The Impact of High Versus Low Levels of Derivation for Mutually and Combinatorially Entailed Relations on Persistent Rule-Following

Colin Harte, Dermot Barnes-Holmes, Yvonne Barnes-Holmes and Ciara McEnteggart

Department of Experimental, Clinical and Health Psychology, Ghent University, Belgium

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

The effects of rules on human behaviour have long been identified as important in the psychological literature. The increasing importance of the dynamics of arbitrarily applicable relational responding (AARR), with regards to rules, has come to be of particular interest within Relational Frame Theory (RFT). One feature of AARR that previous research has suggested may differentially impact persistent rule-following is level of derivation. However, no published research to date has systematically explored this suggestion. Across two experiments, the impact of levels of derivation was examined on persistent rule-following at two stages of relational development: mutual entailment (Exp. 1) and combinatorial entailment (Exp. 2). A Training IRAP was used to establish a mutually entailed relational network in Experiment 1 and a combinatorially entailed network in Experiment 2, and to train these networks to different levels of derivation. This was followed by a contingency switching Match-to-Sample (MTS) task to assess rule persistence. Results from both experiments were generally consistent with the suggestion that lower levels of derivation produce more persistent rule-following. Unexpectedly, however, the findings from Experiment 1 also indicated that persistence was moderated by the type of novel word employed. Variations in results across both experiments and their implications for future research are discussed.

KEYWORDS: Levels of derivation; Relational Frame Theory; Rules; Persistent rule-following

1. Introduction

Within the behaviour-analytic literature, human behaviour has often been distinguished from that of non-humans with respect to two key features – instructional control and derived relational responding. However, recent research has highlighted that studies integrating both of these features (i.e. instructional control and derived relations) has been extremely limited (see Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2017; Monestes, Greville, & Hooper, 2017). Instructional control, also known as rule governed behaviour (RGB), was first suggested by B.F. Skinner (1966) in the context of problem solving. Rules were then defined as stimuli that specified reinforcement contingencies which allowed a listener to solve problems without needing to contact contingencies directly. For example, the simple rule "If the juices don't run clear, put the chicken back in the oven" ensures that the listener can learn to properly roast a chicken without directly experiencing sickness by eating undercooked poultry. The concept of derived stimulus relations was formalised in the behaviour-analytic literature in the early 1970s by Sidman (1971) in the context of developing procedures for teaching basic reading skills to individuals with learning disabilities. The basic finding was that having been taught a limited number of word-referent relations a number of novel untaught relations emerged (see Sidman, 1994, for a book-length treatment).

One of the key findings in the literature on instructional control or RGB is that such behaviour is often associated with lack of sensitivity to scheduled reinforcement contingencies (e.g. Catania, Shimoff, & Matthews, 1989). Research on this rule-based insensitivity has examined a wide range of variables that appear to moderate the insensitivity effect, including: the presence or absence of a rule-giver (e.g. Kroger-Costa & Abreu-Rodrigues, 2012); prior experience with following rules (e.g. Martinez-Sanchez & Ribes-Inesta, 1996); instruction accuracy (e.g. Hojo, 2002); and the presence of human

psychological suffering (e.g. Baruch, Kanter, Busch, Richardson, & Barnes-Holmes, 2007; Hayes, 1993; Rosenfarb, Newland, Brannon, & Howey, 1992).

As noted above, the study of RGB, and the associated insensitivity effect, have made little or no connection with the empirical literature on derived stimulus relations. *Conceptually*, however, the link between the two areas has been strong for some decades, particularly within the literature on Relational Frame Theory (RFT, Hayes, Barnes-Holmes, & Roche, 2001), which has emerged as one of the main behaviour-analytic treatments of derived stimulus relations. The basic argument is that the pattern of derived relational responding identified by Sidman, and known as stimulus equivalence, constitutes only one class of generalised operant behaviour. According to RFT, there are many such classes, including arbitrarily applicable relations of similarity, difference, opposition, comparison and hierarchy (see Hughes & Barnes-Holmes, 2016, for a recent extensive review). The important point here is that both Sidman (1994) and Hayes et al. (see also Hayes, 1989) argued that the human capacity for learning to respond in accordance with derived relations may be critical in understanding how rules or instructions come to specify contingencies of reinforcement. Indeed, Hayes at al. drew heavily on RGB, the insensitivity effect and derived relations in developing behaviour-analytic explanations for human psychological suffering and the treatment of that suffering, largely in the form of Acceptance and Commitment Therapy (ACT, see Hayes, Strosahl, & Wilson, 1999, for a book-length treatment).

Some RFT research has suggested that derived relational responding could provide the basis for a technical analysis of instructional control and, indeed, laboratory models of instructional control as derived relational responding have been successfully developed (O'Hora, Barnes-Holmes, Roche, & Smeets, 2004; O'Hora, Barnes-Holmes, & Stewart, 2014), thus, bringing together the research in both areas. For example, O'Hora et al. (2014) trained participants to respond through derived instructions by teaching them to respond in accordance

with novel networks of derived relations. Specifically, novel images were trained to be functionally equivalent to the words “same”, “opposite”, “before” and “after”, and these stimuli were then used to establish relational networks that controlled sequences of responses using nonsense stimuli that functioned in a broadly similar way to the use of rules in natural language. Results also demonstrated that responding in accordance with these derived rules was sensitive to differential consequences and direct contingency control. The authors concluded that derived rule-following is a possible source of behaviour control that must be considered in the context of RGB.

More recently, research has begun to extend this line of work and to examine the impact of derived relations on persistent rule-following or contingency-based insensitivity (Harte, et al., 2017). Across two experiments, Harte et al. sought to determine the extent to which participants would persist in rule-following when the reinforcement contingencies were reversed, and thus following the rule was no longer rewarded. The main objective in the study was to determine if persistence in rule-following would differ between rules that did or did not require derived relational responding. Specifically, across both experiments participants received either a direct rule or a rule that involved a novel derived relational response, followed by a matching-to-sample (MTS) task. The MTS task initially reinforced behaviour that was consistent with the direct or derived rule, before an un-cued contingency switch in the latter part of the task.

In Experiment 1, all participants received 10 trials in which the direct or derived rules were consistent with the MTS task contingencies before the contingency reversal, followed by 50 trials in which the direct or derived rule no longer matched the contingencies. Experiment 2 partially replicated Experiment 1, but participants were provided with 100 trials (rather than only 10) before the contingency reversal. While there were no significant differences in rule persistence between conditions in Experiment 1, the provision of a direct (rather than derived)

rule in Experiment 2 resulted in significantly more persistent rule-following (i.e. only when the opportunity to follow the reinforced rule was relatively protracted). In addition, it was only in the Direct Rule Condition in Experiment 2 that significant correlations were observed between rule compliance and self-reported stress.

One limitation of the Harte et al. (2017) study, which was acknowledged by the authors, was the dichotomy made between direct and derived rule-following. Strictly speaking, for RFT even the direct rule condition involved a certain (low) level of derivation. That is, according to the theory, virtually all behaviours that involve human language and cognition, by definition, comprise some level of derivation in the sense that they are *derived* from a history of arbitrarily applicable relational responding (see Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017, for a detailed discussion). From this perspective, the direct rule did not require a novel derivation within the experiment, but the ability to follow the rule was based on a (distant) history of deriving. In contrast, the derived rule condition involved that distant history, but also required a novel derivation.

The primary purpose of the current study was to determine if levels of derivation (high versus low) within the experiment, rather than relying upon the dichotomy between direct and derived rules employed by Harte et al., would produce differences in persistent rule-following, as observed in the original study. That is, would a condition that involved low levels of derivation produce more persistent rule-following than a condition that involved high levels of derivation? The study also sought to examine the impact of high versus low levels of derivation in terms of mutually versus combinatorially entailed relations. Specifically, Experiment 1 involved deriving a relation between two directly related stimuli (mutual entailment), whereas Experiment 2 involved deriving a relation between two indirectly related stimuli (combinatorial entailment). A range of self-report measures of psychological suffering were used to explore the extent to which derived rule-following may

correlate with self-reported levels of distress in the general population. Finally, the current research differed from that of Harte, et al. (2017) in that a Training version of the Implicit Relational Assessment Procedure (IRAP) was employed here to establish the mutually and combinatorially entailed relations. The primary reason for using the Training IRAP was based on pilot research, which indicated procedural problems in using the original software to manipulate levels of derivation within the experiment. Given the exploratory and relatively inductive nature of the current research, we refrained from making formal predictions.

