Exploring the Impact of Coherence (Through the Presence versus Absence of Feedback) and Levels of Derivation on Persistent Rule-following

Colin Harte[1], Dermot Barnes-Holmes[1,2], Yvonne Barnes-Holmes[1], and Ciara McEnteggart[1]

[1]Department of Experimental, Clinical and Health Psychology, Ghent University, Ghent, Belgium

[2]School of Psychology, Ulster University, Coleraine, Northern Ireland, UK

Corresponding Author:

Colin Harte

Department of Experimental, Clinical, and Health Psychology

Ghent University

Henri Dunantlaan, 2

9000 Ghent

Belgium

Email: Colin.Harte@UGent.be

Authors' 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). Correspondence concerning this article should be sent to Colin.Harte@UGent.be

Abstract

Recent developments in 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 begun to explore the impact of a number of these variables on persistent rule-following, namely, levels of derivation and coherence. However, no research to date has systematically examined the impact of coherence on persistent rule-following at varying levels of derivation. Across two experiments, the impact of coherence (manipulated through the systematic use of performance feedback) was explored on persistent rule-following when derivation was relatively low (Exp. 1) and high (Exp. 2). A training protocol based on the implicit relational assessment procedure (IRAP) was used to establish novel combinatorially entailed relations that manipulated the feedback provided on the untrained, derived relations (A-C) for five blocks of trials in Experiment 1 and one block of trials in Experiment 2. One of these relations was then inserted into the rule for responding on a subsequent contingency-switching match-to-sample 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. These differences, and the subtle complexities that appear to be involved in persistent rule-following in the face of reversed reinforcement contingencies, are discussed.

KEYWORDS: DERIVATION; COHERENCE; PERSISTENT RULE-FOLLOWING; RFT; HDML

Within behavior analysis, two concepts that have been widely acknowledged as important in the study of human language and cognition are rule-governed behavior and derived stimulus relations. The former, rule-governed behavior, was first proposed by B.F. Skinner (1966) in the context of an operant account of human problem solving. Specifically, it was argued that rules specified reinforcement contingencies that had the potential to override the impact of direct contact with environmental contingencies. In this way, the listener could problem solve without having to directly contact reinforcement contingencies. For example, the simple rule "don't take sweets off strangers" given to a child by a parent allows the child to learn important safety skills without having to potentially experience the negative consequences of engaging in such behavior.

Throughout the 1970s and 1980s, a great deal of research sought to investigate the ways in which rules impacted human reinforcement schedule performance (see Hayes, 1989, for an early book-length treatment). An interesting and key phenomenon that emerged from this research was that rule-governed behavior in verbally-able humans often led to performances that did not adapt readily to task contingencies, but remained consistent with a rule or rules provided by an experimenter (e.g., Hayes, Brownstein, Haas, & Greenway, 1986; Shimoff, Catania, & Matthews, 1981). In very recent work, for example, participants were instructed to choose a comparison stimulus that differed most from a sample on a match-to-sample (MTS) task, and points were awarded for following this rule (Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2017). In general, all participants responded in accordance with the rule and the contingencies. However, when the contingencies changed, and points were awarded for choosing the stimulus that differed *least* from the sample, participants sometimes continued to follow the rule even though doing so now led to a loss of points. This tendency for humans to follow rules in the face of competing reinforcement contingencies has sometimes been referred to as rule-based 'insensitivity' or persistent rule-following.

Furthermore, excessive persistent rule-following has sometimes been highlighted as an important feature of human psychopathology (hereafter referred to as human psychological suffering; e.g., Hayes, Strosahl, & Wilson, 1999; Zettle & Hayes, 1982). The basic idea is that when humans persistently follow verbal rules rather than adapting to natural contingencies, this by definition undermines contextual sensitivity, which is associated with psychological suffering.

As noted above, the second concept widely acknowledged as important in the study of human language in behavior analysis is that of derived stimulus relations. The concept first emerged with the work of Sidman (1971), the basic phenomenon of which came to be referred to as stimulus equivalence. The main finding was that reinforcing a number of matching responses in human participants often readily produced a number of unreinforced matching responses. For example, if stimulus relations X-Y and X-Z were trained, derived Y-Z and Z-Y relations were also observed. When such a pattern of emergent and untrained responding occurred, the stimuli involved were said to form an *equivalence class* or *relation*. Furthermore, other unreinforced responses also emerged when a specific function was trained to a stimulus participating in this newly derived relation (e.g., if X, Y, and Z participate in an equivalence relation, and X is paired with a reinforcer, Z may then acquire reinforcing functions in the absence of direct pairing). Crucially, this phenomenon appeared to occur with relative ease in verbally-able humans, but was not readily or reliably observed in humans with severely limited language abilities or in nonhumans (see Sidman, 1994, for a book-length treatment). Indeed, the lack of evidence for even the most basic equivalence responding in nonhumans has persisted (see Dougher, Twohig, & Madden, 2014).

The extension of stimulus equivalence as an important explanatory tool for analyzing the complexities of human learning came with the development of relational frame theory (RFT), a behavior-analytic account of human language and cognition (Hayes, Barnes-Holmes,

& Roche, 2001; Steele & Hayes, 1991). For RFT, stimulus equivalence is but one class of generalized operant behavior, of which many others are possible. Specifically, RFT suggests that there are many generalized relational operants or patterns of relational responding, referred to as *relational frames* including: similarity, difference, opposition, distinction, temporality, hierarchy, and deictic (see Hughes & Barnes-Holmes, 2016, for an extensive review). The generic term or concept, *arbitrarily applicable relational responding* (AARR) is used to label these operant classes and their various and increasingly complex combinations or networks.

While the experimental analyses of rule-governed behavior and derived stimulus relations have only rarely overlapped, a strong conceptual link has long existed between the two. Indeed, some researchers have suggested that complex derived relational responding, involving networks of derived relations, may provide the basis for rule-governed behavior itself. For example, consider the simple instruction, "When the kettle whistles then take it off the hob." "Kettle", "whistle", "hob", and "take it off" all participate in equivalence relations with an actual kettle and hob, whistling sound, and an action, while the words "when" and "then" function as cues for the temporal relations operating among these events (i.e., whistling sound *before* taking the kettle off the hob). While this suggestion has been successfully modeled in the laboratory (O'Hora et al., 2004; O'Hora, Barnes-Holmes, & Stewart, 2014), empirical research linking these two areas in the context of persistent rule-following remains limited.

The first study that attempted to integrate these two areas explored the extent to which a rule that involved a novel derived relation would generate rule persistence on a contingency-switching MTS task (Harte et al., 2017). Specifically, participants were given either a direct rule that specified exactly how to respond (i.e., choose the stimulus that is least like the sample), a rule that involved a derived relation (i.e., an equivalence relation was first established between “least like” and the novel word “beda”, after which “beda” was inserted into the rule in the place of “least like”), or no rule. For the first 100 trials of the MTS task, points were awarded for choosing the least like comparison (i.e., following the rule) after which the contingencies reversed for a further 50 trials. While the direct rule produced most rule persistence, the rule that contained the derived relation produced more persistence than the no rule condition, thus providing the first evidence that rules that involved derived relations could override direct contingencies of reinforcement.

The study reported by Harte, et al. (2017) emerged in parallel with a new conceptual framework for analyzing the dynamics involved in AARR generally, known as the hyper-dimensional, multi-level (HDML) framework (Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2020). The framework focuses on five levels of relational development (i.e., mutual entailment; combinatorial entailment; relational networks; relating relations; relating relational networks) that intersect with four dimensions (i.e., coherence; complexity; derivation; flexibility). The details of the complete framework are beyond the scope of the present article, but specific features of the framework are relevant to research that followed on from the Harte, et al. work. For example, Harte, Barnes-Holmes, Barnes-Holmes, and McEnteggart (2018) explored the impact of derivation at different levels of relational development. Within the HDML framework, and indeed RFT generally, derivation refers, broadly speaking, to how often a particular derived relational response has been emitted in the past. The more a derived relational response is emitted, the less derived it becomes, because it acquires its own history that extends beyond the derivation that was made based on the ‘baseline’ relation. For example, imagine that an individual learns that A is smaller than B, and thus derives that B is bigger than A. The first time that the B>A relation is derived, it is derived ‘directly’ from the A<B baseline relation. However, if the individual subsequently continues to respond to B as bigger than A, that relational response gradually acquires its own

history, irrespective of whether or not it is directly reinforced, rendering it less and less derived from the original baseline relation (i.e., A smaller than B).

