Learning & Behavior (2020) 48:373–391
https://doi.org/10.3758/s13420-019-00399-0

Exploring the potential impact of relational coherence on persistent rule-following: The first study

Colin Harte[1] · Dermot Barnes-Holmes[1] · Yvonne Barnes-Holmes[1] · Ciara McEnteggart[1] · Jinthe Gys[2] · Charlotte Hasler[1]

Published online: 2 January 2020

Abstract
Rule-governed behavior and derived relational responding have both been identified as important variables in human learning. Recent developments in the relational frame theory (RFT) have outlined a number of key variables of potential importance when analyzing the dynamics involved in derived relational responding. Recent research has explored the impact of one of these variables, level of derivation, on persistent rule-following and implicated another, coherence, as possibly important. However, no research to date has examined the impact of coherence on persistent rule-following directly. Across two experiments, coherence was manipulated through the systematic use of performance feedback, and its impact was examined on persistent rule-following. A training procedure based on the implicit relational assessment procedure (IRAP) was used to establish novel combinatorially entailed relations that manipulated the feedback provided on the trained relations (A-B and B-C) in Experiment 1, and on the untrained, derived relations (A-C) in Experiment 2. One of these relations was then inserted into the rule for responding on a subsequent contingency-switching match-to-sample (MTS) task to assess rule persistence. While no significant differences were found in Experiment 1, the provision or non-provision of feedback had a significant differential impact on rule-persistence in Experiment 2. Specifically, participants in the Feedback group resurged back to the original rule for significantly more responses after demonstrating contingency-sensitive responding than did the No-Feedback group, after the contingency reversal. The results highlight the subtle complexities that appear to be involved in persistent rule-following in the face of reversed reinforcement contingencies.

Keywords Rule-governed behaviour · Relational coherence · Derived relations · Persistent rule-following

Introduction

The importance of the impact of rules or instructions on human behavior has long been acknowledged within the behavior-analytic literature (e.g., Michael, 1980). The concept of *rule-governed behavior* was first proposed by Skinner (1966) within the context of an operant account of problem-solving. At this time, rules were defined as contingency-specifying stimuli that allowed a listener to problem-solve without having to contact the relevant reinforcement contingencies directly. For example, the simple rule "Don't eat the berries growing on a Holly tree, they're poisonous" allows the listener to learn to avoid eating toxic berries without directly experiencing sickness.

During the 1970s and 1980s, a wealth of research emerged that focused on the impact of rules on human performance on schedules of reinforcement (for an early book-length review, see Hayes, 1989). One of the key findings that emerged from this work was that for humans with basic language skills, behavior under the control of rules or instructions quite often led to what was described as insensitivity to direct contingencies of reinforcement (e.g., Bentall, Lowe, & Beasty, 1985; Catania, Shimoff, & Matthews, 1989). For example, following an un-cued change in reinforcement contingencies, participants who had initially been responding in accordance with an experimenter-given rule were more likely to persist in following that rule for longer than participants who had not been given such a rule, despite it now resulting in limited access to reinforcers (e.g., Hayes, Brownstein, Haas, & Greenway, 1986; LeFrancois, Chase, & Joyce, 1988; Shimoff, Catania,

✉ Colin Harte
Colin.Harte@UGent.be

[1] Department of Experimental, Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

[2] Department of Psychology, Thomas More Hogeschool, Antwerp, Belgium

& Matthews, 1981). This rule-based insensitivity has since been widely argued to play an important role in human psychological suffering (e.g., Baruch, Kanter, Busch, Richardson, & Barnes-Holmes, 2007; Rosenfarb, Newland, Brannon, & Howey, 1992; Zettle & Hayes, 1982). Specifically, it has been argued that human psychological suffering can be understood in terms of excessive rule-following, which by definition undermines or reduces contact with reinforcers in the natural environment. The literature investigating the role of rule-governed behavior in human psychological suffering, however, is scarce and conflicting in nature. For example, some research suggests that excessive rule-governance may be associated with self-reported depression (McAuliffe et al., 2014), while others have reported that individuals with self-reported depression are in fact *more* sensitive to changes in environmental contingencies, or their behavior is less rule-governed (Baruch et al., 2007; Rosenfarb et al., 1993). Conflicting results such as these highlight the need for further research on excessive rule-following if we are to better understand how it may play a potentially important role in human psychological distress.

While rule-based insensitivity to direct contingencies of reinforcement seemed to be a key feature of human behavior, a second feature that came to light around the same time was that of *derived relational responding*. This concept emerged in the early 1970s with the seminal work of Sidman and colleagues (e.g., Sidman, 1971; Sidman & Tailby, 1982), and the early work in this area focused almost exclusively on a phenomenon that came to be known as *stimulus equivalence* (see Sidman, 1994, for a book-length review). The key finding was that untrained or unreinforced responses often quite readily emerged out of a small number of trained or reinforced responses. For example, if reinforcement was provided for matching two abstract stimuli, one of which was also matched to a third (e.g., A=B and B=C), previously unreinforced responses often emerged (e.g., A=C and C=A). When such a pattern of unreinforced responding occurred, the stimuli involved were said to form an *equivalence class* or *equivalence relation*. Crucially, this phenomenon appeared to occur with relative ease in verbally-able humans, but was not readily or reliably observed with nonhumans or with humans with severely limited language abilities. Indeed, the lack of clear evidence for even the most basic types of equivalence responding in nonhumans has persisted (see Dougher, Twohig, & Madden, 2014).

The extension of stimulus equivalence as a core explanatory tool for analyzing the complexities of human behavior came with the development of relational frame theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001; Steele & Hayes, 1991). RFT is a behavior-analytic account of human language and cognition that views stimulus equivalence as but one class of generalized operant behavior, and posits that many others are possible. Specifically, RFT proposes that there are many generalized relational operants and the generic term or concept *arbitrarily applicable relational responding* (AARR) is used to label these operant classes. According to RFT, extended histories involving many relevant reinforced exemplars serve to create different patterns of relational responding, referred to as *relational frames*, such as: similarity, difference, opposition, distinction, hierarchy, temporal, and deictic (see Hughes & Barnes-Holmes, 2016, for an extensive review).

While the experimental analyses of derived stimulus relations and of rule-governed behavior have only rarely overlapped (e.g., Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2017; O'Hora, Barnes-Holmes, Roche, & Smeets, 2004), conceptually the link between the two has been relatively strong. That is, both Sidman (1994) and Hayes et al. (2001) argued that the human capacity to engage in derived relational responding may be important for understanding the ways in which contingencies of reinforcement come to be specified by rules (see also Hayes & Hayes, 1989). Indeed, some research thereafter has suggested that derived relational responding could provide the basis for a technical analysis of rule-governed behavior, and this has been successfully modelled in the laboratory (O'Hora et al., 2004; O'Hora, Barnes-Holmes, & Stewart, 2014).

More recently, research has begun to extend this work by examining the impact of different features of derived relational responding on rule-based insensitivity or rule persistence (e.g., Harte et al., 2017; Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2018; Monestes, Greville, & Hooper, 2017). For example, the study by Harte et al. (2017) sought to investigate the extent to which participants persisted in rule-following on a contingency-switching match-to-sample (MTS) task. In one condition, participants were provided with a direct instruction that required no derived relational responding within the experiment; in a second condition, following the instruction required that participants derived the meaning of the instruction within the experiment. Specifically, the direct rule involved instructing participants to choose the comparison image that was the least like the sample image. In the derived-rule condition, however, participants were first trained and tested for a derived relation between the phrase "least like" and a novel word "beda." The novel word was then inserted into the instruction for responding on the MTS task (i.e., "Choose the image that is beda the sample image"). The researchers also manipulated the number of opportunities participants had to obtain points for following this rule before the un-cued contingency reversal (Experiment 1, ten opportunities; Experiment 2, 100 opportunities). In each experiment, the contingency reversal was followed by 50 trials. While no differences in persistent rule-following emerged in Experiment 1 between the direct and derived-rule conditions, when the opportunities to follow the reinforced rule were more protracted in Experiment 2, the direct rule produced greater persistence than the derived-rule condition (and a control condition that involved presenting no relevant rule).

In discussing their findings, Harte et al. (2017) noted that they had not manipulated relative *levels* of derivation within the experiments. Specifically, it was assumed that in the direct rule condition derivation would be extremely low relative to the derived-rule condition, because a direct rule did not involve deriving a new relation during the experiment. In a subsequent study (Harte, et al., 2018), therefore, the authors manipulated levels of derivation within the experiment and replicated the basic effect reported previously – lower levels of derivation appeared to generate greater rule-persistence. This was the case for both mutually entailed (Experiment 1) and combinatorially entailed relations (Experiment 2).

Conceptually, the research reported by Harte et al. (2017, 2018) focused on levels of derivation on persistent rule-following, in part because it was closely linked to the development of a new framework for conducting RFT-based research more generally (see Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017). This framework is known as the multi-dimensional, multi-level (MDML) framework and conceptualizes AARR as varying along five levels of relational development and four dimensions. The five levels are seen as increasingly advanced forms of relational development progressing from: (1) mutual entailment, (2) combinatorial entailment, (3) relational networks, and (4) relating relations, to (5) relating relational networks. The framework presents these five levels as intersecting with four contextual dimensions: (1) coherence, (2) complexity, (3) derivation, and (4) flexibility. The MDML framework was designed to have extremely broad scope as a tool for guiding RFT-based research, and thus a detailed treatment of it is beyond the remit of the current article (but see Barnes-Holmes, Finn, McEnteggart, & Barnes-Holmes, 2018, for a recent summary). Nevertheless, the framework was used to offer post hoc interpretations of the differential results found by Harte et al. (2017), and in identifying level of relational development (i.e., mutual vs. combinatorial entailment) as a potentially important variable for experimental analysis in Harte et al. (2018). Thus, it seems important to draw on the framework to justify the rationale for the current research, to which we now turn.

The research by Harte et al. (2017, 2018) involved integrating the work on derived relations and rule-persistence, and focused specifically on the dimension of derivation, across two levels of relational development (mutual and combinatorial entailment), as specified within the MDML framework. Another potentially important variable of interest highlighted within the framework is the dimension of coherence. Within the framework, coherence refers to the extent to which a particular pattern of relational responding is consistent (coherent) with a previously established pattern. For example, if you are told that "A is larger than B," the derived response that "B is smaller than A" would be deemed coherent, but the response "B is the same size as A" would not (unless, of course, the wider context was modified to support an "incoherent" response, such as "Please respond to all questions with a silly answer").

The primary purpose of the current study was to use a similar paradigm to Harte et al. (2018) to train novel relations, but to manipulate coherence instead of derivation, through the systematic use of feedback.[1] That is, would a condition involving higher levels of coherence/more corrective feedback in the newly trained relations produce more or less persistence in rule-following on the same contingency-switching MTS task than a condition involving no feedback on these relations? Experiment 1 involved training participants on novel A-B and B-C relations followed by further retraining of these relations with and without feedback. Experiment 2 involved training participants on the same novel A-B and B-C relations followed by directly testing the novel A-C relations with and without feedback. A range of self-report measures of psychological distress was used to explore the extent to which derived rule-following may correlate with self-reported levels of distress in the general population. Two self-report measures of rule-following were also employed to determine if they would predict actual persistent rule-following. Given the exploratory and relatively inductive nature of the current research, we refrained from making formal predictions.

Experiment 1

As previously noted, Experiment 1 involved manipulating coherence through the provision of differential performance feedback. Specifically, we asked if delivering such feedback for newly trained relations would produce more or less persistence in rule-following on a contingency-switching MTS task than a condition involving no feedback. Experiment 1, therefore, involved training participants on novel A-B and B-C relations followed by further retraining of these relations with and without feedback.

