GÖTEBORGS UNIVERSITET
PSYKOLOGISKA INSTITUTIONEN

The Impact of Coherence on Persistent Rule-Following in the Face of Reversed Reinforcement Contingencies

Ragnar Bern och Tone Persdotter

Examensarbete 30 hp
Psykologprogrammet
PM 2519
Vårtermin 2018

Handledare: Dermot Barnes-Holmes

The Impact of Coherence on Persistent Rule-Following in the Face of Reversed Reinforcement Contingencies

Ragnar Bern och Tone Persdotter

Abstract: Rule-governed behavior has long been associated with generating insensitivity to direct contingencies of reinforcement. This insensitivity to environmental changes has also been implicated in human psychological suffering. From a behavior analytic perspective, rule-following is a verbal behavior, and thus has been suggested to be potentially affected by the level of coherence within the verbal stimuli involved. In this explorative study, 216 participants received a rule that differed in terms of its coherence based on experimental training. The active conditions did not differ significantly in rule persistency, but differed from the control condition that was less rule-persistent. Rule-persistence was found to be significantly correlated with stress and anxiety in the control condition. A post-hoc interpretation of the findings is provided.

Keywords: persistent rule-following, coherence, contingency insensitivity

Since the claim that behavior analysis failed to explain the development and complexity of human language (Chomsky, 1959), a post-Skinnerian take on verbal behavior was developed in an attempt to grapple with this shortcoming. This approach has since unveiled and explained many elements in language and cognition that had been thus far unexplained and unavailable to experimental analysis with operant and respondent conditioning alone (Hayes, Barnes-Holmes, Biglan & Zettle, 2016). The ability to relate and respond to stimuli based on arbitrary properties, such as for example choosing gold over silver, was posited as a core property of human language that made extrapolation from animal studies kittle. Perhaps most importantly was the discovery of what was termed *stimulus equivalence* (e.g. Sidman, 1971; see Sidman, 1994 for a book length review). That is, the phenomenon by which a small number of trained and reinforced responses (e.g. that gold is *more valuable* than silver) was also seen to generate a number of untrained, unreinforced responses (e.g. silver is less valuable than gold) (Sidman & Tailby, 1982). Furthermore, the trained and emergent untrained responses may also be accompanied by a change in the function of the stimuli involved, making gold more desirable than silver for example, and vice versa.

Arbitrarily Applicable Relational Responding

From the growing research on stimulus equivalence emerged a new modern theory of human language and cognition known as *Relational Frame Theory* (RFT; Hayes, Barnes-Holmes & Roche, 2001). RFT suggested that if stimulus equivalence was a type of operant, then perhaps it constituted but one type of generalized operant of arbitrarily applicable relational responding (AARR), an approach that could then more adequately begin to explain the complexity and generativity of human language and give rise to a rich new program of research (O'Connor, Farrell, Munnelly & McHugh, 2017). If this was indeed the case and stimulus equivalence was a generalized operant, then an extended history of relevant reinforced exemplars would serve to give rise to different patterns of relational units, the basic unit of which was defined as the relational frame (Barnes-Holmes & Barnes-Holmes, 2000). RFT put the relational frame as the basic operant unit involved in verbal behavior, i.e.

the behavior of relating stimuli to one another in a certain way based on arbitrary properties (e.g. value), rather than solely non-arbitrary properties (e.g. hardness) (Hayes et al.). More informally, an arbitrary relation could be established between stimuli based on socially determined qualities rather than purely physical qualities. Furthermore, specific kinds of relational frames, such as "more than" or "the same as" constituted different subtypes of AARR (Hayes et al., 2016).

Subsequent experimental investigation empirically supported three core properties of AARRing, that constituted the basic features of verbal behavior (Törneke, 2014): 1) *mutual entailment* (if A is related to B, then B becomes related to A, 2) *combinatorial mutual entailment* (if A is related B and B is related to C, then A becomes related to C, and C becomes related to A) and 3) *transformation of stimulus function* (the functions of the related stimuli are transformed based upon the types of relations into which those stimuli enter). Experimental research has thus far identified several types of relational frames: coordination, distinction, hierarchical, comparison, deistical, causal, opposition, spatial and temporal (e.g. Luciano, Rodrigquez, Manas & Ruiz, 2009; Törneke, 2014; Hayes & Gregg, 2001).

More recent developments within RFT have given rise to a multi-dimensional multi-level (MDML) framework for analyzing the dynamics of AARR (see Barnes-Holmes, Barnes-Holmes, Luciano & McEnteggart, 2017 for a full length treatment of the MDML). This framework puts forward that AAAR can be thought of as occurring across 5 levels of relational development (mutual entailment, relational framing, relational networking, relating relations, and relating relational networks) and 4 dimensions (derivation, complexity, coherence, and flexibility) that interact with each other in a dynamic fashion. Thus, relational responding can be analyzed functionally along these levels and dimensions. Flexibility concerns the degree to which an AARR can be modified by a current contextual cue, such as "it´s the other way around" when someone is being told they are wrong. Complexity concerns the level of detail within a pattern of AARR. For example, A equals B is less complex than A is better than B since the later also implies that B is worse than A (through mutual entailment). Derivation concerns how well trained an AARR is. The first time stimuli are related, derivation is high since that specific relational response has not been derived before and is therefore novel or emergent. The more times this response is made, the lower derivation will become. An example would be the first time someone is learning a word in a novel language, relating a new word to a known word. The more this relation is repeated–relating a novel word to a known word–the more the level of derivation decreases. Finally, coherence refers to how predictable an AARR is based on the prior learning history of the individual. An individual AARRing is considered coherent if the verbal stimuli in the derived relation are related consistently with what was previously learned (Hayes et al., 2016). The higher the coherence of a specific AARR, the higher the probability of that AARR occurring, and vice versa. More informally, coherence refers to how correct something is. Among individuals with a history of relating gold as being more valuable than silver, being told the opposite would be deemed incoherent with that verbal history and might be corrected or punished. Coherence could therefore be seen as central to communication – the words used need to be correct, or "make sense" for language to have any bearing (Hayes et al., 2016).