In the experiments reported by Harte et al. (2018), the opportunities to derive novel mutually entailed (i.e., Least like=Beda; Experiment 1) and combinatorially entailed (i.e., Least like=XXX=Beda; Experiment 2) relations were manipulated. That is, one group of participants had 120 opportunities (Low Derivation) to derive the critical relation, while the other group had only 8 opportunities (High Derivation). These relations were then inserted into the rule for responding on the same MTS task employed by Harte et al. (2017). Results showed that lower levels of derivation generally produced more persistence in rule-following than higher levels.

Subsequent research focused on the dimension of coherence (Harte et al., 2020). Within the HDML framework, coherence refers to the extent to which a particular pattern of relational responding is consistent (coherent) with previously established patterns. For example, if you are told that 'X is larger than Y,' the derived response that 'Y is smaller than X' would be deemed highly coherent because the contextual cues ("larger than" and "smaller than") participate in many other relational networks that have been reinforced, or at least not punished, by the wider verbal community (e.g., "trucks are generally larger than cars, so cars are generally smaller than trucks").

In the study conducted by Harte et al. (2020), coherence was manipulated through the presence versus absence of performance feedback. Combinatorially entailed relations were first established for all participants after which feedback was either provided or not provided on the trained relations (A=B and B=C; Experiment 1), and on the untrained, derived relations (A=C; Experiment 2). As with the previous studies in this line of research, one of these relations was then inserted into the rule for responding on the same contingency-switching MTS task to assess its impact on rule persistence. While no significant differences were found in Experiment 1, the provision or non-provision of feedback differentially impacted upon rule persistence in Experiment 2. Specifically, participants in the Feedback group showed more persistent rule-following than the No Feedback group, following the contingency reversal.

The primary purpose of the current study was to replicate and extend the research reported by Harte et al. (2020). Specifically, the current research attempted to manipulate both coherence and derivation in an effort to explore the dynamics among these two dimensions. The reader should note that a fundamental assumption of the HDML framework is that the dimensions themselves are inherently dynamical, and thus, research based on the framework will necessarily involve exploring these dynamics.[1]

Experiment 1 involved training participants on novel A-B and B-C relations followed by directly testing the novel A-C relations with and without feedback for five blocks (i.e., 160 trials). Experiment 2 partially replicated Experiment 1, but tested the novel A-C relations with and without feedback for only one block (i.e., 32 trials). Within Experiments 1 and 2, coherence was manipulated through the provision versus non-provision of performance feedback, while derivation was explored across experiments by manipulating opportunities to derive the A-C relation. A range of self-report measures of psychological distress were used to explore the extent to which derived rule-following may correlate with self-reported levels of distress. Two other self-report measures of rule-following were also employed to determine if they would predict actual persistent rule-following. Given the relatively inductive nature of the current research, we refrained from making formal predictions.

[1] An inherent assumption of the HDML framework is that changes in one dimension may involve changes in other dimensions. Thus, it is possible, for example, that as derivation reduces coherence may increase. Recognizing the dynamical nature of the dimension of AARRing does not mean, however, that experimental analyses cannot attempt to examine the differential impact of one dimension relative to another. It is entirely reasonable, therefore, to test the impact of coherence by manipulating feedback while recognizing that derivation *per se* is also likely influencing coherence. Ultimately, the utility of the units of analysis specified within the HDML framework will remain an empirical matter. For example, if derivation and coherence cannot be analyzed as separate dimensions, then no differences should be observed between conditions that attempt to separate these dimensions analytically.

Experiment 1

Participants

A total of 67 individuals participated in Experiment 1, 44 females and 23 males. They ranged in age from 18 to 47 years (*M* = 21.63, *SD* = 5.49) and were recruited through random convenience sampling from the online participant system at X University. Thirty-three participants were paid a fixed sum of 10 euros for participation, while 34 received course credit. Participants were randomly assigned to one of two conditions referred to as: Feedback and No Feedback. Paid and course-credit participants were distributed in a roughly equal manner between the two feedback conditions (17 paid and 18 course-credit in Feedback; 17 paid and 15 course-credit in No Feedback). The data from 7 participants (5 from Feedback [2 paid and 3 course-credit] and 2 from No Feedback [2 paid]) were excluded because they failed to meet specific performance criteria on either a Training IRAP or the MTS task (see below), leaving *N* = 60 for analysis, 30 in each condition. In general, 30 participants per condition yielded statistically significant effects (or approaching significance) in previously published studies that employed the type of procedures utilized here, and thus we decided to run 30 as a minimum.

Setting

The experiment was conducted in a cubicle at X University in which participants were seated in front of a standard Dell laptop. The experimenter was present at the beginning of each task to instruct participants, and also while participants completed Stages 1-3 of the Training IRAPs (see below). Participants were alone at all other stages of the experiment.

Materials and Apparatus

The experiment involved three computer-based tasks (a Derivation Pre-training task, the Training IRAP, and an MTS task) and six self-report measures.

The Derivation Pre-training Task. The purpose of the Derivation Pre-training Task (identical to that employed by Harte et al., 2020) 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" or "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).

INSERT TABLE 1 HERE

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, 2020), 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 (B stimuli) 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 comprised one of two phrases “Least Similar” or “Most Similar”, the target stimulus was always “TTT” or “]][[”, and each pair of response options comprised “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 Figure 1).

INSERT TABLE 2 & FIGURE 1 HERE

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 Figure 2).

INSERT FIGURE 2 HERE

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 final Training IRAP presented the untrained A-C relations that could be derived from the mixed A-B and B-C Training IRAPs. 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; Most Similar-Beda; Least Similar-Sarua; and Most Similar-Sarua (see Figure 3).

INSERT FIGURE 3 HERE

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 Figure 4 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 Figure 4). A second comparison was also quite like the sample, but with more variations (see left-hand side of Figure 4), rendering it *less similar to* the sample. Finally, the third comparison was clearly the *least similar to* the sample because it had little or no overlapping features (right-hand side of Figure 4). 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.

INSERT FIGURE 4 HERE

Questionnaires. Experiment 1 involved six self-report questionnaires, four of which were standardized measures (the Depression, Anxiety and Stress Scales, DASS-21; the Acceptance and Action Questionnaire, AAQ-II; the Psychological Flexibility Index, PFI[2]; the Generalized Pliance Questionnaire, GPQ) and the Certainty Likert Scales and Propensity for Rule-Following Scale (PRFS). The first three scales were included as measures of psychological distress because such measures have been related to persistence in rule-following in previous research (e.g., McAuliffe, Hughes, & Barnes-Holmes, 2014). The Certainty Scales were employed as a self-report measure that could be seen as a potential

[2] The PFI employed in the current study was a very early version and to all intents and purposes no longer exists. The PFI is now named the Everyday Psychological Inflexibility Checklist (EPIC; Thompson, Bond, & Lloyd, 2019), which has only 7 items (in contrast to the 80 items in the PFI). We have not, therefore, presented any psychometric properties for the PFI.

additional measure of coherence that might be sensitive to the coherence manipulation (i.e., the provision versus non-provision of performance feedback). The GPQ and PRFS were included as self-report measures of persistent rule-following.

The *DASS-21* comprises three subscales measuring depression, anxiety, and stress across a total of 21 statements, with 7 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. All overall and subscale scores obtained are then doubled, and severity bands are generated accordingly. Specifically, the overall DASS score ranges from 0-126. Higher scores on the overall score 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 was employed in the current experiment, which according to deBeurs, Van Dyck, Marquenie, Lange, and Blonk (2001) has yielded similar sufficient internal consistency. Reliability analyses were conducted on the measures using the current sample (from Experiments 1 and 2) and yielded similar, albeit slightly lower, levels of internal consistency: depression (alpha = 0.86); anxiety (alpha = 0.73); stress (alpha = 0.74); and total DASS (alpha = 0.90).