1. Experiment 1

2.1. Method

2.1.1. Participants

A total of 88 individuals participated in Experiment 1, 62 females and 26 males. They ranged in age from 18 to 38 years old (M = 22.36, SD = 4.12) and were recruited through random convenience sampling from the online participant system at Ghent University. All participants were Caucasian with Dutch as their first language and were paid 10 euros for participation. All were randomly assigned to one of two conditions, referred to as Low versus High Derivation. The data from 28 participants (17 from the Low Derivation Condition and 11 from the High Derivation Condition) were excluded because they failed to meet specific criteria on either the Training IRAP or the MTS task (see below), leaving N = 60 for analysis, 30 in the Low Derivation Condition and 30 in the High Derivation Condition. Initially, we planned to collect data from just 30 participants in each condition, but an unexpected trend towards a significant interaction effect with a procedural variable emerged with this number of participants (details provided below). At this point, it was decided to run a set number of additional participants to determine if the trend continued to significance.

2.1.2. Setting

The experiment was conducted in an experimental cubicle at Ghent 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 the Familiarisation Blocks of the Training IRAP (see section 2.1.4.2.1 below). Participants were alone while completing all other tasks in the experiment.

2.1.3. Materials and Apparatus

Experiment 1 involved two computer-based tasks (a Training IRAP and an MTS task) and four self-report measures. All participants completed all aspects of the experiment on a standard Dell personal computer.

2.1.3.1. The Training IRAP

The Training IRAP contained Dutch words and phrases, but their English translations are used here. All trials presented a label at the top of the screen, with a single target word below. The label stimuli always comprised of one of four phrases: "Least Similar"; "Differs Most"; "Most Similar"; and "Resembles Most" (see Table 1). "Least Similar" and "Differs Most" were defined as synonymous, as were "Most Similar" and "Resembles Most". For economy of expression, we will refer to trials that included the presentation of the first two stimuli as Least Similar and the latter two trials as Most Similar. The target word was always "Beda" or "Sarua" (both words were translated from Sudanese)[1]. Each pair of response options comprised of: "True" versus "False"; "Yes" versus "No"; "Correct" versus "Incorrect"; or "Right" versus "Wrong". These stimuli were combined to generate four Training IRAP trial types (see Figure 1) referred to as: Least Similar-Beda, Most Similar-Beda, Least Similar-Sarua and Most Similar-Sarua. Half of the participants were required to

[1] It should be noted that while no formal check was made to ensure participants did not speak Sudanese, this language was deemed to be relatively obscure for this sample of participants (as it had been for Harte et al., 2017). Similarly, no participant made any indication throughout the experiment, or in the debriefing afterwards, that they had any knowledge of Sudanese or knew what Beda or Sarua meant.

confirm that Beda was coordinate with Least Similar (e.g. by selecting one of the confirmatory response options, Yes, True, Right, Correct), and distinct from Most Similar (e.g. by selecting one of the dis-confirmatory response options, No, False, Wrong, Incorrect); these participants were also required to confirm that Sarua was coordinate with Most Similar and distinct from Least Similar (again using the same response options as for Beda). The remaining participants were required to respond in the opposite pattern (e.g. to confirm that Sarua was distinct from Most Similar and coordinate with Least Similar). The Training IRAP software program automatically recorded response accuracy (based on the first response emitted on each trial) and response latency (time in ms. between trial onset and emission of the first correct response).

Table 1

Stimuli employed within the Training IRAP as labels, targets and response options.

Labels Stimuli	Target Stimuli	Response Options	
Least Similar	Beda	Yes	No
		True	False
	Sarua	Right	Wrong
		Correct	Incorrect
Differs Most	Beda	Yes	No
		True	False
	Sarua	Right	Wrong
		Correct	Incorrect
Most Similar	Beda	Yes	No
		True	False
	Sarua	Right	Wrong
		Correct	Incorrect
Resembles Most	Beda	Yes	No
		True	False
	Sarua	Right	Wrong
		Correct	Incorrect

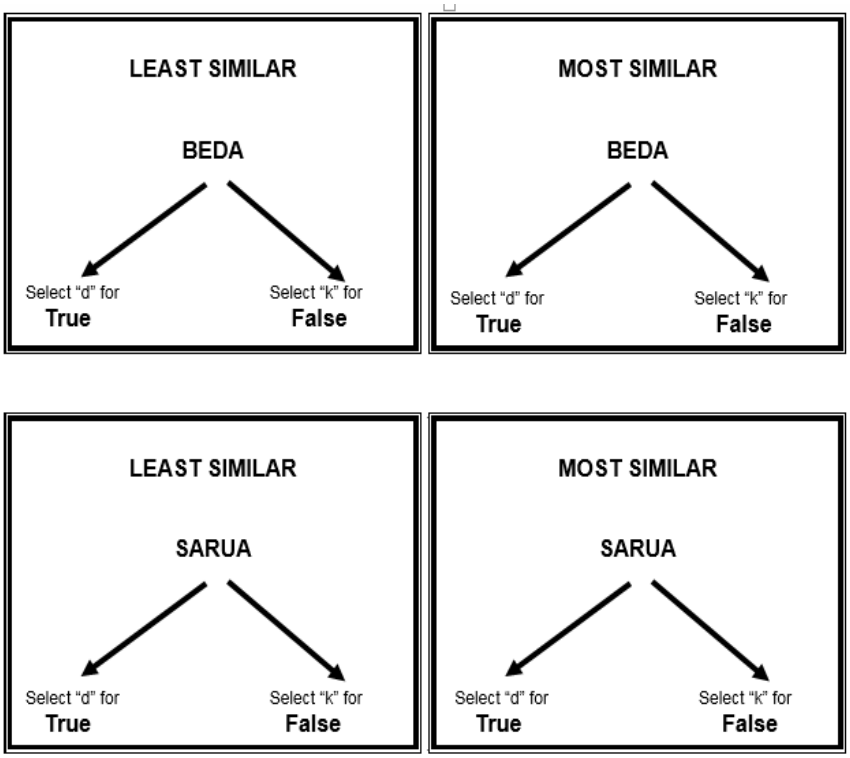


Figure 1. Diagrammatic representation of the four IRAP trial types. Arrows did not appear on screen. The four IRAP trial types were denoted as: *Least Similar-Beda, Most Similar-Beda, Least Similar-Sarua, Most Similar-Sarua.*

2.1.3.2. 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 2 for an illustration of a trial). Each comparison varied in its similarity to the sample. That is, one comparison was clearly the *most similar to the sample* (same basic shape with minor variations, see right-hand side of Figure 2). Another comparison was also clearly like the sample, but had more variations in shape (see left-hand side of Figure 2), rendering it *less similar to* the sample. Finally, the third comparison was clearly the *least similar to* the sample because it comprised a different shape, with little or no overlapping features (middle of Figure 2). 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.

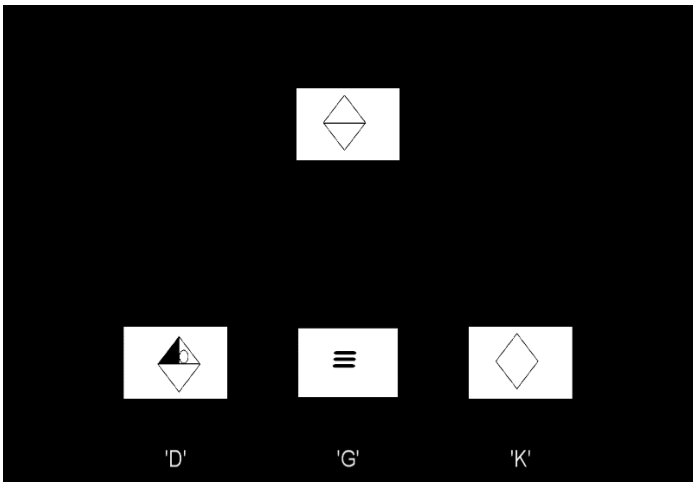


Figure 2. An example of a single trial and single stimulus set presented in the MTS task.

2.1.3.3. Questionnaires

Experiment 1 involved four self-report questionnaires, three standardised measures (the Depression, Anxiety and Stress Scales, DASS-21; the Acceptance and Action Questionnaire, AAQ-II; and the Scale for Personality Rigidity, SPR) and one developed for current purposes, referred to as the Propensity to Rule-Following Scale (PRFS).