Participants

A total of 98 individuals participated in Experiment 1, 74 females and 24 males. They ranged in age from 18 to 26 years (*M* = 20, *SD* = 2.16) and were recruited through random convenience sampling from the online participant system at

[1] Coherence may be defined as both an operation and as a process. In the current study, the provision versus non-provision of feedback should be seen as an attempt to define coherence as an operation. Coherence as a behavioral process, however, is an inference that is made based on the observation of specific behavioral effects or changes that arise from the operation. This is entirely consistent with the definition of reinforcement as an operation (a contingency) and as a process (a change in behavior as a result of that contingency; Catania, 1979).

Ghent University. All participants spoke Dutch as their first language and were paid a fixed sum of 10 euros. Participants were assigned to one of two conditions, referred to as Feedback and No Feedback; the sequence in which participants signed up for the study determined the condition to which they were assigned (i.e., in general, odd numbered = Feedback; even numbered = No Feedback). The data from 38 participants (20 from No Feedback and 18 from Feedback) were excluded because they failed to meet specific performance criteria on either a Training IRAP or an MTS task (see below), leaving $N = 60$ for analysis, 30 in the Feedback condition, and 30 in the No-Feedback condition.

Setting

The experiment was conducted in a cubicle at Ghent University in which participants were seated in front of a standard Dell laptop. The experimenter was present in the cubicle at the beginning of each stage of the experiment to instruct participants about that stage, but left each participant alone thereafter. The experimenter re-entered the cubicle at various points throughout the experiment; for example, when transitioning from one stage to the next (see below).

Materials and apparatus

Experiment 1 involved three computer-based tasks (a Derivation Pre-Training task, the Training IRAP, and an MTS task) and five self-report measures. The Derivation Pre-Training task and the Training IRAP employed a total of eight stimulus sets, each comprised of three separate stimuli (see below). Participants completed all aspects of the experiment on a standard Dell laptop.

The Derivation Pre-Training Task The purpose of the Derivation Pre-Training Task was to provide participants with a history within the experiment of relating stimuli that were deemed to be semantically similar or dissimilar. The task involved six sets of stimuli, with three stimuli in each set (see Table 1). During the task, the stimuli were presented in pairs in such a way that for some pairs participants should already know the relation between them because they were English and Dutch words (e.g., "hond" and "dog"). For other pairs, the relation between them should be unknown because the pairs contained an Irish word (e.g., "madra" and "dubh") or a nonsense stimulus (e.g., "XXX" or "////"). The remaining pairs contained words that allowed participants to derive a relation between a known Dutch word and a previously unknown Irish word. The general purpose of this pre-training task was to prepare participants for deriving the target relations with completely novel stimuli in the context of persistent rule-following in subsequent stages of the experiment (pilot work had indicated high levels of attrition without this type of pre-training).

The Derivation Pre-Training Task was presented in Microsoft PowerPoint. All trials presented a label stimulus at the top of the screen (e.g., "Hond" the Dutch word for "Dog"), a target stimulus in the middle (the English word "Dog"), and two response options, for example, the Dutch words "Goed" (meaning correct) and "Verkeerd" (meaning incorrect), which appeared at the bottom left and right of the screen.

The Training IRAPs Consistent with Harte et al. (2018), three Training IRAPs were used to establish a relational network involving directly trained relations between known words (A stimuli) and symbols (B stimuli), and between the same symbols and novel words (C stimuli). The IRAPs employed stimuli from Sets 7 and 8 (see Table 2). As such, during training of the A-B relations, Dutch words and phrases were presented (the English translations are used here). All trials presented a label at the top of the screen, with a single target below and two response options. The label stimuli always included one of two phrases "Least Similar" or "Most Similar," the target stimulus was always "TTT" or "]][[," and each pair of response options included "True" versus "False," "Yes" versus "No," "Correct" versus "Incorrect," or "Right" versus "Wrong." These stimuli were combined to generate four A-B trial types referred to as: Least Similar-TTT; Most Similar-TTT; Least Similar-]][[; and Most Similar-]][[(see Fig. 1).

During training of the B-C relations, each trial presented the stimuli "TTT" or "]][[" as labels, the novel words "Beda" and "Sarua" as targets, along with the same response options. Taken together, the four B-C trial types were as follows: TTT-Beda;]][[-Beda; TTT-Sarua; and]][[-Sarua (see Fig. 2).

Table 1 Stimulus sets employed within each cycle of the Derivation Pre-Training Task

Derivation Pre-Training Task Stimuli					
Set 1	*Set 2*	*Set 3*	*Set 4*	*Set 5*	*Set 6*
Hond	Zwart	Hemd	Fles	Boek	Jas
Dog	Black	Shirt	Bottle	XXX	////
Madra	Dubh	Leine	Buideal	Leabhar	Cota

Table 2 Stimulus sets employed within each of the Training IRAPs

Training IRAPs Stimuli	
Set 7	*Set 8*
Least Similar	Most Similar
TTT	]][[
Beda	Sarua

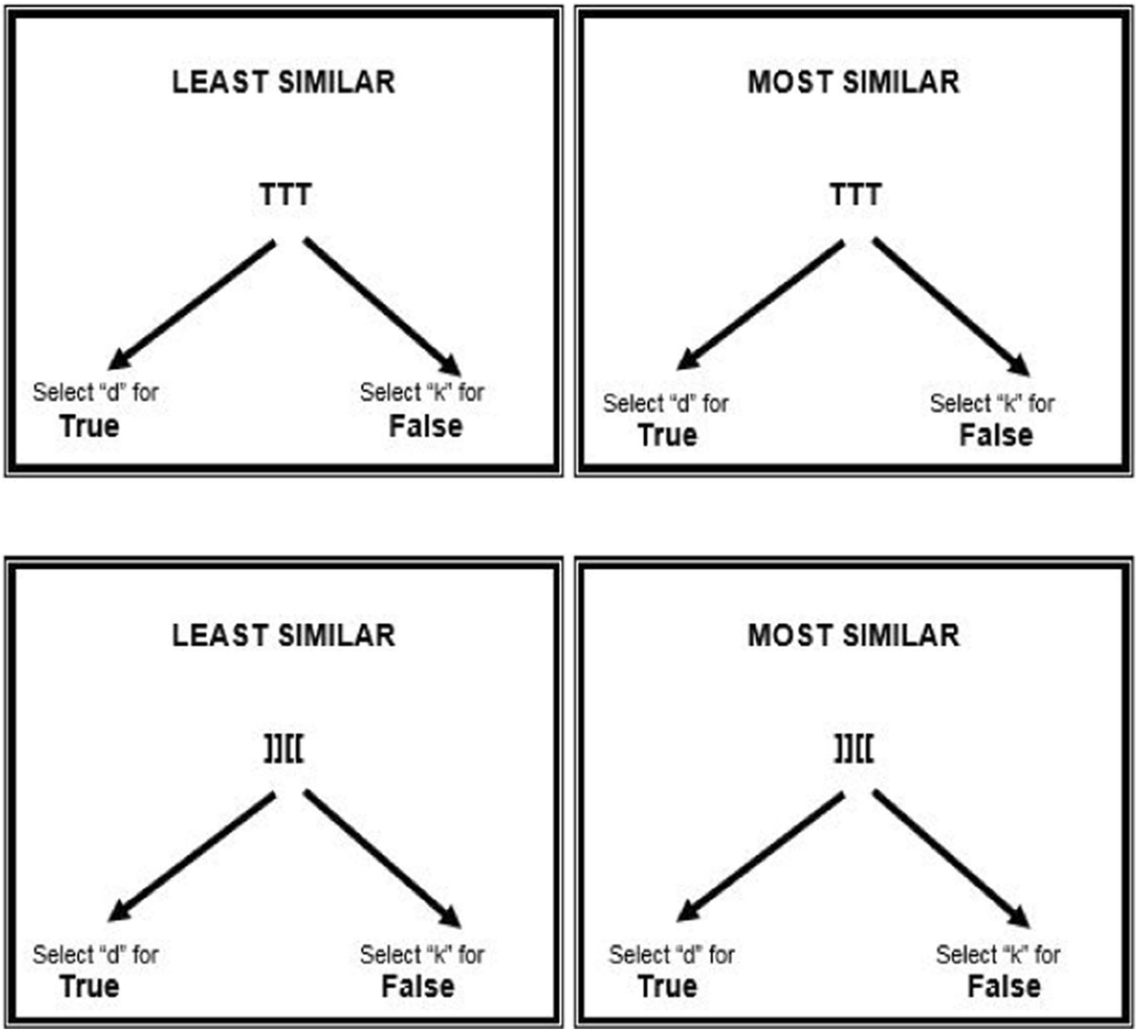

Fig. 1 Diagrammatic representation of the IRAP trial-types that appear in A-B baseline relation familiarization block. Arrows did not appear on-screen. The four IRAP trial-types were denoted as: *Least Similar-TTT, Most Similar-TTT, Least Similar-]][[,* and *Most Similar-]][[*

The mixed A-B/B-C Training IRAP was similar to the A-B and B-C Training IRAPs, except that A-B and B-C relations were presented within each block of training trials, rather than across two separate IRAPs. This created eight trial types, identical to the four A-B trial types and the four B-C trial types listed above.

The MTS task During each MTS trial, a sample stimulus (always a random shape) was presented at the top of the screen, with three comparison stimuli (all random shapes, but none identical to the sample nor to each other) along the bottom (see Fig. 3 for an example of a single trial). Each comparison varied in its similarity to the sample. Specifically, one comparison was clearly the *most similar to the sample* (same basic shape with minor variations, see center of Fig. 3). A second comparison was also quite like the sample, but with more variations (see left-hand side of Fig. 3), rendering it *less similar to* the sample. Finally, the third comparison was clearly the *least similar to* the sample because it had few or no overlapping features (right-hand side of Fig. 3). Each sample and three-comparison combination comprised an individual stimulus set, such that only those comparisons appeared in the presence of that sample. Participants emitted a response by pressing the key (*D, G,* or *K*) directly below the comparison they wished to select. A total of 54 stimulus sets were employed, with each set presented at least once, but no more than three times, across 150 trials.

