Persistent Rule-Following in the Face of Reversed Reinforcement Contingencies: The Differential Impact of Direct Versus Derived Rules

Colin Harte, Yvonne Barnes-Holmes, Dermot 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 third author by the Flanders Science Foundation (FWO). Correspondence concerning this article should be sent to Colin.Harte@UGent.be

Author Biographical Statements

Colin Harte:

Colin Harte is a doctoral researcher at the Department of Experimental, Clinical and Health Psychology at Ghent University since October 2015. His research is interested in how RFT can be used to develop the basic account of rule governed-behavior and the potential effects of derived rules on creating/maintaining maladaptive behaviour.

Prof. Yvonne Barnes-Holmes:

Dr. Yvonne Barnes-Holmes is a Senior Research Fellow and Associate Professor at the Department of Experimental, Clinical and Health Psychology, Ghent University. She has published 120+ articles and book chapters and given 400+ talks and workshops. She is a recognized World Trainer in Acceptance and Commitment Therapy (ACT).

Prof. Dermot Barnes-Holmes:

Dr. Dermot Barnes-Holmes has worked as a tenured professor at the Department of Experimental, Clinical and Health Psychology at Ghent University since October 2015, and is known internationally for the development of RFT. He was the world's most prolific author in the experimental analysis of human behavior between 1980 and 1999.

Dr. Ciara McEnteggart:

Ciara McEnteggart is a postdoctoral researcher at the Department of Experimental, Clinical and Health Psychology at Ghent University since October 2015. She has 13 published articles which centre around the conceptual development of Relational Frame Theory (RFT) and the understanding of human psychological suffering and its alleviation.

Abstract

Rule-governed behavior and its role in generating insensitivity to direct contingencies of reinforcement have been implicated in human psychological suffering. In addition, the human capacity to engage in derived relational responding has also been used to explain specific human maladaptive behaviors, such as irrational fears. To date, however, very little research has attempted to integrate research on contingency insensitivity and derived relations. The current work sought to fill this gap. Across two experiments participants received either a direct rule (Direct Rule Condition) or a rule that involved a novel derived relational response (Derived Rule Condition). Provision of a direct rule resulted in more persistent rule-following in the face of competing contingencies, but only when the opportunity to follow the reinforced rule beforehand was relatively protracted. Furthermore, only in the Direct Rule Condition were there significant correlations between rule compliance and stress. A post-hoc interpretation of the findings is provided.

Behavior analysts have studied human language and cognition as derived relational responding for some decades. The basic paradigm was first demonstrated in the phenomenon of stimulus equivalence (see Sidman 1994, for a book-length review). This basic effect is defined as the emergence of unreinforced or untrained matching responses based on a small set of trained responses. For example, when a person is trained to match two abstract stimuli to a third (e.g., Zid-Paf and Zid-Vek), untrained matching responses frequently appear in the absence of additional learning (e.g., Paf-Vek and Vek-Paf). When such a pattern of unreinforced responses occurs, the stimuli are said to form an equivalence class or relation. Importantly, this behavioral effect appeared to provide a functional analysis of symbolic meaning or symbol-referent specification (Sidman, 1994, p. 561-563). In other words, functionally speaking, a stimulus could only be defined as *specifying* another stimulus if it participated in an equivalence class with that other stimulus.

The relationship between stimulus equivalence and human language, and cognition more broadly, emerged in the form of relational frame theory (RFT, Hayes, Barnes-Holmes, & Roche, 2001). Specifically, the theory argued that equivalence relations constituted just one type of overarching or generalized operant of arbitrarily applicable relational responding (AARR). According to this view, exposure to an extended history of relevant reinforced exemplars serves to establish particular patterns of relational response units, defined as relational frames (Barnes-Holmes & Barnes-Holmes, 2000). For example, a young child would likely be exposed to direct contingencies of reinforcement by the verbal community for pointing to the family dog upon hearing the word "dog" or the specific dog's name (e.g., "Rover"), and to emit other appropriate naming responses, such as saying "Rover" or "dog" when the dog was observed, or saying "Rover" when asked, "What is the dog's name?" Across many such exemplars, involving other stimuli and contexts, eventually the operant class of coordinating stimuli in this way becomes abstracted, such that direct reinforcement

for all of the individual components of naming would no longer be required when a novel stimulus was encountered. Imagine, for example, that the child was shown a picture of an aardvark, the written word "aardvark", and was told its name (the spoken word "aardvark"). Subsequently, the child may say "That's an aardvark" when presented with the relevant picture or word, without any prompting or direct reinforcement for doing so. In other words, the generalized relational response of coordinating pictorial and spoken stimuli, and written words would be established, and directly reinforcing a subset of the relating behaviors "spontaneously" generates the complete set. Once this pattern of relational responding has been established, the generalized relational response could then be applied, given appropriate contextual cues.

Contextual cues were thus seen as functioning as discriminative for particular patterns of relational responding. The cues acquired their functions through the types of histories described above. Thus, for example, the phrase "that is a", as in "*That is a* dog" would be established across exemplars as a contextual cue for the complete pattern of relational responding (e.g., coordinating the word "dog" with actual dogs). Once the relational functions of such contextual cues were established in the behavioral repertoire of a young child, the number of stimuli that may enter into such relational response classes becomes almost infinite (Hayes & Hayes, 1989; Hayes et al., 2001).

The core analytic concept of the relational frame proposed by Hayes et al. (2001) provided a relatively precise technical definition of AARR. Specifically, a relational frame was defined as possessing three properties; mutual entailment (if A is related to B, then B is also related to A), combinatorial mutual entailment (if A is related B and B is related to C, then A is related to C, and C is related to A), and the transformation of functions (the functions of the related stimuli are changed or transformed based upon the types of relations into which those stimuli enter).

