternation} - RT_{Stimulus\ Repetition}$) amounted to sizeable 158 ms and was significantly different from zero, $t(71) = 22.90$, $p < .001$, $d = 2.70$.

10.2.2 Pizza task: One-way violations

A first analysis of the pizza task targeted the distribution of one-way violations across participants (**Fig. 19**). Descriptively, this distribution exhibited two separate modes, one at each end of the scale.

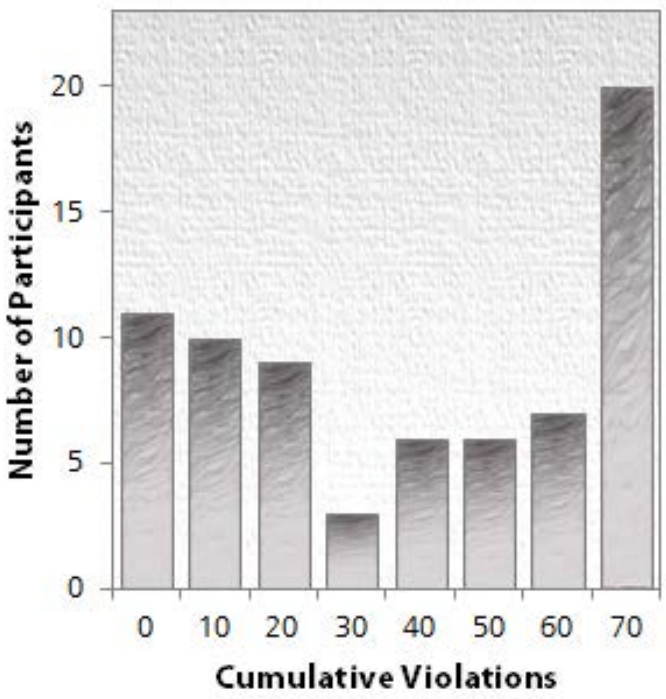

Fig. 19. Histogram of one-way violations across participants for the pizza task of Experiment 6; bins are labelled in terms of their upper boundary.

To quantify this visual impression, I computed two statistics: The bimodality coefficient and Hartigan's dip test (Freeman & Dale, 2013). The bimodality coefficient amounted to $b = .679$, clearly exceeding the cut-off value of $b_{crit} = .555$ that would be expected from a uniform distribution (Knapp, 2007). Furthermore, the dip test for unimodality (Hartigan & Hartigan, 1985) was highly significant, $dip = 0.095$, $p < .001$, indicating a non-unimodal distribution (see **Appendix C** for a more detailed description of both measures).

The following analyses focused on trials that included a violation. Furthermore, I distinguished between the very first violation trial and all following violation trials, because participants did not know what to expect for the first one-way violation. The analyses could thus only be run for participants who committed at least two violations across the experiment and did not produce any missing data during the first violation, e.g., by reversing direction right after entering the one-way. For these participants ($n = 48$), I calculated mean IKIs for four conditions: (1) keystrokes during a violation trial that were unrelated to the violation itself, (2) keystrokes right before entering a one-way in the "wrong" direction, (3) keystrokes initiating the violation, and (4) keystrokes while heading through the one-way in the wrong direction (see **Fig. 20**). IKIs deviating by more than 2.5 standard deviations from their cell mean were considered outliers (3.1%). Importantly, the very first violation of each participant was treated separately so that the IKI data was analysed by a 4 x 2 repeated-measures ANOVA with the factors keystroke type (as described above) and violation order (first vs. following violations).

Most importantly, the described ANOVA revealed a main effect of keystroke type, $F(3, 141) = 12.95$ ($\varepsilon = .51$), $p < .001$, $\eta_p^2 = .22$, driven by slow responses preceding the violation and when initiating the violation on the one hand and fast IKIs while passing through the one-way on the other hand (as compared to violation-unrelated responses). Additionally,

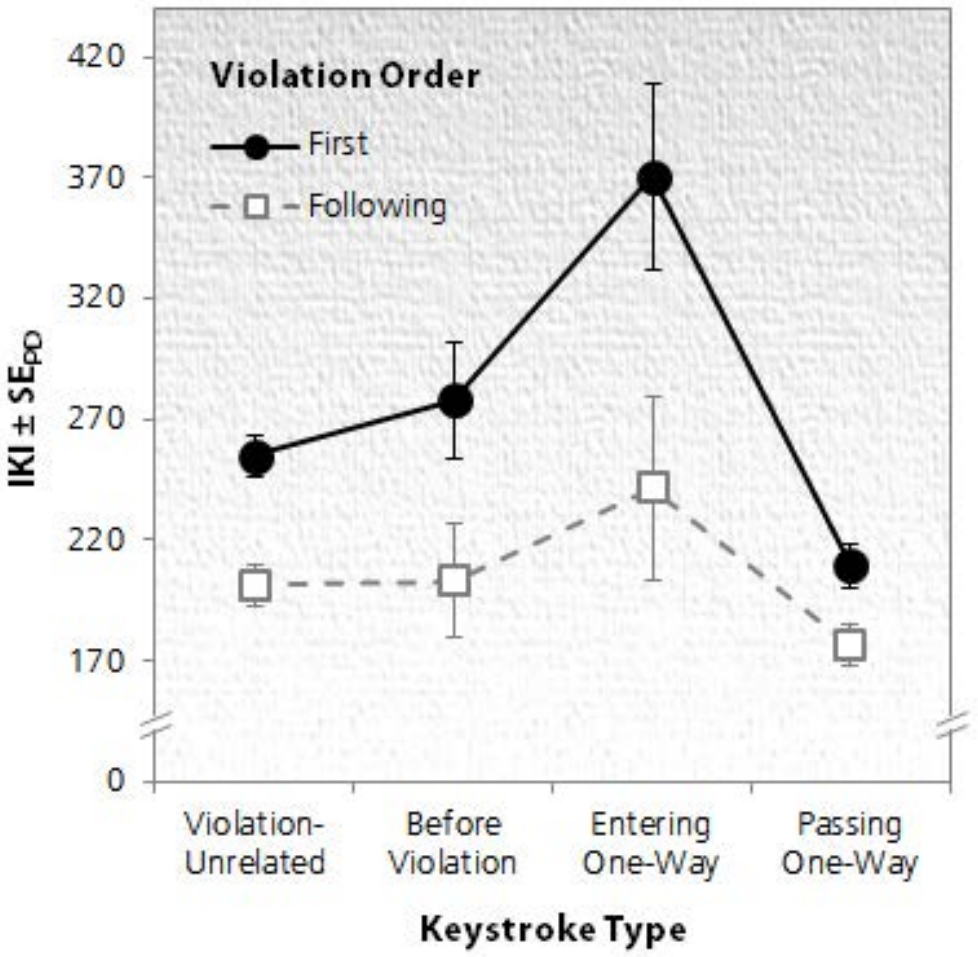

Fig. 20. Inter-keystroke-intervals (IKIs) at four different positions during a violation trial of the pizza task in Experiment 6. First violation data refer to the very first violation committed in the experiment (when participants did not yet know what to expect) whereas the data labelled as "following violations" represents the mean of all subsequent violations. Keystroke types are coded for different responses within a trial in which the participant had committed a violation. Error bars are within-subjects standard errors, computed separately for each keystroke type.

keystrokes during the first violation trial were overall slower than those of the remaining trials, $F(1, 47) = 33.75$, $p < .001$, $\eta_p^2 = .42$, and the effect of keystroke type was stronger for the first violation than for the remaining violations, $F(3, 141) = 3.23$ ($\varepsilon = .59$), $p = .048$, $\eta_p^2 = .07$.

Separate pairwise comparisons indicated that the IKI when entering the one-way was significantly longer than violation-unrelated IKIs for the first violation (Δ = 115 ms), $t(47) = 2.90$, $p = .006$, $d = 0.42$, as well as for the following violations (Δ = 40 ms), $t(47) = 5.94$, $p < .001$, d = 0.86. Similarly, IKIs while passing the one-way were significantly shorter than unrelated ones for the first violation (Δ = -49 ms), $t(47) = -3.31$, $p = .002$, d = -0.48, and the following violations (Δ = -25 ms), $t(47) = -11.10$, $p < .001$, $d = -1.60$. The difference between unrelated IKIs and IKIs right before the violation did not approach significance for either comparison (*p*s > .320).

A potential confound of the above analysis, however, is the different contribution of key switches and key repetitions to the four keystroke types: The first three types – unrelated, before violation, entering one-way – comprise both, response repetitions and response switches, whereas the keystrokes while passing through the one-way were always response repetitions (due to the map design). Consequently, I repeated the above pairwise comparisons for response repetitions only and dropped the factor violation order, yielding a total of $n = 51$ participants for analysis (see **Fig. 21**; 2.6% of the IKIs were identified as outliers by the same criterion as for the previous analysis).

Mean IKIs for response repetitions unrelated to violations amounted to 183 ms whereas IKIs preceding a violation were now significantly longer (189 ms), $t(50) = 2.80$, $p = .007$, d = 0.39. Even though the IKIs when entering the one-way were descriptively even longer (198 ms), the difference to unrelated IKIs was only marginally significant, $t(50) = 1.73$, $p = .089$, d = 0.24. Crucially, however, IKIs when passing the one-way (178 ms) were still significantly faster than unrelated IKIs, $t(50) = -4.67$, $p < .001$, d = -0.65.

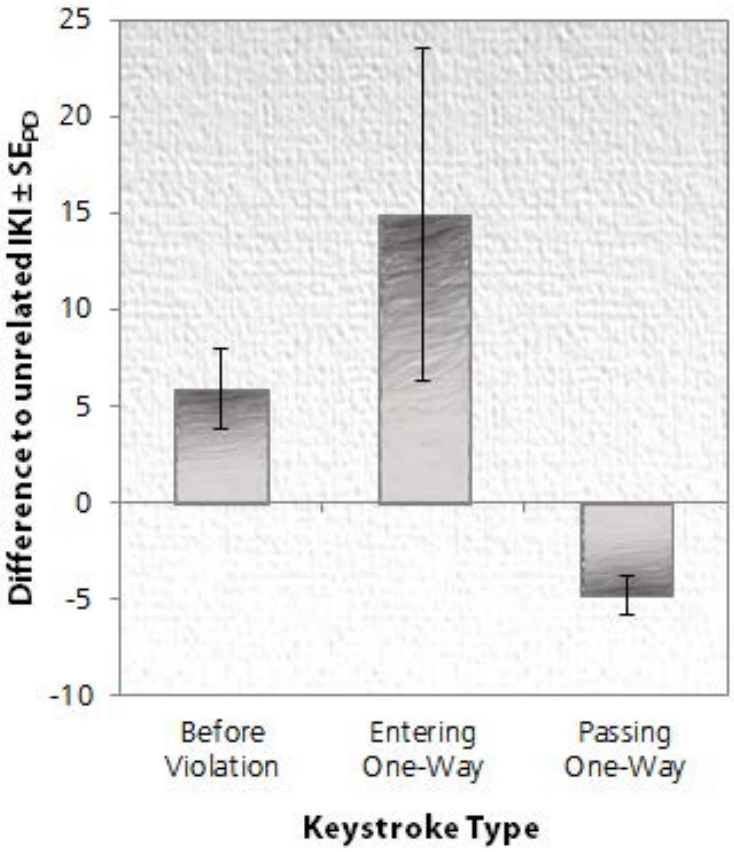

Fig. 21. Control analysis of the pizza task of Experiment 6: Differences between the mean IKI for violation-unrelated response repetitions (183 ms) and response repetitions preceding a violation, when entering a one-way, and when passing through the one-way. Error bars are within-subjects standard errors, computed separately for each keystroke type.

10.2.3 Cross-task correlations

The goal of the cross-task correlations was to predict the overall number of violations that a participant committed. To this end, I evaluated five predictors via pairwise correlations (**Tab. 3**). The first two predictors were derived from the celebrity task. Here, I used the repetition benefit (in ms; as described in Section 10.2.1) as a first predictor and, additionally, a repetition index (in %) that corrected for the overall RT level of the participants (repetition index = repetition benefit / mean RT * 100). The remaining three predictors were the ratings for the three questions in the ad-hoc questionnaire (in % of the visual analogue scale).

Tab. 3. Pairwise correlations of the celebrity task data (repetition benefit, repetition index), the ad-hoc-questionnaire administered after the experiment, and the number of violations committed by a participant (violation count). The upper diagonal of the table lists correlation coefficients (significant correlations are in italics) whereas the lower diagonal gives the corresponding *p*-values. All correlations are based on the entire sample of $n = 72$ participants.

	Repetition Benefit	Repetition Index	Feeling Guilty	Violation Frequency	Condemn Others	Violation Count
Repetition Benefit		*0.95*	*-0.25*	0.17	-0.05	-0.09
Repetition Index	.000		*-0.28*	0.16	-0.01	0.00
Feeling Guilty	.037	.015		*-0.49*	*0.30*	*-0.25*
Violation Frequency	.158	.169	.000		*-0.28*	0.10
Condemn Others	.663	.901	.010	.017		-0.07
Violation Count	.450	.981	.034	.395	.560	

The only significant predictor of the number of violations was the subjective guilt when committing violations, $r = -0.25$, $t(70) = -2.16$, $p = .034$, with a regression line equating to $\hat{y} = -0.26 \cdot x + 49.65$ (**Fig. 22**). Accordingly, participants committed less one-way violations the more they rated themselves to generally feel guilty after having violated a rule.[7] Interestingly, neither the self-rated general frequency of violations, $r = 0.10$, $t(70) = 0.84$, $p = .395$, nor the repetition index, $r = 0.00$, $t(70) = -0.02$, $p = .981$, were significantly related to the number of one-way violations.

[7] The correlation coefficient for the self-reported guilt is only modest in size. Given the low reliability of such single-item tests, however, much higher coefficients should not be expected.

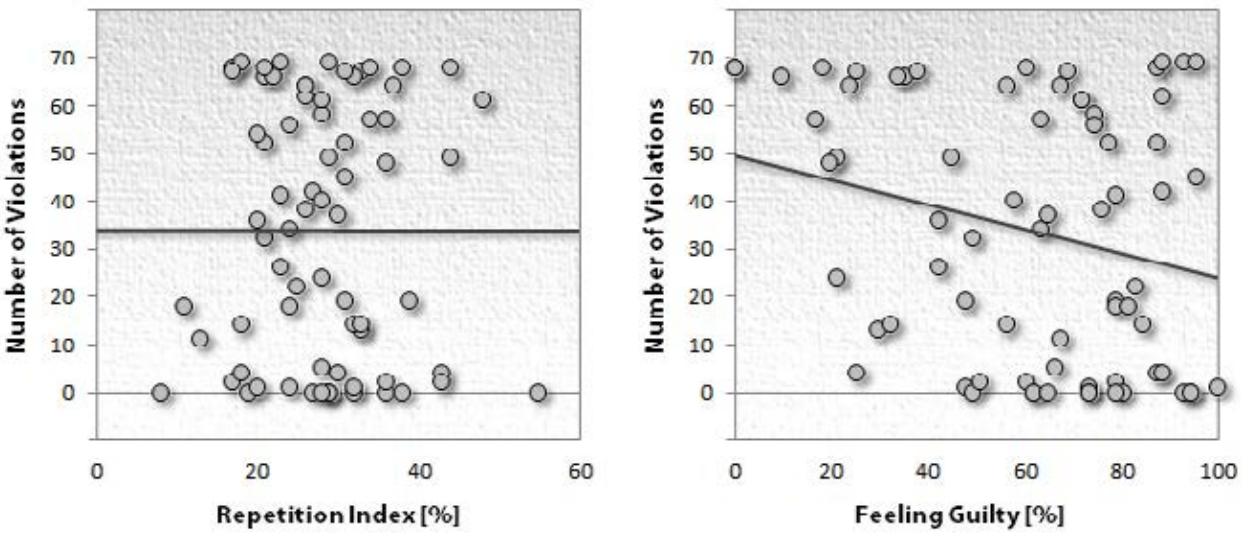

Fig. 22. The number of one-way violations in the pizza task of Experiment 6 could not be predicted by the repetition index of the celebrity task (with repetition index = repetition benefit / mean RT * 100; left panel). Yet, the subjective rating of how guilty one feels after violating a rule did predict the number of one-way violations (right panel).

10.3 Discussion

Experiment 6 yielded three key findings. First, rule violations caused cognitive conflict even though participants deliberately decided to violate the rule, without being prompted or encouraged by the experimenter. Cognitive conflict during rule violations thus also occurs in more externally valid settings (as compared to Experiment 1-5). Secondly, the tendency to violate rules in the employed pizza task seems to be independent of cognitive categorization shortcuts as measured in the celebrity task. Thirdly, participants who reported less guilt when violating rules in the post-experimental questionnaire tended to violate more rules in the pizza task.

The latter two observations suggest that deliberate decisions to violate rules might not be traced back to very basic cognitive shortcuts. Thus, at least for the current operationalization of rule violations and cognitive shortcuts, it does not seem as if the deliberate decisions leading to an

optimising violation (Reason, 1990, 1995) drew on rather automatic cognitive shortcuts that bypass certain categorization processes (Pashler & Baylis, 1991; Tan & Dixon, 2011). Instead, the subjective assessment of possible consequences that might result from rule violations (i.e., felt guilt) seems to determine whether an optimizing violation is committed or not.

Asides from methodological problems of interpreting non-significant results, it seems worthwhile to consider a possible alternative explanation that might also explain the present null-correlation between the number of rule violations and the repetition benefit as an index of cognitive shortcuts. Clearly, the number of rule violations is a rather discrete measure of how often a decision process converged on one or the other option (resulting in a rule violation or rule-based behaviour). By contrast, the repetition benefit seems to be a continuous, performance-based measure because it is based on differences between mean RTs in two conditions. The null-correlation could thus partly be driven by different information captured by each measure (the outcome of a process vs. the speed of a process). This conclusion seems premature, though. Rather, Pashler and Baylis (1991) argue that the repetition benefit does not indicate a genuine speedup of response selection but rather an shortcut that actually skips response selection processes (for similar views, see Dehaene, 1996; Smith, 1968; Smith, Chase, & Smith, 1973; Tan & Dixon, 2011).

Even though repetition effects are likely to entail additional components (e.g., Soetens, 1998; Sommer, Leuthold, & Soetens 1999), the assumed shortcut would imply a rather discrete mechanism that either takes place (creating a repetition benefit in a given trial) or not. Following this logic, differences in repetition benefits across participants can be seen as a measure of how often a shortcut it used (cf. Pfister, Schroeder, & Kunde, 2013, for a similar argument). The applied correlation analysis thus seems to be methodologically sound and the non-significant result might indeed suggest independent processes.

As the most important result of Experiment 6, however, cognitive conflict during rule violation still ensued even though participants were not explicitly encouraged to violate a rule but rather chose to do so because it allowed for faster task completion. This finding clearly suggests that any type of rule violation causes measurable cognitive conflict. In turn, anticipating this conflict might also influence the decision whether to violate a rule or not. In other words: In everyday life, agents obviously weigh the possible benefits of violating a rule against the possible negative consequences (Blanton & Christie, 2003; Reason, 1995, 2000) and the present experiment suggests that cognitive conflict is a reliable and inevitable consequence of this behaviour. Accordingly, the mere anticipation of this conflict might generate an a priori bias towards following rules irrespective of any other expectations.