The *AAQ-II* measures acceptance of negative private events across 7 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.). Again, the Dutch version of the scale was employed currently, which according to Bernaerts, De Groot, and Kleen (2012) has yielded a Cronbach's alpha of 0.85. Once again, reliability analyses were conducted on the measures using the current sample (from Experiments 1 and 2) and a yielded similar, albeit slightly higher, Cronbach's alpha of 0.90.

The *PFI* is 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.

The *Certainty Scales* aimed to attain a self-report measure of participants' certainty of the derived relations involved within the derived A-C network (i.e., the four trial-types involved in the A-C IRAP: Least Similar-Beda-True; Least Similar-Sarua-False; Most Similar-Beda-False; Most Similar-Sarua-True). The study thus involved four individual 7-point scales, one for each trial-type. Participants were presented with a screenshot of each IRAP trial-type as it was presented within the IRAP (i.e., label at the top of the screen, target in the middle of the screen, and two response options at the bottom left- and right-hand sides of the screen) and were asked to rate from 1 (*Extremely uncertain*) to 7 (*Extremely certain*) how certain they were that the answer that they gave on this trial was correct. Along with generating certainty scores for each individual trial-type, a total certainty score was calculated by summing each individual trial-type score. This yielded a maximum certainty score of 28 and a minimum of 4.

The *GPQ* is designed to measure generalized pliance (Ruiz, Suárez-Falcón, Barbero-Rubio, & Flórez, 2019) across a total of 18 statements (e.g., "My decisions are very much influenced by other people's opinions"). All items 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. Reliability analyses were conducted using the current sample from both experiments and yielded an alpha coefficient of 0.93, comparable to the English version of the GPQ which has produced alpha coefficients of .93, .95, and .97 in undergraduate, general, and clinical populations, respectively (Ruiz et al.).

The *PRFS* was created by Harte et al. (2018) to assess propensity to rule-following across 6 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. Reliability analyses were conducted across the samples of both experiments in the current study yielding a Cronbach's alpha score of .65.

Procedure

Experiment 1 comprised 5 stages (see Figure 5). 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 mixed A-B and B-C relations Training IRAP, in which A-B and B-C relations were mixed randomly within each block of trials. Phase 4 presented a fixed number of five blocks of previously untrained A-C trials. Half of the participants continued to receive feedback on each trial, 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., Certainty Scales, GQP, and PRFS).

INSERT FIGURE 5 HERE

Stage 1: DASS-21, AAQ-II, PFI, and DART. All participants completed the DASS-21, the AAQ-II, 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., 2017, 2018), 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 and provided corrective feedback on each response. Once a trial had finished, the next trial was then presented immediately. 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.

INSERT TABLE 3 HERE

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 presented “Hond” and “Madra”; and the fourth 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 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 verbally 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 center of the screen, and the next trial appeared 400ms 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 (≤3000 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: 5 Blocks of Derived A-C Relations with or without Feedback. Phase 4 presented participants with five blocks of previously untrained A-C relations. Half of the participants continued to receive feedback on every trial and at the end of each block, while the other half no longer received feedback at any point. At the beginning of this phase, 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 block and then immediately to Stage 4, once Phase 4 was complete. All participants were advised that during this stage some of the stimuli 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.

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 one point for each correct response and deducted one 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, one 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 3s. Emitting an incorrect response resulted in the loss of one 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: Certainty Scales, GPQ, and PRFS. After the MTS task, participants completed the Certainty Scales, GPQ, and the PRFS in that order.

Results and Discussion

For the purposes of analysis, exclusion criteria were applied to the blocks involved in Phase 4 of the Training IRAPs. The data from 5 participants were removed because they failed to maintain ≥75% accuracy per trial-type in these blocks (4 in Feedback and 1 in No Feedback, $N = 62$ remaining). Consistent with Harte et al. (2020), no response latency criterion was applied to these blocks because the target relations were novel (i.e., they were not preceded by direct training). A strict accuracy criterion was also applied to the MTS task and required correct responding on at least 8 of the first 10 trials, as well as 80 of the first 100 trials in Phase 1. This MTS task criterion was consistent with Harte et al. (2017, 2018, 2020), 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 2 participants were removed on this basis (1 in Feedback and 1 in No Feedback, $N = 60$ remaining).

Certainty Scales and IRAP Data

In order to assess whether participants' self-reported certainty in the derived A-C relations differed between the Feedback and No-Feedback groups, the mean scores on each trial-type and on the overall score were compared. The means and standard deviations for each trial-type's certainty score, and the mean overall certainty score for the Feedback and No Feedback conditions are presented in Table 4 (top). Independent *t*-tests confirmed that none of these scores between the Feedback and No Feedback groups differed significantly from each other (all *p*s > .38).

INSERT TABLE 4 HERE

The mean number of blocks required by participants in each condition in Stages 1-3 of the Training IRAPs were also compared. The mean number of blocks and their standard deviations for each stage for the Feedback and No Feedback conditions are presented in Table 4 (bottom). Independent *t*-tests confirmed that none of these differences were significant (all *p*s ≥ .65, without correction for multiple tests). Thus, any subsequent differences that emerged among the groups during the Training IRAPs 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

The data from the 50 trials in Phase 2 of the MTS task presented after the contingency reversal were analysed in the same three ways as Harte et al. (2020): 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. *Contingency sensitivity* was defined as a pattern that comprised at least 3 consecutive responses that were not in accordance with the original instruction, and at least 1 of these must accord with the reversed contingency. Finally, *rule resurgence* was defined as the percentage of responses consistent with the initial rule that occurred after a participant had demonstrated contingency sensitivity (visual inspection of the data indicated that for the vast majority of participants, all three or more consecutive responses were in accordance with the reversed contingencies).

Rule Compliance. Figure 6 presents the group means for rule compliance for both Feedback and No Feedback groups and shows a marginal difference between them. Specifically, the No Feedback group made a greater number of responses (*M* = 18.33, *SD* = 17.45) in accordance with the original rule in the face of the reversed feedback contingencies than did the Feedback group (*M* = 14.93, *SD* = 13.15). An independent *t*-test revealed that this effect was not significant, *t*(58) = .852, *p* = .40.

INSERT FIGURE 6 HERE

Contingency Sensitivity. Figure 7 presents the group means for contingency sensitivity, and once again shows a marginal difference between the Feedback and No Feedback conditions. That is, the No Feedback group emitted a greater number of responses (*M* = 17.07, *SD* = 16.16) in accordance with the original rule before demonstrating contingency sensitive responding than did the Feedback group (*M* = 14.73, *SD* = 12.56). An independent *t*-test revealed, however, that this difference was not significant, *t* (58) = .627, *p* = .53.

INSERT FIGURE 7 HERE

Rule Resurgence. Figure 8 shows levels of rule resurgence among participants (i.e., there were no exclusions made on the basis of absence of contingency sensitivity). There is little visual evidence for differential levels of resurgence between these groups. For example, in the Feedback group, 6 participants resurged for over 10% of responses, compared with 4 in the No Feedback condition. Given that the data were severely skewed, a Mann Whitney U-test was employed. Results confirmed that there was no significant difference between the groups (*p* = .70).

INSERT FIGURE 8 HERE

Correlations

Pearson's correlational analyses were conducted between the rule compliance and contingency sensitivity measures of rule persistence and the self-report measures. For the rule resurgence measure, Spearman's rank order correlational analyses were conducted. Given that neither condition differed significantly on any measure of rule persistence, correlational analyses were conducted with the data collapsed across Feedback and No Feedback groups. Out of a possible 26 correlations in the rule compliance and contingency sensitivity measures, only two reached significance. Both rule compliance (*r* = -.220, *p* = .04) and contingency sensitivity (*r* = -.281, *p* = .03) correlated negatively with the PFI, such that participants who reported lower levels of psychological flexibility were more likely to persist with rule-following on the MTS task on both of these measures (all other *p*s > .19). Out of a possible 13 correlations for the rule resurgence measure, no correlations reached significance (all *p*s > .23)

Summary

The results of this first experiment indicated that the presence versus absence of feedback did not differentially influence rule compliance, contingency sensitivity, or rule resurgence. In effect, attempting to increase relational coherence of the derived A-C relations with the use of feedback appeared to have limited impact on persistent derived rule-following. The current finding could be seen, therefore, as a failure to replicate an effect reported by Harte, et al. (2020), in which higher levels of resurgence were found when feedback was provided for the derived A-C relations. On balance, the previous study presented only two blocks of A-C testing, whereas the current experiment presented five blocks. As such, it could be argued that the level of derivation was lower in the current experiment (because derivation reduces with additional testing), and perhaps as derivation reduces the impact of coherence, manipulated via feedback, has less influence. Or more informally, the more participants were allowed to practice a particular behavior, the more impervious to the feedback that behavior became. Indeed, it is exactly this type of interpretation of the results reported here that highlights the potentially highly dynamic nature of AARRing itself.