As an account of human language and cognition, RFT provided what is still considered today to be the basic operant unit involved in verbal behavior – the relational frame. However, the seminal text on RFT (Hayes, et al., 2001) also used this unit to provide functional-analytic accounts of specific domains of human language and cognition, and rule-governed behavior was one of these domains. According to RFT, a rule or instruction may be considered a network of relational frames typically involving coordination and temporal relations with contextual cues that transform specific behavioral functions. The simple instruction, "If the light is green, then go" involves frames of coordination between the words "light", "green" and "go" and the actual events to which they refer. In addition, the words "if" and "then" serve as contextual cues for establishing a temporal relation between the actual light and the act of actually going (i.e., first light then go). And the relational network as a whole serves to transform the functions of the light itself, such that it now controls the act of "going" whenever an individual who has been presented with the rule observes the light being switched on. RFT thus provides a way of understanding rule-governed behavior or instructional control in terms of multiple stimulus relations and the transformation of functions. On balance, empirical research on this conceptual analysis of rule-governed behavior remains extremely limited.

In contrast, particularly during the 1970's and 1980's, much work in behavior analysis was conducted on rule-following and the extent to which rule-governed behavior leads to insensitivity to direct contingencies of reinforcement. For example, differences in the behavior of humans and non-humans when they were exposed to schedules of reinforcement (Bentall, Lowe, & Beasty, 1985; Lowe, Beasty, & Bentall, 1983; Weiner, 1969), suggested that the development of human language created important species differences (Lowe, 1979). The basic argument was that some form of precurrent behavior, typically conceptualized as verbal, impacted upon responding on the reinforcement schedule (e.g., Catania, Matthews, &

Shimoff, 1989), and rendered human behavior less sensitive to a schedule's reinforcement contingencies. Often, the so-called insensitivity effect observed with human schedule performance was attributed to the impact of *verbal rules* that were generated by human participants as they interacted with the scheduled contingencies (e.g., Vaughan, 1989). Insofar as non-humans did not possess the capacity for generating such rules, their behavior was seen as being directly controlled by, or entirely sensitive to, reinforcement schedules.

From an RFT perspective, persistent rule-following, or insensitivity to reinforcement contingencies, might help to explain certain features of, for example, the behaviors typically referred to as 'depression'. Consider two individuals, both of whom report feeling significantly depressed. As part of therapy, both have started exercising in a local gym, and both have experienced the positive effects of behavioral activation. For one individual, contacting the contingencies for behavioral activation leads them to maintain a frequent exercise regime and attend the gym on those days even when they feel little motivation to do so. The other individual, however, although experiencing the benefits of behavioral activation, follows a rule that may be described as 'Only go to the gym when you feel motivated to do so'. In one case, rigidly following a rule 'Only exercise when motivated' may serve to maintain depressed-like behaviors, but in the other case, reacting to the reinforcing effects of exercise may help to produce a better therapeutic outcome.

At the current time, there is a relatively rich and growing literature on derived relational responding and a separate substantive literature on rule-governed behavior and its role in generating contingency insensitivity. To date, there has been little or no research that has attempted to bring these two research areas together. For example, while the basic RFT concept of rule-governed behavior has been successfully modeled in laboratory research (e.g., O'Hora et al., 2014; 2004), there has been no attempt to examine the extent to which rule-governed behavior that involves at least some element of derived relational responding

impacts upon insensitivity to contingencies. This is perhaps surprising given that AARR and excessive rule-following or contingency insensitivity have been implicated so widely in human psychological suffering (Hayes, Zettle, & Rosenfarb, 1989). However, there have been some studies that support the basic prediction that contingency insensitivity is implicated in human suffering.

For example, a recent study by McAuliffe, Hughes, and Barnes-Holmes (2014) presented adolescents with self-reported high versus low depression with a match-to-sample (MTS) task, with one visual sample stimulus and three comparisons. At the beginning of the experiment, participants were informed that the aim on the MTS task was to select the comparison that was *most like* the sample, and points per trial were awarded for selection responses that were in accordance with this rule. However, after two blocks of trials, the contingencies for gaining points were reversed (points were now awarded for selecting the comparison that was *least like* the sample and points were *deducted* for selecting the comparison that was *most like* the sample). McAuliffe et al. had predicted that participants with high depression would persist in following the original rule, hence showing contingency insensitivity. Their results supported this hypothesis when participants believed that their rule-following was being monitored by the researchers (but see Baruch, Kanter, Busch, Richardson, & Barnes-Holmes, 2007).

Given the potentially important link between persistent rule-following or contingency insensitivity and human suffering, there may be considerable value in exploring potential links between AARR and persistent rule-following. For example, is rule-following that requires deriving a novel stimulus relation more or less likely to produce contingency insensitivity than a rule that does not require that derivation? And if so, what critical variables might be involved in moderating this effect, including levels of human suffering itself? The primary purpose of the current study was to begin to address this gap in the literature.