Support for this interpretation comes from recent findings on the capability of human agents to generate random sequences of events. Here, if participants are simply asked to generate a random series of responses, e.g., zeros and ones, most participants tend to produce more alternations than is expected by chance (hence, an alternation bias; e.g., Nickerson, 2002; Rapoport & Budescu, 1997; Wagenaar, 1972). This finding is robust and reliable not only for generation tasks but also for judgements of the randomness of physical events that can be construed as sampling without replacement (gambler's fallacy; e.g., Ayton & Fischer, 2004; Kahneman & Tversky, 1972). Yet, this pattern of results completely *reverses* when participants are asked to randomly alternate between two different tasks (Arrington & Logan, 2004, 2005). In such settings – often labelled as voluntary task-switching – participants show a clear bias towards task repetitions, and switches are accompanied by considerable costs just as for traditional task switching experiments (cf. also Liefooghe, Demanet, & Vandierendonck, 2010; Kessler, Shencar, & Meiran, 2009; Vandierendonck, Demanet, Liefooghe, & Verbruggen, 2012).

Moreover, increasing switch costs experimentally causes participants to repeat tasks more often (Yeung, 2010) and the inter-individual variation of switch costs partly predicts repetition frequencies (Mayr & Bell, 2006). Finally, repetition biases are also increased by concurrent working memory load (Demanet, Verbruggen, Liefooghe, & Vandierendonck, 2010). Taken together, these findings clearly suggest that anticipated costs or conflict does affect decisions about what to do. The operationalization of conflict as it is used in experiments on voluntary task switching is further well in line with the concept of conflict used in the present experiments. Accordingly, anticipated conflict might indeed be a driving force behind decisions about whether to follow or to violate a rule. I will come back to this intriguing possibility in the Concluding Remarks. Beforehand, however, two final experiments seek to complement the present series.

.

PART 4: THE ELECTROPHYSIOLOGICAL SIGNATURE OF RULE VIOLATIONS

The following two experiments aim at complementing the behavioural approach of Experiment 1-6 by an electrophysiological perspective on rule violations. To this end, I will start with a short overview of previous electrophysiological research on unintended failures to obey a rule, i.e., errors.

The investigation of such erroneous responses has a long tradition in psychology and neuroscience with their main electrophysiological correlate being the *ERN* / N_E (Falkenstein et al. 1990; Gehring et al., 1993; cf. Section 3.1 for an introduction to this component). One of the most prominent theoretical accounts for this ERP component links the ERN to the concept of prediction errors (the reinforcement learning approach; Holroyd & Coles, 2002; Holroyd, et al., 2005). This theory is motivated by results pointing towards ERN-like waveforms occurring not only after response errors but also after negative feedback (Miltner, Braun, & Coles, 1997; cf. also Folstein & van Petten, 2008) or even after unexpected arbitrary consequences of own actions (Band, van Steenbergen, Ridderinkhof, Falkenstein, & Hommel, 2009; Knolle, Schröger, & Kotz, 2013). The ERN triggered by such feedback seems to originate from the same neural sources as the ERN following errors (Gehring & Willoughby, 2002; Holroyd & Coles, 2002; Holroyd et al., 2004). Moreover, particularly strong ERN responses to feedback emerge

when negative feedback is unexpected or if positive feedback is omitted unexpectedly (Holroyd, Nieuwenhuis, Yeung, & Carter, 2003).

These findings clearly suggest that the ERN signals deviations from the agent's expectations. This view allows for a first clear hypothesis for the following experiments: Because successful rule violations are perfectly predictable events for the agent, they should not give rise to ERN-like waveforms in the ERP. In other words, what counts "is not what is correct or an error in the eyes of the experimenter, but rather what is deemed correct or an error by the brain of the subject. These are not identical, and some confusion in the literature arises from the assumption that they are." (Gehring et al., 2012, p. 241). The following experiments set out to confirm this first hypothesis, i.e., that rule violations do not evoke an ERN-like waveform as it is typically observed for unintended failures to obey a rule (see also Stemmer, Witzke, & Schönle, 2001).

But is there also reason to expect a specific electrophysiological signature that would set rule violations apart from normal, rule-based behaviour? Current theorizing on a different component of the ERP – the P300 – does indeed suggest a positive answer to this question (Nieuwenhuis, Aston-Jones, Cohen, 2005; Verleger, 1997; Verleger, Jaśkowski, & Wascher, 2005).[8] More precisely, the P300 component might reflect processes that translate a given stimulus into the appropriate response, possibly in terms of a "prepared reflex" (Verleger et al., 2005, p. 179, referring to De Jong, 1993). The connotation of the concept of prepared reflexes as used in this theory, however, is somewhat different from the view described in Section 2.2 (cf. Hommel, 2000; Woodworth, 1938). Rather than focusing on the role of intentions for generating automatic S-R associations, Verleger and colleagues stress that certain

[8] An alternative term for the ERP component described here is "P3b" which is distinct from the stimulus-driven processes related to detection and attention as indexed by "P3a" (e.g., Polich, 2007).

stimuli might be linked closely enough to corresponding responses to be given in a reflex-like manner (cf. also Frith & Done, 1986). This strong coupling between stimuli and following responses, their theory suggests, is indexed by the P300 component of the ERP.

Evidence for this claim comes from a study in which participants performed a simple classification task (Roche & O'Mara, 2003). Participants first trained the mapping of a particular stimulus to the corresponding response; in a following test block, this particular stimulus triggered an enlarged P300 response as compared to the remaining stimuli. Moreover, the P300 response to the trained stimulus occurred earlier than for the control stimuli. Both findings are compatible with the view of the P300 as a marker for relatively automatic processes mediating between stimuli and associated responses (Verleger et al., 2005).

Importantly, this view yields direct predictions regarding the electrophysiological signature of rule violations, because rule violations can be construed as the very opposite of responding with a canonical response to a stimulus. Accordingly, rule violations should be characterized by an attenuated P300 waveform with either lower amplitudes and/or longer peak latencies as compared to normal, correct responses.

This prediction can also be reconciled with the *context updating hypothesis*, the major alternative theory concerning the functional significance of the P300 component (Donchin, 1981; Donchin & Coles, 1988; Polich, 2007). According to this framework, the P300 component reflects memory processes that guide behavioural decisions (but see Verleger, 2008). Accordingly, P300 amplitudes were found to be reduced during memory load (e.g., Kotchoubey, Jordan, Grözinger, & Westphal, 1996; Pratt, Willoughby, & Swick, 2011) and, similarly, P300 was shown to be affected by response complexity (prolonged latencies in more complex tasks; Hoffman, Simons, & Houck, 1983; see also Kok, 2001). Conceptualising the P300 component as being related to memory retrieval

thus also yields that rule violations should give rise to reduced P300 amplitudes and/or longer latencies compared to normal, rule-based responses.

In sum, the following experiments tested two hypotheses. The first hypothesis targeted differences between rule violations and unintended errors and I expected an ERN response only for errors but nor for violations. The second hypothesis targeted possible differences between rule violations and normal, rule-based behaviour, and I expected rule violations to show attenuated P300 responses in terms of reduced amplitude and/or prolonged latency.

11 Experiment 7: Of chickens, eggs, and yolk

The design of Experiment 7 was similar to the design of Experiment 4. Participants were again asked at the beginning of each trial whether they were going to follow a rule for an upcoming task or whether they intended to violate the rule. Yet, I introduced several changes to optimize the design for electrophysiological recordings.

As a first change, I had participants perform a simple keypress task instead of the mouse classification task of Experiment 4 to minimize overt movements. Secondly, I changed the framing of the story to motivate the participants throughout the session. To this end, the experiment was designed as a computer game in which participants operated an "egg factory" by placing egg cups under a chicken that was soon going to lay an egg (**Fig. 23A**). To ensure smooth operation of the factory, the egg had to be placed under the chicken's rear whereas wrong placements would destroy the egg.

The two hypotheses were tested by examining (a) the stimulus-locked ERP elicited by the onset of the chicken stimulus (probing for changes of the P300 component) and (b) the response-locked ERP when placing the egg cup (probing for a possible ERN).

11.1 Method

11.1.1 Participants

Sixteen volunteers were recruited; the data of two participants had to be replaced due to technical difficulties. All participants of the final sample (mean age: 22.1 years, 14 female, 1 left-hander, 1 ambidextrous) reported normal or corrected-to-normal vision and received either course credit or monetary compensation.

11.1.2 Apparatus and stimuli

Participants sat in front of a 17'' monitor and responded with their left and right index finger on the A and the Hash (#) key of a standard German QWERTZ keyboard. The keys were marked with small stickers.

As described in the introduction, the task was embedded in a game-like setting which made the participants operate an egg factory (cf. **Fig. 23A**). The screen consisted of a conveyor belt spanning the horizontal midline right below the screen center (19.8 cm x 2.0 cm). Two tubes were displayed to the left and right of the screen, extending from the virtual ceiling (6.8 cm x 9.1 cm each). The upper center of the screen either featured a shutter (6.0 cm x 9.4 cm) or the image of a chicken looking to the left or to the right (approximately 4.7 cm x 6.3 cm).

Further stimuli that appeared on screen during a trial were a warning sign (4.4 cm x 4.7 cm) that was superimposed on the shutter, an egg (1.3 cm x 1.5 cm), as well as a red and a blue egg cup (1.5 cm x 1.7 cm).

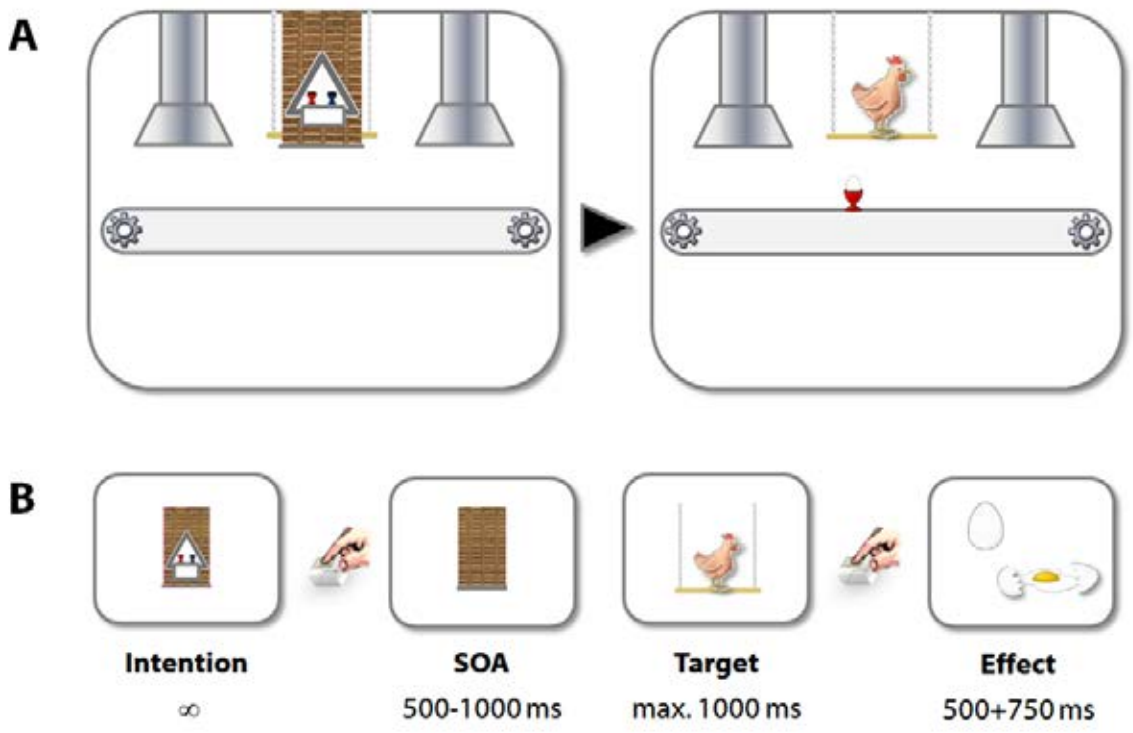

Fig. 23. Design and procedure of Experiment 7. **(A)** Participants first announced whether they wanted to adhere to the factory rules (perform correctly) or whether they wanted to violate these rule and commit an error by intention. These intentions implied that participants placed an egg cup either in a position to catch a falling egg or in a position where the egg would not be caught and destroyed. **(B)** The intention response was given at leisure and was followed by a variable SOA. The critical events for all analyses were target onset and the corresponding response. This response was followed by an animated effect that illustrated the egg's fate.

11.1.3 Instructions

Participants received verbal instructions, supported by exemplar stimuli on screen. They were first informed that, to ensure smooth operation of the factory, they would have to place a cup to the left or to the right of the chicken in each trial by pressing the left or the right key. Correct cup placement ensured that an egg could fall down in the cup, be carried away on the belt and absorbed in the tubes. Wrong cup placement resulted in the egg hitting the conveyor belt and shattering to pieces. Both effects were demonstrated as on-screen animations.

Afterwards, participants were introduced to the crucial element of the task which was a compliance prompt similar to Experiment 4. This compliance prompt consisted of the warning sign which showed two cups: A red and a blue one standing next to each other. Below these cups appeared the letters R (for German "richtig", correct) and F (for German "falsch", wrong). Participants pressed the left or the right key to indicate whether they wanted to stick to the factory regulations (perform correctly) or whether they wanted to violate them (and commit an error by intention). Each intention was indicated by a constant cup colour which was counterbalanced across participants.

11.1.4 Procedure

Trials started with the compliance prompt that stayed on screen until the participant gave the corresponding response. Afterwards, the warning sign disappeared and the empty shutter was displayed for a variable SOA (500 ms vs. 750 ms vs. 1000 ms). Then the shutter disappeared and gave way to a chicken looking to the left or right (the target stimulus). The chicken waited for up to 1000 ms to lay an egg which appeared below the chicken's rear. This interval served as response deadline for the participant who had to place an egg cup to the left or right by pressing the corresponding key.

The cup appeared immediately after the response and its colour depended on the participant's announced intention. Whereas the events up to this point were static and discrete to allow for undisturbed EEG recording, the remaining procedure was now animated and showed the egg falling down and either landing safely in the cup (if the cup stood below the chicken's rear) or shattering on the conveyor belt (overall duration: 500 ms). Then, the belt started moving and transported the egg to the nearest tube which started to absorb the contents of the belt (750 ms). Finally, the shutter went down again and the empty factory was displayed for an ITI of 1000 ms.

Participants completed a training block of 48 trials and 10 experimental blocks of 48 trials each. Individual sessions lasted between 1.5 hours and 2 hours including preparation of the EEG electrodes.

11.1.5 Electrophysiological recordings

EEG data was recorded throughout the session from the following electrodes of the extended 10-20 system: FP1, FP2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10. I used average reference and recorded the EEG signal at a sampling rate of 500 Hz, low-pass filtered at 100 Hz. The signal was amplified by a BrainVision QuickAmp amplifier with active electrodes (actiCAP; Brain Products, Germany) and impedances were below 10kΩ at the start of the experiment.

Ocular movements were recorded with passive electrodes on the outer canthi of both eyes and above and below the left eye (electrooculogram; EOG). Participants were encouraged to reduce eye movements and blinks, especially between target onset and response to minimize artefacts in the EEG data.

11.1.6 Data treatment

Unintended errors in terms of wrong keypresses occurred only rarely (2.0%), as did accidentally correct responses (after having announced to violate the factory rules; 2.0%). Procedural errors – double keypresses for intention or target response, responses during the SOA or during feedback– occurred in additional 3.9% of the trials. These data were excluded from the RT analysis as were trials following such errors. Furthermore, I corrected for outliers by excluding trials with RTs that deviated by more than 2.5 standard deviations from the corresponding cell mean, calculated separately for each participant and condition (2.8%).

For EEG analyses, I only used trials in which participants responded correctly as intended, violated the factory rules as intended, or committed

an unintended error by pressing the wrong key in response to the chicken. As for RT analyses, I excluded trials following any errors and performed the same outlier correction. I segmented the remaining trials into stimulus-locked epochs around the chicken onset (500 ms pre-stimulus to 1000 ms post-stimulus) response-locked epochs around the corresponding response (1000 ms pre-response to 500 ms post-response). Data was preprocessed via FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) and custom Matlab scripts. I first applied filters – 1 Hz highpass, 70 Hz lowpass, [47.5 Hz; 52.5 Hz] bandstop – and eliminated trials with artefacts by using the FieldTrip outlier detection mechanism based on z-scores (z = 20). Ocular artefacts were addressed via independent component analysis (ICA; Makeig, Bell, Jung, & Sejnowski, 1996) and I removed components that correlated with at least one EOG channel (r > .40; Flexer, Bauer, Pripfl, & Dorffner, 2005). Finally, baseline correction and filters were re-applied.

11.2 Results

11.2.1 Behavioural data

Participants showed a highly consistent preference for adherence to the factory rules (62%) which clearly deviated from chance level, $t(15) = 16.9$, $p < .001$, d = 4.23. Furthermore, correct responses were significantly faster than rule violations (461 ms vs. 492 ms), $t(15) = 3.44$, $p = .004$, d = 0.86.

11.2.2 Stimulus-locked ERPs

The stimulus-locked ERPs at the electrode locations Cz and Pz are plotted in the left panels of **Figure 24**. A clear P300 response was present for both, correct response trials and rule violation trials. Descriptively, the P300 component peaked considerably earlier in correct response trials than in rule violation trials (Cz: 420 ms vs. 470 ms; Pz: 404 ms vs. 448 ms) and the P300 in correct response trials also appeared slightly more pronounced than in rule violation trials.

To qualify these observations, I first extracted the time to maximum amplitude in a time window of 300 to 600 ms post-stimulus for each participant to probe for P300 latency for both electrode sites. The time window was chosen with respect to the grand-average waveform and by considering previous studies on the P300 component (e.g., Bennington & Polich, 1999; O'Connell et al., 2012; for similar ranges, see also Duncan-Johnson & Kopell, 1981; Ilan & Polich, 1999). The resulting latency scores were then submitted to a 2 x 2 repeated-measures ANOVA with the factors rule compliance (correct response trials vs. rule violation trials) and electrode (Cz vs. Pz). This analysis yielded a significant main effect of rule compliance, $F(1, 15) = 7.75$, $p = .014$, $\eta_p^2 = .34$, whereas neither the main effect of electrode, $F(1, 15) = 1.60$, $p = .226$, $\eta_p^2 = .10$, nor the interaction approached significance ($F < 1$).

To assess changes in P300 amplitude, I extracted the mean amplitude from 300 to 450 ms post-stimulus where a clear P300 topography was visible (**Fig. 24**, top-right panel). The corresponding 2 x 2 ANOVA yielded a main effect of rule compliance[9], $F(1, 15) = 10.57$, $p = .005$, $\eta_p^2 = .41$, whereas a significant main effect indicated higher P300 amplitudes at Pz as compared to Cz, $F(1, 15) = 12.79$, $p = .003$, $\eta_p^2 = .46$. The interaction was not significant, $F(1, 15) = 1.41$, $p = .254$, $\eta_p^2 = .09$.

[9] The present 2 x 2 ANOVA on the individual channels has rather limited power. Even stronger results could be achieved by extending the analysis to more electrodes at posterior locations, either by adding these electrodes to the analysis reported here or by (temporo-)spatial principal component analysis (PCA; Dien, 2012; Dien & Frishkoff, 2006). Yet, I chose to remain conservative for two reasons. First, such additional analyses would represent a clear a-posteriori test, and secondly, Pz is a candidate location for P300 (e.g., Polich, 2007) and would thus be expected to show the experimental effects in any case.