Coherence

Following the tradition of behavior analysis, the coherence of an AARR is understood from a specific speaker and its context rather than in itself. For example, a certain group may have a narrative that may be coherent in their context, but that is incoherent with the narratives of other groups (Hammack, 2011), as when different groups have different stories of the creation, all being coherent in themselves, but incoherent in between. In a similar way, a person may have a self-narrative that is coherent with the network itself (e.g. coherent with the relational network in which it participates for that individual), but incoherent with how others speak about that person. For example, a person with social anxiety may be viewed by others as being socially skilled, while the socially anxious person themselves may think of him- or herself as being a social failure.

While coherence has been suggested to be a key dimension/aspect of verbal behavior, few studies have investigated it per se. However, previous research has indicated that individuals tend to revert back to previously coherent ways of relating stimuli when faced with situations in which coherent responding is not possible (Leonhard & Hayes, 1991; Pilgrim & Galizio, 1995; Stewart, Stewart & Hughes, 2016; Wilson & Hayes, 1996). In other words, coherence may function as a form of heuristic, or a confirmatory bias that guides people towards responding in a way that is consistent with previous information or behavior (Quinones & Hayes, 2014; Chater & Loewenstein, 2016). This suggests that coherence may function as a generalized reinforcer for AARR in and of itself (Brodieri et al., 2015; Wray et al., 2012). In support of this claim, the related concept of *sense-making* (defined broadly as finding meaning and coherence within and between events) seems to emerge in the absence of external reinforcement (Chapman & Chapman, 1967; Holyoak & Simon, 1999; Kelley, 1973; Skinner, 1936; Starr & Katlin, 1969; Harrison & Green, 1990; Peterson & Seligman, 1984). Informally speaking, sense-making in itself seems to be rewarding (Villatte et al., 2017). This is perhaps not surprising given that in the social environment, language need to make sense to have any bearing (Blackledge, Moran & Ellis, 2009). Over time, individuals are rewarded for speaking coherently about events and punished for speaking incoherently. Also, there seem to be a preference for contexts in which coherent responding is possible (Bordieri et al., 2015). Since coherence seem to be established as being rewarding in itself, this may explain why coherent AARRing may persist even though it its aversive (Wray et al., 2012; Hayes, Strosahl & Wilson, 1999), which has been suggested to influence and maintain psychological suffering (Villatte et al.). Furthermore, it has been argued that it is more so that non-coherent responding is aversive (Roche et al., 2001).

Perhaps because finding coherence (i.e. identifying relevant relations among a series of events) can increase effective behavior such as problem solving and prevention of harm (Stewart et al., 2016), generating coherence, i.e. engaging in sense making, has been found to be associated with positive psychological outcomes (Bird & Reese, 2006; McAdams, Reynolds, Lewis, Patten & Bowman, 2001; see Mineka & Henderson, 1985 for a review). But despite its positive effects, it has also been associated with some downsides (Borkovec, Robinson, Pruzinsky & Depree, 1983; see Watkins, 2008, for a review). This may again be explained with coherence being a generalized reinforcer that maintains behavior despite its aversive outcomes. In other words, an individual may prioritize coherence over whether coherence actually produces desirable outcomes (Villatte et al., 2017).

Several other theories have also thus far attempted to deal with the subject of coherence. For example, consistency theory argues that humans have a preference for consistency, for example between behaviors and values (Grawe, 2007; Simon & Holyoak, 2002). Inconsistency is experienced as aversive and steps are taken to re-establish consistency. One of the more prominent inconsistency theories, cognitive dissonance theory, explains how people adjust either behavior or attitudes to restore consistency, which also has been conceptualized from behavior analysis: "Festinger and colleagues' (1957) classic studies on cognitive dissonance provide more directly applied and well-controlled evidence for the notion that incoherent framing can function as an aversive, and coherent framing as a reinforcer." (Blackledge et al. 2009, p. 245). Also, research on the concept of *confabulation* displays the propensity to make up seemingly coherent explanations for things that cannot possibly be explained, without the intention to lie (e.g. Fotopoulou, 2008).

Rule-Governed Behavior

Rule-governed behavior (RGB) has traditionally been conceptualized as behavior controlled by antecedent verbal stimuli, without any apparent shaping contingencies (Törneke et al., 2008). This can explain why humans can engage in behavior that has not been previously reinforced. The original treatment of RGB within behavior analysis was provided by BF Skinner (1966) and dealt with rules as a method of problem solving. When facing a problem, defined as a situation demanding a behavior that has not previously been reinforced, a person could solve the problem by generating and following a rule. As mentioned, RFT has strived to develop a functional analytic account of human language and cognition, one area of which is RGB (Hayes et al., 2001). From an RFT standpoint, a rule refers to a "relational network", that is, a bundle of interconnected stimuli that typically specify the context, temporal antecedent and topography of a specific behavior, the consequence that will be delivered, and when those consequences are to be delivered (Hayes et al., 2016). For example: "if I train hard for the upcoming competition, I might win the gold medal". RGB enables people to learn without having to contact direct contingencies of reinforcement. As an extension, this type of behavior has also been seen to produce human schedule insensitivity, or what has often been termed the 'insensitivity effect'. That is, behaving in a manner that deviates from the current direct contingencies of reinforcement, a behavior not observed in non-human animals. While this can be extremely beneficial (e.g. enduring physical pain when training, given that winning a gold medal has reinforcing functions for that individual; inhibiting aggression; future planning) it also has dark sides that have been implicated in human psychopathology (Hayes, Zettle & Rosenfarb, 1989; Törneke, Luciano & Salas, 2008; McAuliffe, Hughes & Barnes-Holmes, 2014).