Questionnaires Experiment 1 involved five self-report questionnaires, three of which were standardized measures (the Depression, Anxiety and Stress Scales, DASS-21; the Acceptance and Action Questionnaire, AAQ-II; the Generalized Pliance Questionnaire, GPQ), the fourth of which, the Psychological Flexibility Index (PFI), is currently still under development, and the fifth of which, the Propensity to Rule-Following Scale (PRFS), was created by Harte et al. (2018) to measure rule-persistence. The first three scales were included as measures of psychological distress because such

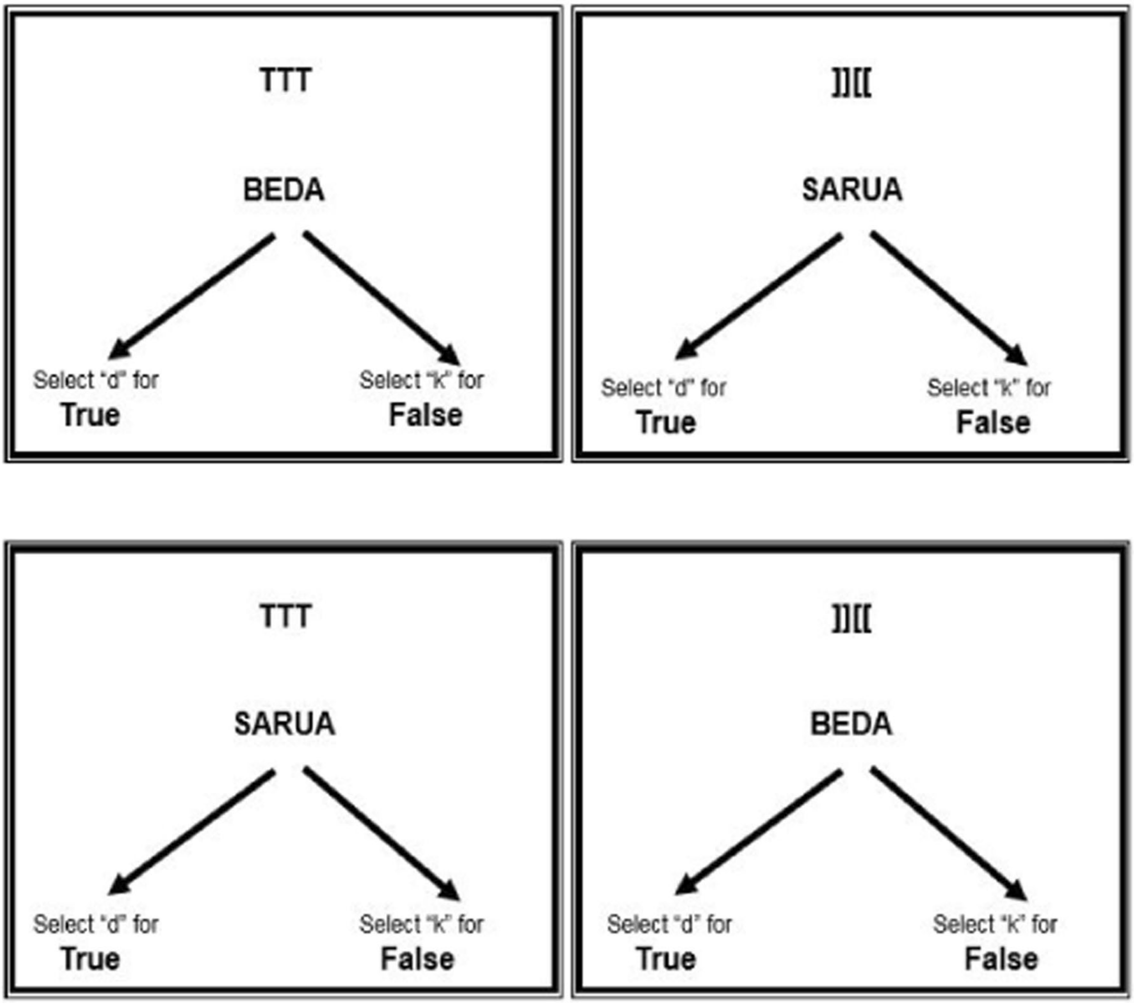

Fig. 2 Diagrammatic representation of the IRAP trial-types that appear in B-C baseline relation familiarization block. Arrows did not appear on-screen. The four IRAP trial-types were denoted as: *TTT-Beda,]][[-Beda, TTT-Sarua,* and *]][[-Sarua*

measures have been related to persistence in rule-following in previous research (e.g., McAuliffe, Hughes, & Barnes-Holmes, 2014). The final two scales were included as self-report measures of persistent rule-following.

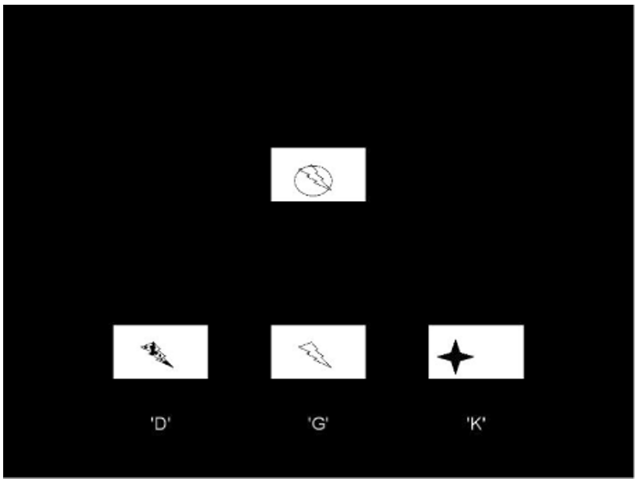

Fig. 3 An example of a single trial and single stimulus set presented in the MTS task

The *DASS-21* comprises three subscales measuring depression, anxiety, and stress across a total of 21 statements, with seven statements per subscale (e.g., an item from the anxiety subscale was "I found it hard to wind down"; Lovibond & Lovibond, 1995). All items were rated in terms of participant experiences within the last week on a 4-point scale from 0 (*Did not apply to me at all*) to 3 (*Applied to me very much or most of the time*). An overall DASS score is calculated by summing all 21 items. However, all overall and subscale scores obtained must be doubled, and severity bands are generated accordingly. Specifically, the overall DASS score ranges from 0 to 126. Higher scores on the overall scale and on each subscale indicate greater psychological distress. The measure has demonstrated excellent internal consistency (Henry & Crawford, 2005): depression (alpha = 0.88); anxiety (alpha = 0.82); stress (alpha = 0.90); and total DASS (alpha = 0.93). The Dutch version of the scale, which according to deBeurs, Van Dyck, Marquenie, Lange, and Blonk (2001) has yielded similar sufficient internal consistency, was employed in the current experiment.

The *AAQ-II* measures acceptance of negative private events across seven statements (e.g., "My painful memories prevent me from having a fulfilled life"; Bond et al., 2011). All items were rated on a 7-point scale from 1 (*Never true*) to 7 (*Always true*), yielding a minimum score of 7 and a maximum of 49. High scores indicate *low* acceptance, while low scores indicate *high* acceptance. The measure has demonstrated adequate internal consistency with alpha coefficients ranging from 0.78 to 0.88 (Bond et al., 2011). Again, the Dutch version of the scale was employed in the present experiment. According to Bernaerts, De Groot, and Kleen (2012), this version has yielded a Cronbach's alpha of 0.85.

The *PFI* is a measure currently under development and designed to measure psychological flexibility (Bond et al., 2017) across a total of 80 statements (e.g., "Even when I am uncertain of what to do, I can still do what is right for me"). All items were rated on a Likert scale from 1 (*Disagree strongly*) to 6 (*Agree strongly*) and the measure yields a total score (based on the summation of all items), with a minimum of 80 and a maximum of 480. High scores indicate high flexibility, while low scores indicate low flexibility. All items were translated into Dutch using the backward-forward method. As the measure is still in development, there are currently no published validity or reliability data.

The *GPQ* is designed to measure generalized pliance (Ruiz, Suárez-Falcón, Barbero-Rubio, & Flórez, 2019). Pliance was originally defined as rule-governed behavior that is controlled mainly by consequences mediated by the speaker for correspondence between the rule and behavior (e.g., when a young child follows the rule "Eat your vegetables and you can have ice-cream"; Zettle & Hayes, 1982). All 18 items on the GPQ (e.g., "My decisions are very much influenced by other people's opinions") were rated on a Likert scale from 1 (*Never true)* to 7 (*Always true*) and the measure yields a total score (based on the summation of all items), with a minimum of 18 and a maximum of 126. High scores indicate high pliance, while low scores indicate low pliance. Due to the fact there is no Dutch translation available, all items were again translated into Dutch using the backward-forward method. This translation has no reliability data. However, the English version has demonstrated adequate internal consistency in undergraduate, general, and clinical populations with alpha coefficients of .93, .95, and .97, respectively (Ruiz et al., 2019).

The *PRFS* was created by Harte et al. (2018) to assess propensity to rule-following across six statements (i.e., "I would describe myself as someone who follows rules"; "If someone gives me a rule to follow, I do my best to follow that rule"; "I break rules often"; "When I break rules I feel uncomfortable"; "Rules are made to be broken"; and "If I was given a rule to follow and the rule proved to be incorrect, I would abandon the rule"). All items were rated on a Likert scale from 1 (*Always agree*) to 5 (*Always disagree*), yielding a minimum score of 6 and a maximum of 30. Items 3, 5, and 6 were reverse scored. High scores indicate low propensity for rule-following, while low scores indicate high propensity for rule-following. No evidence on the psychometric properties of the PRFS is available.

Procedure

Experiment 1 comprised five stages (see Fig. 4). Stage 1 presented the three initial questionnaires (i.e., DASS-21, AAQ-II, and PFI). Stage 2 presented the Derivation Pre-Training Task, which comprised three cycles, each made up of three phases: Phases 1 and 2 always comprised four trials, while Phase 3 always comprised six trials. In Phases 1 and 2, the relation between the two stimuli was always one of similarity, whereas in Phase 3, the relation was always one of difference. Stage 3 involved the Training IRAPs, which comprised four phases: Phase 1 presented the A-B relations Training IRAP; Phase 2 presented the B-C relations Training IRAP; Phase 3 presented the A-B/B-C relations Training IRAP, in which A-B and B-C relations were mixed randomly within each block of trials. Phase 4 was similar to Phase 3, except that half of the participants continued to receive feedback on each trial of the A-B/B-C Training IRAP, whereas the other half did not. Stage 4 involved the MTS task, with rule-consistent contingencies in Phase 1 and rule-inconsistent contingencies in Phase 2. Finally, Stage 5 presented the remaining questionnaires (i.e., GQP and PRFS).

Stage 1: DASS-21, AAQ-II, and PFI Participants completed the DASS-21, the AAQ-II, and the PFI, in that order, and proceeded immediately to Stage 2.

Stage 2: The Derivation Pre-Training Task The aim of the Derivation Pre-Training Task was to minimize the attrition observed in previous studies using this paradigm (e.g., Harte et al., 2018, 2017) by providing participants with the opportunity to derive relations of sameness and difference between two stimuli based on a single "mediating" third stimulus. A total of 42 trials were presented, and on each trial, the experimenter read aloud the two on-screen stimuli (e.g., "Hond" with "Dog" or "Hond" with "Black") and asked participants to respond to the question "Do these two stimuli have the same meaning?" by stating, for example, "Yes" or "No," which appeared on the bottom left- and right-hand sides of the screen.

The experimenter recorded each response and provided corrective feedback on every response. The next trial was presented immediately after. The Derivation Pre-Training Task comprised three separate cycles of training (see Table 3). Each cycle contained the same three phases and the same training trials; only the stimulus sets differed across the three cycles (see Table 1). Participants progressed immediately from one phase to the next and from one cycle to the next.