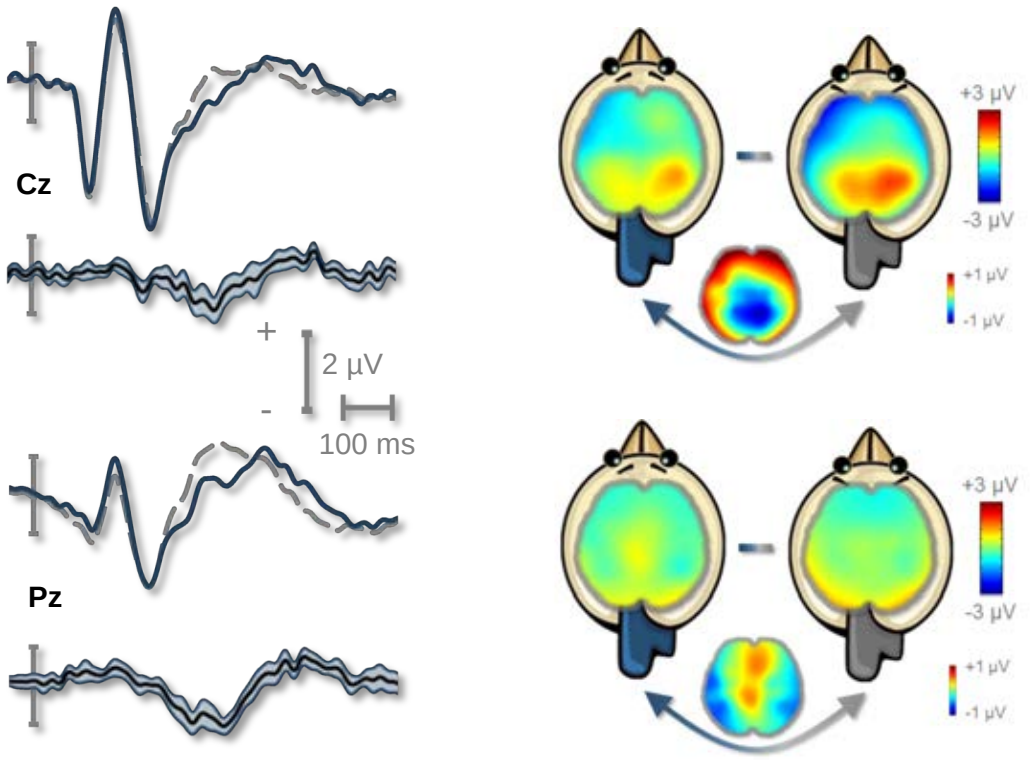

Fig. 24. Results of the stimulus-locked analysis of Experiment 7. **(Left Panels)** ERP results at the two electrode locations Cz and Pz. The upper plots for each electrode show the resulting ERP for correct response trials (dashed gray line) and for violation trials (solid blue line). The lower plots for each electrode show the difference wave ($voltage_{Violation}$ – $voltage_{Correct}$; solid black line) ± 1 standard error of paired differences, computed separately for each data point (coloured area). **(Right Panels)** Mean voltage distributions across the scalp relating to the first and the second half of the P300 interval (top: 300-450 ms; bottom: 450-600 ms). Correct response trials are plotted to the right (gray heads) whereas violation trials are plotted to the left (blue heads). The distribution of the difference wave is plotted in between the heads ($voltage_{Violation}$ – $voltage_{Correct}$).

An additional noteworthy observation relates to the scalp distribution of the P300 in the two conditions as shown in the right panels of **Figure 24**. For the first interval (300-450 ms post-stimulus), the scalp distribution showed a central, posterior positivity as would be expected for the P300 component. The distribution of the difference wave further indicated an attenuation of this component in its typical posterior region. In contrast, for

the second interval (450-600 ms post-stimulus), no clear response was evident for correct response trials whereas a central positivity emerged for rule violation trials. Accordingly, the difference waveform showed a medial, fronto-central positivity that does not seem to resemble the distribution observed in the first interval.

11.2.3 Response-locked ERPs

To probe for response-related effects of rule violations, I compared the ERP for the electrodes Fz and Cz – two candidate electrodes for ERN and P_E (**Fig. 25**, left panels). As can be seen from the figure, differences between rule violations and rule-based responses were small and unreliable across the entire epoch. Still, a first planned analysis targeted a time interval that would capture the ERN in case of unintended errors (0 ms to 100 ms post-response; Rodríguez-Fornells, Kurzbuch, & Münte, 2002) and I subjected the mean amplitudes in this interval to a 2 x 2 ANOVA with the factors rule compliance (correct responses vs. rule violations) and electrode (Fz vs. Cz). Most importantly, the main effect of rule compliance did not approach significance, $F(1, 15) = 0.14$, $p = .714$, $\eta_p^2 = .01$. A significant main effect of electrode further indicated overall lower voltages at Fz as compared to Cz, $F(1, 15) = 10.34$, $p = .006$, $\eta_p^2 = .41$, whereas the interaction was not significant ($F < 1$). A similar analysis of the mean amplitudes in a time-window suitable for capturing a possible P_E (100 ms to 300 ms post-response) did not show any effect to be significant ($ps > .179$, $\eta_p^2 < .12$).

The right panels of **Figure 25** show the corresponding scalp distributions in both time windows. Corroborating the above analyses, no apparent differences emerged for the distribution of correct responses and rule violations. The same was true for the response-locked lateralized readiness potentials (LRPs; cf. **Appendix D**).

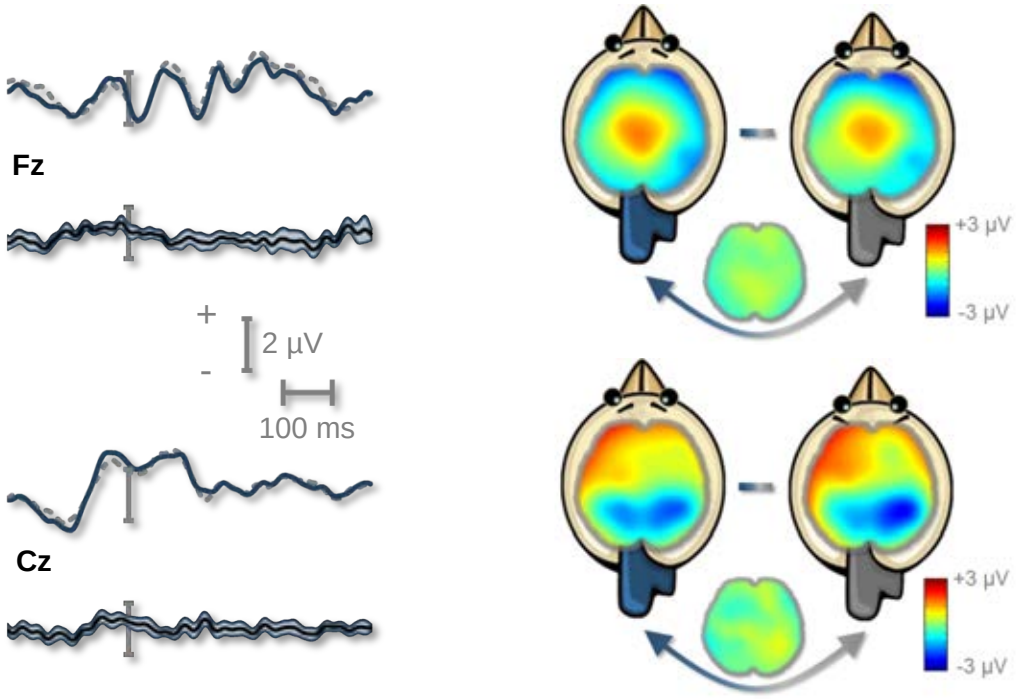

Fig. 25. Response-locked ERPs for correct responses and rule violations. **(Left Panels)** ERP results at Fz and Cz. The upper plots for each electrode show the resulting ERP for correct responses (dashed gray line) and rule violations (solid blue line). The lower plots for each electrode show the difference wave ($voltage_{Violation}$ – $voltage_{Correct}$; solid black line) ± 1 standard error of paired differences, computed separately for each data point (coloured area). **(Right Panels)** Mean voltage distributions at 0-100 ms post-response (upper plot) and 100-300 ms post-response (lower plot). Correct responses are plotted to the right (gray heads) whereas rule violations are plotted to the left (blue heads); the difference wave is plotted in between the heads.

To confirm that the present setup is per se able to detect ERN-like responses, I also compared correct responses and unintended errors for Fz and Cz. A pronounced ERN was present at both electrodes as shown in **Figure 26** (left panels). This observation is supported by a 2 x 2 repeated-measures ANOVA with the factors response type (correct responses vs. unintended error) and electrode (Fz vs. Cz) on the mean voltages between 0 and 100 ms post-response. The analysis showed a significant main effect of response type, $F(1, 15) = 17.39$, $p < .001$, $\eta_p^2 = .54$. Additionally, overall voltages were lower at Fz as compared to Cz, $F(1, 15) = 4.56$, $p = .050$, $\eta_p^2 = .23$, whereas the interaction was not significant ($F < 1$).

Similarly, I probed for P_E by comparing mean voltages at Fz and Cz in the time window of 100-300 ms post-response. The corresponding 2 x 2 ANOVA yielded a significant main effect of condition, $F(1, 15) = 10.47$, $p = .006$, $\eta_p^2 = .41$, with higher voltages after errors than after correct responses. Neither the main effect of electrode ($F < 1$) nor the interaction were significant, $F(1, 15) = 1.89$, $p = .189$, $\eta_p^2 = .11$.

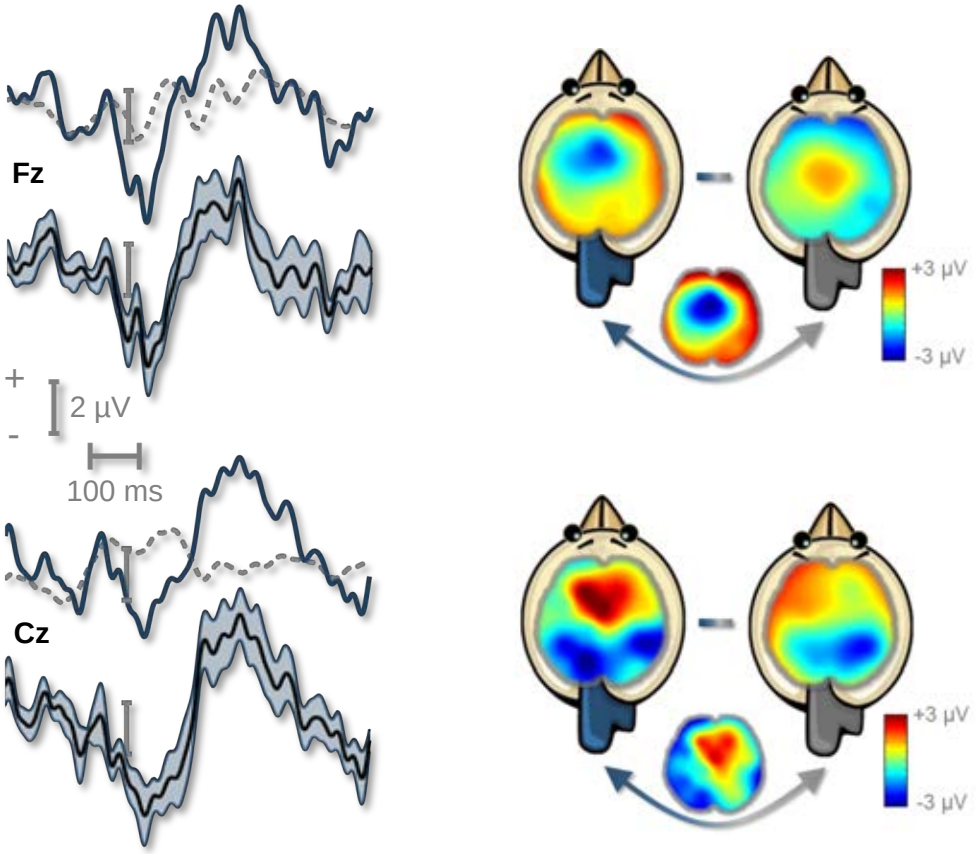

Fig. 26. Response-locked ERPs for correct responses and unintended errors. **(Left Panels)** ERP results at Fz and Cz. The upper plots for each electrode show the resulting ERP for correct responses (dashed gray line) and errors (solid blue line). The lower plots for each electrode show the difference wave ($voltage_{Error}$ – $voltage_{Correct}$; solid black line) ± 1 standard error of paired differences, computed separately for each data point (coloured area). **(Right Panels)** Mean voltage distributions 0-100 ms post-response (upper plot) and 100-300 ms post-response (lower plot). Correct responses are plotted to the right (gray heads) whereas unintended errrors are plotted to the left (blue heads). The distribution of the difference wave is plotted in between the heads ($voltage_{Error}$ – $voltage_{Correct}$).

11.3 Discussion

Experiment 7 tested two hypotheses on the electrophysiological signature of rule violations, both of which were supported by the data. First, such violations do not elicit any ERN-like waveforms that are routinely observed for unintended errors (e.g., Gehring et al., 2012; Holroyd & Coles, 2002).[10] Secondly, the peak latency of the P300 component was delayed for rule violations as compared to rule-based responses whereas the mean amplitude of this component was reduced.

Before drawing any conclusions from these results, I strived to address an obvious alternative explanation. As described above, the ERN was typically measured in paradigms in which errors occurred relatively rarely and were thus unexpected (for reviews, see e.g., Falkenstein et al., 2000; Gehring et al., 2012). The absence of ERN-like responses for the present rule violations might thus at least partly be driven by the comparatively high frequency of such violations (in about 4 of 10 trials).

Several findings speak against this issue. For one, the ERN was found to be primarily related to error significance (Hajcak, Moser, Yeung, & Simons, 2005; Maier, Steinhauser, & Hübner, 2008), which should be rather independent of error frequency. Moreover, ERN responses have also been reported for error rates that match the frequency of rule violations in Experiment 7 (e.g., Vocat, Pourtois, & Vuilleumier, 2008). Still, an empirical test of this alternative explanation seems to be in order.

[10] The results reported in Section 11.2.2 were based on direct comparisons of either rule violations or errors to normal, rule-based responses. Strictly speaking, these data do not address potential differences between violations and errors even though the ERN was significant for the latter comparison but not for the former. Yet, I refrained from reporting such tests because violations and rule-based responses were not only similar but virtually *identical* in all aspects of the response-locked analysis.

12 Experiment 8: Controlling for violation frequencies

Experiment 8 employed a simple and straightforward design to control for the frequency of rule violations. To this end, participants responded to the identity of a letter stimulus by pressing either a left or a right response key. The target stimuli were white in the majority of trials (5/6) and called for normal, rule-based responding whereas they were red in on 1 out of 6 trials which called for rule violations.

12.1 Method

Sixteen new participants were recruited (mean age: 22.4 years, 3 male, all right-handed). They reported normal or corrected-to-normal vision and received either course credit or monetary compensation

The task simply required participants to respond to the identity of a centrally presented letter with the same keys as in Experiment 7. Two letter identities were used – X and H – and letter-response assignment was counterbalanced across participants. The letters were displayed against a black background in 20 pt Arial font. In 5/6 of the trials, the letter was white and required normal responses. In the remaining 1/6 of the trials, the letter was red and participants were instructed to violate the original mapping rule in this case. Trials started with a fixation cross (1000 ms) followed by the target letter which was displayed until a response was given (max. 1000 ms). The next trial started after an ITI of 1000 ms. I did not provide any error feedback, but trials were aborted after response anticipations (during fixation) or if more than one response was given in a single trial.

Participants worked through a training block of 30 trials and 9 experimental blocks of 60 trials each. The whole session took about 1.5 hours to complete, including preparation of the EEG electrodes. Electrophysiological recordings and data processing were the same as in Experiment 7.

12.2 Results

12.2.1 Behavioural data

As in Experiment 7, Unintended errors in terms of wrong keypresses occurred rarely (1.9%), as did accidentally correct responses (in violation trials; 2.0%). Procedural errors – response anticipations and omissions or double keypresses – occurred in additional 1.7% of the trials. These data were excluded from the RT analysis as were trials following such errors. Another 2.1% of the trials were excluded as outliers.

Correct responses were again significantly faster than rule violations (429 ms vs. 610 ms), $t(15) = 17.92$, $p < .001$, d = 4.48. These results, however, cannot be taken as a pure measure for the impact of rule violations because they are clearly confounded with the effect of stimulus frequency (especially because the occurrence of violation trials could not be predicted). For the same reason, I did not perform any more sophisticated analyses to probe for sequential effects on RTs but concentrated on the electrophysiological data.

12.2.2 Stimulus-locked ERPs

The stimulus-locked ERPs at the electrode locations Cz and Pz are plotted in the left panels of **Figure 27**. As in Experiment 7, I first extracted the time to maximum amplitude in a time window of 300 to 600 ms post-stimulus for each participant to probe for P300 latency for both electrode sites. The latency scores were then submitted to a 2 x 2 repeated-measures ANOVA with the factors rule compliance (correct response trials vs. rule

violation trials) and electrode (Cz vs. Pz). The analysis yielded a marginally significant main effect of rule compliance, $F(1, 15) = 3.42$, $p = .084$, $\eta_p^2 = .19$, that was qualified by a significant interaction of rule compliance and electrode, $F(1, 15) = 9.97$, $p = .006$, $\eta_p^2 = .40$. The main effect of electrode was not significant, $F(1, 15) = 2.00$, $p = .178$, $\eta_p^2 = .12$. Tested separately, the difference between correct responses and rule violations was significant only for Pz (322 ms vs. 402 ms), $t(15) = 3.04$, $p = .008$, $d = 0.76$, but not for Cz (397 ms vs. 387 ms), $t(15) = -0.48$, $p = .641$, $d = -0.12$.

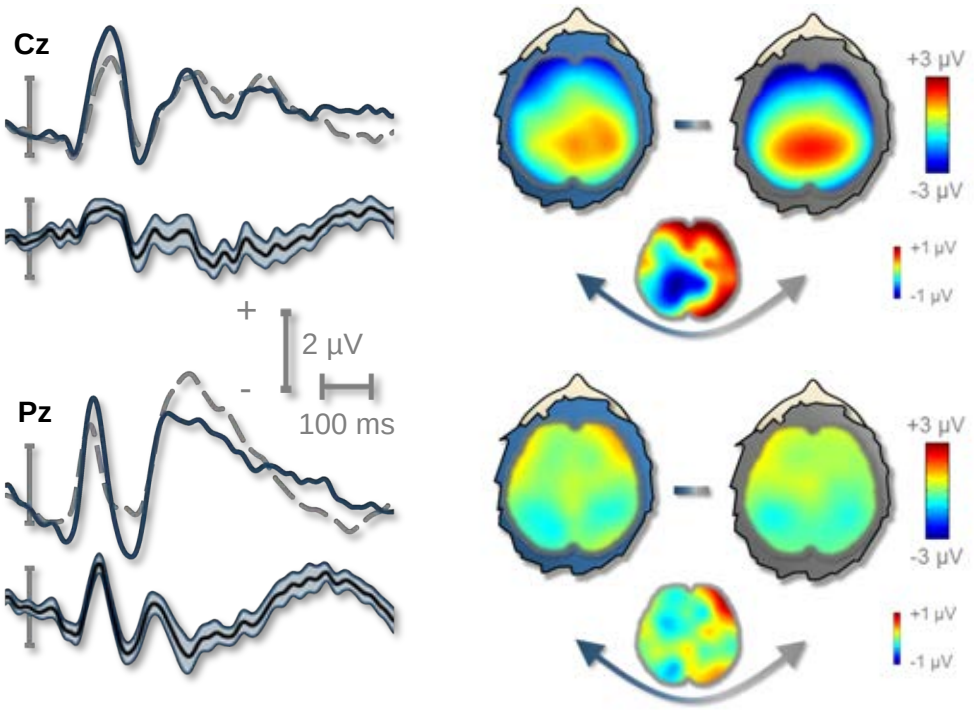

Fig. 27. Results of the stimulus-locked analysis of Experiment 8. **(Left Panels)** ERP results at Cz and Pz. The upper plots for each electrode show the resulting ERP for correct response trials (dashed gray line) and for violation trials (solid blue line). The lower plots for each electrode show the difference wave ($voltage_{Violation}$ – $voltage_{Correct}$; solid black line) ± 1 standard error of paired differences, computed separately for each data point (coloured area). **(Right Panels)** Mean voltage distributions across the scalp relating to the first and the second half of the P300 interval (top: 300-450 ms; bottom: 450-600 ms). Correct response trials are plotted to the right (gray heads) whereas violation trials are plotted to the left (blue heads). The distribution of the difference wave is plotted in between the heads ($voltage_{Violation}$ – $voltage_{Correct}$).