Across two experiments, we manipulated the amount of opportunities to follow a rule that either did or did not require a novel derived relational response. A matching task similar to that employed by McAuliffe et al. (2014) was employed. In Experiment 1, all participants received 10 trials in which the rules matched the task contingencies before the contingencies were reversed and participants were required to complete an additional 50 MTS trials. The key purpose of the experiment was to determine if participants persisted in rule-following in the face of reversed contingencies and if this rule-following differed for rules that required, versus did not require, derived relational responding[1]. Experiment 2 partially replicated Experiment 1, but participants were provided with *100* trials before the contingencies were reversed. Furthermore, a control (no rule) condition was added for comparison with the two rule conditions. Finally, we used a range of self-report measures of psychological suffering to determine if rule-following or contingency insensitivity[2] correlated with these measures. Before proceeding, we should emphasize that the current study was largely exploratory and Experiment 1 could be seen as a type of pilot study. However, we have chosen to report the results of both experiments here because the contrasting results that we obtained across them appear to have some bearing on the on-going conceptual development of RFT and its potential for the analysis of human psychological suffering. We shall return to this issue in the

[1] It is important to note that all instances of rule-governed behavior may be conceptualized as involving derived relational responding. Indeed, we would argue that rules that have been followed and directly reinforced many times should be considered as instances of AARR. However, the level of derivation involved in following a well-established previously reinforced rule would be lower than the level of derivation involved in following a rule that required relatively novel derived relational responding (see Barnes-Holmes, Barnes-Holmes, Hussey, & Luciano, 2016, for a detailed discussion). In the current study, however, we will distinguish between the two Rule conditions as Direct versus Derived because we do not manipulate levels of derivation directly as an experimental variable.

[2] Persistent rule-following and contingency insensitivity should not be conflated. For example, contingency insensitivity may be used to refer to situations in which individuals simply fail to discriminate some subtle change in a reinforcement contingency, such as switching from a variable-ratio (VR) 20 to a VR 25 schedule. The current study was not concerned with this type of insensitivity. Rather, we were concerned with situations in which individuals may continue to follow a rule even when a change in the contingencies has been directly contacted. That is, we examined the extent to which rule-following continued when the reinforcement contingencies on a MTS task were reversed completely, and thus it was impossible that rule-following could persist simply because participants failed to contact (i.e., notice) the reversed contingencies.

Discussion. Given the exploratory nature of the current research, we refrained from making specific predictions.

Experiment 1

Participants

A total of 67 individuals participated, 49 females and 18 males. These ranged in age from 18 to 38 years old (M = 22.67, SD = 4.14), and were recruited through random convenience sampling from the online participant system at Ghent University. The majority of participants recruited through this system were bachelors or masters level students. All participants were Caucasian with Dutch as their first language. All were randomly assigned to one of two conditions, referred to as the Direct Rule Condition and the Derived Rule Condition. The data from 29 participants (22 from the Derived Rule Condition and 7 from the Direct Rule Condition) were excluded because they failed to meet specific performance criteria described subsequently (leaving N = 38 for analysis, 20 in the Derived Rule Condition, 16 females and 4 males, and 18 in the Direct Rule Condition, 12 females and 6 males).

Apparatus and Materials

The experiment involved two computer-based tasks programmed in PsychoPy (version 1.8), a Derivation Task, and a Match-to-Sample (MTS) Task. The aim of the former task was to allow participants to derive the critical part of the correct rule for completing the MTS task. The experiment also involved two questionnaires. The first of these was the Depression Anxiety and Stress Scale-21 (DASS-21, Lovibond & Lovibond, 1995); the Dutch version of this scale was employed in the current experiment (de Beurs, Van Dyck, Marquenie, Lange, & Blonk, 2001). The DASS comprises 3 subscales that measure depression, anxiety, and stress. All items (e.g., "I found it hard to wind down") are rated 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*). Subscales are

scored independently as follows: depression 0-28+, anxiety 0-20+, stress 0-34+. Higher scores indicate poorer mental health. The English version has demonstrated excellent internal consistency (Henry & Crawford, 2005): depression (alpha = 0.82); anxiety (alpha = 0.90); and stress (alpha = 0.93). The Dutch translation has yielded similar sufficient internal consistency (de Beurs et al.).

The second questionnaire employed was the Acceptance and Action Questionnaire II (AAQ-II 7-item version; Bond et al., 2011). Again, the Dutch version was employed (Bernaerts, De Groot, & Kleen, 2012). The AAQ measures acceptance of negative private events (e.g., "My painful memories prevent me from having a fulfilling life"). All items are rated on a 7-point scale from 1 (*Never true*) to 7 (*Always true*), yielding a minimum of 7 and a maximum of 49. High scores indicate *low* acceptance, while low scores indicate *high* acceptance. The English version has demonstrated adequate internal consistency with alpha coefficients ranging from 0.78 to 0.88; the Dutch translation has yielded similar psychometric strength (Bernaerts et al.).

Procedure

The experiment involved three stages: the Derivation Task, the MTS task, and the questionnaires, always conducted in this order.

The Derivation Task.

Derived Rule Condition. The aim of the derivation task in this condition was to allow participants to derive the critical part of a rule for completing the subsequent MTS task (i.e., to choose the comparison least like the sample). The derivation task comprised 3 trial-types. Two of these trial-types were filler trials and had no relevance to the MTS task. The third trial-type was directly relevant to the MTS task.

All 3 trial-types comprised 3 short statements, a question, and 2 response options. The task-relevant trial-type is presented on the left-hand side of Figure 1. This trial was denoted as

task-relevant because it enabled participants to abstract the meaning of the phrase "least like*"* from words in foreign languages, deemed to be obscure for Belgian participants[3], and to then use this abstraction to respond correctly on the MTS task. In the first statement, "least like*"* was coordinated with the Irish word "eagsula*"*; "*eagsula"* was then made opposite to the Welsh word *un; un* was in turn made opposite to the Sudanese word "beda*";* hence participants could derive that "beda*"* was coordinated (meant) "least like*"*. To respond correctly, participants were required to select the *least like* response option (rather than *most like*), by pressing the relevant key, when asked *"What does beda mean?"*. This was task-relevant because *"*beda*"* was subsequently presented in the MTS task and participants were required to respond to "beda*"* as if it meant "least like*"*.