The potential of RGB to make individuals insensitive to changes in their direct environment has been readily observed experimentally for some decades (e.g. Vaughan, 1989). As mentioned previously, there is a known difference between humans and non-humans in this regard (e.g. Bentall, Lowe & Beasty, 1985; Lowe, Beasty & Bentall, 1983), indicating that human verbal behavior creates an important species difference (Catania, Shimoff & Matthews, 1989) giving rise to this insensitivity effect - an effect that has also been implicated in clinically relevant behavior, such anxious avoidance (Dymond, Bennett, Boyle, Roche & Schlund, 2017; Törneke, 2014). For example, after hearing about the public humiliation of a friend, someone may derive and follow rules on how to avoid social anxiety provoking situations. For example, "if I avoid speaking in front of others, I won't be humiliated". Even though the rule may be accurate insofar as speaking in front of others may cause some anxiety, it may not cause as much anxiety as expected and may also come with positive consequences, such as praise. More informally, rules describe consequences that are unknown and possibly incorrect, but the imagined consequences still exert influence over behavior.

RGB and Psychopathology

There are some previous studies on RGB and psychopathology. For example, Rosenfarb and colleagues (1993) compared depressed and non-depressed individuals in RGB. They found that depressed individuals were less contingency sensitive than non-depressed. A tentative conclusion was that individuals with depression are less weary about doing a favorable impression on others (as in following their rules), and that they've historically received less reinforcement from their social context. Another study investigating the interaction between depressive symptomatology and RGB was conducted by Baruch and colleagues (2007). Dysphoric individuals displayed greater contingency sensitivity than individuals with non-dysphoric participants. Contrary, McAuliffe and colleagues (2014) found that individuals with high self reported depression symptoms was more rule-governed compared to those with low self reported depressive symptoms when given an inaccurate instruction. However, the somewhat different designs in between these studies make it difficult to do draw any conclusions.

While RGB is thought to contribute to important aspects of psychopathology, and derived stimulus relations have already been used to explain some aspects of psychopathology, there is a lack of basic experimental research combining these two areas (Harte et al, 2017). However, one recent study aimed to begin to address this gap through investigating whether a directly instructed rule versus one that involved a novel experimental derivation impacted persistent rule-following (Harte et al., 2017). Across two experiments, 140 subjects were either given a rule in which the key aspect of the rule (i.e. "least like") was derived from foreign words (Derived Rule Condition), or where this was directly instructed without any need to derive it from foreign words (Direct Rule Condition). After being given the rule, the subjects proceeded to a MTS task in which the "least like" instruction or derivation was necessary for completing the task. In this task, participants had to match an arbitrary symbol (sample stimulus) to one of three other arbitrary symbols (comparison stimuli) using the "least like" instruction or derivation, a procedure repeated 10 times (Experiment 1) or 100 times (Experiment 2). Correct responding resulted in the provision of one point, while incorrect responding resulted in the loss of one point. A control group was included where the subject received similar training in derivation but was not given any rule for the MTS task except the rule to gain as many points as possible. After the first 10 or 100 trials, the contingencies switched unbeknownst to the participants, meaning that the subjects now received points for choosing the comparison stimuli that was *most* like the sample stimuli, and lost points for choosing any other stimuli across a final 50 trials. Rule persistence was measured by assessing how long subjects persisted with the previously reinforced rule whilst now losing points. Results showed that participants in the Direct Rule condition persisted significantly longer than the Derived Rule Condition, but only when there were more opportunities to follow the reinforced rule before the contingency switch (100 trials vs. 10). Also, the persistent rule following in the Direct Rule Condition was associated with significantly higher stress levels. Taken together, this suggest that derivation affects rule persistency, but only when coherence is high.

Rules may also be analyzed regarding its coherence, that is, the stimuli in a rule network may be more or less coherent with an individual's previous learning history. For example, telling someone that smoking will make them long lived will probably be deemed incoherent from the listener's point of view, and would probably not lead to an increase in smoking (Törneke, 2014). On the other hand, a more coherent rule, that accurately describes a problem, could facilitate problem solving behaviors (Stewart et al., 2016). Therefore, finding coherence within a problem, i.e. accurately describing the relations among its different parts, could predict efficient problem solving and therefore reinforcement (see Mineka & Henderson, 1985).

The Present Study

The current study aimed to extend the limited research bringing together RGB and RFT by further investigating the potential links between dimensions of AARRing and persistent rule-following, specifically with respect to the coherence dimension and how this may be linked to human psychological suffering. Based on the research conducted by Harte et al. (2017), the key word in the rule that was provided for completing a contingency switching MTS task was manipulated by varying levels of coherence. That is, this key word was either part of a relational network that was coherent, or a part of a network that was partially incoherent. A MTS task similar to that employed by Harte et al. was applied. 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-persistency differed for high coherent versus low coherent rules. In addition, a control condition, in which participants received an irrelevant rule, was added for comparison with the two rule conditions. Lastly, a measure of general psychological distress was employed to examine stress, anxiety and depression and to examine whether persistent rule-following correlated with these dimensions of human suffering. Due to the exploratory nature of this study, specific predictions were not made.