Experiment 2

At this point it was decided to run a second experiment, similar to Experiment 1, but in which level of derivation was relatively high. This was achieved by allowing participants to derive the A-C relations across only a single block of test trials. If the foregoing interpretation is correct, then the presence versus absence of feedback should impact upon persistent rule-following as was reported in the Harte et al. (2020) study. The current experiment is not a direct replication, however, because participants were provided with only one block of A-C trials (rather than two). Thus, if the foregoing interpretation is correct, the impact of feedback should be readily observed because derivation will be even higher than was the case when feedback appeared to impact upon rule persistence. Experiment 2 also included the Dutch Adult Reading Test (DART; Schmand, Bakker, Saan, & Louman, 1991) in order to increase the length of time participants spent in the experiment, thus rendering it similar to Experiment 1. Using the DART also allowed us to address concerns that any differences found in this and preceding studies using similar preparations (e.g., Harte et al., 2018, 2020) could be due to variations in participant IQ (note, the DART has been used to predict IQ).

Participants

A total of 72 individuals participated in Experiment 2, 46 females and 26 males. They ranged in age from 18 to 42 years (*M* = 20.52, *SD* = 4.05) and were recruited through random convenience sampling from the online participant system at X University. Thirty-three participants were paid a fixed sum of 10 euros for participation, while 39 received course credit. Participants were randomly assigned to one of two conditions, again referred to as: Feedback and No Feedback. Once again, paid and course-credit participants were distributed in a roughly equal manner between the two feedback conditions (16 paid and 22 course-credit in Feedback; 17 paid and 17 course-credit in No Feedback). The data from 12 participants (8 from Feedback [3 paid and 5 course-credit] and 4 from No Feedback [3 paid and 1 course-credit]) were excluded because they failed to meet specific performance criteria on either a Training IRAP or the MTS task (see below), leaving *N*=60 for analysis, 30 in each condition.

Setting

The setting was similar to Experiment 1, except that the Experimenter also now remained in the room to administer the DART.

Materials and Apparatus

The experiment involved the same three computer-based tasks and six self-report measures as Experiment 1. All participants were also required to complete the DART. Participants completed all aspects of the experiment on a standard Dell laptop. The DART (Schmand, et al., 1991) consists of 50 words that are considered irregular in terms of grapheme-phoneme correspondences. Participants are asked to read aloud and pronounce each word correctly, with accurate responding based on correct pronunciation. Participant total errors are tallied, with higher errors suggesting lower IQ. The error score is then converted into a predicted WAIS Full Scale score, verbal IQ score, and performance IQ score.

Procedure

Experiment 2 again comprised 5 stages. Stage 1 presented the three initial questionnaires and the DART (i.e., DASS-21, AAQ-II, PFI, and the DART). Stage 2 presented the Derivation Pre-training Task, which comprised the same three cycles as Experiment 1. Stage 3 involved the Training IRAPs, which again comprised four phases, similar to Experiment 1: Phases 1-3 were identical to Experiment 1, while Phase 4 now presented participants with only one further block of previously untrained A-C trials after baseline relation training (i.e., one block instead of five). As in Experiment 1, half of the participants received programmed feedback for their responses, while the remaining half did not. Stage 4 involved the same MTS task, with rule-consistent contingencies in Phase 1 and rule-inconsistent contingencies in Phase 2, while Stage 5 again presented the remaining questionnaires (i.e., Certainty Scales, GQP, and PRFS).

Results

The same exclusion criteria that applied in Experiment 1 were applied here. With respect to Phase 4 of the Training IRAPs, data from 1 participant from the No Feedback condition were removed because they failed to maintain ≥75% accuracy per trial-type in these blocks (*N*=71 remaining). With respect to the MTS task, the data from 10 participants were removed (7 in Feedback and 3 in No Feedback; *N*=61 remaining) because they failed the inclusion criteria, as employed in Experiment 1. The data for one participant in the Feedback condition was also removed because they failed to meet the performance criteria on the A-B Training trials (*N*=60 remaining).

DART, Certainty Scales, and IRAP Data

Prior to conducting the primary analyses, participants' predicted full scale IQ, verbal IQ, and performance IQ (as measured by the DART) were compared (see top of Table 5). Independent *t*-tests confirmed that participants' predicted IQ scores did not differ significantly between the two groups (all *p*s > .26). Thus, any subsequent differences that emerged among the groups during the Training IRAPs or the MTS task would not likely be due to differences in participant IQ.

INSERT TABLE 5 HERE

In order to assess whether participants' self-reported certainty in the derived A-C relations differed between the Feedback and No-Feedback groups, the mean scores of each trial-type and the mean overall score were compared (see Table 5, centre). Independent *t*-tests confirmed that none of these certainty scores between the Feedback and No Feedback groups differed significantly from each other (all *p*s > .10).

The mean number of blocks required in each condition in Stages 1-3 of the Training IRAPs were compared (see Table 5, bottom). Once again, independent *t*-tests confirmed that none of these differences were significant (all *p*s > .054).

Measures of Rule Persistence

The data from the 50 trials in Phase 2 of the MTS task presented after the contingency reversal were analysed in the same three ways as in Experiment 1.

Rule Compliance. Figure 9 presents the group means for rule compliance. Participants in the No Feedback group made a greater number of responses (*M* = 20.33, *SD* = 18.38) in accordance with the original rule in the face of the reversed feedback contingencies than did those in the Feedback group (*M* = 13.03, *SD* = 11.82). An independent *t*-test revealed, however, that this effect did not reach significance, *t*(58) = -1.829, *p* = .07.

INSERT FIGURE 9 HERE

Contingency Sensitivity. Figure 10 presents the group means for contingency sensitivity. The No Feedback group made a greater number of responses (*M* = 20.33, *SD* = 18.38) in accordance with the original rule before demonstrating contingency-sensitive responding than did the Feedback group (*M* = 10.07, *SD* = 4.09). An independent *t*-test confirmed that this difference was significant, *t*(58) = -2.99, *p* = .004.

INSERT FIGURE 10 HERE

Rule Resurgence. Figure 11 presents differential levels of rule resurgence among all participants in both conditions (i.e., there were no exclusions made on the basis of absence of contingency-sensitivity). The data show some suggestion of greater resurgence in the Feedback condition than in the No Feedback condition. Given that the data were again severely skewed, a Mann Whitney U-test was employed, which revealed a significant difference between the conditions (Feedback, *Md* = 6.56%, No Feedback, *Md* = 5.00%, *U* = 302.50, *z* = -2.181, *p* = .03).

INSERT FIGURE 11 HERE

Correlations

Correlational analyses were conducted between the three measures of rule persistence and the self-report measures. Parametric analyses (Pearson's *r*) were conducted for rule compliance and contingency sensitivity, while non-parametric analyses (Spearman's *rho*) were conducted for rule resurgence.

Given that the conditions did not differ significantly on the rule compliance measure, correlational analyses were conducted with the data collapsed across groups. Out of a possible 13 correlations among the rule compliance measure of rule persistence and the self-report measures, only one reached significance. That is, rule compliance correlated with the Trial-Type 2 Certainty Scale (*r* =.220, *p* = .04), such that participants who reported greater certainty that Least Similar and Sarua did not have the same meaning were more likely to persist with rule-following on the MTS task (i.e., choosing the Beda/Least similar stimulus).