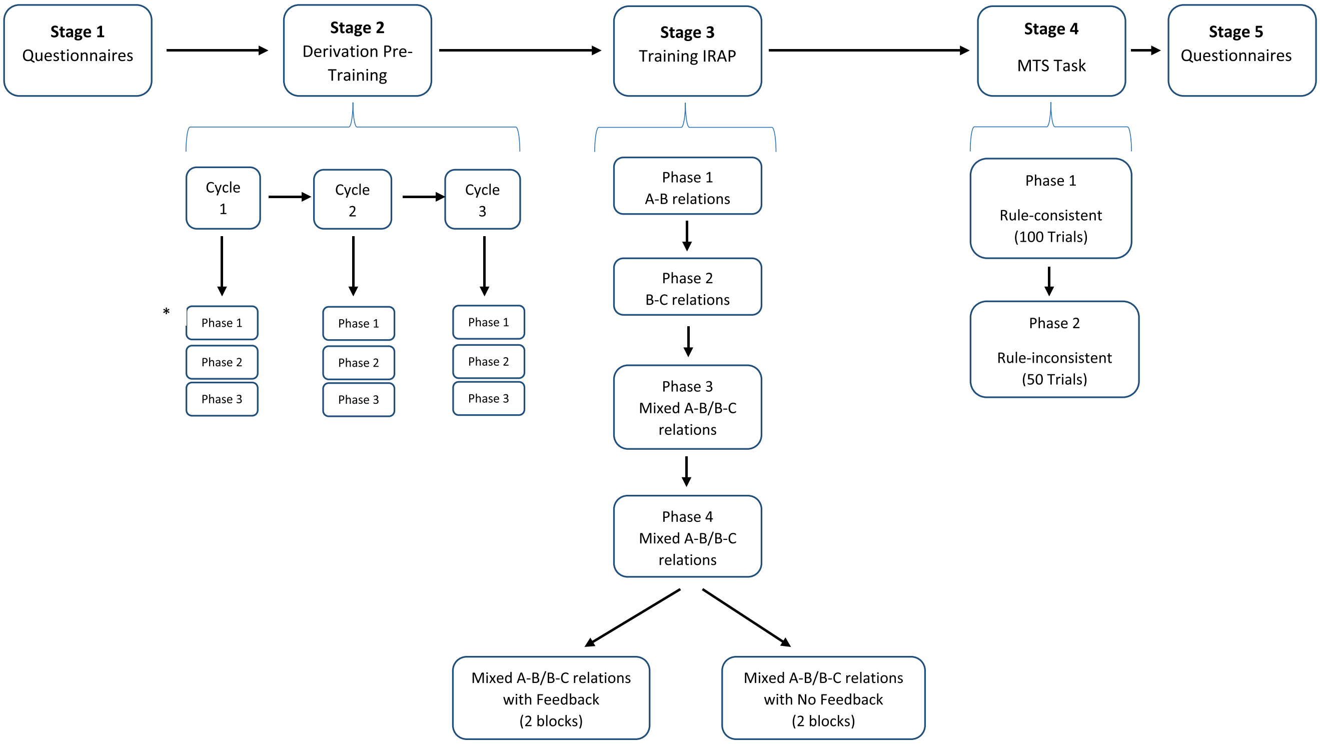

Fig. 4 A flow chart showing the experimental sequence of Experiment 1. * See Table 2 for the specific stimulus set sequencing presented within each phase of each cycle involved in Stage 2

Phase 1: Co-ordination relations I Phase 1 consisted of four trials involving stimulus Set 1. The first trial presented the stimuli "Hond" and "Dog" (feedback was provided after all trials); the second trial presented "Dog" and "Madra"; the third trial presented "Hond" and "Madra"; and the fourth trial presented the stimuli from the third trial but in the reversed order, "Madra" and "Hond." Correct responding involved relating all of these stimuli as the same.

Phase 2: Co-ordination relations II Phase 2 also consisted of the same four trials, but involving the stimuli from Set 2. Again, the first trial presented "Hemd" and "Shirt"; the second presented "Shirt" and "Leine"; the third presented "Hemd" and "Leine"; and the fourth presented the stimuli from the third trial but in the reversed order, "Leine" and "Hemd." Correct responding involved relating all of these stimuli as the same.

Phase 3: Distinction relations Phase 3 consisted of six trials that combined the relations established above. The first trial presented "Hond" and "Black"; the second presented "Zwart" and "Dog"; the third presented "Dog" and "Dubh"; the fourth presented "Black" and "Madra"; the fifth presented "Hond" and "Dubh"; and the sixth presented "Zwart" and "Madra". Correct responding involved relating all of these stimuli as different.

Cycles 2 and 3 were identical to Cycle 1, except that new stimulus sets were employed. Specifically, Cycle 2 employed Set 3 ("Hemd," "Shirt," "Leine") and Set 4 (Fles, Bottle, Buideal) and Cycle 3 employed Set 5 (Boek, XXX, Leabhar) and Set 6 ("Jas", "////", "Cota"). As noted above, Sets 5 and 6 contained both words and symbols. At the end of the third cycle, participants proceeded immediately to Stage 3.

Stage 3: The Training IRAPs Participants were initially instructed orally on how to complete the Training IRAP. That is, they were advised that each trial would present a phrase at the top of the screen with a symbol in the center, and that their task was to relate these together using one of the two response options as accurately as possible across each block (i.e., pressing *D* for the left option or *K* for the right option). This stage involved three Training IRAPs presented across four phases, and participants were required to reach the mastery criteria on each phase before proceeding to the next.

Phase 1: A-B Relations Training IRAP Phase 1 consisted of a block of 24 trials involving "Least Similar" and "TTT" from Set 7, and "Most Similar" and "]][[" from Set 8. There were four trial-types: Least Similar-TTT; Least Similar-]][[; Most Similar-]][[; and Most Similar-TTT. Correct responding was as follows: Least Similar-TTT/True; Most Similar-TTT/False; Least Similar-]][[/False; and Most Similar-]][[/True. There were six exposures to each trial-type, presented quasi-randomly within each block of 24 trials. Given that this was a Training IRAP, if a correct response was emitted the word "Right!" appeared immediately in the centre of the screen, and

Table 3 Stimulus combinations employed within each block of trials in each cycle of the Derivation Pre-Training task

Relation Type	Cycle 1		
	Phase 1	*Phase 2*	*Phase 3*
	Set 1	*Set 2*	*Sets 1 + 2*
Known Relations	Hond = Dog	Zwart = Black	Hond ≠ Black
			Zwart ≠ Dog
Trained Relations	Dog = Madra	Black = Dubh	Dog ≠ Dubh
			Black ≠ Madra
Derived Relations	Hond = Madra	Zwart = Dubh	Hond ≠ Dubh
	Madra = Hond	Dubh = Zwart	Zwart ≠ Madra
	Cycle 2		
	Phase 1	*Phase 2*	*Phase 3*
	Set 3	*Set 4*	*Sets 3 + 4*
Known Relations	Hemd = Shirt	Fles = Bottle	Hemd ≠ Bottle
			Fles ≠ Shirt
Trained Relations	Shirt = Leine	Bottle = Buideal	Shirt ≠ Buideal
			Bottle ≠ Leine
Derived Relations	Hemd = Leine	Fles = Buideal	Hemd ≠ Buideal
	Leine = Hemd	Buideal = Fles	Fles ≠ Leine
	Cycle 3		
	Phase 1	*Phase 2*	*Phase 3*
	Set 5	*Set 6*	*Set 5*
Trained Relations	Boek = XXX	Jas = ////	Boek ≠ ////
			Jas ≠ XXX
Trained Relations	XXX = Leabhar	//// = Cota	XXX ≠ Cota
			//// ≠ Leabhar
Derived Relations	Boek = Leabhar	Jas = Cota	Boek ≠ Cota
	Leabhar = Boek	Cota = Jas	Jas ≠ Leabhar

Each cell represents an individual trial

the next trial appeared 400 ms later. If an incorrect response was emitted, a red X appeared until a correct response was emitted. Participants received automated feedback on their overall accuracy and latency performances at the end of the first block of trials. If they had failed to achieve a mean accuracy (≥ 80%) and/or a mean latency (≤ 3,000 ms) *per trial-type* during Phase 1, they were re-exposed to Phase 1 until these criteria were reached, at which point they could proceed to Phase 2.

Phase 2: B-C Relations Training IRAP Phase 2 consisted of a block of 24 trials involving "TTT" and "Beda," and "]][[" and "Sarua." The four trial-types were: TTT-Beda; TTT-Sarua;]][[-Sarua; and]][[-Beda . Correct responding was as follows: TTT-Beda/True;]][[-Beda/False; TTT-Sarua/False; and]][[-Sarua/True. Again, there were six exposures to each trial-type and all other aspects of Phase 2 were identical to Phase 1.

Phase 3: Mixed A-B and B-C Relations Training IRAP Phase 3 consisted of a block of 32 trials involving all of the stimuli from Sets 7 and 8, presented in the same manner in which they had been presented in Phases 1 and 2, all within the same block. Each of the four trial-types from Phase 1 and each of the four from Phase 2 were presented four times each, quasi-randomly. All other aspects of Phase 3 were identical to Phases 1 and 2. Participants could not proceed to Stage 4 until they had reached the mastery criteria on all three phases of Stage 3. It is important to emphasize that all participants received feedback on each trial throughout Phases 1–3 of the Training IRAP.

Phase 4: Mixed A-B and B-C Relations with or without Feedback Phase 4 was similar in format to Phase 3, with the only exception that half of the participants (Feedback condition) received feedback on every trial and at the end of each block (as they had done previously), while the other half (No-Feedback condition) no longer received feedback at any point, across two identical blocks of trials. Instead, at the beginning of the Training IRAP, participants in the No-Feedback condition were explicitly instructed that they would no longer receive feedback at any point, but that it was still possible to get all trials correct. No performance criteria applied in Phase 4; thus, all participants proceeded through each of the two blocks and then immediately to Stage 4 once Phase 4 was complete.

Stage 4: MTS task At the beginning of the MTS task, participants were instructed to "Respond by selecting the shape that is *Beda* to the sample stimulus." It is important to recall that "Least Similar" had been trained as coordinate with "TTT," and "TTT" was trained as coordinate with "Beda"; hence, based on that training it was now assumed that participants could correctly derive that "Least Similar" was coordinate with "Beda." They were then instructed that each trial would present a shape at the top of the screen with three shapes on the bottom. Participants were advised that they would be awarded 1 point for each correct response and deducted 1 point for each incorrect response, and that their total score would appear after each trial. All participants were explicitly instructed to try to accrue as many points as possible. The total MTS task comprised 150 trials, 100 trials presented in Phase 1 and 50 trials presented in Phase 2.

Phase 1: Rule-consistent contingencies During the 100 trials that comprised Phase 1, all participants were required to select the comparison that was *least similar* to the sample. When a correct response was emitted, 1 point was awarded, and the screen cleared immediately to present the total number of points accrued thus far (in large red text in the center of the screen) for 3 s. Emitting an incorrect response resulted in the loss of 1 point, again followed by a display of the total number of points. These feedback contingencies were thus consistent with the instruction to select the comparison that was least similar to the sample.

Phase 2: Rule-inconsistent contingencies At precisely the 101st trial, the task contingencies were reversed *without warning*. That is, the contingencies for correct and incorrect responding switched for the 50 trials that comprised Phase 2. Therefore, correct responding now involved selecting the comparison that was physically most similar to the sample, rather than least similar.

Stage 5: GPQ and PRFS After the MTS task, participants completed the GPQ and the PRFS in that order.

Results and discussion

For the purposes of analysis, exclusion criteria were applied to both blocks of Phase 4 of the Training IRAPs. In each condition, the data from one participant were removed because they failed to maintain ≥ 75% accuracy and ≤ 3,500ms response latency criteria per trial-type in both training blocks (N = 96 remaining). As an aside, the mastery criteria for training trials was set at 80% and 3,000 ms, but the exclusion criteria for the test trials were set at 75% and 3,500 ms. The slightly relaxed criteria for testing were designed to reduce potential attrition, particularly in the No-Feedback condition, and this was consistent with the exclusion criteria employed in Harte et al. (2018). A strict accuracy criterion was also applied to the MTS task and required correct responding on at least 8 of the first 10, as well as 80 of the first 100 trials in Phase 1. This MTS task criterion was consistent with Harte et al. (2017, 2018), and again was designed to reduce the likelihood that participants learned to respond correctly and match the stimuli on the basis of trial and error. The data from 36 participants were removed when these participants failed to achieve this criterion, 17 from the Feedback condition and 19 from the No-Feedback condition (N = 60 remaining). Although this strict criterion (at least eight out of the first ten trials correct) led to the removal of many datasets, it was deemed essential that participants in both conditions performed equally well from the very beginning of the MTS task, because a potential difference at this point might indicate that one group learned to respond more through trial and error on the initial MTS trials than through derivation that was based on the previous IRAP training.