Method

Participants

216 undergraduate and graduate students (138 females, 70 males, 5 others, and 3 preferred not to answer) were recruited through random convenience sampling at the University of Gothenburg, Sweden. Their ages ranged from 18 to 57 years old (*M* range = 22 - 25 years) and 94,9 % had Swedish as their first language.

All participants were randomly assigned into one of three conditions: High Coherence Condition, Low Coherence Condition or Control Condition. The Control Condition was further divided into two conditions for counterbalancing. The data from 63 participants (27 from the High Coherence Condition, 24 from the Low Coherence Condition and 12 from the two control groups) were excluded because they failed to meet the specific performance criteria described subsequently. This left an *N* of 149 for the analysis (47 in the High Coherence Condition, 29 females, 17 males, and 1 prefer not to respond; 50 in the Low Coherence Condition; 28 females, 19 males, and 3 other; 23 in the Control Condition Faster, 17 females, and 6 males; and 28 in the Control Condition Slower, 20 females, 6 males, and 2 other).

Materials & Apparatus

The experiment involved one self report measure of psychological distress and two computer-based tasks made in Qualtrics: a Coherence Task, and an Match-to-Sample Task (MTS), and the Depression Anxiety and Stress Scale -21 (DASS-21; Lovibond & Lovibond, 1995).

Coherence Task. The aim of the Coherence Task was to train the participants in a new relational network (A=B=C=D=E=F) and to allow them to derive the critical part of the rule that would be needed subsequently for completing the MTS task. The Coherence Task also differed in the degrees to which feedback was coherent for participants between conditions. It consisted of 8 trials comprised of a task relevant trial-type (appearing 6 times) and two task irrelevant trial-types (appearing one time each). One open-text question followed the fourth trial-type, the response to which participants had to manually input. This sequence was as follows: 1. task relevant trial-type; 2. task irrelevant trial-type; 3. task relevant trial-type; 4. task irrelevant trial-type; open-text question; 5. task relevant trial-type;

6-8. task relevant trial-type in which feedback was manipulated between conditions. Each trial-type comprised two or three short statements, a question, and two or three response options. The trial-types were either task relevant or task irrelevant. The task relevant trial-types involved the key phrase that would be used in the rule (i.e. "LEAST LIKE"), provided subsequently in the MTS task (see Figure 1, left-hand side). The task irrelevant trial-types involved phrases that would not be necessary for completing the MTS task (see Figure 1, right-hand side). All trial-types in the Control Condition were task irrelevant. Each trial was also followed by a 7-point Likert scale in which participants were asked to grade how certain they were about their answer, ranging from 1 (very uncertain) to 7 (very certain).

KROS is the same as ZID. ZID is the same as LEAST LIKE. What does KROS mean? "MOST LIKE" "LEAST LIKE" "LEAST LIKE"	SAM is younger than TOM. TOM is younger than PAT. Which is the oldest one? "PAT" "SAM" "TOM"

Figure 1. The task-relevant trial-type presented to the High and Low Coherence Condition (left-hand side) and an example of a task-irrelevant trial-type (right-hand side) presented to all participants in the Coherence Task.

The first three task relevant trial-types included the first part of the network (A=B=C) that participants were being trained on, and the last three task relevant trial-types included the second part of this network (C=D=E=F; see Figure 2 for an illustration of this complete network per condition). The statements, question, and response options that comprised the task relevant trial-type in the first part of the network (A=B=C) are presented in Figure 1 (left-hand side). This trial-type was denoted as task relevant because it enabled participants to derive the meaning of the phrase "LEAST LIKE" from nonsense words which would then be necessary to respond correctly in the subsequent MTS task. In the first statement, the word "KROS" (C) was coordinated with the word "ZID" (B), and "ZID" was then coordinated with "LEAST LIKE" (A). Hence, participants could derive that "KROS" meant "LEAST LIKE". Participants could select the options "LEAST LIKE", "MOST LIKE" or "SAME", when asked "What does KROS mean?". This was task-relevant because "KROS" was subsequently presented in the MTS task.

	A	**=**	**B**	**=**	**C**	**=**	**D**	**=**	**E**	**=**	**F**	**(= D)**
High Coherence	LEAST LIKE	=	ZID	=	KROS	=	VEK	=	JUM	=	PO M	= VEK
Low Coherence	LEAST LIKE	=	ZID	=	KROS	=	VEK	=	JUM	=	PO M	≠ VEK
Control	FASTER/ SLOWER THAN	=	ZID	=	KROS	=	VEK	=	JUM	=	PO M	= VEK

Figure 2. The relational network with the variation in different conditions defined.

One of the task irrelevant trial-types is presented in Figure 1 (right-hand side). This trial-type was denoted as task irrelevant because nothing derived from it could be used to inform responding on the MTS task. In the first task irrelevant trial-type, "SAM" was said to be younger than "TOM", and "TOM" was said to be younger than "PAT". The participants were then asked which one is the oldest, and had a choice between the three names. The second irrelevant trial-type was similar, except that the relations between the three stimuli varied along the dimension of strength instead of age. The two task irrelevant trial-types were included to allow the researchers to check that participants were deriving the relationships accurately, rather than responding by trial and error.

The last three trial-types were task relevant, but were comprised of the last four stimuli of the network (i.e. C=D=E=F). This trial-type was denoted as task relevant because it enabled participants to coordinate the nonsense word “KROS”, previously coordinated with the word “LEAST LIKE”, with the rest part of the network (D=E=F), and to then use this derivation to respond correctly on the MTS task. The feedback given in this last trial-types were used as the manipulation in the experiment. In the first statement of the three last trial-types, the nonsense word “KROS (C) ” was coordinated with the nonsense word “VEK (D)”, “VEK” was then coordinated with “JUM (E) ”, and finally “JUM” was coordinated with “POM (F)”. Hence, participants could derive that “JUM” was coordinated with “KROS", which was previously derived to mean “LEAST LIKE”. To respond correctly, participants were required to select the “yes” response option, when asked “Are VEK and POM the same?”.