INSERT FIGURE 1 HERE

One of the *task-irrelevant* trial-types is presented on the right-hand side of Figure 1. This trial was denoted as task-irrelevant because nothing abstracted from it could be used to inform responding on the MTS task. In the first statement, *hot* was coordinated with the Irish word "te"; "te" was then made opposite to the Welsh word "oer"*; "*oer" was in turn made opposite to the Sudanese word "panas"*;* hence participants could derive that "oer" was coordinated (meant) "cold". To respond correctly, participants were required to select the *cold* response option (rather than "hot") when asked *"What does oer mean"?* The second irrelevant trial-type was similar, except that the words referred to up and down, rather than hot and cold. None of these words in the filler trial-types was subsequently presented in the instructions for the MTS task or anywhere thereafter in the experiment. Filler trials were included to allow the researcher to check that participants were deriving the relationships as predicted, rather than responding by chance (see below).

[3] It should be noted that while no formal check was made to ensure participants did not speak Irish, Welsh or Sudanese, these languages were deemed to be relatively obscure for this sample of participants. Similarly, no participant made any indication at any stage throughout the experiment, or in the debriefing afterwards, that they had any sort of proficiency in any of these languages or knew any of the words used.

Participants received 8 presentations of the fillers (4 exposures to each of 2 trial-types) and 4 presentations of the task-relevant trial-type. The sequence of the presentation of the trial-types was fixed: filler trial-type 1, filler trial-type 2, and then the task-relevant trial-type. This sequence was presented 4 times in that order to ensure that the task-relevant trial-type exposures were not noticeably different from the filler trial-types. An accuracy criterion was applied that required correct responding in the last 10/12 responses in the Derivation Task. The criterion also required correct responses on all 4 exposures to the task-relevant trial-type. If participants did not reach the criterion on the first exposure, they repeated the task up to three times. All participants proceeded to the MTS task when they had achieved the criteria or completed 3 exposures to the Derivation Task. Data from participants who failed to reach the criteria on the third exposure were not included in subsequent analyses.

Direct Rule Condition. The Direct Rule Condition was similar to the Derived Rule Condition, except that a third filler trial-type replaced the task-relevant trial-type. This trial-type employed words that meant "black" and "white" in Irish, Welsh, and Sudanese. All three trial-types were thus irrelevant to the rule that participants subsequently received on the MTS task.

MTS Task. During each trial, a sample stimulus (random shape) was presented at the top of the screen, with three comparison stimuli (all random shapes, but none identical to the sample nor each other) along the bottom (see Figure 2). Each comparison varied in its similarity to the sample presented. That is, one comparison was clearly the *most like 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 like* the sample than the previous comparison. Finally, the third comparison was clearly the *least like* 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. A total of 54 stimulus sets were used in the experiment with each being presented at least once and no more than twice.

INSERT FIGURE 2 HERE

All participants were explicitly advised to try to gain as many points as possible in the MTS task, but additional instructions varied across conditions. All participants were instructed that "In the next part of the experiment you will be presented with a sample stimulus at the top of the screen and three target stimuli at the bottom of the screen". In the Direct Rule Condition, participants were explicitly instructed to "Respond by selecting the target stimulus that is *LEAST LIKE* the sample stimulus". In the Derived Rule Condition, participants were instructed to "Respond by selecting the target stimulus that is *BEDA* (i.e., least like) the sample stimulus".

The MTS task comprised 60 trials. On all trials, participants emitted a response by pressing the key (*D, G,* or *K*) directly below the comparison they wished to select (see Figure 2). *For all participants, the task on the first block of trials involved selecting the comparison that was least like the sample.* When a correct response was emitted, one point was awarded, and the screen cleared immediately to present the total number of points achieved thus far (in large red text in the center of the screen) for 3secs. Emitting an incorrect response resulted in the loss of one point, again followed by a display of the total number of points so far. For both conditions, the feedback contingencies for the first 10 trials were consistent with the directly instructed or derived rule. After these 10 trials, the task contingencies were reversed. That is, for all participants, the contingencies for correct and incorrect responding switched, without warning for the final 50 trials. *During these trials, therefore, correct responding now involved selecting the comparison that was most like the sample, while incorrect responding involved selecting either of the two remaining comparisons*.

Questionnaires. After the MTS task, participants completed the DASS-21 and the AAQ-II.

Results

For analysis, participants in both conditions were subject to a strict accuracy criterion that required correct responding on at least 8 out of the first 10 trials in the MTS task, thus reducing the likelihood that participants learned to match, based purely on trial and error (rather than by the direct rule or its derived version). The data from participants who did not meet this accuracy criterion in the initial trials were excluded.

Insofar as the primary aim of the experiment was to compare performances between the Direct and Derived Rule Conditions, the data from the 50 trials presented after the contingency reversal were analyzed in two related ways. We employed a measure of *rule compliance*, defined as the total number of responses (out of 50) that were consistent with the ("least like"/"BEDA") rule or derivation, but were inconsistent with the reversed contingencies for those 50 trials. The mean total number of responses (out of 50) that were consistent with the rule/derivation (i.e., level of rule compliance) for each of the two conditions were similar; Direct Rule Condition $M = 19.78$, $SD = 16.47$; Derived Rule Condition $M = 17.55$, $SD = 15.98$. An independent *t*-test confirmed that this small difference was non-significant ($p = .68$).