Given the significant group differences recorded on the contingency sensitivity measure, separate correlational analyses were conducted for the Feedback and No Feedback groups and all self-report scales. Out of a possible 26 correlations, 2 proved to be significant. In the No Feedback group, contingency sensitivity correlated positively with the Trial-Type 2 certainty score (r = .43, p = .01), suggesting that the more certain participants were that Least Similar did not have the same meaning as Sarua, the longer they would persist with the original rule (i.e., Beda has the same meaning as Least Similar). In the same condition (No Feedback), contingency sensitivity also correlated positively with participants' overall certainty score (r = .40, p = .03), such that higher levels of certainty were associated with more persistent rule-following.

Given the significant group differences recorded above for rule resurgence, separate correlational analyses were conducted for the Feedback and No Feedback groups and all self-report scales. Only one correlation reached significance; in the No Feedback condition, greater resurgence was associated with higher compliance as measured by the GPQ (*rho* = .390, p = .03).

Summary

The findings from Experiment 2 suggested that manipulating the presence versus absence of feedback for the novel derived A-C target relations when derivation was high (i.e., only 1 block) influenced the three measures of rule persistence, significantly for contingency sensitivity and resurgence, with marginal significance for the rule compliance measure. Furthermore, the correlational analyses yielded a small number of significant effects, all of which were in the intuitively correct direction (e.g., increased rule persistence was associated with increased levels of self-reported compliance on the GPO).

General Discussion

The current study sought to extend recent research exploring the behavioral dynamics involved in persistent rule-following, focusing on two potentially key variables: levels of derivation and coherence. Coherence was manipulated through the provision or non-provision of feedback on novel, derived A-C relations while derivation was low (5 blocks of trials in Experiment 1) and subsequently while derivation was high (1 block in Experiment 2). The results indicated that the presence versus absence of feedback differentially impacted upon rule persistence when derivation was high, but not when it was low. Specifically, when derivation of the novel A-C relations was high in Experiment 2, the presence versus absence of feedback significantly impacted upon both the contingency sensitivity and rule resurgence measures. It should be noted, that a similar effect for rule resurgence was found when derivation was defined as relatively high (Harte et al., 2020). Unlike that previous study, however, an effect for contingency sensitivity was also found in Experiment 2 of the current study. On balance, participants in the current study received only one further block of A-C test trials whereas those in the previous study received two. In principle, therefore, derivation could be considered as even higher in the current study than in that of Harte et al., and perhaps this explains why contingency sensitivity also yielded a significant effect between the feedback and no feedback conditions. Certainly, future research could examine this suggestion more systematically. In any case, the current findings, along with those of the previous study, provide support for the suggestion that as derivation reduces, coherence (via feedback) has less influence on rule persistence; but when derivation is high, the impact of feedback seems to be far greater. In general, the results highlight the highly complex and dynamic nature of persistent rule-following, and indeed, of AARRing itself.

An interesting pattern emerged in the results of Experiment 2 that should be noted. Specifically, the contingency sensitivity measure indicated that the No Feedback group persisted for longer than the Feedback group. In simple terms, the former group persisted for longer in following the rule before checking to see if points were available for choosing the opposite (Most Like) stimulus. For the resurgence measure, however, the Feedback group returned to rule-following more readily than the No Feedback group. Thus, although the *absence* of feedback appeared to generate more persistence in terms of simply continuing with rule-following, the *presence* of feedback caused more participants to return to rule-following after they had contacted the reversed contingencies. This finding highlights the need to be relatively precise in defining exactly what we mean by the term "contingency sensitivity" or "rule-persistence" and the likely impact of contextual variables (such as the presence versus absence of feedback) on these different measures. It is also worth noting that the differential effect for feedback was only observed in Experiment 2, in which there was only one block of A-C testing (i.e., derivation was high), which further complicates the analyses.

In a previous study using a similar preparation, the authors suggested that perhaps differences in intelligence between the groups could be a contributing factor to the level of rule persistence between groups (Harte, et al., 2018). Indeed, previous studies did not check or control for intelligence (except through random sampling). On balance, the lack of differences across conditions in the number of blocks taken by participants to complete the IRAP training suggested that intelligence was unlikely to have played a significant role (because performance on the IRAP has been shown to correlate with measures of intelligence; O'Toole & Barnes-Holmes, 2009). The same check for differences in IRAP training blocks was made in the current study, and Experiment 2 also included the DART -- a measure that has been shown to predict intelligence scores on the WAIS. The results showed no significant differences in the predicted intelligence scores (from the DART), and thus again it seems unlikely that intelligence played a significant role in determining differences in performance between Feedback and No Feedback conditions. Nonetheless, the DART is not a formal measure of intelligence and thus we should be cautious in drawing strong conclusions concerning the absence of any role for levels of intelligence in explaining the current findings.

Although the correlations between the three measures of rule persistence and the various self-report measures were very few in number, and should thus be interpreted with caution, those that did emerge may be worthy of further research. For example, a significant positive correlation emerged between the GPQ and the rule resurgence measure for No Feedback participants, a correlation that was found in the previous Harte et al. (2020) study. It appears, therefore, that when coherence for a derived rule is somewhat low (i.e., in the absence of feedback), self-reported compliance is more likely to predict rule persistence on the MTS task. In other words, coherence may moderate the relationship between the self-reported tendency to engage in rule-following and actual rule-following itself. It is also interesting that in the No Feedback condition, higher scores on the certainty scales predicted higher contingency sensitivity scores (i.e., increased certainty correlated with greater rule persistence). Finally, significant correlations were also found in Experiment 1 between the rule compliance and contingency sensitivity measures and the PFI, indicating that limited psychological flexibility predicted greater persistence in rule-following.

In reflecting upon the potential implications of these correlational analyses, it is worth noting that the concept or definition of coherence was restricted to the derived relation that was contained within the rule (e.g., between *Least Similar* and *Beda*). However, the concept of coherence may also be applied to the relationship between the (derived) rule itself and the contingencies contacted during the MTS task. More informally, during initial exposure to the MTS task the rule fully cohered with the task, but following the contingency switch coherence between the rule and the task was completely undermined. At the present time it remains unclear if and how these two types of coherence/incoherence may have interacted but it could be an important area for future research to pursue. Consider, for example, that when coherence for a derived rule was low (i.e., no feedback), self-reported compliance predicted rule persistence on the MTS task; this was not the case in the Feedback condition. Perhaps the presence of feedback for the derived relation reduced the functional overlap between this relational responding in the experiment and in the natural environment (in the 'real world' verbal relations are rarely reinforced in a continuous and highly programed basis). As a result, the 'unnatural' level of feedback undermined the extent to which subsequent performance on the MTS task could predict actual rule-following (as measured using a self-report measure).

Another finding worth noting is the fact that the certainty scales did not differentiate between the Feedback and No Feedback conditions, which might be deemed a counter-intuitive result. Specifically, one might expect certainty to increase given programed feedback. Upon reflection, this counterintuitive finding could be due to the fact that the certainty scales were presented *after* participants had completed the contingency switching MTS task. In effect, all of the participants had experienced a spontaneous and 'unexplained' reversal in feedback contingencies before completing the scales, and perhaps this undermined the impact of the earlier feedback for the derived relations on the certainty measure. More informally, the contingency reversal in the MTS task undermined participants' trust in the feedback as a reliable source of information. On balance, a small number of intuitively sensible correlations with the certainty scales emerged in Experiment 2. For example, as noted earlier, in the No Feedback condition higher levels of self-reported certainty (for the A-C relations) was associated with greater persistence in rule-following, in terms of the contingency sensitivity measure. In any case, it seems important that a future study would ask participants to complete the certainty scales before the MTS task, so that the potential impact of the spontaneous contingency reversal would be removed from the experimental sequence.

At a more general level, it is also worth bearing in mind that the absence of an "expected" significant correlation(s) could be related to lack of power. Indeed, the current findings may be useful in designing future studies because the relevant effects sizes reported here could be used as the basis for increasing the sample sizes. On balance, even if a higher number of correlations were detected with larger samples, it would still be important to explain any differences in the strength of the significant correlations that were obtained in terms of the variables (e.g., coherence, derivation, etc.) that we have identified here. In other words, if significant correlations between rule-persistence and measures of psychological suffering were obtained with larger samples, but the relative strength of these correlations were modified by the types of variables identified in the HDML, this would call for a more sophisticated theoretical analysis of the link between excessive rule-following and psychological suffering (see Harte, Barnes-Holmes, Barnes-Holmes, & Kissi, 2020, for a more extensive discussion).