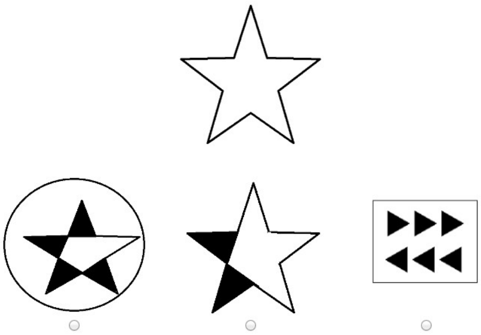


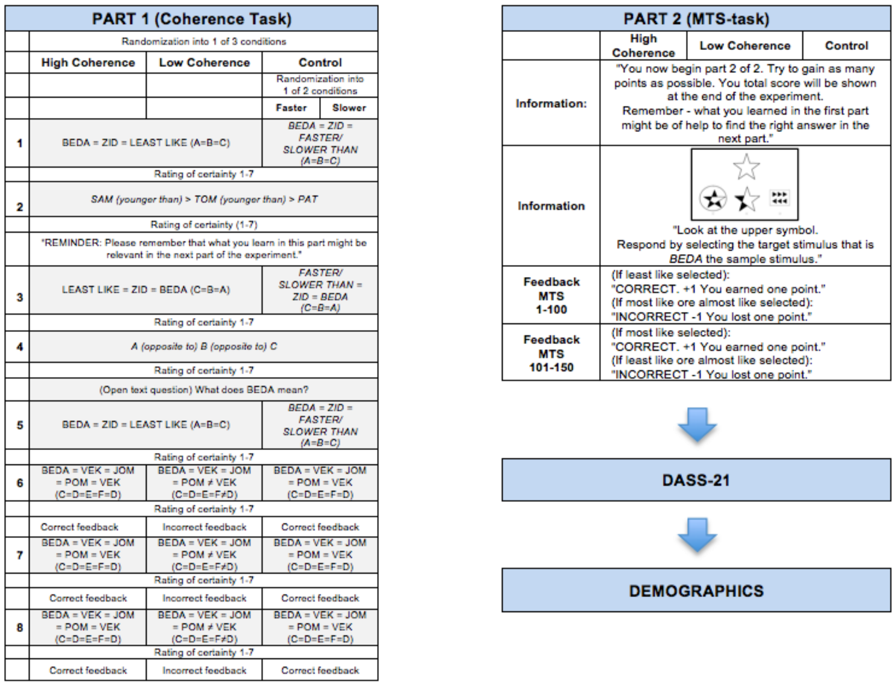
Figure 3. An example of a single trial and single stimulus set presented in the MTS task.
Note. MTS = match-to-sample.

MTS. The MTS task consisted of 150 trials. Each trial consisted of a sample stimulus (random shape), presented at the top of the screen, and three comparison stimuli below (all random shapes, but none identical to the sample stimulus or to each other; see Figure 3). Each comparison stimulus varied in its similarity to the sample stimulus presented, but one comparison stimulus was clearly most like the sample stimulus (see middle of Figure 3). Another comparison stimulus was also clearly like the sample stimulus, but had more variations in shape (see left-hand side of Figure 3), 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 (see right-hand side of Figure 3). Each sample stimulus and three-comparison stimuli in combination comprised an individual stimulus set, such that only those comparisons stimuli appeared in the presence of a specific sample stimulus. A total of 54 stimulus sets were used in the experiment with each being presented at least once and no more than twice.

DASS-21.The Swedish version of the DASS-21 (Alfonson et al., 2017) is a measure of general psychological distress and is comprised of three subscales measuring depression, anxiety and stress. All items 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”). Higher scores indicate poorer mental health. The English version has demonstrated high internal consistency (Henry & Crawfoord, 2005): depression (a =.82), anxiety (a = .90), and stress (a = .93). The Swedish translation has yielded similar sufficient internal consistency (Alfonson et al., 2017).

Procedure

The experiment involved three stages: the Coherence Task, the MTS task, the DASS-21 questionnaire and demographics, and were always conducted in this order (see Figure 4). The procedure is mainly a replication of the study made by Harte et al. (2017), with the exception of the Coherence Task and variations in the MTS task.



PART 1 (Coherence Task)				
	Randomization into 1 of 3 conditions			
	High Coherence	Low Coherence	Control	
			Randomization into 1 of 2 conditions	
			Faster	Slower
1	BEDA = ZID = LEAST LIKE (A=B=C)		*BEDA = ZID = FASTER/ SLOWER THAN (A=B=C)*	
	Rating of certainty 1-7			
2	*SAM (younger than) > TOM (younger than) > PAT*			
	Rating of certainty (1-7)			
	"REMINDER: Please remember that what you learn in this part might be relevant in the next part of the experiment."			
3	LEAST LIKE = ZID = BEDA (C=B=A)		*FASTER/ SLOWER THAN = ZID = BEDA (C=B=A)*	
	Rating of certainty 1-7			
4	*A (opposite to) B (opposite to) C*			
	Rating of certainty 1-7			
	(Open text question) What does BEDA mean?			
5	BEDA = ZID = LEAST LIKE (A=B=C)		*BEDA = ZID = FASTER/ SLOWER THAN (A=B=C)*	
	Rating of certainty 1-7			
6	BEDA = VEK = JOM = POM = VEK (C=D=E=F=D)	BEDA = VEK = JOM = POM ≠ VEK (C=D=E=F≠D)	BEDA = VEK = JOM = POM = VEK (C=D=E=F=D)	
	Rating of certainty 1-7			
	Correct feedback	Incorrect feedback	Correct feedback	
7	BEDA = VEK = JOM = POM = VEK (C=D=E=F=D)	BEDA = VEK = JOM = POM ≠ VEK (C=D=E=F≠D)	BEDA = VEK = JOM = POM = VEK (C=D=E=F=D)	
	Rating of certainty 1-7			
	Correct feedback	Incorrect feedback	Correct feedback	
8	BEDA = VEK = JOM = POM = VEK (C=D=E=F=D)	BEDA = VEK = JOM = POM ≠ VEK (C=D=E=F≠D)	BEDA = VEK = JOM = POM = VEK (C=D=E=F=D)	
	Rating of certainty 1-7			
	Correct feedback	Incorrect feedback	Correct feedback	