The second way in which the data were analyzed was designed to indicate the point at which participants stopped following the initial rule and began to respond in accordance with the reversed contingencies (i.e., contingency sensitivity). Our concern here was that some participants may have shown contingency sensitivity relatively early in the 50 trials by switching their responding, but subsequently reverted back to rule-following after only a small number of trials. Overall therefore, participants may have shown relatively rapid contingency sensitivity, but because the overall number of rule-consistent responses would

have remained high, the metric would not have accurately reflected sensitivity *per se*. On balance, to control for the occasional or "random" rule-inconsistent response, contingency sensitivity was defined as the point at which responding in accordance with the reversed contingency emerged and did not return reliably to a rule-consistent pattern. Contingency sensitivity, therefore, was defined as 3 or more consecutive responses in accordance with the reversed contingency followed by no more than 4 consecutive rule-consistent responses thereafter. More informally, contingency sensitivity referred to the point at which participants tended to respond in accordance with the reversed contingency and continued to do so across most of the remaining trials. The point at this occurred was again similar across the two conditions; Direct Rule Condition = 21.33, *SD* = 16.32; Derived Rule Condition = 19.45, *SD* = 16.27). An independent *t*-test again confirmed that this small difference was non-significant ($p = .72$).

Correlational analyses were also conducted between the DASS and the AAQ scores and the scores for rule compliance and contingency sensitivity for both conditions. Of the 20 correlations possible, none proved significant (all *p*s >.23).

Experiment 2

Although there was little evidence for the impact of the Direct versus Derived Rule manipulation in Experiment 1, in the study reported by McAuliffe et al. (2014) participants were exposed to two blocks of 40 MTS trials before the contingencies switched (i.e., 80 trials). At this point, therefore, we decided to replicate Experiment 1 but increase the number of MTS trials participants were required to complete before the feedback contingencies switched. In addition, we also included a control condition in which participants did not receive a formal rule for the MTS task.

Participants

A total of 140 individuals participated, 106 females and 34 males. Participants ranged from 18 to 49 years (M = 22.04, SD = 4.14), and were again recruited through random convenience sampling from the online participant system at Ghent University. Again, the majority of participants recruited through this system were bachelors or masters level students. All participants were Caucasian with Dutch as their first language. All, except 25, were randomly assigned to one of two conditions, referred to, again, as the Direct Rule Condition and the Derived Rule Condition, with the remaining 25 participants assigned to the Control Condition. The data from 63 participants (46 from the Derived Rule Condition, 5 from the Direct Rule Condition, and 12 from the Control Condition) were excluded because they failed to meet specific performance criteria described subsequently (leaving N = 77 for analysis, 30 in the Derived Rule Condition, 24 females and 6 males, 34 in the Direct Rule Condition, 24 females and 10 males, and 13 in the Control Condition, 8 females and 5 males).

Apparatus and Materials

All materials and apparatus were similar to Experiment 1, with the exception of a minor change to the MTS task. The experiment now involved a total of 56 stimulus sets, each presented at least once and no more than three times, for a maximum of 150 trials.

Procedure

The procedure for Experiment 2 was similar to the previous experiment, with the exception of two modifications to the MTS task. First, the MTS task now comprised 150 trials (rather than 60). Of these, the first 100 trials (rather than 10) involved a match between the task contingencies and the directly instructed or derived rule. Second, participants in the Control Condition received only instructions that highlighted the need to acquire points for correct responding (i.e., there was no explicit reference to any rule for matching).

Results

The strict accuracy criterion that required correct responding on the first 8/10 MTS trials, as in Experiment 1, remained in place for the Direct Rule and Derived Rule Conditions in Experiment 2. It was not feasible to apply this criterion to the Control Condition because very few participants would meet it (i.e., they had no rule, neither direct nor derived, to follow during their initial exposure to the MTS task). In fact, none of the Control participants emitted 8/10 correct responses in the first 10 trials. Participants from all three conditions, however, were required to achieve 80/100 correct responses prior to the switch in feedback contingencies. The assumption here was that an acceptable number of Control participants would have adapted to the contingencies across the first 20 trials and would thus achieve the 80/100 criterion. The data from participants who did not meet this accuracy criterion were excluded. All analyses were identical to the previous experiment.

In terms of rule compliance, the mean total numbers of responses (out of 50) consistent with the initial rule for all three conditions are presented in Figure 4 (left-hand side). In contrast to Experiment 1, the means differed considerably across conditions. Participants in the Direct Rule Condition emitted more responses (M = 31.09, SD = 17.95) in accordance with the original rule than both the Derived Rule Condition (M= 19.27, SD = 14.57) and the Control Condition (M = 7.92, SD = 2.84). A one-way ANOVA proved to be significant, $F(2, 74) = 12.227$, $p < .001$, $\eta_p^2 = .25$, with post-hoc tests (Fisher's PLSD) indicating significant differences between: the Direct Rule and Derived Rule Conditions (p = .003); the Direct Rule and Control Conditions ($p < .001$); and the Derived Rule and Control Conditions (p = .03).

INSERT FIGURE 3 HERE

The mean total number of responses for contingency sensitivity for the three conditions are presented in Figure 3 (right-hand side). Participants in the Direct Rule Condition required more trials (M = 32.74, SD = 18.51) to respond in accordance with the

new contingencies than both the Derived Rule (M = 21.03, SD = 15.81) and Control Conditions (M = 8.62, SD = 2.53). A one-way ANOVA proved to be significant, $F(2, 74) = 11.79$, $p < .001$, $\eta_p^2 = 0.24$, with post-hoc tests (Fisher's PLSD) indicating significant differences between: the Direct Rule and Derived Rule Conditions ($p = .004$); the Direct Rule and Control Conditions ($p < .001$); and the Derived Rule and Control Conditions ($p = .02$). Both sets of analyses, therefore, indicated more persistent rule-following with the provision of a directly instructed rule than with a rule that involved some form of derivation. Both rule conditions, however, produced more persistent rule-following than the *no rule* Control Condition.