IRAP data

Before conducting the primary analyses, we compared the mean number of blocks in Phases 1–3 of the Training IRAPs required by participants in each condition. On the A-B relations, Feedback participants required a mean of 1.73 blocks (SD = 1.34), while No-Feedback participants required 1.61 (SD = 0.88). On the B-C relations, Feedback participants required 1.33 blocks (SD = .55), while No-Feedback participants required 1.23 (SD = .50). Finally, on the mixed A-B and B-C relations, Feedback participants required 1.30 blocks (SD = .92), while No-Feedback participants required 1. Overall therefore, the mean number of training blocks required by the Feedback group was 4.37 (SD = 2.17), while the No-Feedback group required 3.84 (SD = .93). Independent t-tests confirmed that none of these differences were significant (all p's > .07, without correction for multiple tests). Thus, any subsequent differences that emerged between the two groups during the experimental feedback manipulation in Phase 4 of the Training IRAP or the MTS task are not likely due to differences in the ability to learn how to respond on the IRAP per se.

Measures of rule persistence

Insofar as the primary aim of Experiment 1 was to compare performances between the Feedback and No-Feedback

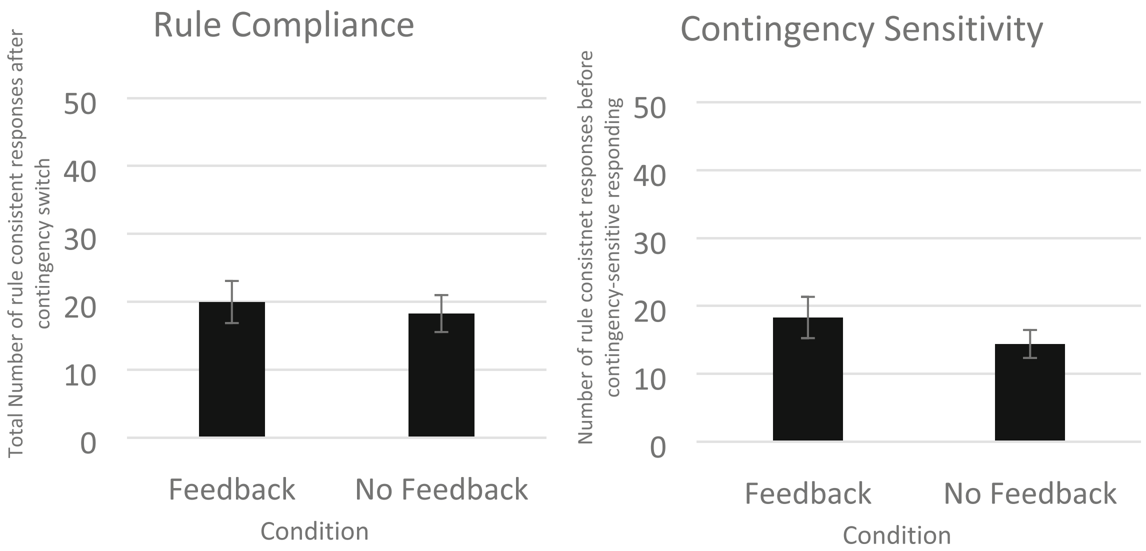

Fig. 5 Experiment 1: Mean rule compliance scores (**left panel**) and contingency sensitivity scores (**right panel**), with standard error bars for the Feedback and No-Feedback conditions

conditions, the data from the 50 trials in Phase 2 of the MTS task presented after the contingency reversal were analyzed in three ways (based on Harte et al., 2018). The three types of analyses that examined rule persistence are referred to as: rule compliance, contingency sensitivity, and rule resurgence.

Rule compliance was defined as the total number of responses (out of 50) that were consistent with the initial instruction "Respond by selecting the shape that is *Beda* [Least Similar] the sample stimulus," but were inconsistent with the reversed contingencies on the last 50 trials. Figure 5 (left-hand side) presents the group means for rule compliance and shows little difference between the conditions. That is, the Feedback group emitted almost the same number of responses (*M* = 19.97, *SD* = 17.004) in accordance with the original instruction as the No-Feedback group (*M* = 18.267, *SD* = 14.90). An independent *t*-test confirmed that this difference was not significant, *t*(58) = .41, *p* = .68.

Contingency sensitivity was defined as a pattern that consisted of at least three consecutive responses that were not in accordance with the original instruction, and at least one of these must accord with the reversed contingency. In principle, therefore, a participant could stop following the instruction and choose instead the stimulus that lost points (i.e., the stimulus that was "mid-way" between most like and least like the sample), but could only do this for two of the three responses. Including this requirement ensured that the term "contingency sensitivity" was appropriate, given that a participant must obtain at least 1 point when they ceased rule-following. However, a post hoc analysis of the data at the individual participant level indicated that all participants selected the most similar comparison across all three responses (gaining 3 points), hence showing contingency sensitivity.

Mean contingency sensitivity scores are presented in Fig. 5 (right panel) and show a small difference between the conditions. Specifically, the Feedback group completed more trials (*M* = 18.33, *SD* = 16.63) before responding in accordance with the new contingencies than the No-Feedback group (*M* = 14.40, *SD* = 11.29). However, an independent *t*-test indicated that this difference was not significant, *t*(58) = 1.06, *p* = .29.

Rule resurgence attempted to capture responding that was consistent with the initial rule (i.e., percentage of responses), but that occurred after a participant had emitted three consecutive responses that were in accordance with the reversed contingencies (hence the term "resurgence"). This measure was designed to supplement contingency sensitivity, which fails to capture when responding reverted in this way to the original rule-consistent pattern of responding. In order to illustrate individual resurgence most clearly, Fig. 6 (left-hand side) presents the density and range of participant data in each condition. The data indicate modest levels of resurgence in both conditions, with some suggestion that resurgence may have been greater in the No-Feedback group. Given that the data were severely skewed, a Mann-Whitney U-test was employed, and the difference proved not to be significant (No-Feedback condition: *Md* = 7.44, Feedback condition: *Md* = 5.53, *U* = 404, *z* = -.680, *p* = .50, *r* = .08).

When analyzing the individual resurgence data, we noted that participants who did not abandon the original rule (and thus did not at any point emit three consecutive responses in a contingency-sensitive direction) could not by definition demonstrate resurgence. The right-hand panel of Fig. 6 thus contains the data from only those participants who could in principle show resurgence (*N*=25 Feedback and *N*=28 No Feedback). When the data were re-analyzed with a Mann-Whitney U-test, the analysis confirmed once again that the difference was not significant (No Feedback: *Md* = 8.95, Feedback: *Md* = 7.75, *U* = 349, z = -.018, p = .99, *r* = .002).

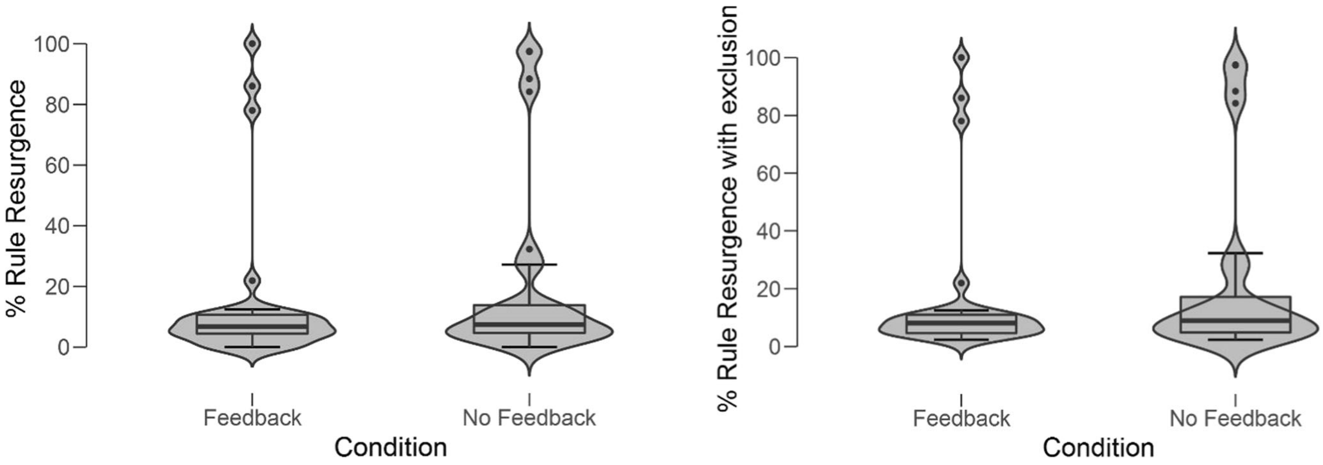

Fig. 6 Experiment 1: Box plots with a violin element illustrating the distribution and density of participant rule resurgence scores with all participants included (**left panel**) and excluding participants who did not demonstrate contingency sensitive responding (**right panel**) for the Feedback and No-Feedback conditions

Correlations

Given that the conditions did not differ significantly on any measure of rule-persistence, correlational analyses were conducted with the data collapsed across groups. Out of a possible 24 correlations among the three measures of rule-persistence and the eight self-report measures, only one reached significance. That is, contingency sensitivity positively correlated with the PRFS (r =.217, p = .04), such that participants who reported higher levels of rule-following were, perhaps counterintuitively, less likely to persist with rule-following on the MTS task. Given that only one relatively weak correlation was significant, this result should be interpreted with extreme caution.

Summary

The current work, as far as we are aware, was the first to investigate the potential impact of coherence on persistent-rule-following involving derived relations, by manipulating the presence versus absence of feedback during repeated exposures to the baseline training. The findings from Experiment 1 suggested that the presence or absence of feedback did not differentially influence rule compliance, contingency sensitivity, or rule resurgence. In effect, increasing relational coherence with the use of feedback (for the baseline relations) appeared to have virtually no impact on persistent derived rule-following. On balance, Harte et al. (2017, 2018) argued (in their *Discussion* sections) that a combination of high coherence and low derivation might facilitate more persistence in rule-following, but they did not manipulate coherence directly. In principle, in Experiment 1 of the current study level of derivation for the untrained A-C relations could be defined as relatively high because participants were not required to derive these relations until they entered the MTS rule-following task. In Experiment 2, participants were required to derive the A-C relations *before* entering the MTS task, thus reducing levels of derivation relative to Experiment 1.

Experiment 2

In Experiment 2, we again manipulated coherence through the presence versus absence of differential feedback, but the manipulation was now applied to a Training IRAP that required participants to derive the A-C relations. In effect, we sought to reduce levels of derivation relative to Experiment 1, while manipulating coherence for the derived relations through the presence versus absence of feedback.

Participants

A total of 115 individuals participated in Experiment 2, 95 females and 20 males. Participants ranged in age from 18 to 42 years (M = 20.39, SD = 3.09), and were recruited through random convenience sampling from the online participant system at Ghent University. All participants spoke Dutch as their first language. Twenty participants (ten in each condition, see below) received course credit for participation, while the remaining participants were paid 10 euro for participation. All participants were randomly assigned to one of two conditions, again referred to as Feedback and No Feedback. The data from 25 participants (12 from Feedback and 13 from No Feedback) were excluded, because they failed to meet either the IRAP performance criteria or the MTS task criteria (leaving N = 90 for analysis, 45 in Feedback and 45 in No Feedback). Initially, we collected data from 30 participants in each condition, after which a strong trend toward significance emerged on the rule-resurgence measure (details provided below). At this point, we made an *a priori* decision to collect data from an additional 15 participants in each condition in order to determine if the trend continued to a significant level. It is important to note, however, that we did not continue to add participants to each condition until we reached significance. Rather, given the strong trend observed with 30 participants, and the lack of previous work to draw upon in order to conduct a power analysis, a set number of 15 additional participants in each condition was decided upon, and analyses were only conducted once this final dataset was complete.

Setting

The setting was identical to that in Experiment 1.

Materials and apparatus

Experiment 2 again involved three computer-based tasks (a Derivation Pre-Training task, the Training IRAP, and an MTS task; the Derivation Pre-Training task and MTS task were identical to Experiment 1), as well as five self-report measures. The only difference in materials between Experiments 1 and 2 was in the configuration of the stimuli presented in the final phase of the Training IRAP.