Mean P300 amplitude was assessed as in Experiment 7, and the corresponding 2 x 2 ANOVA again yielded a significant main effect of rule compliance, $F(1, 15) = 6.25$, $p = .025$, $\eta_p^2 = .29$. Additionally, a significant main effect indicated higher P300 amplitudes at Pz as compared to Cz, $F(1, 15) = 9.25$, $p = .008$, $\eta_p^2 = .38$, and the effect of rule compliance was stronger at Pz than at Cz, $F(1, 15) = 5.35$, $p = .035$, $\eta_p^2 = .26$.

Moreover, the scalp distribution of the P300 in the two conditions (**Figure 27**, right panels) replicated the findings of Experiment 7 for the first interval (300-450 ms post-stimulus). Here, the scalp distribution again showed a central, posterior positivity as would be expected for the P300 component and the distribution of the difference wave indicated an attenuation of this component in its typical posterior region. For the second interval (450-600 ms post-stimulus), no clear response was evident for correct response trials again. Contrary to the findings of Experiment 7, rule violation trials showed a largely similar effect as correct responses for this interval.

12.2.3 Response-locked ERPs

Again, I first compared the ERP for the electrodes Fz and Cz (**Fig. 28**, left panels) for the time interval that would capture the ERN in case of unintended errors (0 ms to 100 ms post-response). The corresponding 2 x 2 ANOVA with the factors rule compliance (correct responses vs. rule violations) and electrode (Fz vs. Cz) did not show any effects of rule compliance, $F(1, 15) = 0.30$, $p = .594$, $\eta_p^2 = .02$. The main effect of electrode was not significant, $F(1, 15) = 2.70$, $p = .121$, $\eta_p^2 = .15$, even though overall voltages were again smaller at Fz as compared to Cz. The interaction was not significant either ($F < 1$).

Interestingly, the analysis of the second time window (100-300 ms post-response) showed a significant main effect of rule compliance, $F(1, 15) = 8.29$, $p = .011$, $\eta_p^2 = .36$, driven by more positive voltages for violations

than for correct responses. Furthermore, mean amplitudes were overall higher at Fz than at Cz, $F(1, 15) = 18.69$, $p < .001$, $\eta_p^2 = .55$, whereas the interaction was not significant, $F(1, 15) = 1.94$, $p = .184$, $\eta_p^2 = .11$.

The right panels of **Figure 28** show the corresponding scalp distributions in both time windows. Corroborating the above analyses, no apparent differences emerged for the first time interval whereas a pronounced negativity emerged for correct responses in the second window that was not present for rule violations.

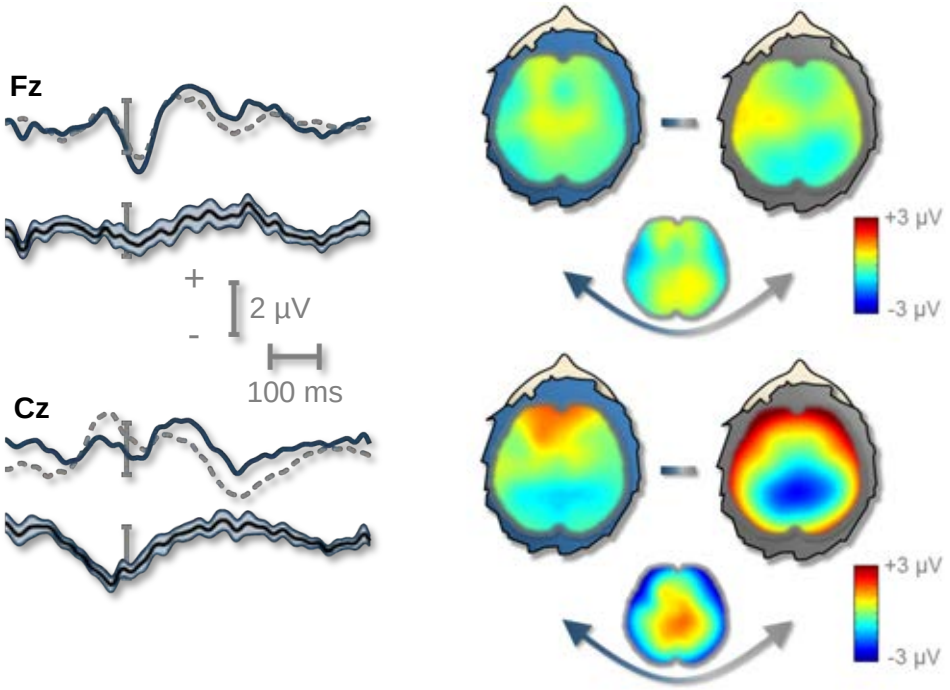

Fig. 28. Response-locked ERPs for correct responses and rule violations. **(Left Panels)** ERP results at the two electrode locations Fz and Cz. The upper plots for each electrode show the resulting ERP for correct responses (dashed gray line) and rule violations (solid blue line). The lower plots for each electrode show the difference wave ($voltage_{Violation}$ – $voltage_{Correct}$; solid black line) ± 1 standard error of paired differences, computed separately for each data point (coloured area). **(Right Panels)** Mean voltage distributions across the scalp, 0-100 ms post-response (upper plot) and 100-300 ms post-response (lower plot). Correct responses are plotted to the right (gray heads) whereas rule violations are plotted to the left (blue heads). The distribution of the difference wave is plotted in between the heads ($voltage_{Violation}$ – $voltage_{Correct}$).

To confirm that an ERN also emerged for the changed experimental setup, I compared the ERPs of correct responses and unintended errors for the electrodes Fz and Cz. As in Experiment 7, a pronounced ERN was present at both electrodes as shown in **Figure 29** (left panels).

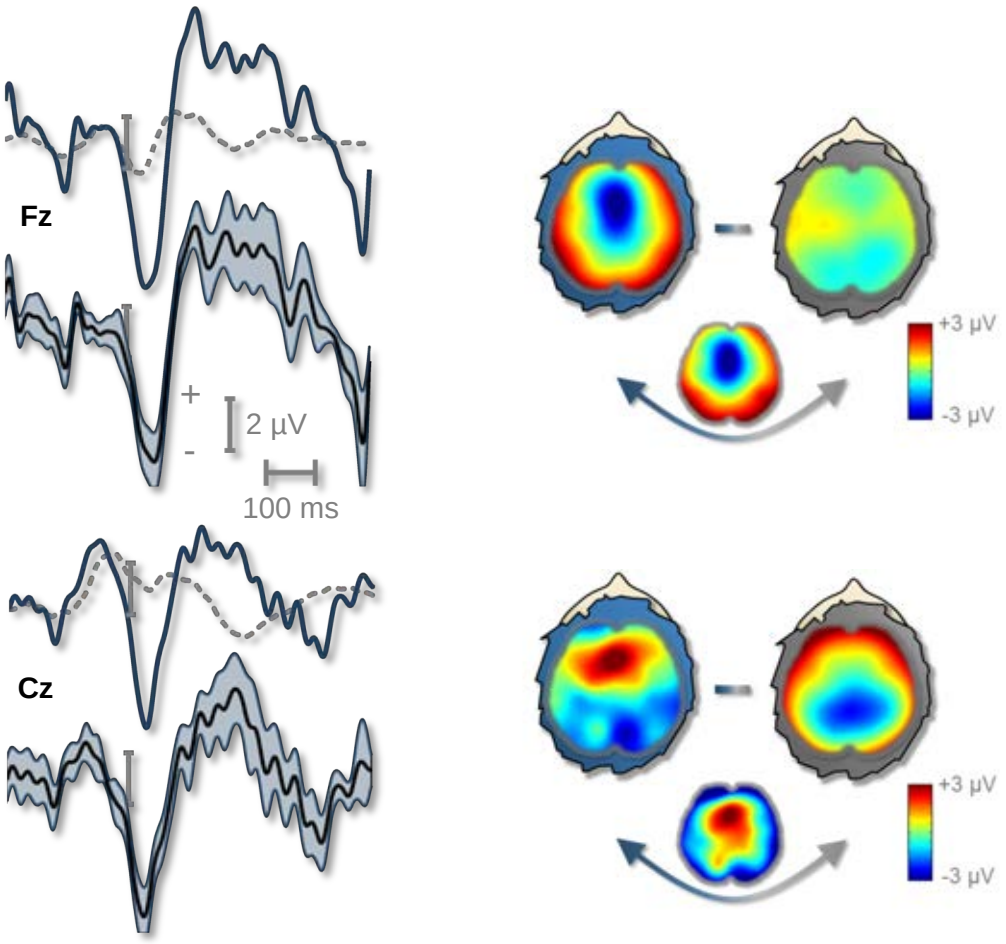

Fig. 29. Response-locked ERPs for correct responses and unintended errors. **(Left Panels)** ERP results at Fz and Cz. The upper plots for each electrode show the resulting ERP for correct responses (dashed gray line) and errors (solid blue line). The lower plots for each electrode show the difference wave ($voltage_{Error}$ – $voltage_{Correct}$; solid black line) ± 1 standard error of paired differences, computed separately for each data point (coloured area). **(Right Panels)** Mean voltage distributions across the scalp relating 0-100 ms post-response (upper plot) and 100-300 ms post-response (lower plot). Correct responses are plotted to the right (gray heads) whereas rule violations are plotted to the left (blue heads). The distribution of the difference wave is plotted in between the heads ($voltage_{Error}$ – $voltage_{Correct}$).

A 2 x 2 repeated-measures ANOVA with the factors response type (correct responses vs. unintended error) and electrode (Fz vs. Cz) on the mean voltages between 0 and 100 ms post-response again showed a highly significant main effect of response type, $F(1, 15) = 20.37$, $p < .001$, $\eta_p^2 = .58$. Neither the main effect of electrode not the interaction was significant (*ps* > .271, $\eta_p^2 < .08$).

Similarly, I probed of P_E by comparing the mean voltages at Fz and Cz in the time window of 100 to 300 ms post-response. The corresponding 2 x 2 ANOVA showed the main effect of condition to be significant, $F(1, 15) = 13.98$, $p = .002$, $\eta_p^2 = .48$, with higher voltages after errors than after correct responses. A significant main effect of electrode was driven by overall higher voltages at Fz than at Cz, $F(1, 15) = 7.23$, $p = .017$, $\eta_p^2 = .03$, whereas the interaction was not significant ($F < 1$).

12.3 Discussion

Experiment 8 replicated both key findings of Experiment 7. Again, rule violations did not elicit any ERN-like waveforms that characterise unintended errors (e.g., Gehring et al., 2012; Holroyd & Coles, 2002) and, secondly, the peak latency of the P300 component was delayed for rule violations as compared to rule-based responses and the mean amplitude of this component was attenuated for rule violations.

Most importantly, this was true even though violations now occurred only rarely throughout the experiment, a condition that would arguably work against both effects. As outlined in Section 11.3, ERN responses to unintended errors were mainly observed with rather infrequent errors (e.g., Holroyd & Coles, 2002) – decreasing the frequency of rule violations should thus favour ERN-like responses. The same is true for rule violations: Rendering the relevant stimuli infrequent (i.e., turning them into oddballs) might have also increased the P300 amplitude (e.g., Polich, 2007). Still,

both effects occurred as in Experiment 7 with an absent ERN for rule violations and a prolonged P300 with and decreased rather than increased amplitude.

Asides from painting a first picture of the electrophysiological signature of rule violations, the present findings also support reinforcement-learning accounts of the ERN (Holroyd & Coles, 2002; Holroyd et al., 2005). These accounts link the ERN to the detection of unexpected events whereas rule violation responses were clearly predictable for the participants. Thus even though the behaviour did not match the original mapping rule, the ERN depended entirely on the match of intention and behaviour, not on the match of behaviour and rule (cf. Gehring et al., 2012).

The present results also have direct implications for theories about the functional significance of the P300 component. More precisely, it supports accounts that link the P300 component to the process of translating a stimulus to a given response (Nieuwenhuis et al., 2005; Verleger, 1997; Verleger et al., 2005). According to these accounts P300 might be seen as an index of direct and reflex-like response tendencies to a stimulus that had been tightly associated with a canonical response (Roche & O'Mara, 2003; Verleger et al., 2005). Interestingly, Experiment 7 found differential P300 responses for one and the same stimulus depending on the agent's current intentions: When an agent was going to perform according to the instructed mapping rule, P300 responses occurred earlier and with higher amplitude as compared to a situation in which the agent was going to violate the mapping rule. In other words: It seems as if a direct translation from the target stimulus to the intended motor response occurred only for rule-based responses whereas such direct translation was not possible for violation responses. This pattern of results also fits with studies that completely decoupled stimuli and responses in free choice tasks in which the stimuli did not convey any response-relevant information (Keller et al.,

2006; Waszak et al., 2005). These studies reported decreased P300 responses in free choice settings as compared to forced choice settings.

Assuming that rule violations do not allow for a direct translation from stimulus to (violation) response, however, also raises the question whether changed instructions would yield similar results (as discussed in the context of Experiment 3-5). Here, the previously mentioned study by Schroder et al. (2012) clearly suggests a negative answer (cf. the introduction to Part 2 on p. 34, for a description of the experimental design). These authors did not report any changes of the P300 response in blocks that used changed instructions as compared to blocks employing the original mapping rule. Thus, the reported modulation of the P300 component does indeed seem to represent an electrophysiological marker of intended rule violations – complementing the behavioural perspective of Experiment 1-6.

In sum, the present findings provide a detailed view on rule violations and allow for a more explicit understanding of the processes governing this type of behaviour. These implications are discussed in the following sections.

PART 5: A NEW LOOK ON RULE VIOLATIONS

The present study aimed at a first characterization of the cognitive effects of intended rule violations. To this end, I conducted three experimental series that focused on different aspects of the phenomenon (cf. **Fig. 30**).

Experiment 1-3 employed alternating choice reaction tasks as a first, controlled approach. Committing violations required more time than responding according to an original mapping rule and this was even true for repeated rule violations, i.e., when participants were tuned towards violating. Crucially, these effects exceeded the impact of instructing the same behaviour in terms of a reversed mapping rule. By contrast, after-effects of violations on subsequent violation-unrelated behaviour were comparable to typical task-switching effects (i.e., there was no violation-specific slowing).

Experiment 4-5 then focused on the immediate effects of violating a rule. The results clearly suggest that the original mapping rule cannot be suppressed completely when an action is seen as a violation; and the continued influence of the mapping rule was much stronger when an action was labelled as rule violation as compared to a reversed mapping rule. Interestingly, reliable effects for rule violations also emerged when participants could choose whether to violate a rule or not in Experiment 4 and 6.

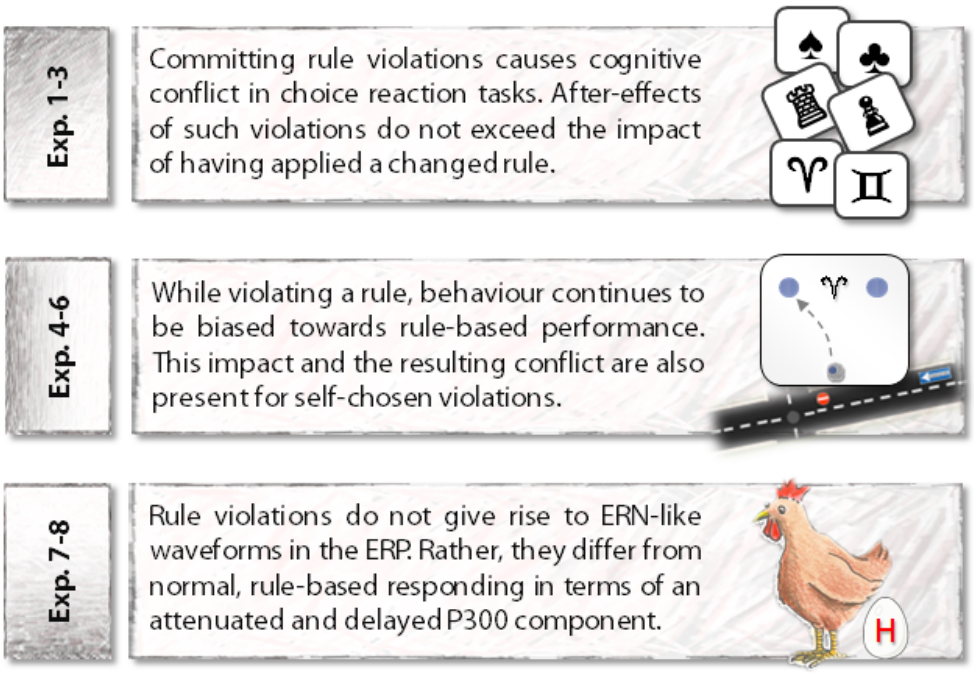

Fig. 30. Summary of the conducted experiments and their main findings. In addition to attesting cognitive conflict as a reliable consequence of rule violations, the results also allow for several speculations about the mechanisms underlying the observed effects as outlined below.

Experiment 7-8 finally probed for the electrophysiological signature of rule violations. Here, I did not observe any pronounced differences between the immediate after-effects of rule violations as compared to correct performance. By contrast, rule violations and correct performance seem to differ in how stimuli are processed that trigger one or the other action. This was particularly evident in the P300 component that occurred later and with reduced amplitude for rule-violations than for normal, rule-based responses. These findings show that a stimulus is not translated as directly into rule-violation behaviour as is the case for rule-based responses. Rather, they suggest additional processing steps to be necessary to successfully violate a rule.

13 Cognitive mechanisms underlying intended rule violations

When describing the impact of rules on cognitive processes (cf. Section 2.2), I focused on research demonstrating the rather automatic effects of task rules after only limited practice (e.g., Eriksen & Eriksen, 1974; Logan, 1988) or even without any practice, given appropriate instructions (e.g., Cohen-Kdoshay & Meiran, 2007; Kunde et al., 2003; Hommel, 2000; Wenke, Gaschler, & Nattkemper, 2007). While being impressive in their own right, these findings and the corresponding theoretical accounts would have been also able to explain the observed differences between rule violations and normal, rule-based behaviour – were it not for the pronounced difference between rule violations and reversed-rule responses as evident in several instances.

This fly in the (explanatory) ointment calls for additional mechanisms to account for the present findings. In the following, I will outline two speculations to fill this gap: The motivational conflict hypothesis and the negation hypothesis. The first hypothesis draws on the learning history of the agent and assumes a mixture of motivational influences that deteriorate performance by inducing an avoidance tendency. By contrast, the negation hypothesis draws on the possible mental representation of rules and their violation. This hypothesis predicts the effects of rule violations to be due to the difficulty of human agents to represent negations in general.

Each hypothesis might account for a different proportion of the data and, on a larger scale, might be more or less relevant for different types of rule violations. Also, both hypotheses make intriguing predictions that go beyond the effects investigated with the present experiments.

13.1 Expecting the unexpectable: The motivational conflict hypothesis

The present experiments were deliberately controlled and artificial to study rule violations in the absence of punishment, retaliation, or otherwise negative consequences that may occur outside the laboratory. However, such consequences are tightly linked to rule violations in everyday life and participants might perhaps continue to be affected by corresponding expectations even when told that these consequences would not occur.