Correlational analyses were also conducted among each condition, the DASS scores, and AAQ scores. Of the total 30 correlations possible, two were significant. In both cases, these correlations were found only in the Direct Rule Condition. That is, greater rule compliance predicted lower stress on the DASS (r = -.34; p = .05), as did less contingency sensitivity (r = -.36, p = .04). These analyses suggest that participants in the Direct Rule Condition, who showed reduced rule-compliance, reported more stress.

Discussion

The data from the current study failed to produce a statistically significant difference between the Direct and Derived Rule Conditions in Experiment 1. In Experiment 2, however, significant differences emerged amongst all three conditions, with the Direct Rule Condition producing the most persistent rule-following and the Control Condition producing the least. The different outcomes across the two experiments raise some interesting questions, but before addressing those, it seems important to consider a procedural issue relevant to both experiments.

In conducting the current study, there was a considerable difference in the number of participants who failed to reach the 8/10 accuracy criterion in the MTS task. For example, in

Experiment 1, 19/42 (45%) failed to meet this criterion in the Derived Rule Condition, compared with 2/26 (8%) in the Direct Rule Condition. Any difference recorded between the groups in terms of performance on the MTS task should, therefore, be interpreted with caution. However, no difference actually emerged on the key variables. Interestingly, a similar level of attrition was observed in Experiment 2, in that 38/76 (50%) failed to meet this criterion in the Derived Rule Condition, compared with 2/46 (4%) in the Direct Rule Condition. However, in this case, a significant difference between the groups did emerge on both of the key variables. Of course, this difference in attrition should be acknowledged when interpreting the between-group difference, but attrition *alone* cannot be used to explain the difference found in Experiment 2, because a similar difference should therefore have been observed in Experiment 1.

At the present time, it remains unclear why the attrition rates were so different across the conditions, but a likely explanation is that the derived relational responding required in the Derived Rule Condition failed to transfer, for some reason, to the MTS task. Perhaps numerous differences in the topographical stimulus properties of the two tasks served to undermine the transfer that was required. Or more informally, participants simply failed to see the connection between the two tasks in terms of the single common word "Beda". Seeing the connection between the two tasks was completely unnecessary for participants in the Direct Rule Condition because no novel derivation was required. Perhaps future studies could include some procedural instruction that would increase the required transfer, such as encouraging participants to apply what they learned in the Derivation Task to the MTS task.

As noted in the Introduction, a recent study found that participants with high depression persisted in rule-following, when they believed that their rule-following was being monitored by the researchers (McAuliffe et al., 2014). A direct comparison of this study with the current findings is difficult because the studies differed in many ways. For example,

McAuliffe et al. employed adolescent pupils who were categorized into groups of high versus low depression, whereas the current study employed randomly sampled bachelors and masters level students. Furthermore, the condition in which a difference in persistent rule-following was observed in the McAuliffe et al. study between high and low depressed individuals placed a considerable emphasis on the fact that participants were being monitored by the experimenter, whereas in the current study this was not the case. The primary manipulation in the current study was direct versus derived rule-following, thus it remains unclear to what extent participants felt they were being closely monitored for (direct or derived) rule-following by the experimenter. Perhaps future studies could examine the interaction between direct versus derived rule-following and level of monitoring by the experimenter.

The key finding in the current study was the differences recorded across the three conditions in Experiment 2. Although the research was exploratory, it seems important to offer some form of RFT-based explanation, albeit post-hoc, for the current findings. As noted in the Introduction (footnote 1), we have distinguished between direct and derived rule-following as if they were completely dichotomous conditions, but strictly speaking, for RFT, even the Direct Rule Condition involved a certain level of derivation. From this theoretical perspective, it appears that lower levels of derivation (Direct Rule Condition) may have produced more persistent rule-following than higher levels of derivation (Derived Rule Condition). This interpretation is certainly consistent with the suggestion that relational flexibility (in rule-following) may vary as a function of levels of derivation (see Barnes-Holmes, et al., 2016). Or more informally, we may "give up on a rule" more readily when it no longer works for us if the rule requires some recent derivation in terms of understanding its meaning.

Of course, this interpretation of the current findings does not address the fact that levels of derivation appeared to have little impact on relational flexibility in Experiment 1

(i.e., because there was no obvious difference between the Direct Rule Condition and Derived Rule Conditions). On balance, Barnes-Holmes et al. (2016) also argued that relational coherence may interact in a dynamic fashion with levels of derivation and relational flexibility. If this was the case, then perhaps the contrasting results of Experiments 1 and 2 may be explained readily by RFT. Specifically, it could be argued that relational coherence between the rule and the contingencies in Experiment 1 was considerably lower than in Experiment 2 (because participants completed only 10 versus 100 trials, respectively, before the contingencies switched). Or, more informally, participants may have been a great deal more certain that the rule was correct (i.e., coherent with the contingencies) in Experiment 2 than in Experiment 1. If this interpretation is correct, then it suggests that the relationship between levels of derivation and relational flexibility (in rule-following) is moderated by levels of relational coherence. Or more precisely, level of derivation impacts more on relational flexibility when relational coherence is high rather than low.