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 may be calculated by summing all 21 items. However, it is important to note that according to Lovibond and Lovibond, all overall and subscale scores obtained on the DASS-21 must be doubled, and severity bands are generated accordingly. That is, the overall DASS

score ranges from 0-126 (where the latter is a score of 63 doubled). The severity bands for the overall scores are as follows: Normal: 0-30; Mild: 31-40; Moderate: 41-59; Severe: 60-79; Extremely Severe: 80+. The severity bands for the depression subscale are as follows: Normal: 0-9; Mild: 10-13; Moderate: 14-20; Severe: 21-27; Extremely Severe: 28+. The severity bands for the anxiety subscale are as follows: Normal: 0-7; Mild: 8-9; Moderate: 10-14; Severe: 15-19; Extremely Severe: 20+. The severity bands for the stress subscale are as follows: Normal: 0-14; Mild: 15-18; Moderate: 19-25; Severe: 26-33; Extremely Severe: 34+. Higher scores on the overall score and 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.

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.; 2011). 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.

The *SPR* measures personality rigidity across 37 statements (e.g. "My painful memories prevent me from having a fulfilling life"; Rehfisch, 1958). All items were rated on a "Yes" or "No" basis, yielding a minimum score of 0 and a maximum of 37. High scores indicate high rigidity, while low scores indicate low rigidity. Two items from the original 39-

item SPR were removed (i.e. Items 23 “I do not like to see women smoke” and 35 “Many of the girls I knew in college went with a fellow only for what they could get out of him”) on the basis that they were deemed irrelevant in contemporary life and potentially offensive. In the absence of an available Dutch translation of the measure, all 37 items were forward-backward translated from English to Dutch. That is, they were first translated into Dutch by a native bilingual Dutch speaker and subsequently translated from Dutch back to English by a different bilingual native Dutch speaker. This latter translation was then compared to the original English version and differences between the two versions were resolved and checked for fluency.

The *PRFS* was created for current purposes to assess propensity to rule-following across 5 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”; “If I was given a rule to follow and the rule proved to be incorrect, I would abandon the rule”). All items are rated on a 5-point scale from 1 (*Always agree*) to 5 (*Always disagree*), yielding a minimum score of 5 and a maximum of 25. High scores indicate high propensity for rule-following, while low scores indicate low propensity for rule-following. The PRFS was simply created as a bespoke instrument for the current study, and thus no formal psychometric properties were derived for the scale.

2.1.4. Procedure

Experiment 1 comprised of four stages that commenced with two questionnaires in Stage 1; the Training IRAP in Stage 2, with familiarisation blocks in Phase 1 and training blocks in Phase 2; the MTS task in Stage 3, with rule-consistent contingencies in Phase 1 and rule-inconsistent contingencies in Phase 2; and the remaining questionnaires in Stage 4.

2.1.4.1. Stage 1: DASS-21 and AAQ-II

Participants completed the DASS-21 and the AAQ-II in that order.

2.1.4.2. Stage 2: The Training IRAP

2.1.4.2.1. Phase 1: Familiarisation Blocks

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 word in the centre, and that their task was to relate these together using one of the two response options as accurately as possible across each block. All participants completed at least one initial familiarisation block of trials. The Training IRAP consisted of blocks of 36 trials, with each of the 4 trial types presented 9 times within each block. Participants selected a response option by pressing *D* (for the left option) or *K* (for the right option). If a correct response was emitted, the screen cleared and the next trial appeared 400ms later. If an incorrect response was emitted, a red X appeared until a correct response was emitted. If participants failed to achieve accuracy (≥ 80%) and/or latency criteria (≤3000 ms) on each of the four trial types during the initial familiarisation block, they received automated feedback and additional blocks were presented until the criteria were reached.

2.1.4.2.2. Phase 2: Training Blocks

The training blocks were identical to the familiarisation blocks and only commenced when the mastery criteria had been reached on the latter. The actual number of training blocks to which any participant was exposed depended upon the condition to which they were assigned. Specifically, participants in the High Derivation Condition received only one training block, while participants in the Low Derivation Condition received 15 training blocks. During the training blocks, no performance criteria applied to any participant.

2.1.4.3. Stage 3: MTS Task

At the beginning of the MTS task, participants were instructed that the aim of the task was to accrue as many points as possible. They were then instructed on the basic format of

each trial in terms of the presentation of a shape at the top of the screen and three shapes on the bottom of the screen. The next instruction depended upon whether during the Training IRAP a participant had learned to coordinate "Beda" or "Sarua" with "Least Similar". Specifically, participants who had previously been trained to coordinate "Beda" with "Least Similar" were now instructed to "Respond by selecting the shape that is *Beda* the sample stimulus". Conversely, participants who had previously been trained to coordinate "Sarua" with "Least Similar" were now instructed to "Respond by selecting the shape that is *Sarua* the sample stimulus". The total MTS task comprised of 150 trials, 100 trials presented in Phase 1 and 50 trials presented in Phase 2.

2.1.4.3.1. 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 centre 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.

2.1.4.3.2. 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.

2.1.4.4. Stage 4: SPR and PRFS

After the MTS task, participants completed the SPR and the PRFS in that order.

2.2. Results

For the purposes of analysis, exclusion criteria were applied to the training blocks of the IRAP. In the High Derivation Condition, the data from two participants were removed because they failed to maintain ≥75% accuracy and ≤3500ms response latency per trial-type across the single training block to which they were exposed (N = 39 remaining). In the Low Derivation Condition, the data from eight participants were removed because they failed to maintain these criteria across the final 10 of the 15 training blocks to which they were exposed (N = 39 remaining). A strict accuracy criterion was also applied to the MTS task, which required correct responding on at least 8 of the first 10 and 80 of the first 100 trials in Phase 1, aimed at reducing the likelihood that participants learned to match based on trial and error (18 participants were removed on this basis, 9 from the High Derivation Condition and 9 from the Low Derivation Condition, N = 60 remaining). Although the relatively strict criteria led to the removal of many participants, it was deemed very important that participants in both the High and Low Derivation Conditions performed equally well from the very beginning of the MTS task (i.e. at least 8 out of the first 10 MTS trials correct). Any difference between the two conditions at the beginning of the MTS task might indicate that one group learned to respond more through trial and error on the MTS task itself, than through derivation based on the previous IRAP training.

Before conducting the primary analyses, the number of familiarisation blocks (Phase 1) that participants received before they progressed to the training blocks (Phase 2) of the Training IRAP for each condition was compared. Participants in the Low Derivation Condition took an average of 2.33 (SD = .80), while participants in the High Derivation Condition took an average of 2.77 (SD = 1.41) blocks. An independent t-test confirmed that this difference was not significant, $t(58) = -1.47$, $p = .15$. Thus, any subsequent differences that emerged between the two groups during the training blocks of the 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.

Insofar as the primary aim of Experiment 1 was to compare performances between the Low and High Derivation Conditions, the data from the 50 trials in Phase 2 of the MTS task presented after the contingency reversal were analysed in three ways. These three types of analyses 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/Sarua* [Least Similar] the sample stimulus", but were inconsistent with the reversed contingencies on the last 50 trials. Figure 3 (left-hand side) presents the data for rule compliance for Low and High Derivation Conditions, divided according to whether the novel word Beda or Sarua was trained to "Least Similar." Unexpectedly, the descriptive statistics revealed a clear difference between conditions in rule compliance for Beda, but not for Sarua. Specifically, the Low Derivation Condition produced higher levels of rule compliance than the High Derivation group, but only for Beda. A 2x2 (condition x novel word) between group analysis of variance (ANOVA) produced a marginally significant interaction effect, $F(1, 56) = 3.81$, $p = .056$, $\eta_p^2 = .07$. At this point, we decided to recruit a set number of additional participants (15 in each condition, thus allowing for potential attrition due to "no shows" and failure to meet performance criteria) to determine if the interaction became statistically significant. When the additional participants were included in the analyses the interaction did indeed reach significance, but in the interests of statistical fidelity only the data collected from the original 60 participants were analysed and presented here.