The Training IRAPs The Training IRAPs again employed stimuli from Sets 7 and 8, but now the A-C relations, which could be derived from the mixed A-B and B-C Training IRAPs, were presented. Specifically, each trial presented the stimulus "Least Similar" or "Most Similar" as labels, with the novel words "Beda" and "Sarua" as targets, along with the same response options as before. Taken together, the four A-C trial types were as follows: Least Similar-Beda;

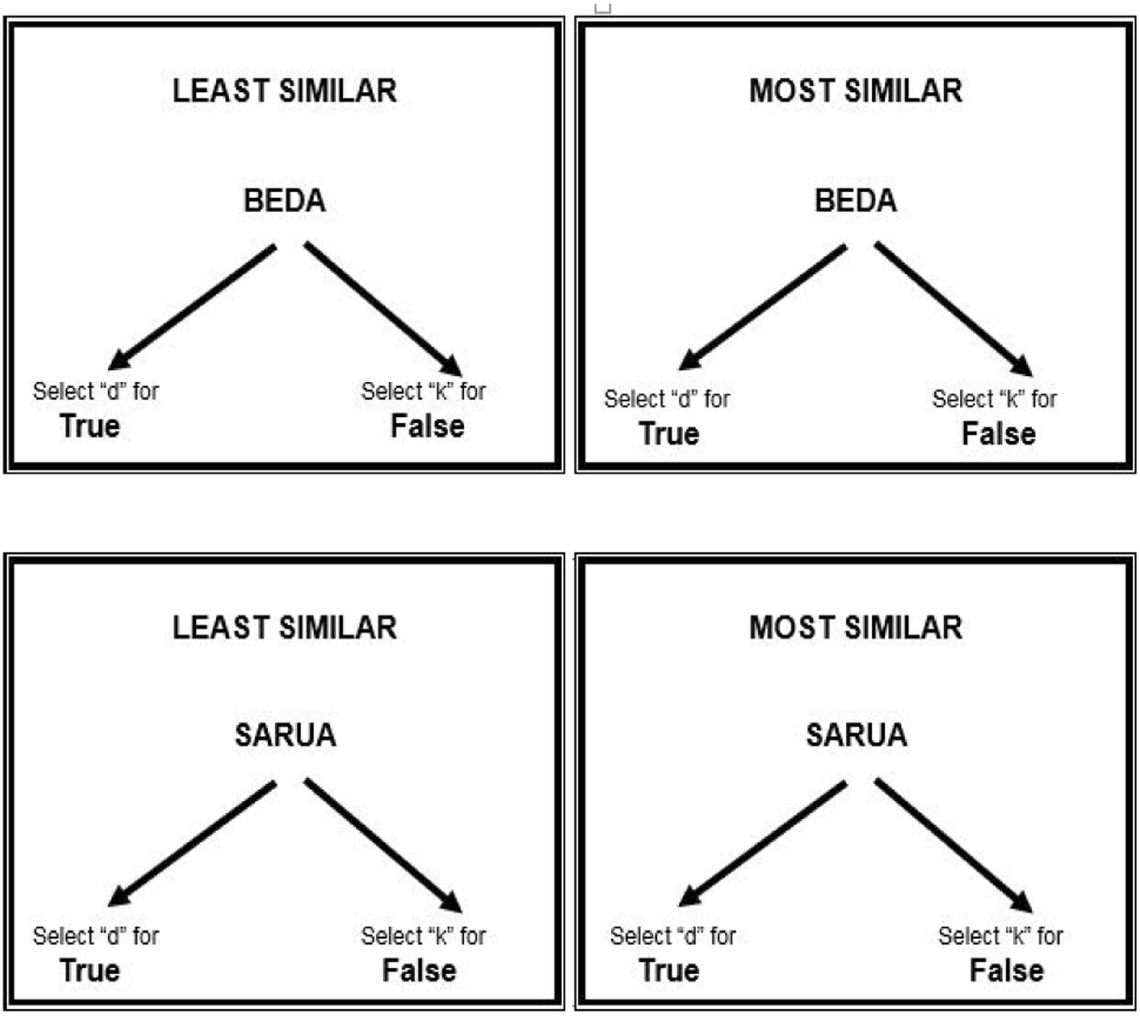

Fig. 7 Diagrammatic representation of the four IRAP trial-types that appear in the A-C relation test blocks. Arrows did not appear on-screen. The four IRAP trial-types were denoted as: *Least Similar-Beda, Most Similar-Beda, Least Similar-Sarua,* and *Most Similar-Sarua*

Most Similar-Beda; Least Similar-Sarua; and Most Similar-Sarua (see Fig. 7).

Procedure

The only procedural difference between Experiments 1 and 2 concerned the addition of trials presenting A-C relations in the Training IRAP.

Stage 3: Training IRAPs Phases 1, 2, and 3 of the Training IRAPs in Experiment 2 were identical to those in Experiment 1. Phase 4 in Experiment 2 differed from Experiment 1, however, in that it no longer presented two blocks of the mixed A-B and B-C relations from Phase 3, but now replaced these with two blocks of previously untrained A-C relations. As in Experiment 1, coherence was again manipulated via feedback such that participants in the Feedback condition received feedback on every A-C trial and at the end of each block, while participants in the No-Feedback condition received no feedback on these trials, or at the end of each block. All participants were advised that during this stage some of the stimuli that they had seen previously would be presented again, but in combinations that they had not seen before. Participants were also explicitly instructed not to worry about speed of responding (because the target relations were novel) but to focus on accuracy (i.e., "Because this block will involve presenting the stimuli you have learned about before in combinations that you *haven't* seen them in before, don't worry about speed. Instead, take your time, and focus on getting them all right").

Results and discussion

The same accuracy criterion (≥75%), as applied in Experiment 1 to the Training IRAP, resulted in data from eight participants being removed (five in Feedback and three in No Feedback, $N = 107$ remaining). Unlike Experiment 1, the response latency criterion was *not* applied in Experiment 2, again because the target relations were novel (i.e., not preceded with direct training). The same MTS accuracy criterion applied again, resulting in the data from 17 participants being removed (eight in Feedback and nine in No Feedback, $N = 90$ remaining).

IRAP data

Again, prior to conducting the primary analyses, we compared the mean number of blocks required in each condition in Stages 1–3 of the Training IRAP. On the A-B relations, Feedback participants required a mean of 2.20 blocks (SD = 1.46), while No-Feedback participants required 2.04 (SD = 1.17). On the B-C relations, Feedback participants required 1.69 blocks (SD = .82), while No-Feedback participants required 1.40 (SD = .65). Finally, on the mixed A-B and B-C relations, Feedback participants required 1.47 blocks (SD = 1.04), while No Feedback participants required 1.27 blocks (SD = .50). Overall, therefore, the mean number of blocks required was 5.36 (SD = 2.13) for the Feedback group and 4.71 (SD = 1.62) for the No-Feedback group. Independent t-tests confirmed that none of these differences were significant (all p's > .07, without correction). Thus, any subsequent differences that emerged between the two groups during the Training IRAP or the MTS task would not likely be due to differences in the ability to learn how to respond on the IRAP per se.

Measures of rule persistence

Rule compliance scores are presented in Fig. 8 (left panel) and showed only a small difference between the conditions (Feedback: M = 18.40, SD = 14.99, No Feedback: M = 15.56, SD = 14.43). An independent t-test confirmed that this effect was non-significant $t(88) = .917$, $p = .36$.

Contingency sensitivity scores are presented in Fig. 8 (right panel) and again show only a small difference between conditions (Feedback, M = 14.64, SD = 11.48; No Feedback, M = 16.09, SD = 13.98). An independent t-test again confirmed that this difference was not significant, $t(88) = -5.36$, $p = .59$.

Figure 9 (left-hand side) presents differential levels of *rule resurgence* among all participants in each condition (i.e., there were no exclusions made on the basis of absence of contingency sensitivity). The data show modest resurgence and a greater range in participant-resurgence scores in the Feedback condition but not in the No-Feedback condition. For example, a cluster of participants resurged for over 90% of responses in the Feedback condition, while no participants resurged to this degree in the No-Feedback condition. Given that the data were once again severely skewed, a Mann-Whitney U-test was employed, which confirmed that the difference was significant (Feedback, Md = 5.41, No Feedback, Md = 4.65, U = 723.5, z = -2.332, p = .02, r = .25).

Once again, in order to more closely examine only those participants who had the opportunity to resurge (i.e., those participants who demonstrated contingency-sensitive responding for three consecutive responses after the contingency reversal), we included only those datasets in right-hand panel of Fig. 9 (N=42 Feedback and N=39 No Feedback). Once again, there was modest resurgence in the Feedback group but not in the No-Feedback group. When the data were re-analyzed with a Mann-Whitney U-test, the difference remained significant (Feedback, Md = 6.43, No Feedback, Md = 5.00, U = 594.50, z = -2.122, p = .03, r = .24).

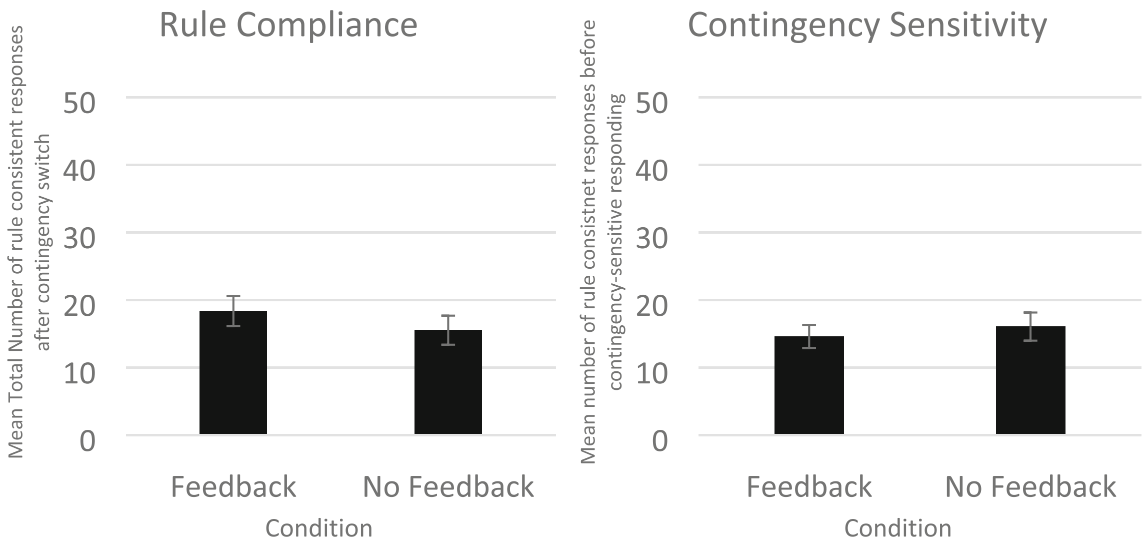

Fig. 8 Experiment 2: Mean rule compliance scores (**left panel**) and contingency sensitivity scores (**right panel**), with standard error bars for the Feedback and No-Feedback conditions

Correlations

Given that the conditions did not differ significantly on the rule compliance and contingency sensitivity measures, correlational analyses were conducted with the data collapsed across groups among these two measures of rule persistence and the self-report scales. Out of a possible 16 correlations, only one reached significance: contingency sensitivity positively correlated with the PRFS (*r* = .213, *p* = .04), such that participants who reported a low propensity for rule-following were more likely to persist on the MTS task. Given the significant group differences recorded on the rule resurgence measure, separate correlational analyses were conducted for each condition between resurgence and the self-report scales. Two separate sets of analyses were conducted, the first involving all participants (in each condition), the second excluding participants (from each condition) who did not demonstrate contingency-sensitive responding. In the first analysis, one correlation was significant, in the No-Feedback condition. Specifically, rule resurgence positively correlated with the GPQ (*r* = .40, *p* = .006), such that greater rule resurgence predicted higher compliance. In the second analysis, the same correlation remained significant (*r* = .37, *p* = .02). No other correlations were significant (all other *ps* > .08).