At this point it seems important to acknowledge that the foregoing post-hoc interpretation is consistent with a multi-dimensional, multi-level (MDML) framework for analyzing the dynamics of AARR as presented in Barnes-Holmes, et al. (2016), but the experiments were not based directly upon it. Rather, the current research and framework have co-evolved in an inductive manner within our research group. At the present time, therefore, the foregoing interpretation must remain highly speculative, although potentially instructive in terms of what variables might be manipulated in future studies, such as levels of derivation and/or coherence.

Another potentially interesting finding arising from the current study was the fact that lower levels of persistent rule-following predicted higher levels of stress on the DASS, but only in the Direct Rule Condition in Experiment 2. Again, an explanation for this finding must remain tentative at the present time, but it is worth noting that the DASS was presented

to all participants *after* completing the experimental tasks, and thus the self-reported stress levels may have been influenced by those very tasks. If this was the case, then abandoning a rule when coherence is relatively high, and derivation is relatively low, may increase levels of stress. More informally, the more participants felt they were disobeying a clear and well-established rule, the more stress they experienced. If this interpretation is correct, perhaps increased relational flexibility (in rule-following), in the context of high coherence and low derivation, may come at the cost of increased stress levels. Or, in other words, disobeying a clear and well-established rule, even when it no longer works, creates greater stress, particularly when the rule is abandoned relatively quickly.

Leaving aside the potential stress-inducing impact of the task in the current study, it may be useful to consider the implications of the findings for psychological suffering more generally. The current study suggests that persistent rule-following occurs when relational coherence is high and levels of derivation are low. In so far as maladaptive persistent rule-following may be involved in depressed behavior, as outlined in the Introduction, it may be useful to consider the extent to which the assessment and treatment of depression focuses on these variables. For example, when an individual presents in therapy as depressed, a therapist may explore the extent to which specific rules are being followed which may undermine attempts at behavioral activation (e.g., 'only exercise when you feel motivated'). More informally, this may involve discussing with the client how firmly they believe that such rules are indeed true or accurate (i.e., coherent) and how long they have been following them (i.e., level of derivation). Doing so may provide some insight into the potential reasons why a program of behavioral activation succeeds with one client but fails with another (Addis, Truax, & Jacobson, 1996).

Two limitations to the current study should be noted. First, the sample from which participants were drawn was relatively narrow, thus limiting the generalizability of the

findings. Second, as noted above, the DASS measure was taken after participants completed the experimental tasks, and thus it is possible that the difference in stress levels were present before the study and this may explain, at least in part, the difference in persistent rule-following. On balance, in a partial replication of the current study in which we directly manipulated differences in derivation, similar results were obtained as in the current study, even though no differences in stress levels were recorded at baseline.

The current study constitutes the first attempt to analyze the impact of derived relational responding on persistent rule-following, and thus brings together two areas of research that have often been drawn upon in developing conceptual analyses of human psychological suffering. As noted previously, the current work is exploratory and much of the foregoing (post-hoc) theorizing must remain speculative at the present time. Nevertheless, the results, particularly from Experiment 2, appear to be quite compelling and are thus certainly worthy of further investigation.

References

Addis, M.E., Truax, P., & Jacobson, N.S. (1996). Why do people think they are depressed: The reasons for depression questionnaire. *Psychotherapy, 32*, 476-483.

Barnes-Holmes, D. & Barnes-Holmes, Y. (2000). Explaining complex behavior: Two perspectives on the concept of generalized operant classes. *The Psychological Record, 50*, 251-265.

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I., & Luciano, C. (2016). Relational Frame Theory: Finding its historical and intellectual roots and reflecting upon its future development: An introduction to part II. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley handbook of Contextual Behavioral Science* (pp. 117-128). West Sussex: John Wiley & Sons, Ltd.

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.

Bentall, R. P., Lowe, C. F., & Beasty, A. (1985). The role of verbal behavior in human learning: II. Developmental differences. *Journal of the Experimental Analysis of Behavior*, *43*(2), 165-181.

Bernaerts, I., De Groot, F., & Kleen, M. (2012). De AAQ-II, een maat voor experiëntiële vermijding: Normering bij jongeren [The AAQ-II, a measurement of experiential avoidance: Standardization amonog young people]. *Gedragstherapie, 4*, 389-399.

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.

Catania, A. C., Shimoff, E., & Matthews, B. A. (1989). An experimental analysis of rule-governed behavior. In S. C. Hayes (Ed.), *Rule-governed behavior: 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 [The DASS: A questionnaire fort he measurement of depression, anxiety and stress]. *Gedragstherapie, 34*, 35-53.

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

Hayes, S.C., Zettle, R.D., & Rosenfarb, I. (1989). Rule Following. In S.C. Hayes (Ed.), *Rule-governed beahvior: Cognition, contingencies, and instructional control* (pp. 191-220). New York, NY: Plenum.

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

Lowe, C. F. (1979). Determinants of human operant behavior. In M. Zeiler & P. Harzem (Eds.), *Advances in the analysis of behaviour: Volume 1. Reinforcement and the organisation of behaviour* (pp. 159-192). Chichester, England: Wiley.

Lowe, C. F., Beasty, A., & Bentall, R. P. (1983). The role of human verbal behavior in human learning: Infant performance on fixed-interval schedules. *Journal of the Experimental Analysis of Behavior, 39*, 157-164.

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.

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.

Sidman, M. (1994). *Stimulus equivalence: A research story*. Boston, MA: Authors Cooperative.

Vaughan, M. (1989). Rule-governed behaviour in behaviour analysis: A theoretical and experimental history. In S. C. Hayes (Ed.), *Rule-governed behavior: Cognition, contingencies, and instructional control* (pp. 97-118). New York: Plenum.

Weiner, H. (1969). Controlling human fixed-interval performance. *Journal of the Experimental Analysis of Behavior, 12*, 349-373.