Overall, the current study, and the previously published studies in this line of research, have indicated quite clearly that the study of rule persistence in the face of conflicting contingencies requires a relatively sophisticated analytic approach. The highly dynamic nature of the interaction among variables, such as the impact of levels of coherence and derivation on flexibility in rule-following, needs to be factored into any analysis of excessive rule-following generally, and particularly into any interpretive analysis of this behavior as a marker, or partial explanation for, human psychological distress (Zettle & Hayes, 1982; see also Kissi, Harte, Hughes, De Houwer, & Crombez, 2020, for a recent systematic review). Although the type of findings presented here, and in the previously published studies, somewhat complicate the narrative in this area, they also may help us to better understand the nature of the relationship. Indeed, this general approach could certainly be seen as more consistent with recent calls for a process-based approach to understanding and treating human psychological suffering (e.g., Hayes et al., 2019; Hofmann & Hayes, 2018).

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., & McEnteggart, C. (2020). Updating RFT (more field than frame) and its implications for process-based therapy. *The Psychological Record.* doi: 10.1007/s40732-019-00372-3

Bernaerts, I., De Groot, F., & Kleen, M. (2012). De AAQ-II, een maat voor experiëntiële vermijding: Normering bij jongeren. *Gedragstherapie, 45*, 389-400.

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: 10.1016/j.beth.2011.03.007

de Beurs, E., Van Dyck, R., Marquenie, L. A., Lange, A., & Blonk R. W. B. (2001). De DASS: een vragenlijst voor het meten van depressie, angst en stress. *Gedragstherapie, 34*, 35-53.

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, D., Barnes-Holmes, Y., & Kissi, A. (2020). The study of rule-governed behavior and derived stimulus relations: Bridging the gap. *Perspectives on Behavior Science, 43*(2), 361-385. doi: 10.1007/s40614-020-00256-w

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: 10.1016/j.beproc.2018.08.005.

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: 10.1177/0145445517715871.

Harte, C. Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C., Gys, J., & Hasler, C. (2020). Exploring the potential impact of relational coherence on persistent rule-following: The first study. *Learning and Behavior.* doi: 10.3758/s13420-019-00399-0

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: 10.1901/jeab.1986.46-137

Hayes, S.C., Hofmann, S.G., Stanton, C.E., Carpenter, J.K., Sanford, B.T., Curtiss, J.E., & Ciarrochi, J. (2019). The role of the individual in the coming era of process-based therapy. *Behavior Research and Therapy, 117,* 40-53. doi: 10.1016/j.brat.2018.10.005

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

Henry, J.D. & Crawford, J.R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 44*(2), 227-239. doi: 10.1348/014466505X29657

Hofmann, S.G. & Hayes, S.C. (2019). The future of intervention science: Process-based therapy. *Clinical Psychological Science, 7,* 37-50. doi: 10.1177/2167702618772296

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.

Kissi, A., Harte, C., Hughes, S., De Houwer, J., & Crombez, G. (2020). The rule-based insensitivity effect: A systematic review. *Peer J.*

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: 10.1177/0145445514521630

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: 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: 10.1002/jeab.95

O'Toole, C. & Barnes-Holmes, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligence test: The importance of relational flexibility. *The Psychological Record, 59*, 119-132. doi: 10.1007/BF03395652

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: 10.1016/j.jcbs.2018.03.003

Schmand, B., Bakker, D., Saan, R., & Louman, J. (1991). The Dutch Reading Test for Adults: A measure of premorbid intelligence level. *Tijdschrift Voor Gerontologie en Geriatrie, 22*(1), 15-19.

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: 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: 10.1044/jshr.1401.05

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

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: 10.1901/jeab.1991.56-519

Thompson, M., Bond, F.W., & Lloyd, J. (2019). Preliminary properties of the Everyday Psychological Inflexibility Checklist. *Journal of Contextual Behavioral Science, 12,* 243-252. doi: 10.1016/j.jcbs.2018.08.004

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

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

Table 3

Stimulus combinations employed within each block of trials in each cycle of the Derivation Pre-training task. Each cell represents an individual trial.

	Cycle 1		
Relation Type	***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
	Madra = Hond
		Zwart = Dubh
		Dubh = Zwart
			Hond ≠ Dubh
			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
	Leine = Hemd
		Fles = Buideal
		Buideal = Fles
			Hemd ≠ Buideal
			Fles ≠ Leine

	Cycle 3		
	Phase 1	***Phase 2***	***Phase 3***
	Set 5	*Set 6*	*Set 5 + 6*
Trained Relations	Boek = XXX	Jas = ////	Boek ≠ //// Jas ≠ XXX
Trained Relations	XXX = Leabhar	//// = Cota	XXX ≠ Cota //// ≠ Leabhar

Derived Relations	Boek = Leabhar
	Leabhar = Boek
		Jas = Cota
		Cota = Jas
			Boek ≠ Cota
			Jas ≠ Leabhar

Table 4

Group means and standard deviations for participant certainty scores per trial-type on the certainty scales and overall certainty score (top), and for the mean number of blocks taken per phase of the Training IRAP (bottom) in Experiment 1.

Certainty Scale	***Feedback condition***		***No Feedback Condition***	
	M	*SD*	*M*	*SD*
Trial-Type 1	6.68	.59	6.60	.72
Trial-Type 2	6.19	1.62	6.27	1.26
Trial-Type 3	6.29	1.22	6.13	1.14
Trial-Type 4	6.74	.51	6.60	.77
Overall Certainty Score	25.90	2.84	25.60	3.04
Training IRAP	***Feedback condition***		***No Feedback Condition***	
	M	*SD*	*M*	SD
A-B relations	2.07	1.03	2.20	1.27
B-C relations	1.77	.72	1.70	.88
Mixed A-B/B-C relations	1.39	.67	1.43	.57
Total number of training blocks	5.23	1.69	5.33	1.52

Table 5

Group means and standard deviations for 1. Participant predicted IQ scores as measured by the DART 2. participant certainty scores per trial-type on the certainty scales and overall certainty score (centre), and for 3. the mean number of blocks taken per phase of the Training IRAP (bottom) in Experiment 2.

DART	***Feedback Condition***		***No Feedback Condition***	
	M	*SD*	*M*	*SD*
Predicted Full Scale IQ	121.90	3.06	120.60	5.22
Predicted Verbal IQ	119.30	2.95	118.20	4.90
Predicted Performance IQ	120.00	2.67	118.80	4.90
Certainty Scale	***Feedback Condition***		***No Feedback Condition***	
	M	*SD*	*M*	*SD*
Trial-Type 1	6.33	.80	6.03	1.45
Trial-Type 2	6.10	.93	5.55	2.06
Trial-Type 3	6.23	.77	5.66	1.76
Trial-Type 4	6.13	1.33	6.10	1.35
Overall Certainty Score	24.80	3.13	23.34	5.78
Training IRAP	***Feedback Condition***		***No Feedback Condition***	
	M	*SD*	*M*	*SD*
A-B relations	2.37	1.25	1.83	.81
B-C relations	1.60	.68	1.62	.68
Mixed A-B/B-C relations	1.43	.68	1.59	.87
Total number of training blocks	5.4	1.59	5.03	1.43