In fact, sanctioning of rule violations is an elementary experience that even drives the developing understanding of rules and social norms (e.g., Colby et al., 1973; Hoffman, 2001; Kohlberg, 1971; Eisenberg, 2000). In turn, adults show a consistent bias towards punishing others who have violated a rule (e.g., Darley, 2009), even when the punisher is not directly affected by the violation behaviour (Fehr & Fischbacher, 2004). Continued experiences of ensuing punishment or at least possible punishment might thus not only bias decisions whether to violate a rule or not (e.g., Blanton & Christie, 2003; Reason, 1990), but they might also bias participants towards expecting punishment even when it is unlikely or even impossible to occur.

Such latent expectancies may have several side-effects relevant for committing a rule violation. First and foremost, labelling an action as rule violation might induce a significant degree of avoidance motivation concerning possible actions and action goals (see Eder & Rothermund, 2008, concerning the power of labelling for inducing motivational tendencies). This view might explain both, genuine differences between rule violations and rule-based behaviour as well as differences between rule-violations and reversed rule responses by behavioural inhibition or increased response caution (cf. also, Carver, 2006; Gray, 1990; Gray & McNaughton, 2000; Higgins, 1997, 2012).

More specifically, assuming that labelling an action as rule violation triggers avoidance motivation, the resulting motivational conflict is likely to decrease general response speed as I have observed throughout the present experiments. Motivational conflict is further likely to foster a continued activation of the original mapping rule. This latter interpretation is especially relevant for the trajectory analyses of Experiment 4-5 which, according to the motivational conflict hypothesis, might indicate an on-going struggle to overcome avoidance tendencies triggered by the "wrong" target stimulus (cf. Dignath, Pfister, Eder, Kiesel, & Kunde, in press).

Less clear are the predictions of the motivational conflict hypothesis regarding the observed effects in Experiment 7-8. Concerning the effects on the P300 component, only diffuse predictions based on differential involvement of attentional processes seem to be backed up by the literature (e.g., Kok, 2001). Instead, the differential activation of approach or avoidance motivation has been linked to more sustained frontal asymmetries of alpha power in the EEG (Coan & Allen, 2002; Davidson, 1993, 1998; Davidson, Ekman, Saron, Senulis, & Friesen, 1990).

In any case, the motivational conflict hypothesis seems to be able to account for a considerable proportion of the current results. And it also generates clear predictions for different settings. For instance, rule violations should be evaluated more negatively than rule-based responses, even in the absence of negative consequences as in the present experiments (cf. Aarts, De Houwer, Pourtois, 2012, for a suitable experimental approach to this question). This prediction, however, is not unique for the motivational conflict hypothesis and might also be based on the simple assumption that cognitive conflict is aversive (Dreisbach & Fischer, 2013).

A more distinctive prediction of the motivational conflict hypothesis relates to situations in which negative consequences *can* occur. Here, cognitive effects of violating a rule should be enhanced because the

assumed motivational conflict underlying the effects should be increased by the expected negative outcome. Accordingly, one could predict the described enhancement to be a parametric function of both, expected likelihood and severity of negative outcomes. The present results do not allow for any firm conclusions regarding this prediction which clearly awaits further research.

13.2 Unaccepting the unacceptable: The negation hypothesis

An alternative speculation about the mechanisms underlying the observed effects can be deduced from a careful analysis of the employed instructions. For all experiments except for Experiment 6, rule violations were instructed as doing the opposite of what the mapping rule prescribed for the target stimuli. In contrast, reversed rule responses were instructed by introducing two distinct tasks (with opposite mapping rules). Thus, reversed rule responses were instructed in terms of two mappings whereas violations were instructed in terms of one mapping and its *negation*.

On any account, the negation of a rule lies at the very heart of any rule violation behaviour – but psycholinguistic research suggests that a negation is processed uniquely on its own right (Just & Carpenter, 1976; Hasson & Glucksberg, 2006; Kaup, 2001; Wason & Johnson-Laird, 1972). In fact, producing a negation response was shown to require more time and resources than producing an affirmative response (Wason, 1959), and so does responding with a response opposite to the learned S-R association for a given stimulus (Schroder et al., 2012; Seymour, 1977). Further, responses opposite to a learned mapping are not facilitated by features of a target stimulus that are congruent or even semantically identical with the required opposite response (Seymour, 1977).

On a larger scale, intentions referring to negations are known to trigger ironic effects by promoting exactly the unwanted thoughts and actions when under mental load (Wegner, 1994, 2009). For instance, intending not to activate a given stereotype can actually promote stereotyping (Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008) and, similarly, intending to suppress anxiety-related thoughts may actually increase anxiety (Koster, Rassin, Crombez, & Näring, 2003). The same is true for intentions to reverse an unloved habit where planning what *not* to do has only limited chances of success (Adriaanse, van Oosten, de Ridder, de Wit, & Evers, 2011).

Importantly, the mentioned examples are driven directly by difficulty caused by negations (cf. Deutsch, Gawronski, & Strack, 2006). For instance, stereotypying can be reduced efficiently by activating the counter-stereotype (Gawronski et al., 2008) and unloved habits can be countered efficiently by forming implementation intentions specifying what actually to do in which situation (Adriaanse, Gollwitzer, de Ridder, de Wit, & Kroese, 2011; Gollwitzer, 1999; Gollwitzer & Sheeran, 2006).

But what causes the ineffectiveness of intentions that build on negations? A first, elaborate step towards answering this question is Gilbert's (1991) comparison of two philosophical approaches to the nature of mental representations. Dating back at least to Descartes, the *Cartesian approach* holds that mental representations draw on two separate processes: comprehension and assessment. These processes occur serially and are thought to be independent; an idea is understood in a first, purely semantic step (comprehension) and then judged to be right or wrong (assessment). By contrast, the *Spinozan approach* holds that the comprehension of an idea necessarily entails accepting the idea as true – comprehension and acceptance are thus seen as a single, conjoint process. In this view, wrong statements can only be identified as such by an effortful tagging process that *unaccepts* the idea. The unacceptance of an idea thus requires time and ressources, and the human mind does indeed seem to

work by and large as described in the Spinozan approach (Gilbert, 1991; Gilbert, Krull, & Malone, 1990; Petty & Cacioppo, 1986; Wegner, Coulton, & Wenzlaff, 1985).

These considerations are directly relevant for the concept of negations because negations are assumedly represented as a proposition that is tagged to be false (Clark & Chase, 1972, 194; Just & Carpenter, 1976; Kaup, 2001). When retrieving a negated concept, the proposition itself is thus accessed rather automatically whereas the falsity tag is only retrieved if enough time and resources are available (Strack & Deutsch, 2004; Wegner, 1994, 2009).

These described mechanisms would indeed be able to explain most, if not all findings of the present experiments. That is, negations as used in the violation instructions could cause general performance decrements as compared to reversed rule instructions (Wason, 1959). They could bias movement trajectories accordingly (Dale & Duran, 2011), and might also affect the ERPs as observed in the present experiments (Lüdtke, Friedrich, de Filippis, & Kaup, 2008).

Still, at least two observations do not seem to fit the negation hypothesis. First, ironic effects of negated intentions manifest mostly when an agent's resources are actively depleted by dual-task load (Gilbert, 1991; Wegner, 1994, 2009) whereas, in the present experiments, participants were able to focus entirely on the task at hand. Secondly, Experiment 6 did not employ any specific instructions relating to rule violations but cognitive conflict still emerged in this setting. The negation hypothesis might thus not fully account for the present findings even though it can accommodate a considerable share of the data.

Assuming that the observed effects were indeed driven by negations and their increased processing demands (at least to a considerable degree) opens up another question, however. This question is: Why did the participants stick to the negated instructions instead of reversing them

beforehand? Violation-unrelated negations were indeed found to have only limited impact if the negated concept had a relevant meaning on its own (Hasson, Simmons, & Todorov, 2005; Fillenbaum, 1966; Mayo, Schul, & Burnstein, 2004) – a condition that is obviously the case for the present violation instructions.

Perhaps, such continued effects of negations might be driven by the context of tasks and tasks rules. Research in this domain, e.g., using the Wisconsin Card Sorting Test (Berg, 1948), indicated that adapting to changing rules is a particularly effortful task (Monchi, Petrides, Petre, Worsley, & Dagher, 2001) and participants might thus have preferred not to mentally mess with the rules they were going to violate. Validating this speculation, as well as pinpointing the exact contributions of the negation hypothesis vis-à-vis the motivational conflict hypothesis will require additional, refined experimentation.

To complement the view on rule violations advanced so far, I will instead briefly comment on deception and lying as related phenomena (Chapter 14). Finally, I will highlight possible connections of the investigated processes to certain individual difference variables (Chapter 15).

14 Related phenomena: A dishonest detour

Modern society provides countless opportunities to violate rules, social norms and even laws: Cheating, deceiving, bribing, stealing, faking, free riding, blackmailing, and plagiarising are some but certainly not exhaustive examples. Arguably, the effects studied in the present experiments are relevant for understanding most, if not all of the above actions.

This becomes apparent when considering the phenomenon of deception and lies. Lying obviously violates a social norm of responding honestly. Additionally, lying always involves knowledge of some kind and a response that runs counter to this knowledge. Not surprisingly, several studies found lying to be cognitively more demanding than responding honestly (e.g., Debey, Verschuere, & Crombez, 2012; Langleben et al., 2002; Spence et al., 2001; Walczyk, Mahoney, Doverspike, & Griffith-Ross, 2009) and this increased cognitive demand seems to entail the inhibition of the honest response (Buller & Burgoon, 1996; McCornack, 1997; Nuńez, Casey, Egner, Hare, & Hirsch, 2005; Spence et al., 2004). Just as in the present Experiment 4 and 5, the inhibition of the honest answer does not seem to work perfectly, however, and a residual activation of the honest response biases the liar's behaviour (Duran, Dale, & McNamara, 2010).

Automatic activation of the honest response also is a cornerstone of the Activation-Decision-Construction model of deception (Walczyk et al., 2009; Walczyk, Roper, Seemann, & Humphrey, 2003) even though this model highlights two additional processes: Deciding whether to lie or not (cf. also Spence, Kaylor-Hughes, Farrow, & Wilkinson, 2008) and constructing a plausible lie afterwards (Undeutsch, 1967; Vrij, Edward, Roberts, & Bull, 2000). Especially the latter process reminds of the additional effort of rule violations documented in the present experiments

in that the explicit construction of a lie detaches the response from the prompting stimuli or situations. In line with this notion, the electrophysiological signature of active lying resembles that of rule violations as documented in Experiment 7 and 8 (cf. Johnson, Barnhardt, & Zhu, 2003, 2005; Pfister, Foerster, & Kunde, under revision; Rosenfeld et al., 1999).

The outlined commonalities of rule violations and lying seem to suggest that both phenomena might be two sides of the same coin. Interestingly though, the study of active lying has, to my knowledge, never adopted control conditions akin to the reversed rule groups of the present experiments. Such conditions could be constructed easily, e.g., by instructing participants to use some form of code to communicate with others ("Person X is going to understand the exact opposite of what you say"). Using these or similar conditions would certainly help to clarify commonalities and differences between rule violations and lying. In turn, this strategy might also help to arrive at a clearer understanding of the cognitive mechanisms underlying the production of deceptive responses.

15 What makes a good rule-breaker?

Previous research on rule violations considered the characteristics of individual agents – cognitive processes and inter-individual differences alike – only relevant as far as these variables allowed for predicting whether or not a rule violation will occur in a given setting (e.g., Reason, 1995; see Section 3.2). This focus is especially prominent for the prediction of driving violations (Forward, 2009; González-Iglesias, Gómez-Fraguela, & Luengo-Martín, 2012; Johnson, Newstead, Charlton, & Oxley, 2011; Yagil, 1998), but it is also present elsewhere, for example in medical contexts (Booth & Carruth, 1998; Delgado, 2002; Parker & Lawton, 2003; Phipps et al., 2008), or organisational settings (Berry, Ones, & Sackett, 2007).

Though these studies have considerable practical value, the aim of predicting whether or not a violation will occur is only partly relevant for the present experiments. I will thus refrain from summarizing the above literature; instead, I will argue for several promising connections between individual difference variables and the cognitive effects of rule violations studied here. In other words: What counts for the following argument is not the prediction of who is more or less likely to violate a rule; rather, I will focus on those variables that might determine the ease or difficulty of actually committing a rule violation.

The psychological traits determining ease or difficulty of rule violations are conceivably diverse and heterogeneous (cf. Elms & Milgram, 1966, for a related discussion in the context of obedience to authority); looking for the equivalent of a "moral character" (Narváez & Rest, 1995; Rest, 1986) may thus be a scientific blind alley. Instead, it seems worthwhile to look for concepts that imply processes similar to those assumed to underlie intended rule violations.

One class of personality traits pertinent to rule violations are traits that draw on an individual's ability to resist situational factors influencing perception and action. That is, the more a person is able to detach own actions from situational influences, the easier this person should be able to violate rules in general. Prima facie, the concept of field independence seems to be such a trait (Witkin & Asch, 1948; Witkin & Goodenough, 1977, 1981). This concept is based on the assumption that individuals differ in responsiveness and reliance on external events and would therefore represent a seemingly direct expression of how easily the individual breaks a rule. Unfortunately, none of the available measures, such as the rod-and-frame test or the embedded figures test (Witkin & Goodenough, 1981), seems to capture precisely this connotation. Instead, these tests focus on merely perceptual abilities with emphasis on flexibility and speed (Widiger, Knudson, & Rorer, 1980; for a recent critique, see Kozhevnikov, 2007).

More promising than the construct of field independence might be the related distinction of an internal versus external locus of control (Rotter, 1966, 1990; see also Krampen, 1988; Levenson, 1972). An internal locus of control implies the belief that certain goals can be reached by own actions quite independent of external events. An external locus of control, by contrast, implies the belief that own success depends on various external factors that cannot be affected by own actions. Clearly, efficient rule-breaking seems to be related to the former concept of an internal locus of control whereas agents with an external locus of control should have more pronounced difficulty with violating rules.

A further interesting prediction can be derived from the negation hypothesis described in Section 13.2. If rule violations do indeed entail additional and rather complex processing steps, agents with generally high processing speed should be particularly good rule-breakers. In other words: The negation hypothesis points to a direct link between the cognitive conflict during rule violations and measures of (fluid) intelligence (e.g., Bowie & Harvey, 2006; Oswald & Roth, 1987; Vernon, 1993). Similarly, the

motivational conflict hypothesis points to a relation of conflict during rule violations to anxiety and worrying because these traits are typically assumed to reflect the sustained anticipation of negative events (e.g., Beck, Epstein, Brown, & Steer, 1988; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983).

In sum, the traits sketched above seem to offer interesting connections to the effects investigated in the present experiments. The list of traits and constructs is of course not exhaustive. Other connections may exist not only with regard to additional common traits but also to domain-specific constructs such as integrity (Marcus, Kiebom, Ashton, 2007) and the work-related "need for rules and supervision" (Lewis & Anderson, 1998, as cited in Sanz, Gil, Carcía-Vera, & Barrasa, 2008). Complementarily, connections might further exist to high-level variables as suggested by research on cultural influences on conformity (Bond & Smith, 1996). Finally, it seems worthwhile to address specific populations relevant for the concept of rule violations. For one, these populations can be defined developmentally in terms of age groups that are typically prone to conflict with rules and norms (i.e., children and adolescents; Ruma & Mosher, 1967; Smetana, 2005; Smetana & Bitz, 1996; Turiel, 1983). On an entirely different note, such populations can also be defined in terms of relevant clinical conditions such as psychopathy and related constructs (Blair, 1995, 1997; Cleckley, 1976; Hare, 1996; Hare, Neumann, & Widiger, 2012; Harenski, Harenski, Shane, & Kiehl, 2010).

Because such investigations would certainly benefit from a clearer understanding of the processes underlying rule violations in the first place, I prefer to refrain from more detailed speculations at this point. In any case, the outlined possible connections of rule violations to certain individual difference variables seem to be a fruitful approach for future enquiry.

CONCLUDING REMARKS

"Look, that's why there's rules, understand?

So that you think before you break 'em."

Terry Pratchett, Thief of Time

Rules are an integral part of human life. They come in many different forms, and the different ways of violating a rule are just as diverse. The present experiments have gathered first evidence for cognitive conflict as a reliable consequence of rule violations even in the absence of any negative consequences such as punishment and retaliation. At a closer look, these findings open up many more intriguing questions. First and foremost, these questions relate to pitting the two discussed hypotheses of motivational conflict and negation against each other. In addition to understanding the mechanics of rule violations, another important question relates to the impact of the documented conflict on the decision whether or not to violate a rule, or, in other words: how to act. According to the "law of least mental effort" (Solomon, 1948; cf. Botvinick, 2007; Botvinick & Rosen, 2009), anticipating this conflict might be a driving force in favour of rule-based performance. Taking this thought one step further one could even pose a provocative chicken-and-egg question: Perhaps – just perhaps – is anticipated cognitive conflict alone *sufficient* to induce rule-based

performance, not only in the controlled settings investigated here but also in settings that breed actual conformity and obedience. There are still many steps between the present experiments and answers to questions like these, but embarking on this way promises a fascinating endeavour.

APPENDICES

Appendix A: Condition means for Experiment 1-3

Tab. 4. Mean RTs for all analysed trial sequences in Experiment 1-3. Correct responses and errors are coded separately for the frequent task (index F) and the infrequent tasks (index IF).

		Trial N-1				
Experiment	**Trial N**	C_F	C_{IF}	E_F	E_{IF}	V
1	C_F	452	522	471	533	530
	C_{IF}	545	470	536	504	553
	V	578	580	604	591	508
2	C_F	452	512	465	521	519
	C_{IF}	535	471	538	525	547
	V	568	582	599	585	499
3	C_F	453	502	461	533	512
	C_{IF}	530	463	539	509	538
	V	562	557	578	556	468

Legend: C = correct, rule-based response, E = error, V = violation (Experiment 1 and 2) or reversed rule response (Experiment 3).

Appendix B: Quantifying mouse trajectories

Experiment 4 and 5 employed a mouse-tracking setup in which participants started to move only after an imperative stimulus had appeared. This setup allows studying effects of certain experimental conditions on movement trajectories that are unlikely to be driven by differences in the time participants need to decide for one or the other target location. Thus, while potentially allowing for a methodologically sound assessment of movement trajectories, the present setup comes with several pitfalls relating to the decision period (i.e., the RT interval).

For common analyses of time-normalized trajectory data (cf. Section 8.1.3), the data points of the decision period clearly need to be stripped off before interpolating the data to a number of time steps. This is imperative because (a) the trajectory data of the decision period does not contain particularly relevant information for most analyses and (b) each data point in the decision interval decreases the temporal resolution available for the actual movement.

Stripping off the data of the decision period, however, creates a new pitfall concerning measures that compare the actual trajectory of a given movement with a straight line from the movement's startpoint to its endpoint (such as AUC and MAD). Whereas it is easy to determine the endpoint of the movement, at least two solutions are possible for the startpoint. A first intuitive possibility would be using the first data point after having stripped off the decision data. In Experiment 4-5, this point would correspond to the cursor leaving the home area (the point in time when RT is measured; other criteria such as reaching a certain percentage

of maximum velocity would yield similar results). Choosing this solution has the obvious advantage that the resulting trajectory is not contaminated by dwell time data assembled during the RT interval. Still, this solution has a serious drawback: As I will outline below, it leads to distorted measures for AUC and MAD, effectively underestimating the true effect size. The better solution might thus be to use the starting coordinates *of the RT interval* even though this point is not part of the trajectory proper. In the following, I briefly illustrate the above points by using a simplified scenario as sketched in **Figure 31**. Based on the nomenclature described in the figure caption, I will refer to the first solution – using the first data point after the decision period as startpoint – as AUC_s, whereas I will refer to the latter solution – using the very first data point of the decision period as startpoint – as AUC_0.