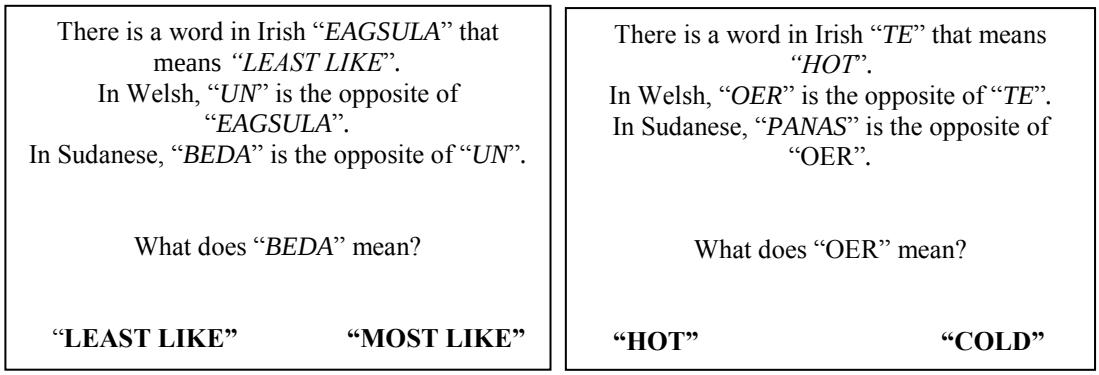

Figure 1. The task-relevant trial-type presented to the Derived Rule Condition (left-hand side) and an example of a filler trial-type (right-hand side) presented to all participants in the Derivation Task.

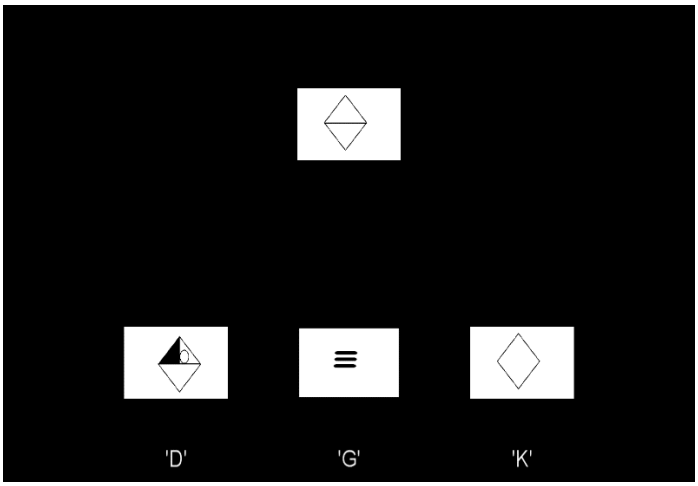

Figure 2. An example of a single trial and single stimulus set presented in the MTS task.

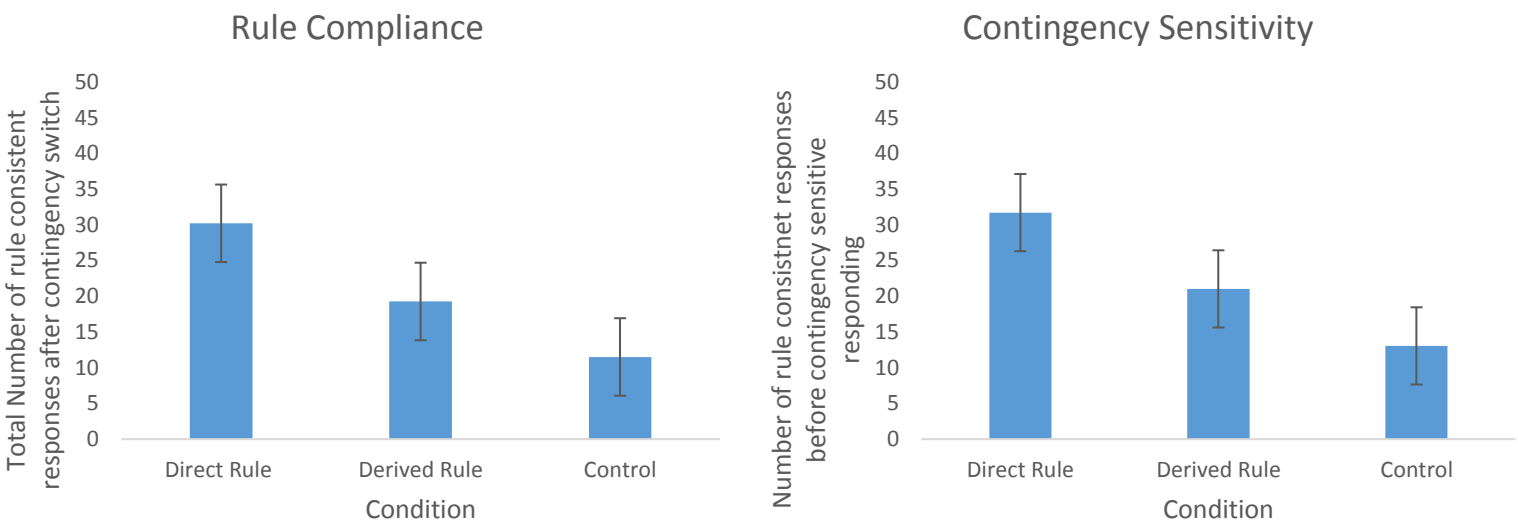

Figure 3. Mean rule compliance scores (left-hand side) and contingency sensitivity scores (right-hand side) with standard error bars for the Direct Rule, Derived Rule, and Control Conditions.