Post-hoc analyses in the form of independent *t*-tests confirmed a significant difference in rule compliance scores between the Low ($M = 30.40$, $SD = 18.43$) and High Derivation Conditions ($M = 15.00$, $SD = 10.54$) when "Beda" was trained to "Least Similar", $t(28) =$

2.81, p = .009, Cohen's d = 1.03. When "Sarua" was trained to "Least-Similar", however, the difference was non-significant; Low (M = 20.60, SD = 16.77) versus High Derivation (M = 21.33, SD = 16.39), t (28) = -.09, p = .93, Cohen's d = .04.

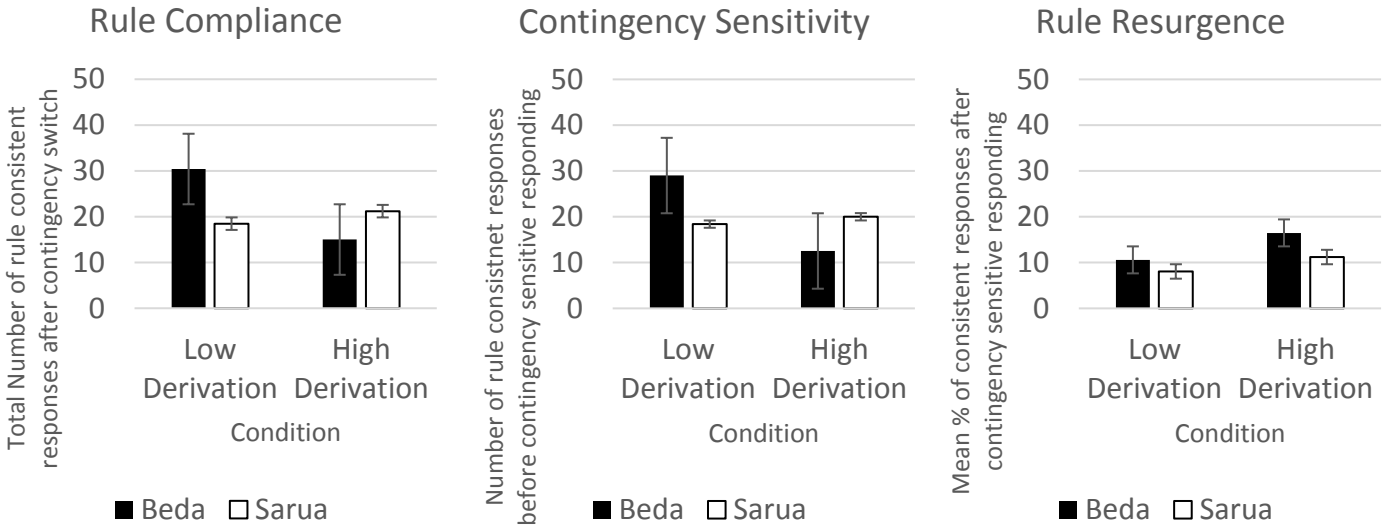


Figure 3. Experiment 1: Mean rule compliance scores (left panel), contingency sensitivity scores (center panel), and rule resurgence scores (right panel) with standard error bars, for the Low Derivation and High Derivation Conditions split for specific word used to mean "Least Similar" within each condition.

Consistent with Harte, et al. (2017), *contingency sensitivity* was defined as a pattern comprising of at least 3 consecutive responses not in accordance with the rule, with at least 1 of these responses being in accordance with the reversed contingency. In principle, therefore, a participant could stop following the rule and choose the stimulus that also lost points (i.e. the stimulus that was "mid-way" between the most like and least like the sample), but could only do this for 2 of the 3 responses. Including this requirement ensured that the term "contingency sensitivity" was appropriate, given that a participant must obtain at least one 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 3 responses (gaining 3 points), hence showing contingency sensitivity.

Figure 3 (center panel) again shows a similar pattern for contingency sensitivity, as was obtained for rule compliance; that is, a potential interaction effect between condition and novel word. Indeed, a 2x2 between group ANOVA again indicated a marginally significant interaction, $F(1, 56) = 4.05$, $p = .051$, $\eta_p^2 = .07$. When the data from the additional participants were included, the interaction became significant, but again only the data from the original 60 participants are reported here. Post-hoc analyses confirmed a significant difference in contingency sensitivity between the Low ($M = 29.00$, $SD = 18.65$) and High Derivation Conditions ($M = 12.53$, $SD = 4.49$) for "Beda", $t(28) = 3.33$, $p = .003$, Cohen's $d = 1.21$, but not for "Sarua"; Low ($M = 20.73$, $SD = 18.12$) and High Derivation ($M = 20.60$, $SD = 17.06$), $t(28) = .02$, $p = .98$, Cohen's $d = .007$.

Rule resurgence attempted to capture responding that was consistent with the initial rule (i.e. percentage of) but which occurred after three consecutive responses that were in accordance with the reversed contingencies (hence the term resurgence). This measure supplemented contingency sensitivity, which did not capture when participants' responding reverted to the pattern required by the initial instruction after sensitivity was shown. Figure 3 (right-hand panel) shows that on rule resurgence, there appeared to be only minor differences between the conditions, and indeed a 2x2 between group ANOVA yielded no significant effects (all p's > .49).

Given the interaction effects recorded on both rule compliance and contingency sensitivity, correlational analyses with the DASS, AAQ, SPR and PRFS were conducted separately for each of the four groups (Low and High Derivation for Beda, and Low and High Derivation for Sarua). Only one marginally significant correlation was obtained for the High Derivation Beda group: rule compliance correlated negatively with the SPR ($r = -.52$, $p = .05$), suggesting that increased compliance predicted lower levels of personality rigidity. There were no significant correlations with rule resurgence (correlations were conducted

across all four groups because the ANOVA yielded no significant effects). Given that only one marginally significant correlations emerged out of a total of 21, these results should be interpreted with extreme caution.

3. Experiment 2

The findings from Experiment 1 were generally consistent with those reported by Harte et al. (2017) and the suggestion that lower levels of derivation may produce more persistent rule-following, particularly when participants have many opportunities to follow a previously reinforced rule. Interestingly, the current findings suggested that persistence in rule-following was moderated by the type of novel word employed in the MTS task, with differential persistence only observed with "Beda" (rather than "Sarua"). A possible explanation for this difference is provided in the Discussion.

As noted in the Introduction, the focus in Experiment 1 was on mutually entailed relations (e.g. Beda directly co-ordinated with Least Similar). In Experiment 2, we sought to replicate the effect observed with Beda, but via combinatorially entailed relations. That is, would Beda, when it was co-ordinated with Least Similar via a mediating node (e.g. a nonsense word), produce greater persistence in the Low versus High Derivation group. Given that a differential effect was only observed with Beda in Experiment 1, Experiment 2 attempted to replicate the effect with Beda, rather than Sarua.

3.1. Method

3.1.1. Participants

A total of 98 individuals participated in Experiment 2, 75 females and 23 males. The aim here was to generate approximately 30 participants in each condition, similar to Experiment 1, allowing for attrition due to failure to meet the relevant performance criteria. Participants ranged in age from 18 to 39 years (M = 22.27, SD = 4.33), and were recruited through random convenience sampling from the online participant system at Ghent

University. All participants were Caucasian with Dutch as their first language and were paid 10 euros for participation. All were randomly assigned to one of two conditions, again referred to as Low and High Derivation. The data from 37 participants (13 from Low Derivation and 24 from High Derivation) were excluded because they failed to meet either the IRAP performance criteria or the MTS task criteria (leaving N = 61 for analysis, 31 in Low Derivation and 30 in High Derivation).

3.1.2. Setting

The setting was identical to that in Experiment 1.

3.1.3. Materials and Apparatus

Experiment 2 involved two computer-based tasks (a Training IRAP and an MTS task; the latter was similar to Experiment 1) and five self-report measures.

3.1.3.1. The Training IRAP

The Training IRAP in Experiment 2 was broadly similar to Experiment 1, except that it was designed to establish a relational network involving combinatorial relations between Beda and Least Similar. In abstract terms, the basic sequence involved training A-B relations, followed by B-C relations, and then a mixture of both A-B and B-C relations. When participants achieved the performance criteria on the familiarisation blocks (i.e. A-B and B-C relations), they received either one more, or 15 more, training blocks of the mixed A-B and B-C relations.