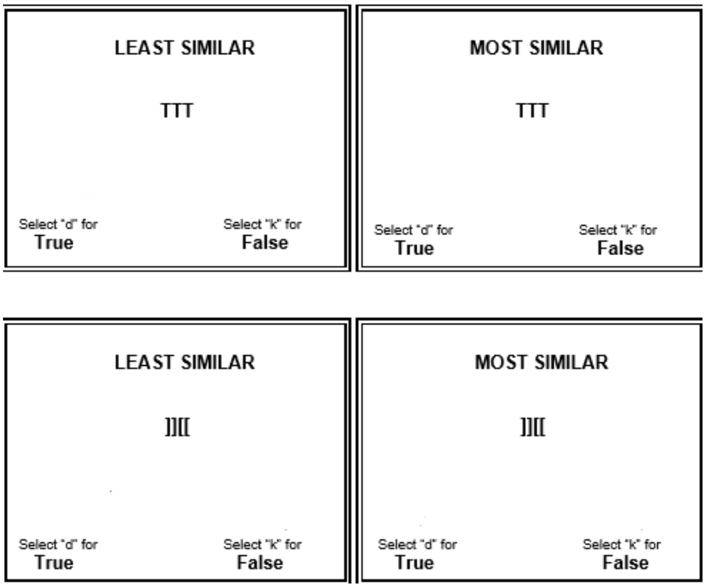


Figure 1. Diagrammatic representation of the IRAP trial-types that appear in the A-B baseline relation training blocks. The four IRAP trial-types were denoted as follows: *Least Similar-TTT; Most Similar-TTT; Least Similar-]][[;* and *Most Similar-]][[.*

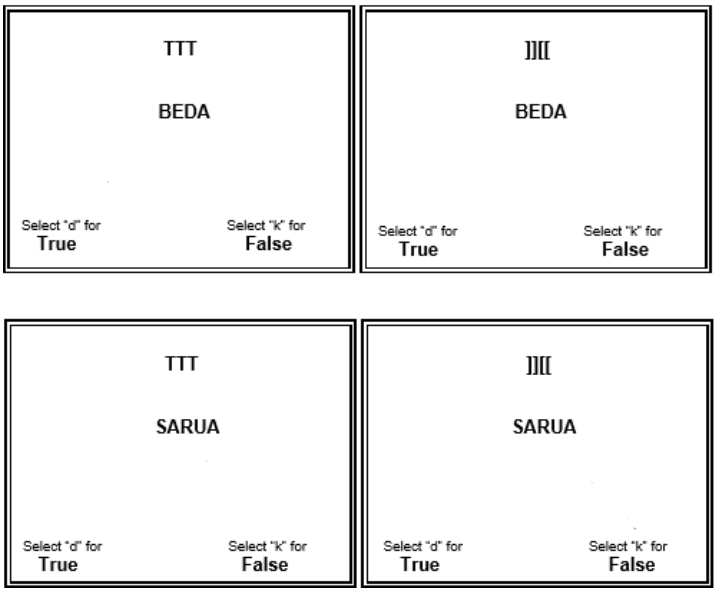
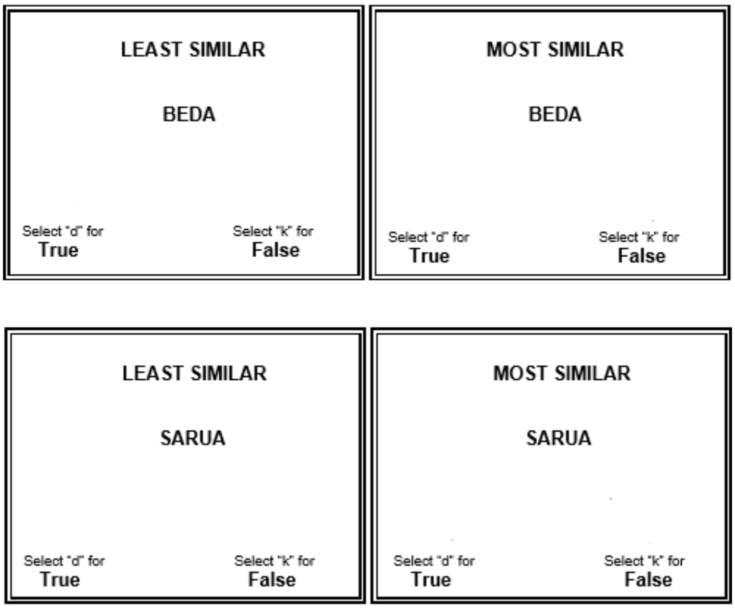


Figure 2. Diagrammatic representation of the IRAP trial-types that appear in the B-C baseline relation training blocks. The four IRAP trial-types were denoted as follows: *TTT-Beda;]][[-Beda; TTT-Sarua;* and *]][[-Sarua.*



LEAST SIMILAR

BEDA

Select "d" for True

Select "k" for False

MOST SIMILAR

BEDA

Select "d" for True

Select "k" for False

LEAST SIMILAR

SARUA

Select "d" for True

Select "k" for False

MOST SIMILAR

SARUA

Select "d" for True

Select "k" for False

Figure 3. Diagrammatic representation of the four IRAP trial-types that appear in the derived A-C relation test blocks. The four IRAP trial-types were denoted as follows: *Least Similar-Beda; Most Similar-Beda; Least Similar-Saru;,* and *Most Similar-Sarua.*

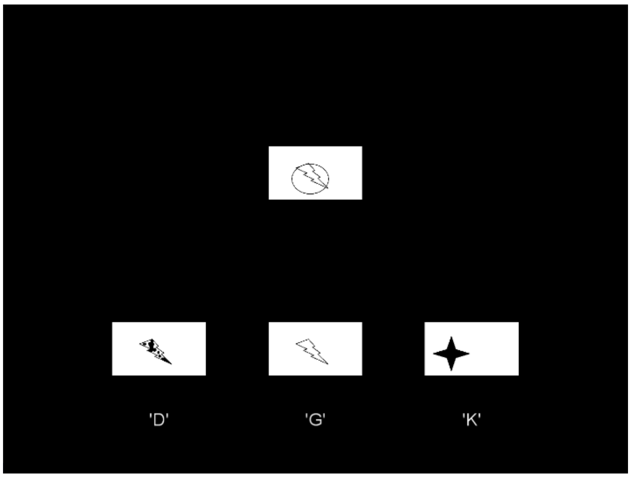


Figure 4. An example of a single trial and stimulus set presented in the MTS task.