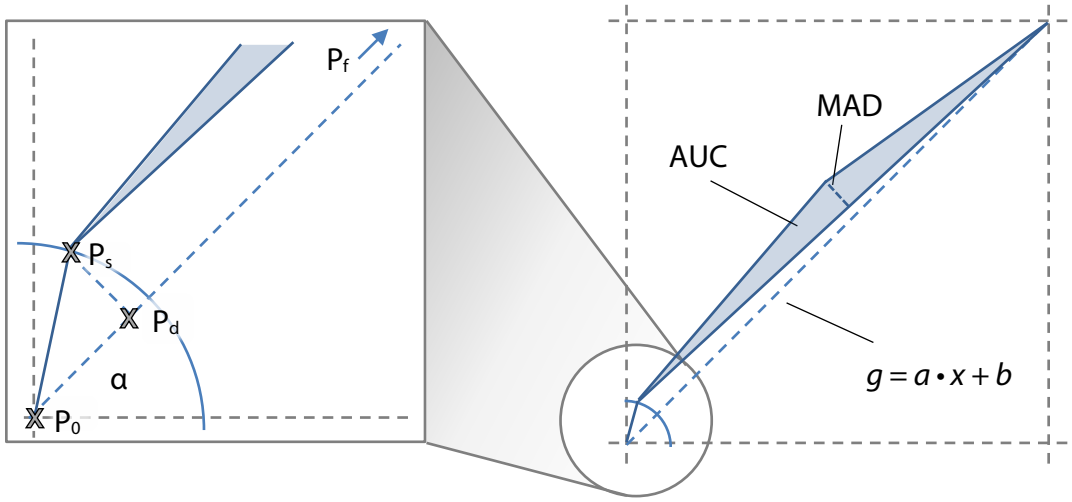

Fig. 31. Simplified scenario for analysing trajectory data. The **right panel** shows an actual trajectory from the bottom-left (corresponding to a home area in Experiment 4 and 5) to the top-right (corresponding to a potential target area). The arch surrounding the startpoint shows the threshold that is used to determine the onset of the movement (RT). The **left panel** shows the four crucial points in this setup: P_0 corresponds to the cursor coordinates at the onset of a target stimulus whereas P_s corresponds to the coordinates when RT is measured (with the movement leaving the home area at angle **α**). P_f corresponds to the coordinates of the endpoint whereas P_d shows the perpendicular dropped from P_s on the line [P_0; P_f].

Several observations follow from **Figure 31**. Most notably, AUC_s (the coloured triangle) seems to be relatively insensitive to the starting angle of the movement. Yet, greater starting angles obviously imply a stronger deviation of the trajectory towards the target in the opposite (left) corner. This deviation is not represented adequately in the resulting value for AUC_s. Rather, greater starting angles increase the area enclosed by the lines $[P_s; P_f]$ and $[P_0; P_f]$. This area is clearly part of the movement's deviation and needs to be included in the AUC, which is the case for AUC_0 but not for AUC_s.

To assess the described impact of different starting angles more thoroughly, the following analysis introduces several simplifying assumptions for the setup sketched in **Figure 31**:

- The startpoint and endpoint of the movement are assumed to have the fixed coordinates $P_0(0, 0)$ and $P_f(1, 1)$.
- Each movement leaves the home area at a given angle α with $45° \leq \alpha \leq 90°$ and $\alpha = 45°$ representing a direct approach towards the target area. The radius of the home area – i.e., $d(P_0, P_s)$ – is 10% of $d(P_0, P_f)$. Accordingly, $d(P_0, P_s) = 0.1 \cdot \sqrt{2}$.
- Each movement deviates from the straight line from P_s to P_f towards the upper-left corner. The trajectory can be approximated by a triangular shape beginning at P_s and ending at P_f. The third point is perpendicular to the point halfway between P_s and P_f. The effective MAD is defined as a fixed percentage of the distance $d(P_s, P_f)$ with $0 < \text{MAD} \leq 0.2 \cdot d(P_s, P_f)$.

These assumptions allow for specifying closed formulas for all coordinates, distances, and areas of the above example. Accordingly, the coordinates of the point P_s are a direct function of the target angle – i.e., $P_s\left(0.1 \cdot \sqrt{2} \cdot \cos(\alpha), 0.1 \cdot \sqrt{2} \cdot \sin(\alpha)\right)$ – and so are the coordinates of P_d – i.e., $P_d\left(0.1 \cdot \cos(\alpha - 45°), 0.1 \cdot \cos(\alpha - 45°)\right)$.

Based on these coordinates, AUC_s is computed as $0.5 \cdot d(P_s, P_f) \cdot \text{MAD}$ whereas AUC_0 is computed as $0.5 \cdot d(P_s, P_d) \cdot d(P_d, P_f)$. Most informative for the present argument is the ratio of AUC_s / AUC_0 which is plotted in **Figure 32**. This ratio roughly indicates how much of the true effect (AUC_0) is captured with the measure AUC_s. As is obvious from the figure, smaller values for MAD – relative to $d(P_s, P_f)$ and especially greater starting angles can yield tremendously reduced effects. I argue to use AUC_0 for any analysis to avoid such distortions. Accordingly, this has been done for all present analyses.

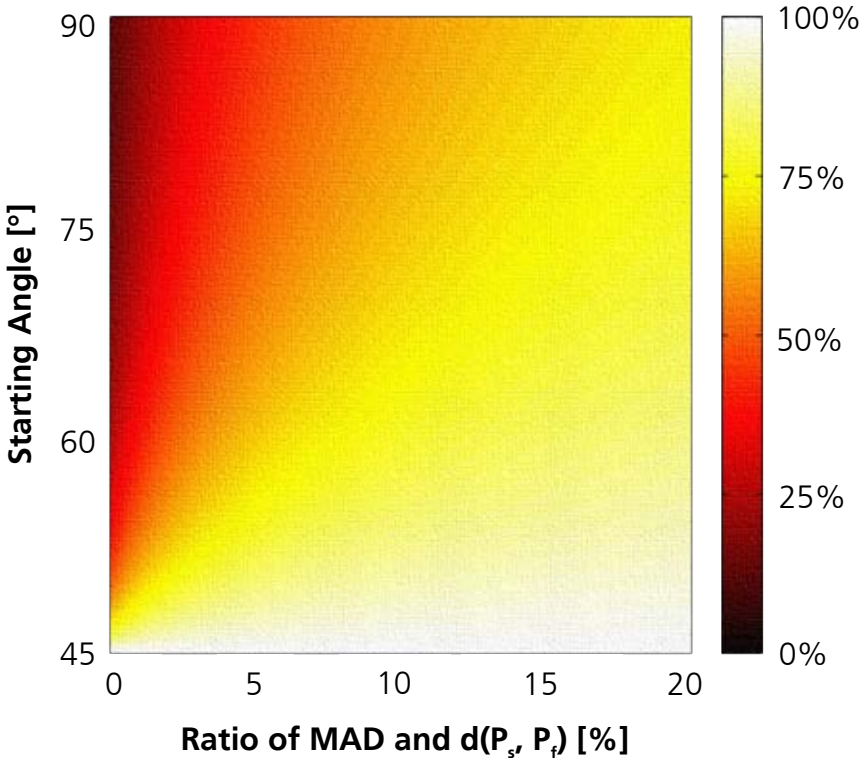

Fig. 32. Percentage of effects on trajectory deviation captured in AUC_s relative to AUC_0. These percentages are plotted in different colours and as a function of MAD and starting angle. Darker colours indicate stronger underestimation of an existing effect. This pitfall can obviously be avoided by using AUC_0 in the first place.

Appendix C: Assessing bimodality

The appropriate way of quantifying and testing empirical distributions for bimodality has been discussed intensively in the statistical literature over the last decades (see Knapp, 2007, for an overview). A seemingly elegant way to quantify the degree of bimodality is the *bimodality coefficient b* that can be easily computed from only two statistics: the skewness and the excess kurtosis of the distribution. It is thus readily available and can serve as a first evaluation for any empirical distribution.

Interestingly, however, wrong or misleading formulas for b can be found in various places ("Bimodal distribution", 2012; Freeman & Dale, 2013). The following paragraphs therefore give some details about how I computed this coefficient for the data of Experiment 6. The appropriate, original formula to compute a value for b was introduced in the SAS User's Guide as

$$b = \frac{\mathrm{m}_3^2 + 1}{\mathrm{m}_4 + 3 \cdot \frac{(n-1)^2}{(n-2)(n-3)}}$$

with m_3 referring to the skewness of the distribution and m_4 referring to the *excess* kurtosis of the distribution (SAS Institute Inc., 1990; cf. Knapp, 2007, for critical remarks about this notation). Using a sample of n values and a sample mean of $\bar{x}$, the sample excess kurtosis m_4 is thus equal to the following formula:

$$m_4 = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} - 3 \quad .$$

Because the above formula is a biased estimator of the population excess kurtosis, however, most programs (including SAS) rely on the adjusted estimator $\hat{m}_4$:

$$\hat{m}_4 = \frac{(n+1)n(n-1)}{(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 \cdot \frac{(n-1)^2}{(n-2)(n-3)}$$

.

Empirical values for b are typically compared to the value of $b_{crit} = 5/9$ which is expected from a uniform distribution given a large enough sample size. Thus, as a rule of thumb, $b > .555$ is taken to indicate a bimodal distribution. The exact probability density function of the bimodality coefficient cannot be derived, however (Knapp, 2007). This is a major drawback because a thorough null-hypothesis significance test is not possible without a density function.

An suitable alternative test for bimodality is the *dip test* (Hartigan & Hartigan, 1985) that probes for deviations from unimodality. The corresponding dip statistic approaches a positive constant for any multimodal distribution whereas it approaches zero for unimodal and/or uniform distributions.

An algorithm that implements the dip test was proposed early after the statistical background of the test was published (Hartigan, 1985). This algorithm has been adopted for modern statistical packes such as R (Maechler, 2012) and MATLAB (Mechler, 2002). Both versions produced converging results for data of Experiment 6 (at the reported precision).

Appendix D: LRP results for Experiment 7-8

The lateralized readiness potentials (LRPs; **Fig. 33**) for Experiment 7-8 were computed as the averaged difference between contra- and ipsilateral electrodes (C3, C4) for left- and right-hand responses, i.e., $LRP = 0.5 \cdot \left[(C4 - C3)_{\text{Left Response}} + (C3 - C4)_{\text{Right Response}} \right]$.

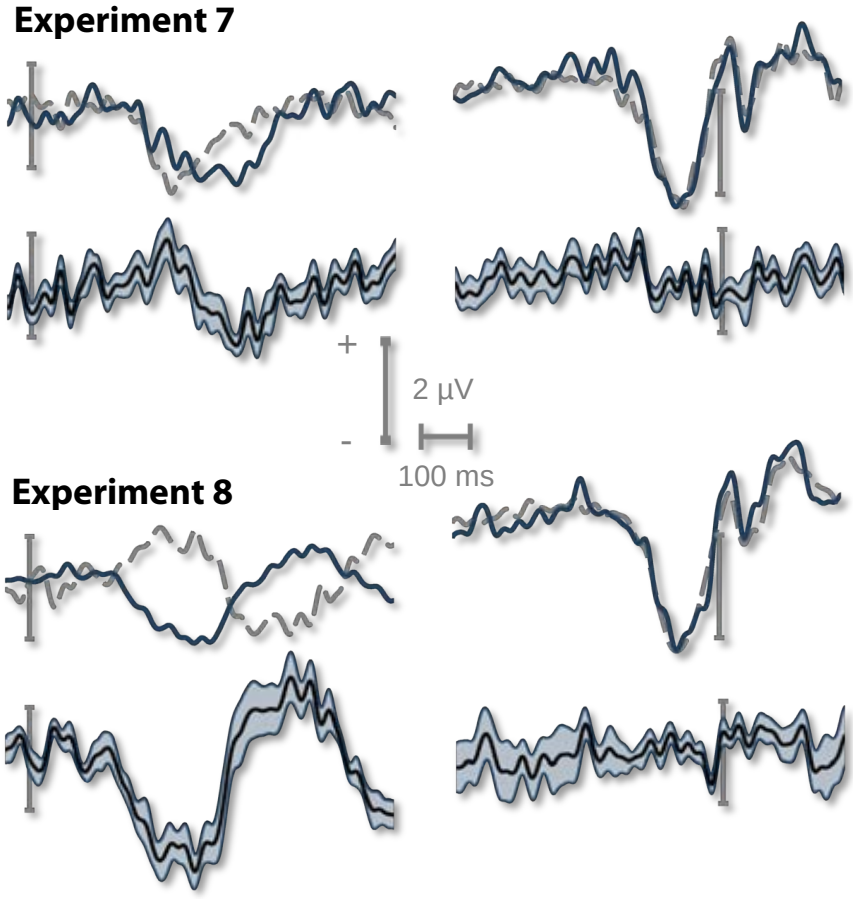

Fig. 33. Stimulus-locked LRPs (left panels) and response-locked LRPs at C3/C4 for Experiment 7-8. The upper plot of each pair shows the raw LRP for violation responses (blue line) and normal, rule-based responses (dashed grey line). The lower plot of each pair shows the difference wave ± 1 standard error of paired differences, computed separately for each data point (coloured area).

IMAGE SOURCES

All artwork included in this thesis was designed by myself and the figures are mostly built from scratch, with two exceptions. The picture of Zoppo Trump on p. 1 – arguably a master of bending and breaking any rules – is based on video footage of the Augsburger Puppenkiste, © Tilde Michels / Augsburger Puppenkiste / Hessischer Rundfunk, 1970; reproduced with permission. The chicken Agathe who served as target stimulus in Experiment 7, thus appearing in Figure 23 and Figure 30, was taken from the online graphics archive NiBiS (Niedersächsicher Bildungsserver), Niedersächsisches Landesinstitut für schulische Qualitätsentwicklung, 2011; reproduced with permission. All other graphical elements are based on non-copyrighted material.

REFERENCES

Aarts, K., De Houwer, J., & Pourtois, G. (2012). Evidence for the automatic evaluation of self-generated actions. *Cognition*, *124*(2), 117-127. doi: 10.1016/j.cognition.2012.05.009.

Adriaanse, M. A., Gollwitzer, P. M., de Ridder, D. T. D., de Wit, J. B. F., & Kroese, F. M. (2011). Breaking habits with implementation intentions: A test of underlying processes. *Personality and Social Psychology Bulletin*, *37*(4), 502-513. doi: 10.1177/0146167211399102.

Adriaanse, M. A., Van Oosten, J. M. F., De Ridder, D. T. D., De Wit, J. B. F., & 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. doi: 10.1177/0146167210390523.

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltŕ, & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 421-452). Cambridge, MA: MIT Press.

Arrington, C. M., & Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological Science*, *15*(9), 610-615. doi: 10.1111/j.0956-7976.2004.00728.x.

Arrington, C. M., & Logan, G. D. (2005). Voluntary task switching: Chasing the elusive homunculus. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 683-702. doi: 10.1037/0278-7393.31.4.683.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177-190). Pittsburgh, PA: Carnegie Press.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1-70.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory and Cognition*, *32*(8), 1369-1378. doi: 10.3758/BF03206327.

Band, G. P., van Steenbergen, H., Ridderinkhof, K. R., Falkenstein, M., & Hommel, B. (2009). Action-effect negativity: Irrelevant action effects are monitored like relevant feedback. *Biological Psychology*, *82*(3), 211-218. doi: 10.1016/j.biopsycho.2009.06.011.

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, *11*(2), 167-203. doi: 10.1177/1088868307301033.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*(6), 893-897. doi: 10.1037/0022-006X.56.6.893.

Bennington, J. Y., & Polich, J. (1999). Comparison of P300 from passive and active tasks for auditory and visual stimuli. *International Journal of*

Psychophysiology, *34*(2), 171-177. doi: 10.1016/S0167-8760(99)00070-7.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. The *Journal of General Psychology*, *39*(1), 15-22. doi: 10.1080/00221309.1948.9918159.

Berns, G. S., Chappelow, J., Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., & Richards, J. (2005). Neurobiological correlates of social conformity and independence during mental rotation. *Biological Psychiatry*, *58*(3), 245-253. doi: 10.1016/j.biopsych.2005.04.012.

Bertelson, P. (1963). S-R relationships and reaction time to new versus repeated signals in a serial task. *Journal of Experimental Psychology*, *65*(5), 478-484. doi: 10.1037/h0047742.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*(2), 410-424. doi: 10.1037/0021-9010.92.2.410.

Bimodal distribution. (n.d.). In *Wikipedia*. Retrieved January 4, 2013, from http://en.wikipedia.org/wiki/Bimodal_distribution.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, *57*(1), 1-29. doi: 10.1016/0010-027(95)00676-P.

Blair, R. J. R. (1997). Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Differences*, *22*(6), 731-739. doi: 10.1016/S0191-8869(96)00249-8.

Blanton, H., & Christie, C. (2003). Deviance regulation: A theory of action and identity. *Review of General Psychology*, *7*(2), 115-149. doi: 10.1037/1089-2680.7.2.115.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, *88*(1), 1-45. doi: 10.1037/0033-2909.88.1.1.

Blass, T. (1991). Understanding behaviour in the Milgram Obedience Experiment: The role of personality, situations, and their interactions. *Journal of Personality and Social Psychology*, *60*(3), 398-413.

Blass, T. (1999). The Milgram paradigm after 35 years: Some things we now know about obedience to authority. *Journal of Applied Social Psychology*, *29*(5), 955-978. doi: 10.1111/j.1559-1816.1999.tb00134.x.

Bond, R. (2005). Group size and conformity. *Group Processes & Intergroup Relations*, *8*(4), 331-354. doi: 10.1177/1368430205056464.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) Line Judgment Task. *Psychological Bulletin*, *119*(1), 111-137. doi: 10.1037/0033-2909.119.1.111.

Booth, D., & Carruth, A. K. (1998). Violations of the nurse practice act: Implications for nurse managers. *Nursing Management*, *29*(10), 35-39.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive Affective and Behavioral Neuroscience*, *7*(4), 356-366. doi: 10.3758/CABN.7.4.356.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624-652. doi: 10.1037/0033-295X.108.3.624.

Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research*, *73*(6), 835-842. doi: 10.1007/s00426-008-0197-8.

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols*, *1*(5), 2277-2281. doi: 10.1038/nprot.2006.390.

Brainerd, C. J. (1973). Judgments and explanations as criteria for the presence of cognitive structures. *Psychological Bulletin*, *79*(3), 172-179. doi: 10.1037/h0033876.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*(3), 203-242. doi: 10.1111/j.1468-2885.1996.tb00127.x.

Carver, C. S. (2006). Approach, avoidance, and the self-regulation of affect and action. *Motivation and Emotion*, *30*(2), 105-110. doi: 10.1007/s11031-006-9044-7.

Chandler, M. J., Greenspan, S., & Barenboim, C. (1973). Judgments of intentionality in response to videotaped and verbally presented moral dilemmas: The medium is the message. *Child Development*, *44*(2), 315-320. doi: 10.2307/1128053.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591-621. doi: 10.1146/annurev.psych.55.090902.142015.

Cialdini, R. B., & Trost, M. (1998). Social influence: Social norms, conformity, and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 2, 4th ed., pp. 151-192). New York: McGraw-Hill.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472-517. doi: 10.1016/0010-0285(72)90019-9.

Clark, H. H., & Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory and Cognition*, *2*(1), 101-111. doi: 10.3758/BF03197499.