3.1.3.1.1. A-B Relations

The Training IRAP contained Dutch words and phrases, but again the English translations are used here. The presentation format of the Training IRAP was similar to Experiment 1. That is, all trials presented a label at the top of the screen, with a single target below. The label stimuli used to train the A-B relations always comprised of one of four phrases: “Least Similar”; “Differs Most”; “Most Similar”; and “Resembles Most”, see Table

2. The target stimulus was always “TTT” or “]][[”. Each pair of response options comprised of: “True” versus “False”; “Yes” versus “No”; “Correct” versus “Incorrect”; and “Right” versus “Wrong”. These stimuli were combined to generate four A-B trial types (see Figure 4) referred to as: Least Similar-TTT, Most Similar-TTT, Least Similar-]][[and Most Similar-]][[.

Table 2

Stimuli employed within the Training IRAP as labels, targets and response options.

Labels Stimuli	Target Stimuli	Response Options	
Least Similar	TTT	Yes	No
		True	False
	]][[	Right	Wrong
		Correct	Incorrect
Differs Most	TTT	Yes	No
		True	False
	]][[	Right	Wrong
		Correct	Incorrect
Most Similar	TTT	Yes	No
		True	False
	]][[	Right	Wrong
		Correct	Incorrect
Resembles Most	TTT	Yes	No
		True	False
	]][[	Right	Wrong
		Correct	Incorrect
TTT	Beda	Yes	No
		True	False
	Sarua	Right	Wrong
		Correct	Incorrect
]][[	Beda	Yes	No
		True	False
	Sarua	Right	Wrong
		Correct	Incorrect

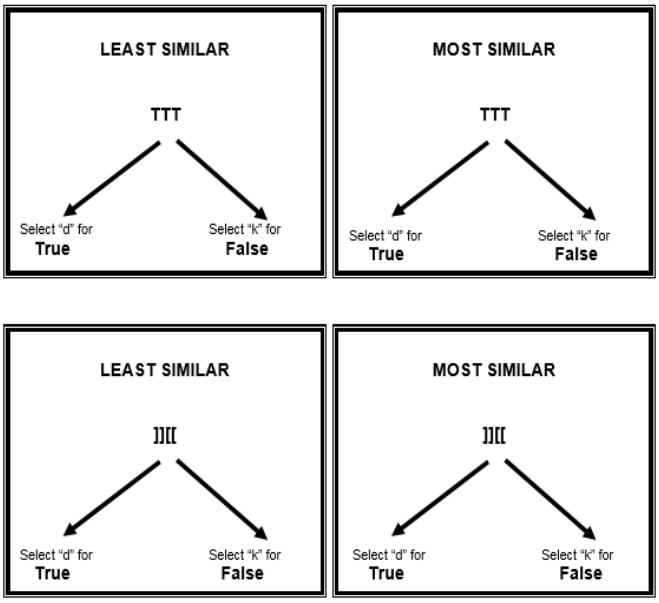


Figure 4. Diagrammatic representation of the IRAP trial-types that appear in A-B baseline relation familiarisation block. Arrows did not appear on screen. The four IRAP trial types were denoted as: *Least Similar-TTT, Most Similar-TTT, Least Similar-]][[, Most Similar-]][[.*

3.1.3.1.2. B-C Relations

During training of the B-C relations, each trial presented "TTT" or "]][[" from the A-B training, but these now appeared at the top of the screen, rather than as target stimuli in the middle. The target stimuli now comprised of the words "Beda" and "Sarua", presented in the middle of the screen, with the same response options as used for the A-B relations (see Table 2). The four B-C trial types were thus as follows: TTT-Beda,]][[-Beda, TTT-Sarua and]][[-Sarua (see Figure 5).

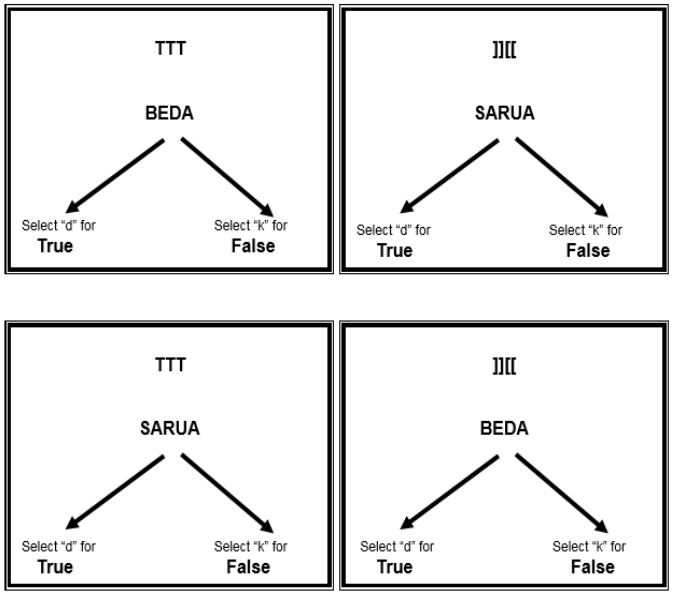


Figure 5. Diagrammatic representation of the IRAP trial-types that appear in B-C baseline relation familiarisation block. Arrows did not appear on screen. The four IRAP trial types were denoted as: *TTT-Beda,]][[-Beda, TTT-Sarua,]][[-Sarua.*

3.1.3.1.3. Mixed A-B and B-C Relations

The mixed training of A-B and B-C relations simply involved presenting A-B and B-C trials, but mixed quasi-randomly within each block.

3.1.3.2. Questionnaires

Experiment 2 involved five self-report questionnaires. The DASS-21, the AAQ-II and the PRF were retained from Experiment 1, while the SPR was excluded. The Psychological Flexibility Index and the Tenacious Goal Pursuit and Flexible Goal Adjustment Questionnaire were now added.

The *Psychological Flexibility Index* (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 are 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. Due to the fact that there is no

Dutch translation available of the PFI, all items were forward-backward translated into Dutch. Because the measure is still in development, there are currently no published validity or reliability data.

The *Tenacious Goal Pursuit and Flexible Goal Adjustment Questionnaire* (TENFLEX) is a 30-item scale with two sub-scales that measure two coping styles, each with 15 statements (Brandstadter & Renner, 1990). Tenacious goal pursuit contained statements such as "When I run up against overwhelming obstacles, I prefer to look for a new goal", while flexible goal adjustment contained statements such as "When I get stuck on something it's hard for me to find a new approach". All items are rated on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree). Each sub-scale was scored separately, with all 15 items in each summed to yield a total sub-scale score, with a minimum of 15 and a maximum of 90 on each sub-scale. High scores on tenacious goal pursuit indicate high tenacity in terms of pursuing goals, while high scores on the flexible goal adjustment subscale indicate high flexibility in terms of adjusting goals. The Dutch version of the TENFLEX was employed here, but there are no reliability data on this version. However, the English version has demonstrated adequate internal consistency with alpha coefficients ranging from 0.80 for the TEN and 0.83 for the FLEX.

3.2. Procedure

Experiment 2 comprised of four stages that commenced in Stage 1 with three questionnaires. In Stage 2, the Training IRAP again had two phases with familiarisation blocks in Phase 1 and training blocks in Phase 2, but Phase 1 now had three sub-phases: Sub-phase 1 presented A-B relations; Sub-phase 2 presented B-C relations; and Sub-phase 3 presented mixed A-B and B-C relations. Stage 3 again contained the MTS task, with rule-consistent contingencies in Phase 1 and rule-inconsistent contingencies in Phase 2. Finally,

Stage 4 presented the remaining two questionnaires (see Appendix A for a flowchart of the experimental sequence).

3.2.1. Stage 1: DASS-21, AAQ-II and PFI.

Participants completed the DASS-21, the AAQ-II and the PFI in that order in Stage 1.

3.2.2. Stage 2: The Training IRAP.

3.2.2.1. Phase 1: Familiarisation Blocks

The instructions, performance criteria and feedback for the familiarisation blocks were identical to Experiment 1. However, the familiarisation blocks now involved three sub-phases, one for each type of trained relation. Thus, participants were required to reach the mastery criteria with the A-B relations before proceeding to the B-C relations, and to reach the criteria with the B-C relations before proceeding to the mixed A-B and B-C relations.