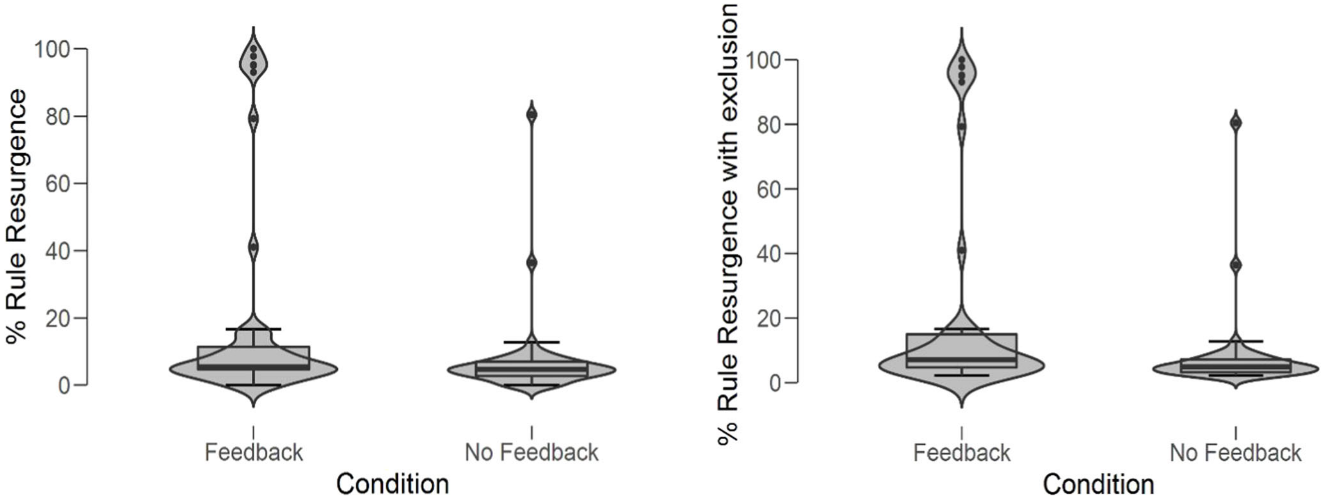

Fig. 9 Experiment 2: Box plots with a violin element illustrating the distribution and density of participant rule resurgence scores with all participants included (**left panel**) and excluding participants who did not demonstrate contingency sensitive responding (**right panel**) for the Feedback and No-Feedback conditions

Summary

The findings from Experiment 2 suggested that manipulating the presence or absence of feedback on the novel derived A-C target relation differentially influenced participant rule persistence, but only on one specific measure. While there seemed to be no impact of the feedback manipulation on the rule compliance or contingency sensitivity measures, significant differential responding did emerge on the rule resurgence measure. Specifically, participants in the Feedback condition resurged for a significantly greater percentage of rule-consistent responses than did participants in the No-Feedback condition. In effect, increasing relational coherence with the use of feedback (on the untrained, derived A-C relation) appeared to impact persistent derived rule-following in terms of participant rule resurgence. Additionally, rule resurgence was correlated positively with scores on the GPQ, but only in the No-Feedback condition. That is, when participants who scored high in self-reported compliance did *not* receive feedback on the derived A-C relation in the Training IRAP, they were more likely to resurge on the MTS task.

General discussion

The current study sought to extend the work reported by Harte et al. (2017, 2018), which had targeted the impact of derivation on persistent rule-following. Specifically, Harte et al. (2017) had shown greater levels of persistence for direct rules (which were assumed to be low in derivation) versus high-derivation rules (i.e., rules that contained a derived relation established within the experiment). In Harte et al. (2018), levels of derivation were manipulated directly within the experiment, and once again low levels of derivation produced greater persistence in rule-following. This effect for derivation was shown for both mutually (Experiment 1) and combinatorially (Experiment 2) entailed relations. In the context of the experimental work reported by these authors, it was noted that coherence may also play an important role in persistent rule-following. For example, Experiment 1 of Harte et al. (2017) did not find a difference in persistent rule-following between direct and derived rules when coherence between the rule and the reinforcement contingencies was low (i.e., participants were only given ten opportunities to follow the rule before the contingencies switched). In contrast, Experiment 2 showed clear differences between direct and derived rules when participants were given 100 opportunities to follow the rule before the contingency switch.

In the research described above, coherence was used in a post hoc manner to interpret unexpected results, and thus it seemed important to examine the impact of coherence directly within the context of studying persistent rule-following. In the current research, we attempted to manipulate coherence through the provision or non-provision of feedback for the trained relations (Experiment 1) or the derived relations (Experiment 2), that were then inserted into the rule for responding on the MTS task. The main question we asked was would a condition involving higher levels of coherence/more corrective feedback, for trained and/or derived relations, produce more or less persistence in rule-following?

In Experiment 1, the results indicated that coherence/feedback had no impact on rule persistence across any of the three measures (rule compliance, contingency sensitivity, or rule resurgence). In Experiment 2, however, feedback significantly impacted upon rule resurgence. That is, participants in the Feedback group resurged back to the original rule for significantly more responses after demonstrating contingency-sensitive responding than did the No-Feedback group. This is the first time that we have seen an effect emerge on the rule resurgence measure in the absence of a significant effect on either rule compliance or contingency sensitivity. Critically, the resurgence effect indicates that participants have successfully identified a change in task contingencies, but then return to and persist with a rule that "no longer works" (i.e., to earn points). More informally, participants clearly know that continuing to follow the rule is costing them points, but they persist with "dysfunctional" rule-following. In this sense, this effect for coherence on rule resurgence may have important implications for the often-cited role of excessive rule-following in process-based accounts of psychological distress (e.g., Hayes, Strosahl, & Wilson, 1999; Zettle & Hayes, 1982).

As noted above, the resurgence effect was observed in Experiment 2, but not in Experiment 1. Whilst an explanation for this disparity must remain speculative, it seems valuable to reflect upon the possible variables involved. Specifically, it appears that coherence in terms of the presence versus absence of feedback in the context of the derived A-C relations (Experiment 2) has an impact (on resurgence) that it does not have in the context of the trained A-B and B-C relations (Experiment 1). If we view this result through the lens of the MDML framework, the coherence of the target A-C relation in Experiment 1 (required for accurate responding on the MTS task) could be considered quite low, even in the Feedback condition, because this relation was not presented during the Training IRAP. In addition, level of derivation for the A-C relation could be considered quite high (i.e., novel and emergent), again because it was not presented during the Training IRAP.

In Experiment 2, however, the coherence for the A-C relation was higher relative to Experiment 1, particularly for the Feedback group (assuming that feedback increases coherence). The level of derivation for the A-C relation could also be considered relatively low, given that participants had many opportunities (unlike Experiment 1) to derive it within the Training IRAP. The critical point here is that the dynamic interaction between coherence and derivation may influence rule persistence in the face of reversed reinforcement contingencies (see Harte et al. 2017, 2018, for similar arguments). However, the nature of this dynamic relationship appears to be complex, in that coherence, as manipulated by the presence versus absence of feedback, only has an impact when it applies to the derived A-C relation (Experiment 2), rather than to the trained A-B and B-C relations (Experiment 1).

Indeed, the complexity is compounded by the fact that differential persistent rule-following was observed only for the resurgence measure in the current study (in Harte et al., 2018, effects were observed for both rule compliance and contingency sensitivity, but not for rule resurgence). If nothing else, this highlights that the functional definition of persistent rule-following is not a trivial matter, and a more complete understanding of this phenomenon will involve explaining why it is observed on one measure in one context but not on another measure in a different context. Once again, the MDML framework may be useful here. Specifically, in Harte et al. (2017, 2018) derivation was argued to be the variable that was manipulated directly, whereas in the current study coherence was manipulated directly (i.e., within experiments). Perhaps, therefore, resurgence, at least in some contexts, is particularly sensitive to shifts in the dimension of coherence. Intuitively this makes sense, if the technical concept of coherence is interpreted as functionally similar to "truth," "correctness," "veracity," or "doing what is right." That is, resurgence as a measure involved participants clearly contacting the reversed reinforcement contingencies and then returning to a pattern of responding that was repeatedly punished by loss of points. Doing so makes sense if the rule controlling that resurgence was deemed to be high in "truth value," "correctness," "veracity," and imbued with a sense of "being the right thing to do" even if the current contingencies are telling you the opposite.[2]

At this point it is worth acknowledging that a possible criticism of the current study is that the resurgence effect could be seen as not particularly strong. On balance, the difference in effect sizes for resurgence between Experiments 1 and 2 seems far from trivial (i.e., .002–.08 vs. .24–.25). That is, the latter effect size was over three times the size of the former. When examining Fig. 9 versus 6, the difference in effect sizes appears to be driven by the fact that a small sub-group of Feedback participants in Experiment 2 approached 100% resurgence responses versus none in the No-Feedback group. In contrast in Experiment 1, both conditions produced only one participant that approached 100% (with two others in each condition above 80%). It appears then that the Feedback in Experiment 2 impacted dramatically on a number of participants by leading them to persist with rule-following for almost the entire session despite having clearly contacted a reinforcement contingency for not doing so. At the current time it remains unclear why only some participants produced this highly persistent pattern. One possible reason is that these participants were particularly sensitive to the increased level of coherence generated by the performance feedback during the A-C test trials. More informally, when presented with a choice to respond in accordance with the feedback presented during the A-C test or the feedback presented during the schedule, these participants simply opted for the former. If relative differences in coherence produced by the feedback during the A-C test versus the schedule is important, then perhaps a future study could test this by attempting to manipulate coherence in some other way. For example, additional verbal feedback could be used to "supercharge" the coherence functions of the A-C feedback. One way to do this might be to use experimenters with potentially high versus low levels of social credibility – such as wearing a white lab-coat and using a clipboard in contrast to a "scruffy" student in jeans and t-shirt – to deliver informal feedback after A-C testing. Would we observe even greater evidence of persistent rule-following in the supercharged coherence condition?

In grappling with the complexities involved in understanding persistent rule-following, it is also worth noting that the number of significant correlations was extremely limited; between the PRFS and contingency sensitivity in Experiments 1 and 2, and between the GPQ and rule-resurgence in Experiment 2 (the latter was restricted to the No-Feedback condition). Given the large number of correlations that were calculated across the two experiments (a total of 64), interpreting these three significant effects should be done with extreme caution, particularly for the PRFS because it is not a standardized measure. On balance, the GPQ is a standardized measure and the observed correlation makes sense intuitively in that higher levels of self-reported pliance predicted increased rule resurgence. Nevertheless, it is interesting that the GPQ correlation was only observed for the No-Feedback condition. Insofar as this correlation is robust, it appears that

[2] It is important to emphasize that the dimensions of coherence and derivation as conceptualized within the MDML are not defined as entirely separable. Indeed, Barnes-Holmes et al. (2017) highlighted that the boundaries among the dimensions and levels within the MDML were "fuzzy" (p.14). Furthermore, the primary focus of the framework was to emphasize the dynamics involved in the various properties of arbitrarily applicable relational responding. The critical point, therefore, is that increases in one dimension may be seen as involving decreases in a second dimension. For example, attempting to increase coherence by providing performance feedback on a block of A-C trials would likely reduce level of derivation simply because responding on those trials itself involves deriving. Nonetheless, providing feedback versus no feedback may be one way in which it is possible to manipulate coherence directly while recognizing that derivation may also be impacted. Although the inherently dynamic and inseparable nature of the units specified within the MDML might be seen as a weakness, it is one shared with many concepts in behavior analysis (e.g., the relationship between the eliciting and reinforcing functions of a stimulus). Ultimately, of course, such distinctions either stand or fall based on their utility within the basic science and its application. And in the context of the current study we have indeed found the distinction between coherence and derivation to be useful.

the relationship between a self-reported tendency to engage in high levels of rule-following and actual rule-following in an experimental task may be moderated by the level of coherence involved in that task. More specifically, when coherence is relatively low for a derived rule (because it has not been reinforced with feedback) self-reported pliance is more likely to predict performance because participants are less certain about the truth or veracity of the rule. Future research will be needed to test this conclusion.