Cleckley, H. (1976). *The mask of sanity* (5th ed.). St. Louis, MO: Mosby.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235-253. doi: 10.1037/0096-3445.120.3.235.

Coan, J. A., & Allen, J. J. B. (2002). The state and trait nature of frontal EEG asymmetry in emotion. In K. Hugdahl & R. J. Davidson (Eds.), *The asymmetrical brain*. Cambridge, MA: MIT Press.

Cohen-Kdoshay, O., & Meiran, N. (2007). The representation of instructions in working memory leads to autonomous response activation: Evidence from the first trials in the flanker paradigm. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1140-1154. doi: 10.1080/17470210600896674.

Cohen-Kdoshay, O., & Meiran, N. (2009). The representation of instructions operates like a prepared reflex: Flanker compatibility effects found in the first trial following S–R instructions. *Experimental Psychology*, *56*(2), 128-133. doi: 10.1027/1618-3169.56.2.128.

Colby, A., Gibbs, J., Lieberman, M., & Kohlberg, L. (1983). *A longitudinal study of moral judgment: A monograph for the society of research in child development.* Chicago, IL: The University of Chicago Press.

Colgan, D. M. (1970). Effects of instructions on the skin resistance response. *Journal of Experimental Psychology*, *86*(1), 108-112. doi: 10.1037/h0030011.

Cook, S. W., & Harris, R. E. (1937). The verbal conditioning of the galvanic skin reflex. *Journal of Experimental Psychology*, *21*(2), 202-210. doi: 10.1037/h0063197.

Czigler, I. (2007). Visual mismatch negativity: Violation of nonattended environmental regularities. *Journal of Psychophysiology*, *21*(3-4), 224-230. doi: 10.1027/0269-8803.21.34.224.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, *35*(5), 983-986. doi: 10.1111/j.1551-6709.2010.01164.x.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15-28. doi: 10.3758/BF03195938.

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*, *2*(233). doi: 10.3389/fpsyg.2011.00233.

Darley, J. M. (2009). Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science*, *5*, 1-23. doi: 10.1146/annurev.lawsocsci.4.110707.172335.

Davidson, R. J. (1993). Cerebral asymmetry and emotion: Conceptual and methodological conundrums. *Cognition & Emotion*, *7*(1), 115–138. doi: 10.1080/02699939308409180.

Davidson, R. J. (1998). Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion*, *12*(3), 307-330. doi: 10.1080/026999398379628.

Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I. *Journal of Personality and Social Psychology*, *58*(2), 330-341. doi: 10.1037/0022-3514.58.2.330.

De Jong, R. (1993). Multiple bottlenecks in overlapping task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(5), 965-980. doi: 10.1037/0096-1523.19.5.965.

Debey, E., Verschuere, B., & Crombez, G. (2012). Lying and executive control: An experi-mental investigation using ego depletion and goal neglect. *Acta Psychologica*, *140*(2), 133-141. doi: 10.1016/j.actpsy.2012.03.004.

Dehaene, S. (1996). The organization of brain activations in number comparison: Event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience*, *8*(1), 47-68. doi: 10.1162/jocn.1996.8.1.47.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error-detection and compensation. *Psychological Science*, *5*(5), 303-305. doi: 10.1111/j.1467-9280.1994.tb00630.x.

Delgado, C. (2002). Competent and safe practice: A profile of disciplined registered nurses. *Nurse Education*, *27*(4), 159-161.

Demanet, J., Verbruggen, F., Liefooghe, B., & Vandierendonck, A. (2010). Voluntary task switching under load: Contribution of top-down and bottom-up factors in goal-directed behavior. *Psychonomic Bulletin and Review*, *17*(3), 387-393. doi: 10.3758/PBR.17.3.387.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629-636. doi: 10.1037/h0046408.

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. doi: 10.1037/0022-3514.91.3.385.

Dien, J. (2012). Applying Principal Components Analysis to event-related potentials: A tutorial. *Developmental Neuroscience*, *37*(6), 497-517. doi: 10.1080/87565641.2012.697503.

Dien, J., & Frishkoff, G. A. (2005). Principal Components Analysis of event-related potential datasets. In T. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 189–208). Cambridge, MA: MIT Press.

Dignath, D., Pfister, R., Eder, A. B., Kiesel, A., & Kunde, W. (in press). Something in the way she moves – Movement trajectories reveal dynamics of self-control. *Psychonomic Bulletin & Review*.

Donchin, E. (1981). Surprise!…Surprise? *Psychophysiology*, *18*(5), 493-513. doi: 10.1111/j.1469-8986.1981.tb01815.x.

Donchin, E., & Coles, M.G.H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(3), 357-374. doi: 10.1017/S0140525X00058027.

Dreisbach, G., & Fischer, R. (2012). Conflicts as aversive signals. *Brain and Cognition*, *78*(2), 94-98. doi: 10.1016/j.bandc.2011.12.003.

Duncan-Johnson, C. C., & Kopell, B. S. (1981). The Stroop effect: Brain potentials localize the source of interference. *Science*, *214*(4523), 938-940. doi: 10.1126/science.7302571.

Duran, N. D., Dale, R., & McNamara, D. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review*, *17*(4), 486-491. doi: 10.3758/PBR.17.4.486.

Dutilh, G., Vanderkerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2010). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics*, *74*(2), 454-465. doi: 10.3758/s13414-011-0243-2.

Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General*, *137*(2), 262-281. doi: 10.1037/0096-3445.137.2.262.

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology*, *51*, 665-697. doi: 0.1146/annurev.psych.51.1.665.

Elms, A. C. (1995). Obedience in retrospect. *Journal of Social Issues*, *51*(3), 21-31. doi: 10.1111/j.1540-4560.1995.tb01332.x.

Elms, A. C., & Milgram, S. (1966). Personality characteristics associated with obedience and defiance toward authoritative command. *Journal of Experimental Research in Personality*, *1*, 282-289.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149. doi: 10.3758/BF03203267.

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, *4*(3), 330-341. doi: 10.3758/BF03210790.

Falkenstein, M., Hohnsbein, J., & Hoormann, J. (1990). Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C. H. M. Brunia, A. W. K. Gaillard, & A. Kok (Eds.), *Psychophysiological brain research* (Vol. 1, pp. 192-195). Tilburg, the Netherlands: Tilburg University Press.

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2-3), 87-107. doi: 10.1016/S0301-0511(99)00031-9.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87. doi:10.1016/S1090-5138(04)00005-4.

Fillenbaum, S. (1966). Memory for gist: Some relevant variables. *Language and Speech*, *9*(4), 217-227. doi: 10.1177/002383096600900403.

Flexer, A., Bauer, H., Pripfl, J., & Dorffner, G. (2005). Using ICA for removal of ocular artifacts in EEG recorded from blind subjects. *Neural Networks*, *18*, 998-1005. doi: 10.1016/j.neunet.2005.03.012.

Folstein, J. R, & van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology*, *45*(1), 152-179. doi: 10.1111/j.1469-8986.2007.00602.x.

Forward, S. E. (2009). The theory of planned behaviour: The role of descriptive norms and past behaviour in the prediction of drivers' intentions to violate. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 198-207. doi: 10.1016/j.trf.2008.12.002.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*(10), 1183-1188. doi: 10.1111/j.1467-9280.2009.02422.x.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247-279. doi: 10.1037/a0022327.

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology: General*, *137*(5), 673-690. doi: 10.1037/a0013875.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*(1), 83-97. doi: 10.3758/s13428-012-0225-x.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*(59). doi: 10.3389/fpsyg.2011.00059.

French, J. R. P., & Raven, B. (1959). The bases of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 150-167). Ann Arbor, MI: University of Michigan Press.

Frith, C. D., & Done, D. J. (1986). Routes to action in reaction time tasks. *Psychological Research*, *48*(3), 169-177. doi: 10.1007/BF00309165.

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. doi: 10.1016/j.jesp.2006.12.004.

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385-390. doi: 10.1111/j.1467-9280.1993.tb00586.x.

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In S. J. Luck, & E. Kappenman (eds.), *Oxford handbook of event-related potential components* (pp. 231-291). New York: Oxford University Press.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279-2282. doi: 10.1126/science.1066893.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107-119. doi: 10.1037/0003-066X.46.2.107.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601-613. doi: 10.1037/0022-3514.59.4.601.

Gollwitzer, P. M. (1999). Implementation intentions. Strong effects of simple plans. *American Psychologist*, *54*(7), 493-503. doi: 10.1037/0003-066X.54.7.493.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, *38*, 69-119. doi: 10.1016/S0065-2601(06)38002-1.

González-Iglesias, B., Gómez-Fraguela, J. A., & Luengo-Martín, M. A. (2012). *Transportation Research Part F: Traffic Psychology and Behaviour*, *15*(4), 404-412. doi: 10.1016/j.trf.2012.03.002.

Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition and Emotion*, *4*(3), 269-288. doi: 10.1080/02699939008410799.

Gray, J. A., & McNaughton, N. (2000). *The neuropsychology of anxiety*. Oxford: Oxford University Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400 (2004). doi: 10.1016/j.neuron.2004.09.027.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108. doi: 10.1126/science.1062872.

Grings, W. W. (1973). Cognitive factors in electrodermal conditioning. *Psychological Bulletin*, *79*(3), 200-210. doi: 10.1037/h0033883.

Habermas, J. (1983). *Moralbewußtsein und kommunikatives Handeln* [Moral consciousness and communicative action]. Frankfurt a. M.: Suhrkamp.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology*, *30*(3), 304-337. doi: 10.1006/cogp.1996.0009.

Haider, H., & Frensch, P. A. (1999a). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 172-190. doi: 10.1037/0278-7393.25.1.172.

Haider, H., & Frensch, P. A. (1999b). Information reduction during skill acquisition: The influence of task instruction. *Journal of Experimental Psychology: Applied*, *5*(2), 129-151, doi: 10.1037/1076-898X.5.2.129.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998-1002. doi: 10.1126/science.1137651.

Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology*, *42*(2), 151-160. doi: 10.1111/j.1469-8986.2005.00270.x.

Hake, H., & Hyman, R. (1953). Perception of the statistical structure of a random series of binary symbols. *Journal of Experimental Psychology*, *45*(1), 64-74. doi: 10.1037/h0060873.

Hare, R. D. (1996). Psychopathy: A clinical construct whose time as come. *Criminal Justice and Behavior*, *23*(1), 25-54. doi: 10.1177/0093854896023001004.

Hare, R. D., Neumann, C. S., & Widiger, T. A. (2012). Psychopathy. In T. A. Widiger (Ed.). *The Oxford handbook of personality disorders* (pp. 478-504). New York, NY: Oxford University Press.

Harenski, C. L., Harenski, K. A., Shane, M. S., & Kiehl, K. A. (2010). Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology*, *119*(4), 863-874. doi: 10.1037/a0020979.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70-84. doi: 10.1214/aos/1176346577.

Hartigan, P. M. (1985). Computation of the dip statistic to test for unimodality. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *34*(3), 320-325.

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? An examination of negated metaphors. *Journal of Pragmatics*, *38*(7), 1015-1032. doi: 10.1016/j.pragma.2005.12.005.

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe it or not. On the possibility of suspending belief. *Psychological Science*, *16*(7), 566-571. doi: 10.1111/j.0956-7976.2005.01576.x.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280-1300. doi: 10.1037//0003-066X.52.12.1280.

Higgins, E. T. (2012). Beyond pleasure and pain: How motivation works. New York, NY: Oxford University Press.

Hodges, B. H., & Geyer, A. L. (2006). A nonconformist account of the Asch experiments: Values, pragmatics, and moral dilemmas. *Personality and Social Psychology Bulletin*, *10*(1), 2-19. doi: 10.1207/s15327957pspr1001_1.

Höffe, O. (2008). *Praktische Philosophie: Das Modell des Aristoteles* (3rd ed.). Berlin: Akademie-Verlag.

Hoffman, J. E., Simons, R. F., & Houck, M. R. (1983). Event-related potentials during controlled and automatic target detection. *Psychophysiology*, *20*(6), 625-632. doi: 10.1111/j.1469-8986.1983.tb00929.x.

Hoffman, M. L. (2001). *Empathy and moral development. Implications for caring and justice*. Cambridge, UK: Cambridge University Press.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679-709. doi: 10.1037/0033-295X.109.4.679.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport*, *14*(18), 2481-2484. doi: 10.1097/01.wnr.0000099601.41403.a5.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497-498. doi: 10.1038/nn1238.

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General*, *134*(2), 163-191. doi: 10.1037/0096-3445.134.2.163.

Hommel, B. (2000). The prepared reflex: Automaticity and control in stimulus-response translation. In S. Monsell, & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 247-273). Cambridge, MA: MIT Press.

Ilan, A. B., & Polich, J. (1999). P300 and response time from a manual Stroop task. *Clinical Neurophysiology*, *110*(2), 367-373. doi: 10.1016/S0168-5597(98)00053-7.

Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Quarterly Journal of Experimental Psychology*, *62*(2), 209-218. doi: 10.1080/17470210802240655.

Johnson, M., Charlton, J., Oxley, J., & Newstead, S. (2013). Why do cyclists infringe at red lights? An investigation of Australian cyclists' reasons for red light infringement. *Accident Analysis and Prevention*, *50*, 840-847. doi: 10.1016/j.aap.2012.07.008.

Johnson, M., Newstead, S., Charlton, J., & Oxley, J. (2011). Riding through red lights: The rate, characteristics and risk factors of non-compliant urban commuter cyclists. *Accident Analysis and Prevention*, *43*, 323-328. doi: 10.1016/j.aap.2010.08.030.

Johnson, R., Barnhardt, J., & Zhu, J. (2003). The deceptive response: Effects of response conflict and strategic monitoring on the late positive component and episodic memory-related brain activity. *Biological Psychology*, *64*(3), 217-253. doi: 10.1016/j.biopsycho.2003.07.006.

Johnson, R., Barnhardt, J., & Zhu, J. (2005). Differential effects of practice on the executive processes used for truthful and deceptive responses: An event-related brain potential study. *Cognitive Brain Research*, *24*(3), 386-404. doi: 10.1016/j.cogbrainres.2005.02.011.

Just, M. A., & Carpenter, P. A. (1976). Relation between comprehending and remembering some complex sentences. *Memory & Cognition*, *4*(3), 318-322. doi: 10.3758/BF03213183.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. doi: 10.1016/0010-0285(72)90016-3.

Kant, I. (1788/2003). *Kritik der praktischen Vernunft* [Critique of practical reason]. Hamburg: Meiner.

Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory & Cognition*, *29*(7), 960-967. doi: 10.3758/BF03195758.

Kekoni, J., Hämäläinen, H., Saarinen, M., Gröhn, J., Reinikainen, K., Lehtokoski, A., & Risto Näätänen (1997). Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. *Biological Psychology*, *46*(2), 125-142. doi: 10.1016/S0301-0511(97)05249-6.

Keller, P. E., Wascher, E., Prinz, W., Waszak, F., Koch, I., & Rosenbaum, D. A. (2006). Differences between intention-based and stimulus-based actions. *Journal of Psychophysiology*, *20*(1), 9-20. doi: 10.1027/0269-8803.20.1.9.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1-24.

Kesselring, T. (1999). *Jean Piaget* (2nd ed.). München: Beck.

Kessler, Y., Shencar, Y., & Meiran, N. (2009). Choosing to switch: Spontaneous task switching despite associated behavioral costs. *Acta Psychologica*, *131*(2), 120-128. doi: 10.1016/j.actpsy.2009.03.005

Knapp, T. R. (2007). Bimodality revisited. *Journal of Modern Applied Statistical Methods*, *6*(1), 8-20.

Knolle, F., Schröger, E., & Kotz, S. A. (2013). Prediction errors in self- and externally-generated deviants. *Biological Psychology*, *92*(2), 410-416. doi: 10.1016/j.biopsycho.2012.11.017.

Kohlberg, L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 151-284). New York: Academic Press.

Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral stages: A current formulation and a response to critics*. Basel: Karger.

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557-577. doi: 10.1017/s0048577201990559.

Koster, E. H. W., Rassin, E., Crombez, G., & Näring, G. W. B. (2003). The paradoxical effects of suppressing anxious thoughts during imminent threat. *Behaviour Research and Therapy*, *41*(9), 1113-1120. doi: 10.1016/S0005-7967(03)00144-X.

Kotchoubey, B. I., Jordan, J. S., Grözinger, B., & Westphal, K. P. (1996). Event-related brain potentials in a varied-set memory search task: A reconsideration. *Psychophysiology*, *33*(5), 530-540. doi: 10.1111/j.1469-8986.1996.tb02429.x.

Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. *Psychological Bulletin*, *133*(3), 464-481. doi: 10.1037/0033-2909.133.3.464.

Krampen, G. (1988). Toward an action-theoretical model of personality. *European Journal of Personality*, *2*(1), 39-55. doi: 10.1002/per.2410020104.

Kunde, W., Kiesel, A., & Hoffmann, J. (2003). Conscious control over the content of unconscious cognition. *Cognition*, *88*(2), 223-242. doi: 10.1016/S0010-0277(03)00023-4.

Kurtines, W., & Greif, E. B. (1974). The development of moral thought: Review and evaluation of Kohlberg's approach. *Psychological Bulletin*, *81*(8), 453-470. doi: 10.1037/h0036879.

Laming, D. R. J. (1968). *Information theory of choice reaction times*. London: Academic Press.

Laming, D. R. J. (1979). Choice reaction performance following an error. *Acta Psychologica*, *43*(3), 199-224. doi:10.1016/0001-6918(79)90026-X.

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., O'Brien, C. P., & Childress, A. R. (2002). Brain activity during simulated deception: An event-related functional magnetic resonance study. *NeuroImage*, *15*(3), 727-732. doi:10.1006/nimg.2001.1003.

Levenson, H. (1972). Distinctions within the concept of internal-external control: Development of a new scale. *Proceedings of the Annual Convention of the American Psychological Association*, *7*(1), 261-262.

Lewicki, P. (1986). Processing information about covariations that cannot be articulated. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 135-146. doi: 10.1037/0278-7393.12.1.135.

Liefooghe, B., Demanet, J., & Vandierendonck, A. (2010). Persisting activation in voluntary task switching: It all depends on the instructions. *Psychonomic Bulletin & Review*, *17*(3), 381-386. doi: 10.3758/PBR.17.3.381.

Liefooghe, B. Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1325-1335. doi: 10.1037/a0028148.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492-527. doi: 10.1037/0033-295X.95.4.492.

Lüdtke, J., Friedrich, C. K., de Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence-picture verification paradigm. *Journal of Cognitive Neuroscience*, *20*(8), 1355-1370. doi: 10.1162/jocn.2008.20093.

Maechler, M. (2012). *diptest: Hartigan's dip test statistic for unimodality – corrected code.* R package version 0.75-4. http://CRAN.R-project.org/package=diptest (retrieved January 4, 2013).

Maier, M., Steinhauser, M., & Hübner, R. (2008). Is the Error-Related Negativity amplitude related to error detectability? Evidence from effects of different error types. *Journal of Cognitive Neuroscience*, *20*(12), 2263-2273. doi: 10.1162/jocn.2008.20159.

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent Compontent Analysis of electroencephalographic data. In D. M. Touretzky, & Hasselmo, M. (Ed.), *Advances in Neural Information Processing Sytems* (pp. 145-151). Cambridge MA.

Marco-Pallarés, J., Camara, E., Münte, T. F., & Rodrígues-Fornells, A. (2008). Neural mechanisms underlying adaptive actions after slips. *Journal of Cognitive Neuroscience*, *20*(9), 1595-1610. doi: 10.1162/jocn.2008.20117.