PART 2 (MTS-task)			
	High Coherence	Low Coherence	Control
Information:	"You now begin part 2 of 2. Try to gain as many points as possible. You total score will be shown at the end of the experiment. Remember - what you learned in the first part might be of help to find the right answer in the next part."		
Information	"Look at the upper symbol. Respond by selecting the target stimulus that is *BEDA* the sample stimulus."		
Feedback MTS 1-100	(If least like selected): "CORRECT. +1 You earned one point." (If most like ore almost like selected): "INCORRECT -1 You lost one point."		
Feedback MTS 101-150	(If most like selected): "CORRECT. +1 You earned one point." (If least like ore almost like selected): "INCORRECT -1 You lost one point."		

DASS-21

DEMOGRAPHICS

Figure 4: Procedure flow chart. The order of the Coherencstask, MTS-task, DASS-21, and demographics presented.
Note: Text in italic marks task irrelevant trial-types

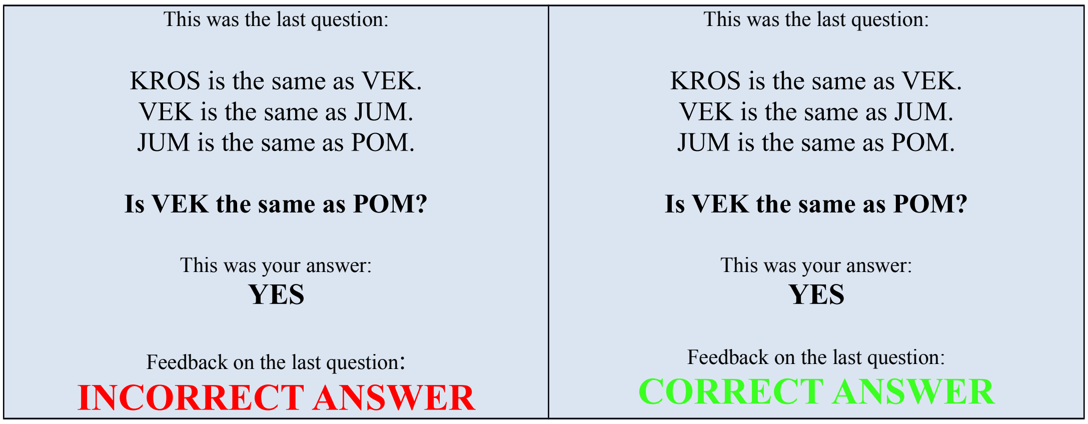
To begin, participants had to fill in a consent form, accepting the conditions of the experiment. Before starting the Coherence Task, the participants were told “You now begin part 1 of 2. What you learn in the first part might be of help in the second part. You will sometimes receive feedback on your answers and sometimes not”.

The Coherence Task. In the High Coherence and Low Coherence Conditions, participants received eight trial-types and one open-text question in the following order: task relevant, task irrelevant, task relevant, task irrelevant, open-text question, task relevant, and lastly, three identical task relevant trial-types. After the first two trial-types, all participants received a reminder that “What you learn in this part might be relevant in the later part of the experiment”. After the second trial-type, an open text box was presented in order to check if the participants had derived the correct meaning of “KROS”, asking them to write in text what KROS meant, based on what they had just learnt. Another purpose of this was to lower the risk of potential attrition. The Control Condition was similar to the active conditions, except that the word “LEAST LIKE” in the task relevant trial-types was replaced with “SLOWER THAN” or “FASTER THAN”. All trial-types in the Control Condition were thus irrelevant to the rule that participants subsequently received on the MTS task.

After each trial-type, all participants were asked to grade on a 7-point scale how certain they were about their answer, ranging from 1 (very uncertain) to 7 (very certain), in order to strengthen the relations between the stimuli within the network.

Feedback was only given on trial-types six through eight, and the type of feedback provided was manipulated between the conditions (described subsequently).

High Coherence Condition. The aim of the Coherence Task in the High Coherence Condition was to allow participants to derive the critical part of a coherent relational network (A=B=C=D=E=F) that would be subsequently used in the rule required for completing the following MTS task (i.e., to choose the comparison stimuli "LEAST LIKE" the sample stimulus). When giving the correct response in the last task, participants in the High Coherence Condition received feedback (i.e. "correct answer"), in order to create a high level of coherence (see Figure 5, right-hand side).



This was the last question:

KROS is the same as VEK.
VEK is the same as JUM.
JUM is the same as POM.

Is VEK the same as POM?

This was your answer:
YES

Feedback on the last question:
INCORRECT ANSWER

This was the last question:

KROS is the same as VEK.
VEK is the same as JUM.
JUM is the same as POM.

Is VEK the same as POM?