Sub-phase 1: A-B Relations. The A-B trials always presented "Least Similar" or "Most Similar" (or their synonyms) as the label stimuli, with TTT and]][[as the target stimuli. Hence, the four trial-types were Least Similar-TTT; Most Similar-TTT; Least Similar-]][[; and Most Similar-]][[. Correct responding was as follows: Least Similar-TTT/True; Most Similar-TTT/False; Least Similar-]][[/False; and Most Similar-]][[/True. The A-B relations were presented in blocks of 24 trials, with 6 exposures to each trial-type, presented quasi-randomly within each block.

Sub-phase 2: B-C Relations. Training the B-C relations was similar in format to training the A-B relations, except that "TTT" or "]][[" were presented as the label stimuli, with "Beda" and "Sarua" as the target stimuli. Hence, the four trial-types were TTT-Beda; TTT-Sarua;]][[-Beda; and]][[-Sarua. Correct responding was as follows: TTT-BEDA/True,]][[-BEDA/False, TTT-SARUA/False and]][[-SARUA/True.

Sub-phase 3: Mixed A-B and B-C Relations. Training the mixed A-B and B-C relations was similar in format to the previous two sub-phases, except that mixed A-B and B-

C relations were presented in blocks of 32 trials, with 4 exposures to each A-B trial type and 4 exposures to each B-C trial type, presented quasi-randomly within each block. Participants could not proceed to Phase 2 until they had reached the mastery criteria on all three sub-phases of Phase 1.

3.2.2.2. Phase 2: Training Blocks

Having reached the mastery criteria in the Familiarisation blocks, participants were then re-exposed to the mixed A-B and B-C trials, but the number of blocks now depended upon the condition to which each participant had been assigned. That is, participants in the High Derivation Condition received only one additional training block of mixed A-B and B-C relations, while the Low Derivation Condition received an additional 15 blocks. As in Experiment 1, no performance criteria were applied in Phase 2 in order to proceed to Stage 3, but performance-contingent feedback was provided at the end of each block to encourage participants to maintain the performance criteria they had achieved in the Familiarisation blocks.

3.2.3. Stage 3: MTS Task.

The MTS task was identical in format to that employed in Experiment 1, with rule-consistent contingencies in Phase 1 and rule-inconsistent contingencies in Phase 2.

3.2.4. Stage 4: Questionnaires.

Finally, participants completed the PFI and the PRFS in that order in Stage 4.

3.3. Results

The exclusion criteria employed in the Training IRAP for the High Derivation Condition ≥75% accuracy and ≤3500ms latency per trial type in the single training block) were identical to Experiment 1, and resulted in the removal of all data from three participants. Given that Experiment 2 involved combinatorial, rather than mutual entailment alone, the criteria were relaxed for the Low Derivation Condition. Specifically, participants were now

required to maintain the criteria across the final 5 (rather than 10) of the 15 training blocks.[2] The accuracy criteria employed in the MTS task were identical to Experiment 1 and the data from 34 participants were removed when these individuals failed to meet the criteria (21 in the High Derivation Condition and 13 in the Low Derivation Condition; $N = 61$ remaining). Taken together, a total of 31 participants in the Low Derivation Condition and 30 participants in the High Derivation Condition were included in the analyses. Although the relatively strict criteria led to the removal of many participants (34), it was deemed very important that participants in both the High and Low Derivation Conditions performed equally well from the very beginning of the MTS task (i.e. at least 8 out of the first 10 MTS trials correct). Any difference between the two conditions at the beginning of the MTS task might indicate that one group learned to respond more through trial and error on the MTS task itself, than through derivation based on the previous IRAP training.

Before conducting the primary analyses, the number of Familiarisation blocks (Phase 1, sub-phases 1, 2 and 3) that participants received before they progressed to the training blocks (Phase 2) of the Training IRAP for each condition was compared. On the A-B relations, participants in the Low Derivation Condition took an average of 2.17 (*SD* = 1.05), while participants in the High Derivation Condition took an average of 2.80 (*SD* = 2.31) blocks. On the B-C relations, participants in the Low Derivation Condition took an average of 1.63 (*SD* = .89), while participants in the High Derivation Condition took an average of 1.57 (*SD* = .63) blocks. Finally, on the mixed A-B and B-C relations, participants in the Low Derivation Condition took an average of 1.13 (*SD* = .35) blocks, while participants in the High Derivation Condition took an average of 1.20 (*SD* = .48) blocks. Overall, across all of

[2] In the Low Derivation Condition, one participant failed to maintain the "relaxed" criteria in the Training IRAP. Specifically, P72 produced an accuracy score of 62.5% on trial type 3 of Block 13. These data were included in the initial analyses and then recalculated with the data removed. Removing the dataset did not change any of the analytical or statistical conclusions, and thus the data from this participant were retained in the analyses reported above.

these sub-phases, participants in the Low Derivation Condition took an average of 4.93 (*SD* = 1.39) blocks, while participants in the High Derivation Condition took an average of 5.57 (*SD* = 2.36) blocks. Independent *t*-tests confirmed that the differences between the performances at each sub-phase, and overall, were not significant (all *p*'s > .18). Thus, any subsequent differences that emerged between the two groups during the training blocks of the IRAP or the MTS task are not likely due to differences in the ability to learn how to respond on the IRAP per se.

Rule compliance scores are presented in Figure 6 (left panel) and show differential levels of rule compliance across conditions. That is, participants in the Low Derivation Condition emitted more responses (*M* = 21.74, *SD* = 18.09) in accordance with the original instruction than the High Derivation Condition (*M* = 13.50, *SD* = 10.82), and an independent *t*-test confirmed this difference to be significant, *t*(59) = 2.15, *p* = .04, Cohen's *d* = .56.

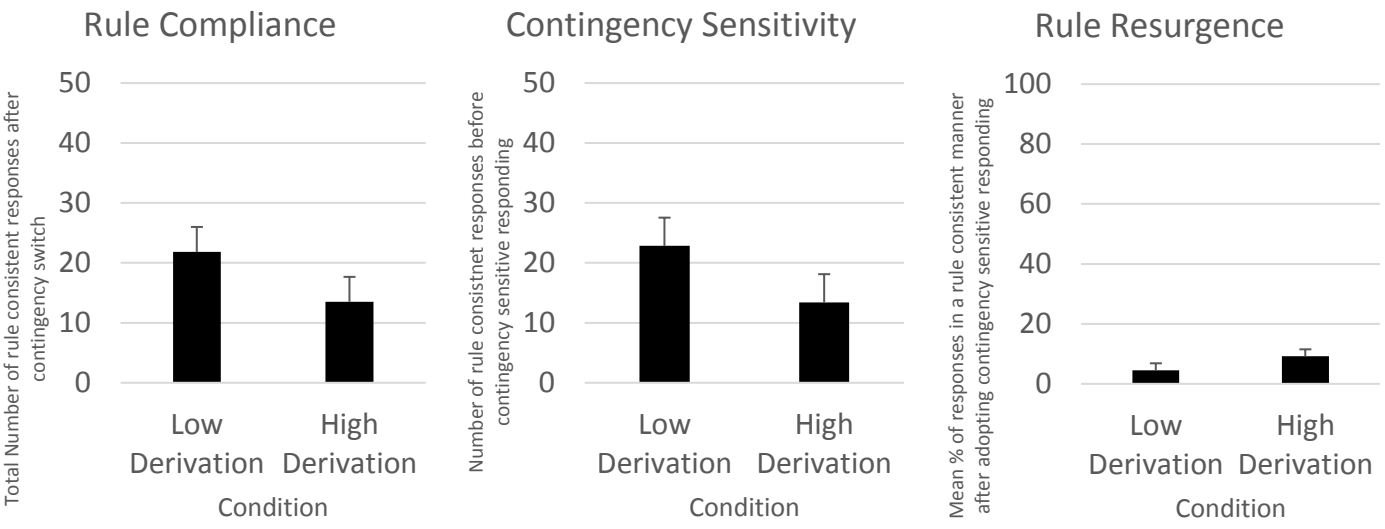


Figure 6. Experiment 2: Mean rule compliance scores (left panel), contingency sensitivity scores (centre panel), and rule resurgence scores (right panel) with standard error bars for the Low Derivation and High Derivation Conditions.

Contingency sensitivity scores are presented in Figure 6 (centre panel) and also show a difference between conditions. Specifically, participants in the Low Derivation Condition

completed more trials (*M* = 22.71, *SD* = 18.33) before responding in accordance with the new contingencies than the High Derivation Condition (*M* = 13.40, *SD* = 9.64). An independent *t*-test again confirmed this difference to be significant, *t*(59) = 2.47, *p* = .02, Cohen's *d* = .64.