Marcus, B., Kiebom, L., & Ashton, M. C. (2007). Personality dimensions explaining relationships between integrity tests and counterproductive behavior: Big five, or one in addition? *Personnel Psychology*, *60*(1), 1-34. doi: 10.1111/j.1744-6570.2007.00063.x.

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs. "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433-449. doi: 10.1016/j.jesp.2003.07.008.

Mayr, U., & Bell, T. (2006). On how to be unpredictable: Evidence from the voluntary task-switching paradigm. *Psychological Science*, *17*(9), 774-780. doi: 10.1111/j.1467-9280.2006.01781.x.

McCornack, S. A. (1997). The generation of deceptive messages. In J. O. Greene (Ed.), *Message production* (pp. 91-126). Mahwah, NJ: Erlbaum.

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, *19*(1), 22-24. doi: 10.1111/j.1467-9280.2008.02041.x.

Mechler, F. (2002). *Hartigan's dip statistic*. http://nicprice.net/diptest/ (retrieved January 4, 2013).

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378. doi:10.1037/h0040525.

Milgram, S. (1974). *Obedience to Authority*. New York: Harper & Row.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a ''generic'' neural system for error detection. *Journal of Cognitive Neuroscience*, *9*(6), 788-798. doi: 10.1162/jocn.1997.9.6.788.

Modgil, S., & Modgil, C. (2011). *Lawrence Kohlberg: Concensus and controversy* (2nd ed.). Abingdon, UK: Routledge.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, *6*(10), 799-809. doi: 10.1038/nrn1768.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *21*(19), 7733-7741.

Näätänen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences*, *13*(2), 201-233. doi: 10.1017/S0140525X00078407.

Näätänen, R., & Alho, K. (1995). Mismatch negativity – A unique measure of sensory processing in audition. *International Journal of Neuroscience*, *80*(1-4), 317-337. doi: 10.3109/00207459508986107.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313-329. doi: 10.1016/0001-6918(78)90006-9.

Narváez, D., & Rest, J. (1995). The four components of acting morally. In W. Kurtines, & J. Gewirtz (Eds.), *Moral Behavior and Moral Development: an introduction* (385-400). New York: McGraw-Hill.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530-542. doi: 10.1016/j.cognition.2005.07.005.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330-357. doi: 10.1037/0033-295X.109.2.330.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus norepinephrine system. *Psychological Bulletin*, *131*(4), 510-532. doi: 10.1037/0033-2909.131.4.510.

Notebaert, W., Houtman, F., Van Opstal, F., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, *111*(2), 275-279. doi:10.1016/j.cognition.2009.02.002.

Nuńez, J. M., Casey, B. J., Egner, T., Hare, T., & Hirsch, J. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. *NeuroImage*, *25*(1), 267-277. doi: 10.1016/j.neuroimage.2004.10.041.

O'Connell, R. G., Balsters, J. H., Kilcullen, S. M., Campbell, W., Bokde, A. W., Lai, R., Upton, N., & Robertson, I. A. (2012). A simultaneous EEG/fMRI investigation of the P300 aging effect. *Neurobiology of*

Aging, *33*(10), 2448-2461. doi: 10.1016/j.neurobiolaging.2011.12.021.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*. doi: 10.1155/2011/156869.

Oswald, W. D., & Roth, E. (1987). Der Zahlen-Verbindungs-Test (ZVT). Ein sprachfreier Intelligenz-Test zur Messung der „kognitiven Leistungsgeschwindigkeit". Handanweisung (2nd ed.). Göttingen: Hogrefe.

Parker, D., & Lawton, R. (2003). Psychological contribution to the understanding of adverse events in health care. *Quality & Safety in Health Care*, *12*(6), 453-457. doi:10.1136/qhc.12.6.453.

Parker, D., Reason, J. T., Manstead, A. S. R., & Stradling, S. G. (1995). Driving error, driving violations and accident involvment. *Ergonomics*, *38*(5), 1036-1048. doi: 10.1080/00140139508925170.

Pashler, H., & Baylis, G. (1991). Procedural learning: 2. Intertrial repetition effects in speeded-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 33-48. doi: 10.1037/0278-7393.17.1.33.

Pazo-Alvarez, P., Cadaveira, F., & Amenedo, E. (2003). MMN in the visual modality: a review. *Biological Psychology*, *63*(3), 199-236. doi:10.1016/S0301-0511(03)00049-8.

Perneger, T. V. (2005). The Swiss cheese model of safety incidents: Are there holes in the metaphor? *BMC Health Services Research*, *5*(71). doi: 10.1186/1472-6963-5-71.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123-205). San Diego, CA: Academic Press.

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. doi: 10.2478/v10053-008-0133-x.

Pfister, R., Foerster, A., & Kunde, W. (under revision). Pants on fire: The electrophysiological signature of telling a lie. *Manuscript under revision*.

Pfister, R., Schroeder, P. A., & Kunde, W. (2013). SNARC struggles: Instant control over spatial-numerical associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *in press*. doi: 10.1037/a0032991.

Phipps, D. L., Parker, D., Pals, E. J. M., Meakin, G. H., Nsoedo, C., & Beatty, P. C. W. (2008). Identifying violation-provoking conditions in a healthcare setting. *Ergonomics*, *51*(11), 1625-1642. doi: 10.1080/00140130802331617.

Piaget, J. (1932/1948). *The moral judgment of the child*. Glencoe, IL: The Free Press.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128-2148. doi: 10.1016/j.clinph.2007.04.019.

Pratt, N., Willoughby, A., & Swick, D. (2011). Effects of working memory load on visual selective attention: Behavioral and electrophysiological evidence. *Frontiers in Human Neuroscience*, *5*(57). doi: 10.3389/fnhum.2011.00057.

Rabbitt, P. M. A. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264-272. doi: 10.1037/h0022853.

Rabbitt, P. M. A., & Rodgers, B. (1977). What does man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology*, *29*(4), 232-240. doi: 10.1080/14640747708400645.

Rapoport, A., & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review*, *104*(3), 603-617. doi: 10.1037/0033-295X.104.3.603.

Reason J. (1990). *Human error*. New York: Cambridge University Press.

Reason, J. (1995). Understanding adverse events: human factors. *Quality in Health Care*, *4*(2), 80-89. doi:10.1136/qshc.4.2.80.

Reason, J. (2000). Human error: Models and management. *British Medical Journal*, *320*(7237), 768-770. doi: 10.1136/bmj.320.7237.768.

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction? *Ergonomics*, *33*(10-11), 1315-1332. doi: 10.1080/00140139008925335.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219-235. doi: 10.1037/0096-3445.118.3.219.

Reisenauer, R., & Dreisbach, G. (2013). The impact of rules on distracter processing: Automatic categorization of irrelevant stimuli. *Psychological Research*, *77*(2), 128-138. doi: 10.1007/s00426-012-0413-4.

Renault, B., Ragot, R., & Lesevre, N. (1980). Correct and incorrect responses in a choice reaction time task and the endogenous components of the

evoked potential. *Progress in Brain Research*, *54*, 547-554. doi: 10.1016/S0079-6123(08)61685-4.

Rest, J. R. (1986). Moral development: Advances in research and theory. New York, NY: Praeger.

Rest, J. R., & Narváez, D. (1994). *Moral development in the professions: Psychology and applied ethics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision making. *Nature*, *461*(7261), 263-266. doi: 10.1038/nature08275.

Reynolds, S. J., & Ceranic, T. L. (2007). The effects of moral judgment and moral identity on moral behavior: An empirical examination of the moral individual. *Journal of Applied Psychology*, *92*(6), 1610-1624. doi: 10.1037/0021-9010.92.6.1610.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443-447. doi: 10.1126/science.1100301.

Roche, R. A. P., & O'Mara, S. M. (2003). Behavioral and electrophysiological correlates of visuomotor learning during a visual search task. *Cognitive Brain Research*, *15*, 127-136. doi: 10.1016/S0926-6410(02)00146-5.

Rodríguez-Fornells, A., Kurzbuch, A. R., & Münte, T. F. (2002). Time course of error detection and correction in humans: Neurophysiological evidence. *Journal of Neuroscience*, *22*(22), 9990-9996.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207-231. doi: 10.1037/0096-3445.124.2.207.

Rosenfeld, J. P., Ellwanger, J. W., Nolan, K., Wu, S., Bermann, R. G., & Sweet, J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *International Journal of Psychophysiology*, *33*(1), 3-19. doi: 10.1016/S0167-8760(99)00021-5.

Rotter, J. B. (1966). Generalized expectations for internal versus external control of reinforcement. *Psychological Monographs*, *80*(1), 1–28. doi: 10.1037/h0092976.

Rotter, J. B. (1990). Internal versus external control of reinforcement. A case history of a variable. *American Psychologist*, *45*(4), 489-493. doi: 10.1037/0003-066X.45.4.489.

Ruma, E. H., & Mosher, D. L. (1967). Relationship between moral judgment and guilt in delinquent boys. *Journal of Abnormal Psychology*, *72*(2), 122-127. doi: 10.1037/h0024444.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110-114. doi: 10.1111/1467-8721.01243.

Sanz, J., Gil, F., Carcía-Vera, M. P., & Barrasa, A. (2008). Needs and cognition/behavior patterns at work and the Big Five: An assessment of the Personality and Preference Inventory-Normative (PAPI-N) from the perspective of the five factor model. *International Journal of Selection and Assessment*, *16*(1), 46-58. doi: 10.1111/j.1468-2389.2008.00408.x.

SAS Institute Inc. (1990). *SAS/STAT user's guide, Version 6* (4th ed.). Cary, NC: Author.

Schroder, H. S., Moran, T. P., Moser, J. S., & Altmann, E. M. (2012). When the rules are reversed: Action-monitoring consequences of reversing

stimulus-response mappings. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(4), 629-643. doi: 10.3758/s13415-012-0105-y.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429-434. doi: 10.1111/j.1467-9280.2007.01917.x.

Schüür, F., & Haggard, P. (2011). What are self-generated actions? *Consciousness & Cognition*, *20*(4), 1697-1704. doi:10.1016/j.concog.2011.09.006.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599-1603. doi: 10.1126/science.275.5306.1599.

Seymour, P. H. K. (1977). Conceptual encoding and locus of the Stroop effect. *Quarterly Journal of Experimental Psychology*, *29*(2), 245-265. doi: 10.1080/14640747708400601.

Smetana, J. G. (1993). Understanding of social rules. In M. Bennett (Ed.), *The development of social cognition: The child as psychologist* (pp. 111-141). New York: Guilford Press.

Smetana, J. G. (2005). Adolescent-parent conflict: Resistance and subversion as developmental process. In L. P. Nucci (Ed), *Conflict, contradiction, and contrarian elements in moral development and education* (pp. 69-91). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Smetana, J. G., & Bitz, B. (1996). Adolescents' conceptions of teachers' authority and their relations to rule violations in school. *Child Development*, *67*(3), 1153-1172. doi: 10.1111/j.1467-8624.1996.tb01788.x.

Smith, E. E., Chase, W. G., & Smith, P. G. (1973). Stimulus and response repetition effects in retrieval from short-term memory. *Journal of Experimental Psychology*, *98*(2), 413-422. doi: 10.1037/h0034398.

Smith, M. C. (1968). The repetition effect and short-term memory. *Journal of Experimental Psychology*, *77*(3), 435-439. doi: 10.1037/h0021293.

Soetens, E. (1998). Localizing sequential effects in serial choice reaction time with the information reduction procedure. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 547-568. doi: 10.1037/0096-1523.24.2.547.

Solomon, R. L. (1984). The influence of work on behavior. *Psychological Bulletin*, *45*(1), 1-40. doi: 10.1037/h0055527.

Sommer, W., Leuthold, H., & Soetens, E. (1999). Covert signs of expectancy in serial reaction time tasks revealed by event-related potentials. *Perception & Psychophysics*, *61*(2), 342-353. doi: 10.3758/BF03206892.

Song, J.-H., & Nakayama, K. (2008). Target selection in visual search as revealed by movement trajectories. *Vision Research*, *48*(7), 853-861. doi: 10.1016/j.visres.2007.12.015.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360-366. doi:10.1016/j.tics.2009.04.009.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport*, *12*(13), 2849-2853. doi: 10.1097/00001756-200109170-00019.

Spence, S. A., Hunter, M. D., Farrow, T. F. D., Green, R. D., Leung, D. H., Hughes, C. J., & Ganesan, V. (2004). A cognitive neurobiological account of deception: Evidence from functional neuroimaging.

Philosophical Transactions of the Royal Society of London Series B, *359*(1451), 1755-1762. doi: 10.1098/rstb.2004.1555.

Spence, S. A., Kaylor-Hughes, C., Farrow, T. F. D., & Wilkinson, I. D. (2008). Speaking of secrets and lies: The contribution of ventrolateral prefrontal cortex to vocal deception. *NeuroImage*, *40*(3), 1411-1418. doi: 10.1016/j.neuroimage.2008.01.035.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393-10398. doi: 10.1073/pnas.0503903102.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, *38*(4), 387-401. doi: 10.1016/0013-4694(75)90263-1.

Stallen, M., De Dreu, C. K. W., Shalvi, S., Smidts, A., & Sanfey, A. G. (2012). The herding hormone: Oxytocin stimulates in-group conformity. *Psychological Science*, *23*(11), 1288-1292. doi: 10.1177/0956797612446026.

Steinhauser, M., & Hübner, R. (2006). Response-based strengthening in task shifting: Evidence from shift effects produced by errors. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 517-534. doi: 10.1037/0096-1523.32.3.517.

Stemmer, B., Witzke, W., & Schönle, P. W. (2001). Losing the error related negativity in the EEG of human subjects: An indicator for willed action. *Neuroscience Letters*, *308*(1), 60-62.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of human behavior. *Personality and Social Psychology Review*, *8*(3), 220-247. doi: 10.1207/s15327957pspr0803_1.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187-1188. doi: 10.1126/science.150.3700.1187.

Tan, S., & Dixon, P. (2011). Repetition and the SNARC effect with one- and two-digit numbers. *Canadian Journal of Experimental Psychology*, *65*(2), 84-97. doi: 10.1037/a0022368.

Turiel, E. (1983). The development of social knowledge: Morality and convention. Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131. doi: 10.1126/science.185.4157.1124.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In U. Undeutsch (Ed.), *Handbuch der Psychologie Vol. 11: Forensische Psychologie* (pp. 26-181). Göttingen, Germany: Hogrefe.

Vandierendonck, A., Demanet, J., Liefooghe, B., & Verbruggen, F. (2012). A chain-retrieval model for voluntary task switching. *Cognitive Psychology*, *65*(2), 241-283. doi: 10.1016/j.cogpsych.2012.04.003.

Verleger, R. (1997). On the utility of P3 latency as an index of mental chronometry. *Psychophysiology*, *34*(2), 131-156. doi: 10.1111/j.1469-8986.1997.tb02125.x.

Verleger, R., Jaśkowski, P., & Wascher, E. (2005). Evidence for an integrative role of P3b in linking reaction to perception. *Journal of Psychophysiology*, *19*(3), 165-181. doi: 10.1027/0269-8803.19.3.165.

Vernon, P. A. (1993). Der Zahlen-Verbindungs-Test and other trail-making correlats of general intelligence. *Personality and Individual Differences*, *14*(1), 35-40. doi: 10.1016/0191-8869(93)90172-Y.

Vincent, C., Taylor-Adams, S., Chapman, E. J., Hewett, D., Prior, S., Strange, P., & Tizzard, A. (2000). How to investigate and analyse clinical incidents: Clinical Risk Unit and Association of Litigation and Risk Management protocol. *British Medical Journal*, *320*(7237), 777-781. doi: 10.1136/bmj.320.7237.777.

Vocat, R., Pourtois, G., & Vuilleumier, P. (2008). Unavoidable errors: A spatio-temporal analysis of time-course and neural sources of evoked potentials associated with error processing in a speeded task. *Neuropsychologia*, *46*(10), 2545-2555. doi: 10.1016/j.neuropsychologia.2008.04.006.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behaviour. *Journal of Nonverbal Behavior*, *24*(4), 239-263. doi: 10.1023/A:1006610329284.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: a critical survey of the literature. *Psychological Bulletin*, *77*(1), 65-72. doi: 10.1037/h0032060.

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009). Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business and Psychology*, *24*(1), 33-49. doi: 10.1007/s10869-009-9090-8.

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: response time as a cue to deception. *Applied Cognitive Psychology*, *17*(7), 755-774. doi: 10.1002/acp.914.

Walther, E., Bless, H., Strack, F., Rackstraw, P., Wagner, D., & Werth, L. (2002). Conformity effects in memory as a function of group size, dissenters and uncertainty. *Applied Cognitive Psychology*, *16*(7), 793-810. doi: 10.1002/acp.828.

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *11*(2), 92-107. doi: 10.1080/17470215908416296.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

Waszak, F., Pfister, R., & Kiesel, A. (2013). Top-down vs. bottom-up: When instructions overcome automatic retrieval. *Psychological Research*, *in press*. doi: 10.1007/s00426-012-0459-3.

Waszak, F., Wascher, E., Keller, P. E., Koch, I., Aschersleben, G., Rosenbaum, D. A., & Prinz, W. (2005). Intention-based and stimulus-based mechanisms in action selection. *Experimental Brain Research*, *162*(3), 346-356. doi: 10.1007/s00221-004-2183-8.

Waszak, F., Wenke, D., & Brass, M. (2008). Cross-talk of instructed and applied arbitrary visuomotor mappings. *Acta Psychologica*, *127*(1), 30-35. doi: 10.1016/j.actpsy.2006.12.005.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, *101*(1), 34-52. doi: 10.1037/0033-295X.101.1.34.

Wegner, D. M. (2009). How to think, say, or do precisely the worst thing for any occasion. *Science*, *325*(5936), 48-50. doi: 10.1126/science.1167346.

Wegner, D. M., Coulton, G., & Wenzlaff, R. (1985). The transparency of denial: Briefing in the debriefing paradigm. *Journal of Personality and Social Psychology*, *49*(2), 338-346. doi: 10.1037/0022-3514.49.2.338.

Wenke, D., Gaschler, R., & Nattkemper, D. (2007). Instruction-induced feature binding. *Psychological Research*, *71*(1), 92-106. doi: 10.1007/s00426-005-0038-y.

Widiger, T. A., Knudson, R. M., & Rorer, L. G. (1980). Convergent and discriminant validity of measures of cognitive styles and abilities. *Journal of Personality and Social Psychology*, *39*(1), 116-129. doi: 10.1037/0022-3514.39.1.116.

Wilson, G. D. (1968). Reversal of differential GSR conditioning by instructions. *Journal of Experimental Psychology*, *76*(3), 491-493. doi: 10.1037/h0025540.

Witkin, H. A. (1949). The nature and importance of individual differences in perception. *Journal of Personality*, *18*(2), 145-170. doi: 10.1111/j.1467-6494.1949.tb01237.x.

Witkin, H. A., & Asch, S. E. (1948). Studies in space orientation IV: Further experiments on perception of the upright with displaced visual fields. *Journal of Experimental Psychology*, *38*(6), 762-782. doi: 10.1037/h0053671.

Witkin, H. A., & Goodenough, D. R. (1977). Field dependence and interpersonal behavior. *Psychological Bulletin*, *84*(4), 661-689.

Witkin, H. A., & Goodenough, D. R. (1981). Cognitive styles: Essence and origins. Field dependence and field independence. New York: International Universities Press.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt, Rinehart and Winston.

Yagil, D. (1998). Gender and age-related differences in attitudes toward traffic laws and traffic violations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *1*(2), 123-135. doi: 10.1016/S1369-8478(98)00010-2.

Yeung, N. (2010). Bottom-up influences on voluntary task switching: The elusive homunculus escapes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 348-362. doi: 10.1037/a0017894.