This was your answer:
YES

Feedback on the last question:
CORRECT ANSWER

Figure 5. Feedback in the two main conditions Low Coherence Condition (left-hand side) and High Coherence Condition (right-side) in the last three task-relevant trial-types in the Coherence Task.

Low Coherence Condition. This condition was similar to the High Coherence Condition, except that the feedback on the last three trial-types was incorrect even when the correct answer was given, and vice versa (see Figure 5, left-hand side). The incorrect feedback to the correct answer on the three last trial-tasks enabled participants to perceive some part of the network as incoherent. However, the coherence of the whole network was not violated since the incorrect feedback was only given to the more peripheral part of the network, resulting in C=D=E=F≠D, with participants still able to correctly derivate that "KROS" is the same as "LEAST LIKE".

Control Condition. In this condition, participants were divided randomly into one of two conditions. The trial-types in these conditions were identical to the High Coherence Condition except that the word "LEAST LIKE" was replaced with either "SLOWER THAN" or "FASTER THAN" respectively. The group that received "SLOWER THAN" was referred to as the Control Condition Slower while the group that received "FASTER THAN" was referred to as the Control Condition Faster. This randomization was to counterbalance for the possible implications within the MTS task based on the derived meaning of "KROS" within the control group. Therefore, all trial-types were task irrelevant as none of them helped participants complete the following MTS task.

MTS. All participants were advised to try to gain as many points as possible in the MTS task. Participants were informed 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." All participants were then instructed to "Respond by selecting the target stimulus that is KROS the sample stimulus." Participants got positive feedback (i.e. "correct answer") if responding by choosing the least like comparison stimuli the first 100 trials. Participants received negative feedback (i.e. incorrect answer") if they chose either the almost like or most like comparison stimulus. Participants were also informed that their total score either increased or decreased by one point, depending on their answer. After 100 trials, from MTS trial 101 to 150, the contingency switched unbeknownst to the participants, and they now received reversed feedback (i.e. positive feedback when they chose the most like sample stimulus, and negative feedback when they chose the least like or almost like sample stimulus).

DASS-21. Finally, participants completed the DASS-21.

Results

The strict accuracy criterions applied in Harte et al. (2017) was also applied to the current analysis. First, participants in the High Coherence Condition and the Low Coherence Condition were required to make at least 8 out of the first 10 responses correct on the MTS task to be included in the analysis. This was to reduce the likelihood that participants learned to match based purely on trial and error, and instead were able to apply the derived rule from the first part of the experiment (the Coherence Task). The data from participants who did not meet this accuracy criterion in the initial trials were excluded. This accuracy criterion was not applied to the Control Condition since it was expected that few participants would meet it (i.e. they had no rule to follow during their initial exposure to the MTS task). Nevertheless, 28,3 % of Control participants emitted at least 8/10 correct responses in the first 10 trials.

Second, as in Harte et al. (2017), participants from all three conditions were required to achieve 80 out of the first 100 responses correct on the MTS task (before the switch in contingencies). This was based on the assumption that an acceptable number of Control participants would have adapted to the contingencies across the first 20 trials (Harte et al.). The data from participants who did not meet this accuracy criterion were excluded from the analyses. Additionally, in order to exclude anyone who did not successfully make the required derivation in the Coherence Task, participants who gave an inaccurate ("NO") response to the last trial-task (i.e. "KROS is the same as VEK. VEK is the same as JUM. JUM is the same as POM. Is VEK the same as POM?") were also excluded from the analysis.

Before conducting the primary analyses, possible differences between the two Control Conditions as a result of the "faster than/slower than" manipulation were assessed. 5 independent *t*-tests were conducted to compare the means of all the relevant variables used in the study (Rule Compliance-scores and Contingency Sensitivity-scores on the MTS described subsequently), and Depression, Stress, Anxiety scores as measured by the DASS-21) between Control Condition Slower and Control Condition Faster (see Table 1). No significant difference in mean scores were found, *p*s > .05. Thus, the two Control Conditions were therefore merged and analyzed as one condition in the following analyses.

Analysis of Coherence

In order to examine whether the primary manipulation resulted in different levels of coherence between the different conditions, the means of each conditions' certainty scores in the last three trial-types in the Coherence Task (i.e. in which participants were given different feedback depending on which condition they were in) were calculated. A one-way between-subject analysis of variance (ANOVA) was conducted with condition (High vs. Low vs. Control) as the independent variable, and Coherence Task mean certainty scores as the dependent variable. There was a significant difference in certainty found between the three groups, *F*(2, 145) = 14.72, *p* < .001, η2 = 0.17 (see Table 2). Post-hoc comparisons using Bonferroni corrections indicated that participants in the Low Coherence Conditions reported significantly lower levels of certainty than both the High Coherence Condition and the Control Condition, *p* < .001. There was no significant difference between the High Coherence and Control Conditions. These results suggest that the manipulation of the feedback resulted in a greater level of uncertainty (i.e. coherence) in the Low Coherence Condition.

Table 2

Effect of Condition on Self-Reported Certainty
Note. Numbers in parentheses are standard deviations.
[a] Self-reported level of certainty after Coherence task 6-8 (1 = "Not certain at all" to 7 = "Very certain").

Analysis of Rule-Compliance and Contingency Sensitivity

Since the aim of the experiment was to compare performances between the High and Low Coherence Conditions, the data from the 50 MTS trials, presented after the contingency switch, were analyzed in two different, but related, ways (as in Harte et al., 2017). These are referred to as RC (i.e. rule-compliance) and CS (i.e. contingency sensitivity).