Rule resurgence scores are presented in Figure 6 (right panel) and again show differences between conditions. Participants in the High Derivation Condition emitted a greater percentage of responses (*M* = 9.18%, *SD* = 12.55) in accordance with the original instruction *after* having been deemed to have switched to contingency consistent responding than the Low Derivation Condition (*M* = 4.68%, *SD* = 3.98). An independent *t*-test proved to be marginally significant, *t*(59) =-1.89, *p* = .06.

Given the significant differences recorded for both rule compliance and contingency sensitivity, and the marginally significant difference for rule resurgence, correlational analyses with the DASS, AAQ, PFI, TENFLEX and PRFS were conducted separately for each group (Low and High Derivation) on each measure. Significant correlations were only found with the High Derivation Condition. Specifically, both rule compliance (*r* = .39, *p* = .03) and contingency sensitivity (*r* = .45, *p* = .01) correlated positively with depression, suggesting that more persistence with the original instruction predicted higher depression. For rule resurgence, a significant negative correlation was found with the PFI (*r* = -.48, *p* = .007), suggesting that increasing resurgence after contingency-sensitive responding predicted less flexibility.

4. Discussion

Harte et al. (2017) highlighted the potential utility of integrating the existing literatures on the contingency insensitivity effect and derived relational responding. In an exploratory study, they sought to determine whether participants would persist in rule-following in the face of reversed contingencies, and whether this rule-following differed between rules that did or did not require derived relational responding based on prior learning within the experiment.

The results of the second experiment demonstrated that the provision of a direct rule (i.e. no novel derivations required within the experiment) resulted in more persistent rule-following than a rule that did require within-experiment derivation. Harte et al. suggested that the difference observed in the study may be based, at least in part, on the levels of derivation involved. Specifically, they speculated that the level of derivation would be relatively low in the direct rule condition and relatively high in the condition that required novel derivations within the experiment. The current study sought to test this suggestion by manipulating levels of derivation within the experiments. That is, both conditions within each experiment required novel derivations in rule-following, but the amount of training involved in establishing those derivations differed across conditions. In addition to manipulating amount of training, the current study explored the impact of derivation, within the rules, for mutual (Experiment 1) and combinatorial (Experiment 2) relations. Overall, the results indicated that rule-following was more persistent when rules were low relative to high in derivation (based on within-experimental training); and this finding applied to both mutual and combinatorial relations.

An unexpected interaction effect emerged in Experiment 1 when rule persistence was observed with "Beda", but not with "Sarua". We did not pursue this effect in Experiment 2, but it seems important to briefly consider it here, although of course any explanation we offer will remain speculative until further empirical work is conducted. Informal post-hoc verbal reports provided by some participants indicated that "Sarua" seemed positive, while "Beda" seemed negative, and similarly that "Most Similar" seemed positive, while "Least Similar" seemed negative (see Spence, 2011 for a review of cross-modal correspondence effects). Thus, participants reported that it was easier to pair "Sarua" with "Most Similar" (i.e. both were perceived as positive) and to pair "Beda" with "Least Similar" (i.e. both were perceived as negative). If this effect was widespread among participants, it may have impacted upon the level of coherence involved in the MTS task (the potential interaction between levels of

derivation and coherence is broadly consistent with points raised in the Discussion section of Harte et al., 2017; see also Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017). Specifically, when "Beda" functioned as "Least Similar" and "Sarua" as "Most Similar", coherence was high, but when the opposite applied, coherence was low. In other words, the task cohered more with the extra-experimental functions of the stimuli during the first 100 trials when "Beda", rather than "Sarua", meant "Least Similar". If this was the case, it suggests that the differential effect we observed in Experiment 1 for levels of derivation was moderated by coherence. That is, derivation impacted upon persistent rule-following only in the context of high, but not low, coherence during the first 100 trials. More informally, when the task appeared less coherent from the beginning (i.e. when "Sarua" meant "Least Similar" and "Beda" meant "Most Similar"), all participants readily changed their responding on the MTS task when the contingencies reversed (i.e. because the task now made more sense). However, when the task appeared relatively coherent from the start (i.e. when "Sarua" meant "Most Similar" and "Beda" meant "Least Similar"), participants' behaviour was more influenced by a difference in derivation than coherence (i.e. because the task made less sense after the contingency switch, prior learning within the experiment had a greater impact on performance).

Although the interaction between level of derivation and novel word was unexpected, it could be seen as quite informative. For example, it could be argued that the difference observed in persistence in rule-following between the high and low derivation conditions was based on perhaps trivial differences between the conditions, such as different levels of boredom, fatigue or distraction. That is, participants in the low derivation condition may have experienced higher levels of boredom, fatigue or distraction simply because they were required to complete many more training blocks than participants in the high derivation condition. If this was the case, however, then there should have been little, if any, impact of

the nonsense word on persistent rule-following. Of course, we cannot conclude that extraneous factors played no role at all and it would be wise for future research to control for these.

The findings from the present study indicate that levels of derivation impact upon persistent rule-following in the context of mutual (Experiment 1) and combinatorial (Experiment 2) entailment. However, it is interesting to note that participants in the low derivation condition in Experiment 1 showed more persistence in rule compliance (M = 30.40) than participants in the low derivation condition in Experiment 2 (M = 21.74). However, no such difference was observed between the two high derivation conditions (Experiment 1 M = 15 and Experiment 2 M = 13.40). A similar difference between Experiments 1 and 2 was also observed in contingency sensitivity. It would not be appropriate to conduct a formal statistical analysis across the two separate experiments. However, the difference is worth noting because it suggests that there may be an interaction effect between level of derivation (low versus high) and level of relational development (mutual versus combinatorial entailment). Perhaps future research could examine this issue directly.

Another issue worth noting is that attrition rates could be considered relatively high in the current study. On balance, participants were required to meet quite stringent performance criteria across both the IRAP and MTS tasks to ensure that differences in participant performances following the contingency switch in the MTS task did not result from individual differences in their relative abilities to perform on both of the key tasks (i.e. the IRAP and the MTS task). For example, if less stringent criteria were applied to avoid high attrition rates, it could be argued that subsequent differences in the persistence of rule-following were influenced by extraneous variables, such as ability or willingness to engage with the tasks. Parenthetically, it is worth noting that the points at which participants failed to reach the various performance-related criteria across the IRAP and MTS tasks did not appear to differ

in any systematic or substantive way between conditions within experiments or even across experiments.

A related issue concerns the fact that in neither experiment did the groups differ significantly in terms of how many familiarisation blocks were needed to progress to the training blocks of the IRAP. Given that performance on the IRAP has been shown to correlate with measures of intelligence (e.g. O'Toole & Barnes-Holmes, 2009), it seems unlikely that any of the differences observed in the persistence of rule-following in the current study were due largely to individual differences in intellectual ability. Of course, future studies may include formal measures of intelligence, but the impact upon participant fatigue in adding more measures to what is already a relatively tiresome procedure would need to be considered.

Another related issue concerning the strict accuracy criteria that were required in the current study was the need to ensure that all participants entering the MTS task did not learn to perform on that task through trial and error, but had more or less made the necessary derivations (e.g. Beda means Least Similar) to complete the task successfully. Interestingly, the results of a recent study by Kissi et al. (in press) highlight how unlikely trial and error learning on the MTS task was in the current research. Specifically, using a similar paradigm, Kissi et al. found that all but one participant spontaneously chose the "Most Similar" comparison stimulus on the first trial on the MTS task when no instruction was provided. In contrast, in the current study out of a total of 121 participants, 89.26% chose the correct "Least Similar" comparison on the first trial and 95.87% within the first two trials. Clearly, therefore, the "natural" bias to pick the "Most Similar" comparison observed in the Kissi et al. study was almost completely absent in the context of a derived rule that specified the "Least Similar" comparison as the correct stimulus.