A potential issue worth noting in the current study is the relatively high level of attrition observed in Experiment 1 (38/98 participants), although this was much reduced in Experiment 2. The exact reason behind the high attrition rate in Experiment 1 remains unclear at the current time, but it may have been due to increased ambiguity with regard to the meaning of Beda relative to Experiment 2 (i.e., in Experiment 1, up until the MTS task, participants had only paired A-B and B-C, but never A-C directly). In any case, levels of attrition in both experiments were more or less equal across groups, and thus the within-experiment difference in rule resurgence observed in Experiment 2 was unlikely due to a difference in attrition. Nonetheless, future research could consider implementing, for example, performance-dependent payment, more trials, or longer inter-trial intervals in an attempt to reduce or circumvent similar levels of attrition.

A related issue that is highlighted by the attrition observed in the current study is the importance of employing such strict accuracy criteria. In the current context, there was a clear need to ensure that all participants entering the MTS task did not learn to perform through trial and error. The primary purpose of these criteria, therefore, was to ensure that participants were instead responding in accordance with the necessary derivations (e.g., Beda means Least Similar) when completing the task. Interestingly, a recent study by Kissi et al. (2019) presented results that highlighted how unlikely trial and error learning on the MTS task was in the current study. Employing a similar paradigm, Kissi et al. found that, when no instructions for responding were provided, all but one participant spontaneously chose the "Most Similar" comparison stimulus on the first trial on the MTS task. In contrast, in the current study out of a total of 120 participants, 82% chose the correct "Least Similar" comparison on the first trial and 89.33% within the first two trials. Clearly, therefore, the seemingly default "Most Similar" response that participants tended towards in the Kissi et al. study was far less evident in the context of a derived rule that specified the "Least Similar" comparison as the correct stimulus.

The current research appears to constitute the first attempt to examine the impact of coherence in the derived relations contained within a rule on persistent rule-following in the face of reversed reinforcement contingencies. The results highlight what appear to be relatively subtle and complex effects. For example, no significant differences between high and low coherence were obtained in Experiment 1, when the trained baseline relations were targeted, but an interesting difference did emerge in Experiment 2 (for the resurgence measure) when the derived relations were targeted. Adopting the current research strategy thus extends beyond simply demonstrating persistent rule following *per se,* or searching for evidence that it is more or less likely in specific populations. Rather, the type of research reported here involved attempting to identify variables that increase or decrease persistent rule-following in the face of reversed reinforcement contingencies. Naturally, many questions remain to be answered in light of the findings reported here, and of course the terms and concepts we have employed in presenting the current work, such as levels of coherence and derivation, could always be questioned. Nevertheless, the effects we found, and failed to find, will need be to be explained if we are to develop a more complete understanding of persistent rule-following.

In closing, we do recognise that some of the findings and conclusions reported in the current research are relatively tentative. Nevertheless, there is much that is novel in the current work, particularly the attempt to explore the dimension of coherence as a property of arbitrarily applicable relational responding in the context of persistent rule-following. Although preliminary, it seems important to share the method and findings reported here with the wider scientific community in the hope that other research groups will be encouraged to pursue similar lines of enquiry so that we may better understand what appears to be an important feature of human behavior.

Author Note This article was prepared with the support of an Odysseus Group 1 grant awarded to the second author by the Flanders Science Foundation (FWO) at the Department of Experimental, Clinical and Health Psychology, Ghent University, Belgium.

Acknowledgements This article was prepared with the support of an Odysseus Group 1 grant awarded to the second author by the Flanders Science Foundation (FWO), Belgium. None of the experiments were preregistered and enquiries about the materials or data may be made by direct correspondence with the first author.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From IRAP and REC model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable relational responding. *Journal of Contextual Behavioral Science, 6*(4), 473-483. doi: https://doi.org/10.1016/j.jcbs.2017.08.001

Barnes-Holmes, D., Finn, M., McEnteggart, C., & Barnes-Holmes, Y. (2018). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *Perspectives* *on Behavioral Science (Special issue on Derived Relations), 41*(1), 155-173. doi: https://doi.org/10.1007/s40614-017-0124-7

Baruch, D.E., Kanter, J.W., Busch, A.M., Richardson, J.V., & Barnes-Holmes, D. (2007). The differential effect of instructions on dysphoric and nondysphoric persons. *The Psychological Record, 57*, 543-554. doi: https://doi.org/10.1007/BF03395594

Bentall, R.P., Lowe, C.F., & Beasty, A. (1985). The role of verbal behavior in human learning: II. Developmental differences. *Journal of the Experimental Analysis of Behavior, 43*, 165-181. doi: https://doi.org/10.1901/jeab.1985.43-165

Bond, F.W., Lloyd, J., Barnes-Holmes, Y., Torneke, N., Luciano, L., Barnes-Holmes, D., & Guenole, N. (2017). A new measure of psychological flexibility based on RFT. Symposium at the Association for Contextual Behavioural Science World Conference 15, 22-25 June 2017, Seville, Spain.

Bond, F., Hayes, S., Baer, R., Carpenter, K., Guenole, N., Orcutt, H., … Zettle, R. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire-II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy*, *42*(4), 676–88. doi: https://doi.org/10.1016/j.beth.2011.03.007

Catania, C. (1979). Learning. Englewood Cliffs: Prentice Hall.

Catania, A.C., Shimoff, E., & Matthews, B.A. (1989). An experimental analysis of rule-governed behavior. In S.C. Hayes (Ed.), Rule-governed behaviour: Cognition, contingencies, and instructional control (pp. 119-150). New York: Plenum.

Dougher, M., Twohig, M.P., & Madden, G.J. (2014). Stimulus-stimulus relations [Special issue]. *Journal of the Experimental Analysis of Behavior*, 101(1).

Harte, C., Barnes-Holmes, Y., Barnes-Holmes, D., & McEnteggart, C. (2017). Persistent rule-following in the face of reversed reinforcement contingencies: The differential impact of direct versus derived rules. *Behavior Modification, 41*(6), 743-763. doi: https://doi.org/10.1177/0145445517715871.

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2018). The impact of high versus low levels of derivation for mutually and combinatorially entailed relations on persistent rule-following. *Behavioural Processes, 157,* 36-46. doi: https://doi.org/10.1016/j.beproc.2018.08.005.

Hayes, S.C. (1989). Rule-governed behavior: Cognition, contingencies, and instructional control. New York: Plenum.

Hayes, S.C., Barnes-Holmes, D., & Roche, B. (2001). Relational frame theory: A post-Skinnerian account of human language and cognition*.* New York: Plenum.

Hayes, S.C., Brownstein, A.J., Haas, J.R., & Greenway, D.E. (1986). Instructions, Multiple Schedules, and Extinction: Distinguishing Rule-Governed from Scheduled-Controled Behavior. *Journal of the Experimental Analysis of Behavior, 46*(2), 137-147. doi: https://doi.org/10.1901/jeab.1986.46-137

Hayes, S.C. & Hayes, L.J. (1989). The verbal action of the listener as a basis for rule-governance. In S.C. Hayes (Ed.), *Rule-governed behavior: Cognition, contingencies, and instructional control* (pp. 153-190).

Hayes, S. C., Strosahl, K., & Wilson, K.G. (1999). Acceptance and Commitment Therapy: An experiential approach to behaviour change*.* New York: Guilford Press.

Hughes, S. & Barnes-Holmes, D. (2016). Relational Frame Theory: The basic account. In R. D. Zettle, S.C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), The Wiley handbook of contextual behavioral science (pp. 129-178). West Sussex, UK: Wiley.

LeFrancois, J. R., Chase, P. N., & Joyce, J. H. (1988). The effects of a variety of instructions on human fixed-interval performance. *Journal of the Experimental of Behavior, 49*(3), 383-393. doi: https://doi.org/10.1901/jeab.1988.49-383

Lovibond, S. H., & Lovibond, P. F. (1995). Manual for the Depression Anxiety Stress Scales (2nd ed.). Sydney: The Psychology Foundation of Australia.

McAuliffe, D., Hughes, S., & Barnes-Holmes, D. (2014). The dark-side of rule governed behavior: An experimental analysis of problematic rule-following in an adolescent population with depressive symptomatology. *Behavior Modification, 38*(4), 587-613. doi: https://doi.org/10.1177/0145445514521630

Michael, J. (1980). Flight from behavior analysis presidential address ABA 1980. *The Behavior Analyst, 3*, 1-22. doi: https://doi.org/10.1007/BF03391838

Monestes, J.L., Greville, W.J., & Hooper, N. (2017). Derived insensitivity: Rule-based to contingencies propagates through equivalence. *Learning and Motivation, 59*, 55-63. doi: https://doi.org/10.1007/s40732-014-0029-8

O'Hora, D., Barnes-Holmes, D., Roche, B., & Smeets, P.M. (2004). Derived relational networks and control by novel instructions: A possible model of generative verbal responding. *The Psychological Record, 54*, 437-460. doi: https://doi.org/10.1007/BF03395484

O'Hora, D., Barnes-Holmes, D., & Stewart, I. (2014). Antecedent and consequential control of derived instruction-following. *Journal of the Experimental Analysis of Behavior, 102* (1), 66-85. doi: https://doi.org/10.1002/jeab.95

Rosenfarb, I.S., Newland, M.C., Brannon, S.E., & Howey, D.S. (1992). Effects of self-generated rules on the development of schedule-controlled behaviour. *Journal of the Experimental Analysis of Behavior, 58*(1), 107-121. doi: https://doi.org/10.1037/0021-843X.102.4.642

Ruiz, F. J., Suárez-Falcón, J. C., Barbero-Rubio, A., & Flórez, C. L. (2019). Development and initial validation of the generalized pliance questionnaire. *Journal of Contextual Behavioral Science, 12*, 189-198. doi: https://doi.org/10.1016/j.jcbs.2018.03.003

Shimoff, E., Catania, A.C., & Matthews, B.A. (1981). Uninstructed human responding: Sensitivity of low-rate performance to schedule contingencies. *Journal of the Experimental Analysis of Behavior, 36*(2), 207-220. doi: https://doi.org/10.1901/jeab.1981.36-207

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech, Language, and Hearing Research, 14*, 5-13. doi: https://doi.org/10.1044/jshr.1401.05

Sidman, M. (1994). Equivalence relations and behaviour: A research story. Boston, MA: Authors Cooperative.

Sidman, M. & Tailby, W. (1982). Conditional discrimination vs. matching to sample: an expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37*(1), 5-22. doi: https://doi.org/10.1901/jeab.1982.37-5

Skinner, B.F. (1966). An operant analysis of problem solving. In B. Keinmuntz (Eds.), Problem-solving: Research, method, and therapy (pp. 225-257). New York: Wiley.

Steele, D.L. & Hayes, S.C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior, 56,* 519-555. doi: https://doi.org/10.1901/jeab.1991.56-519

Zettle, R.D. & Hayes, S.C. (1982). Rule-governed behavior: A potential theoretical framework for cognitive-behavior therapy. In P.C. Kendall (Ed.), Advances in cognitive-behavioral research and therapy (Vol. 1: pp. 73-118). New York: Academic.