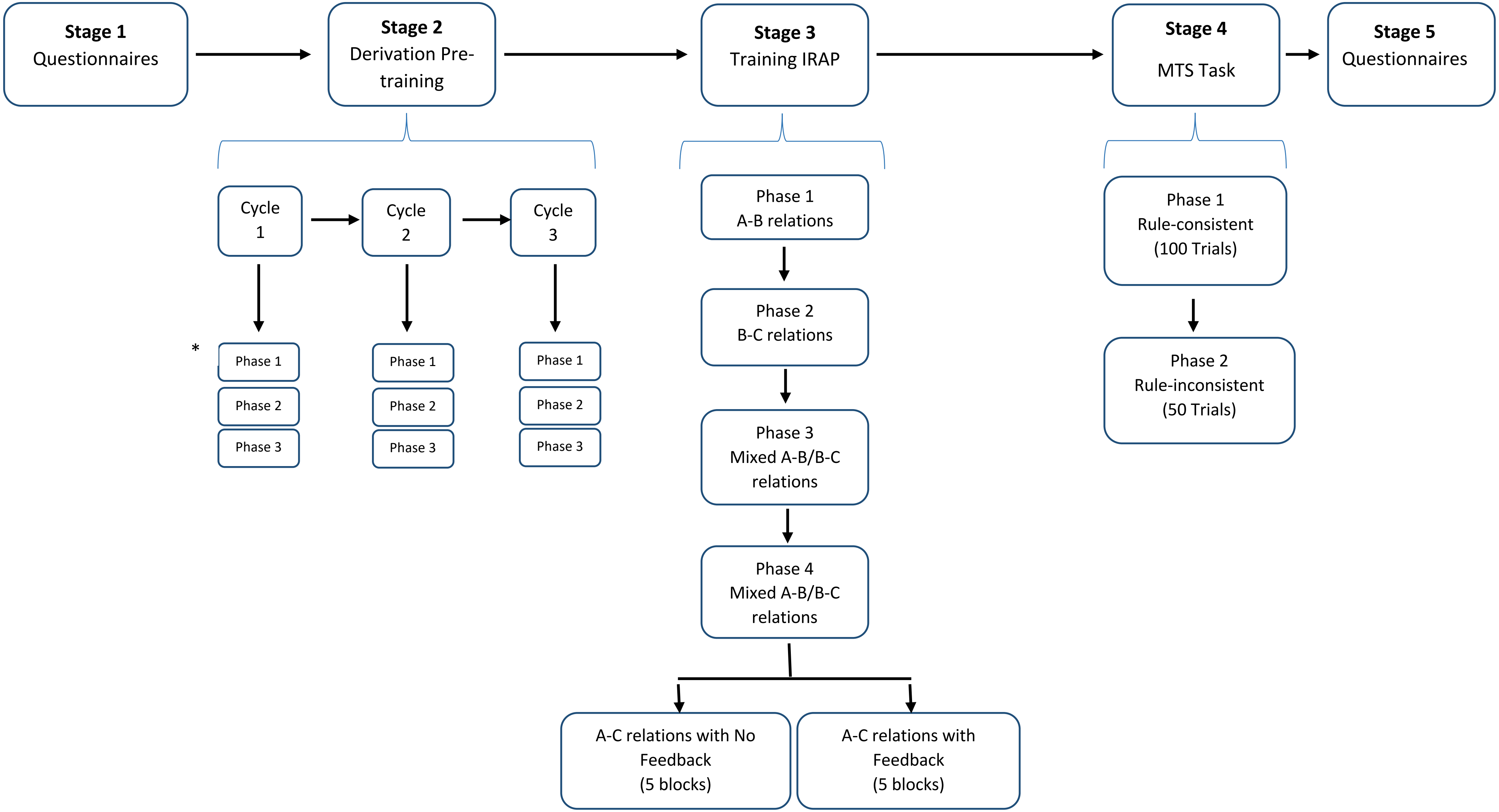


Figure 5. An illustration of the experimental sequence of Experiment 1. The procedure was similar for Experiment 2, except that the DART was included in Stage 1, and the A-C relation blocks in Stage 3 comprised one block instead of five. * See Table 2 for a detailed description of the stimulus set sequencing involved in each phase per cycle in Stage 2.

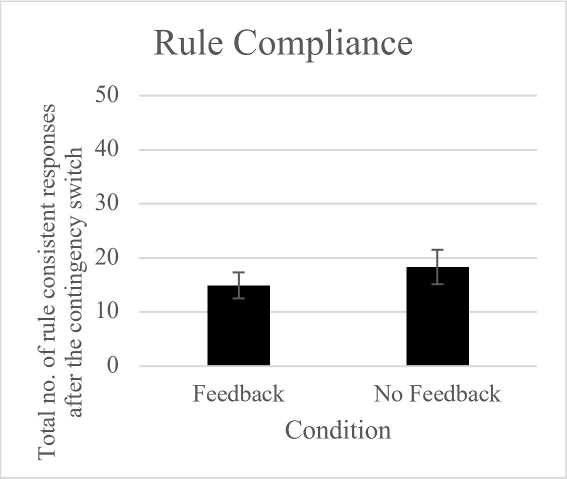


Figure 6. Mean rule compliance scores with standard error bars for Feedback and No Feedback groups in Experiment 1.

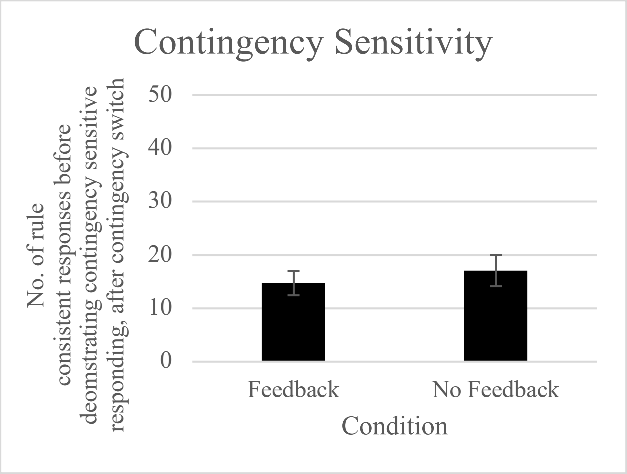


Figure 7. Mean contingency sensitivity scores with standard error bars for Feedback and No Feedback groups in Experiment 1.

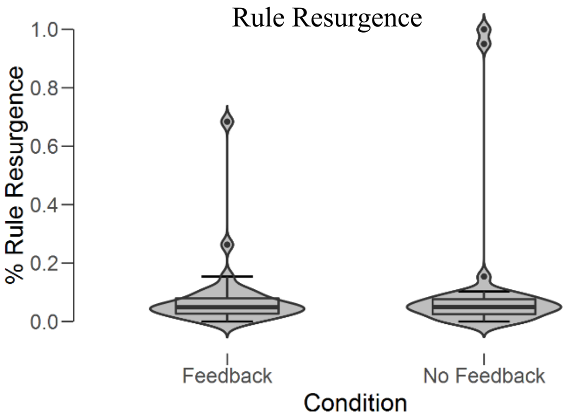


Figure 8. Box plots with a violin element illustrating the distribution and density of participant rule resurgence scores for the Feedback and No Feedback conditions in Experiment 1.

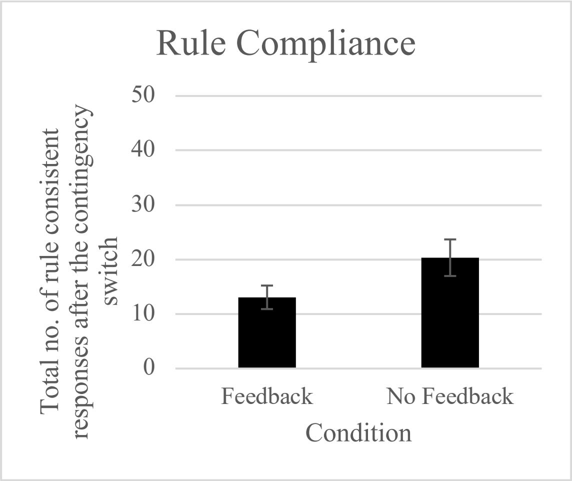


Figure 9. Box plots with a violin element illustrating the distribution and density of participant rule resurgence scores for the Feedback and No Feedback conditions in Experiment 2.

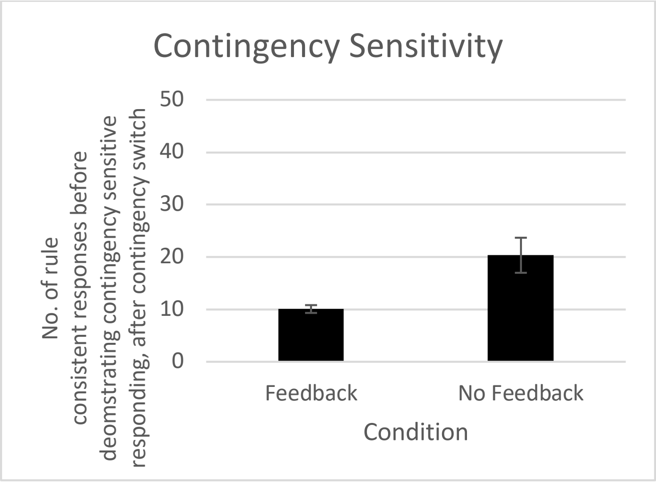


Figure 10. Box plots with a violin element illustrating the distribution and density of participant rule resurgence scores for the Feedback and No Feedback conditions in Experiment 2.

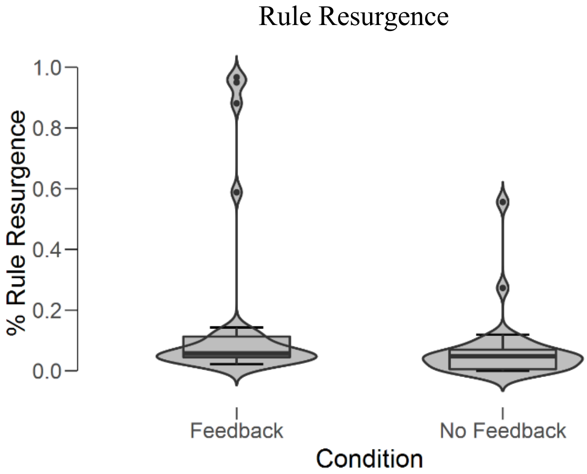


Figure 11. Box plots with a violin element illustrating the distribution and density of participant rule resurgence scores for the Feedback and No Feedback conditions in Experiment 2.