Previous research on rule persistence has frequently been linked to the study of human "psychopathology" and depression in particular (e.g. McAuliffe et al., 2014). Although this focus was not inherent in the current research, participants were required to complete a number of self-report measures related to psychological distress. In Experiment 1, there were no significant correlations between persistent rule-following and the self-reports, and only one marginally significant correlation (1 out of 21), but the *N* used for the correlational analyses was relatively low because the sample was split according to level of derivation and novel word (i.e. because of the unexpected influence of novel word). In Experiment 2, the number of significant correlations remained extremely low (i.e. 3 out of 21), although the *N* was now higher (30-31 per group). Interpreting so few correlations must be done with extreme caution, but it is worth noting that that higher levels of self-reported depression and lower levels of psychological flexibility correlated positively with increased rule persistence, but only in the High Derivation condition. This finding is broadly consistent with the argument that excessive rule-following is associated with psychological distress. On balance, the correlations were restricted to the High Derivation condition, which suggests, if only tentatively, that the relationship between excessive rule-following and distress is more complex than originally thought. Perhaps future research could pursue this matter further.

In closing, the current study has once again demonstrated the impact of level of derivation on persistent rule-following, similar to that reported by Harte et al. (2017). Unlike the previous study, however, level of derivation was manipulated directly within the experiments (i.e. we did not rely upon a "direct rule" condition). On balance, the findings here could be interpreted in terms of the impact of over-training (15 blocks) versus under-training (1 block) the baseline relations for mutual entailment (Experiment 1) and combinatorial entailment (Experiment 2). One way in which it may be possible to address this issue would be to give both groups equal numbers of blocks but have the high derivation group get

exposure to the stimuli of interest in only one block. On balance, and as noted in the Introduction, describing this difference in amount of training as levels of derivation is conceptually consistent with the way the term derivation itself has been used in the RFT literature. Nonetheless, future research may attempt to manipulate levels of derivation in a different manner. For example, an alternative strategy might involve providing participants in one condition with an opportunity to derive the relationship between "Least Similar" and "Beda" before entering the MTS task and then comparing this with a second condition in which participants are simply re-exposed to the training trials. In such a study, the number of training trials completed could be kept constant across conditions, but derivation would be required before the MTS task in the first condition but not in the second. Our research group is currently pursuing this and related lines of inquiry.

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).

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 analysing the dynamics of arbitrarily applicable relational responding. *Journal of Contextual Behavioral Science, 6*(4), 473-483.

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.

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., 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.

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.

Brandtstädter, J., & Renner, G. (1990). Tenacious goal pursuit and flexible goal adjustment: Explication and age-related analysis of assimilative and accommodative strategies of coping. *Psychology and Aging, 5*, 58-67.

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

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.

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.

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

Hayes, S.C. (1993). Rule governance: Basic behavioural research and applied applications. *Current Directions in Psychological Science, 2*, 193-197.

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., 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.

Hojo, R. (2002). Effects of instructional accuracy on a condition discrimination task. *The Psychological Record, 52*(4), 493-506.

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: John Wiley & Sons, Ltd.

Kissi, A., Hughes, S., De Schryver, M., De Houwer, J., & Crombez, G. (In press). Examining the modrerating impact of plys and tracks on the insensitivity effect: A preliminary investigation. *The Psychological Record.*

Kroger-Costa, A., & Abreu-Rodrigues, J. (2012). Effects of historical and social variables on instruction following. *The Psychological Record, 62*(4), 691-705.

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

Martinez-Sanchez, H. & Ribes-Inesta, E. (1996). Interactions of contingencies and instructional history on conditional discrimination. *The Psychological Record, 46*(2), 301-318.

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.

Monestes, J.L., Greville, W.J., & Hooper, N. (2017). Derived insensitivity: Rule-based to contingencies propogates through equivalence. *Learning and Motivation, 59*, 55-63.

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.

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.

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

Rehfisch, J.M. (1958). A scale for personality rigidity. *Journal of Consulting Pyschology, 22*(1), 11-15.

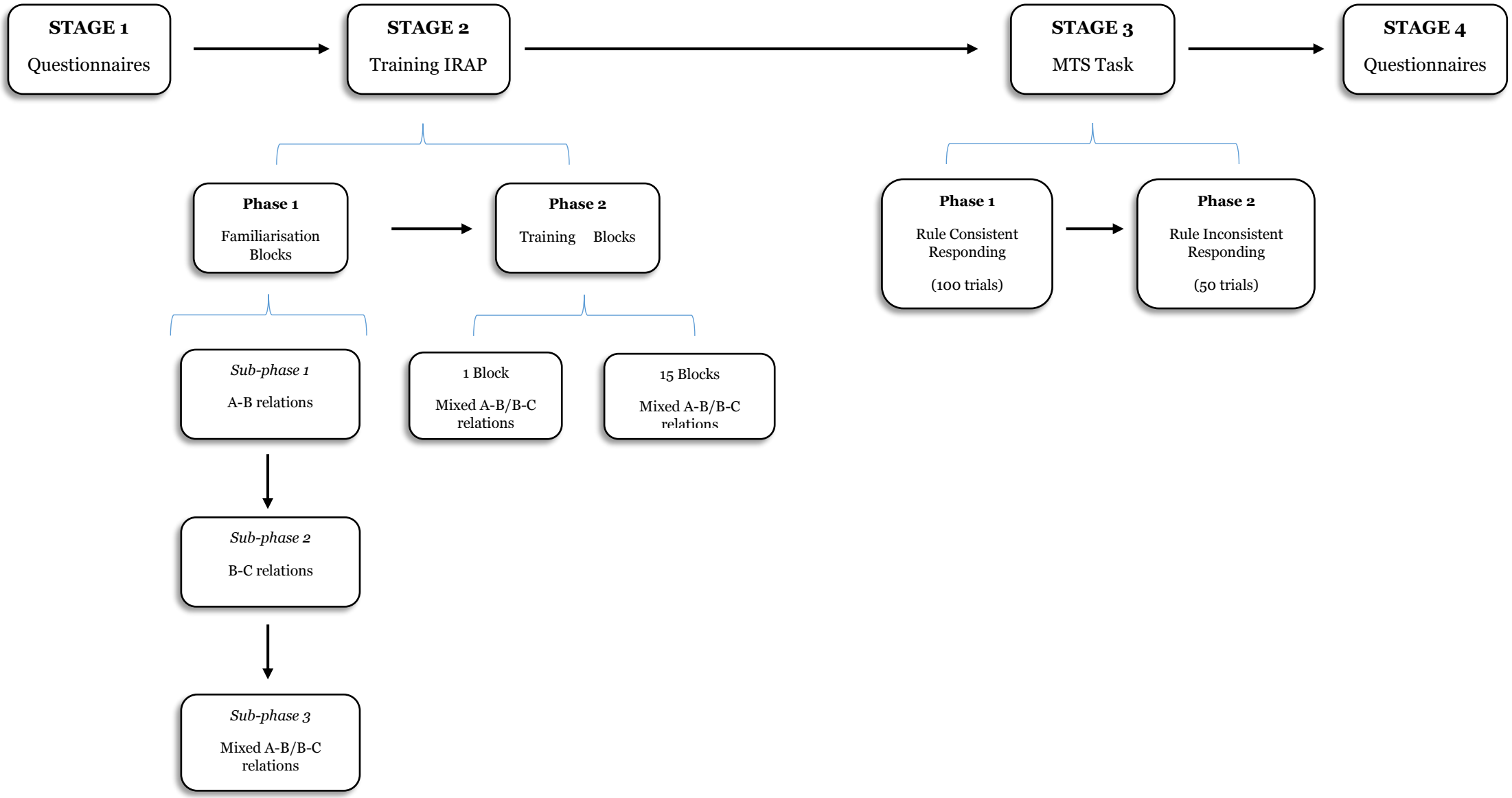
Rosenfarb, I.S., Newland, M.C., Brannon, S.E., & Howey, D.S. (1992). Effects of self-generated rules on the development of schedule-controlled behaviour. *Journal of the Experimental Analysis of Behavior, 58*(1), 107-121.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech, Language, and Hearing Research, 14*, 5-13.

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

APPENDIX A


STAGE 1
Questionnaires
STAGE 2
Training IRAP
STAGE 3
MTS Task
STAGE 4
Questionnaires
Phase 1
Familiarisation Blocks
Phase 2
Training Blocks
Sub-phase 1
A-B relations
Sub-phase 2
B-C relations
Sub-phase 3
Mixed A-B/B-C relations
1 Block
Mixed A-B/B-C relations
15 Blocks
Mixed A-B/B-C relations
Phase 1
Rule Consistent Responding
(100 trials)
Phase 2
Rule Inconsistent Responding
(50 trials)