RC was defined as the total number of responses out of 50 that was consistent with the rule or derivation "LEAST LIKE"/"KROS", but inconsistent with the reversed contingencies for those 50 trials after the contingencies switch. Frigure 6 (left-hand side) presents the mean RC-scores for each condition. A one-way between subjects analysis of variance (ANOVA) was conducted to compare the mean total number of responses (out of 50) that were consistent with the rule/derivation, after the contingency switch between the High Coherence, Low Coherence, and Control Conditions. There was a significant difference in RC-scores between the three different conditions, *F*(2, 147) = 12.88, *p* < .001 ηp2 = 0.15. Post-hoc comparison using Tukey's HSD indicated that the mean RC-score for the Control Condition (*M* = 5.98, *SD* = 6.93) was significantly lower than the High Coherence Condition (*M* = 21.34, *SD* = 20.26) and Low Coherence Condition (*M* = 18.62, *SD* = 18.26), *p*s < .001. However, there was no significant difference in rule-compliance between the two main conditions, *p* = .68.

	Conditions			
	High Coherence (*n* = 47)	Low Coherence (*n* = 50	Control (*n* = 51)	Total (*N* = 148)
Mean Certainty Score[a]	6.36 (.94)	5.43 (1.25)	6.43 (.83)	6.07 (1.12)

CS was defined as the point at which participants stopped following the initial rule and began to respond in accordance with the reversed contingencies. In effect, it would more accurately reflect early CS among participants who subsequently reverted back to rule-consistent responding after a small number of trials. Thus, the overall number of rule-consistent responses would remain high, even though some participants may have shown relatively rapid contingency sensitivity. On balance, some participants might have shown random or occasional rule-inconsistent behavior in their responding. The definition of contingency sensitivity used by Harte et al. (2017, p. 11) and employed here was "the point at which responding in accordance with the reversed contingency emerged and did not return reliably to a rule-consistent pattern". This point was defined as "three or more consecutive responses in accordance with the reversed contingency followed by no more than four consecutive rule-consistent responses thereafter" (Harte et al., p. 11).

Rule Compliance
Total Numer of rule consistent responses after contingency switch
Contingency Sensivity
Number of rule consistent responses before contingency sensitive responses

Frigure 6. Mean Rule-Compliance scores (left-hand side) and Contingency Sensitivity scores (right-hand side) with standard error bars for the High Coherence, Low Coherence, and Control Conditions.

Frigure 6 (right-hand side) presents the mean CS-scores for each condition. A one-way between-subjects analysis of variance (ANOVA) was conducted to compare at which point this occurred between the High Coherence, Low Coherence, and Control Condition. There was a significant difference in CS-scores between the three conditions, *F*(2, 143) = 11.02, *p* < .001 ηp2 = 0.13. Post-hoc comparisons using Tukey's HSD indicated that the mean CS-score for the Control Condition (*M* = 6.88, *SD* = 6.75) was significantly lower than the mean CS-scores in High Coherence Condition (*M* = 21.00, *SD* = 19.60) and Low Coherence Condition (*M* = 18.22, *SD* = 18.46), *p*s ≤ .001. That is, participants in the High and Low Coherence Conditions emitted significantly more responses in accordance with the original rule than the Control Condition in the face of the reversed reinforcement contingencies. However, there was no significant difference on contingency-sensitivity between the two main conditions, *p* = .67 (see Frigure 6).

DASS-21

A Pearson's correlation analysis was calculated among each condition, between reported levels of stress, depression, anxiety, CS-score, and RC-score for each condition. Three correlations were found to be significant, all within the Control Condition. RC correlated positively with stress (*r* = .35 *p* = .01) and anxiety (*r* = .33, *p* = .02) on the DASS-21, suggesting that greater rule compliance predicted higher levels of stress and anxiety. In addition, CS also correlated positively with DASS-21 stress (*r* = .33, *p* = .02), suggesting that less sensitivity predicted higher levels of stress (see Table 4).

Table 4.

Correlations between DASS-21, RC and CS-scores for Control Condition

	Contingency Sensitivity	Stress	Depression	Anxiety

Rule Compliance	.98***	.35*	.08	.33*
Contingency Sensitivity		.33*	-.04	.27
Stress			.64***	.73***
Depression				.43**

Note. * *p* < .05 ** *p* < .01. *** *p* < .001

In order to assess possible effects on stress, depression, and anxiety that each condition might have caused participants, a one-way between-subjects ANOVA was conducted with condition (High vs. Low vs. Control) as the independent variable, and DASS-21 mean scores (stress, depression, and anxiety) as the dependent variables. There was a main effect of condition on levels of stress *F*(2, 145) = 3.54, *p* < .05, η2 = 0.05. Post-hoc comparison using Bonferroni corrections revealed that participants in the Control Condition reported significantly higher levels of stress (*M* = 1.20, *SD* = .63), compared to the Low Coherence Condition (*M* = .88, *SD* = .60), *p* = .03. The High Coherence Condition did not significantly differ from the other two conditions (*M* = .99, *SD* = .59), *p* > .05. There was no significant differences found between the three conditions on reported levels of depression and anxiety, *p*s > .33. This suggests that the amount of self-reported depression and anxiety the experiment itself might have caused the participants were equal in all three conditions. That is, the significant relationship between increased rule-compliance and higher levels of anxiety, found in the Control Condition, could not be explained by the absence of a rule that was helpful.

Discussion

The primary purpose of the current research was to explore how different levels of coherence in a rule may impact persistence in rule-following. The study failed to produce a statistically significant difference with regard to whether different levels of coherence leads to differences in rule persistence. However, the main conditions (High and Low Coherence) persisted significantly more in accordance with the rule than did the Control Condition. While the coherence manipulation did not differentially impact persistence in rule-following, the participants self-reports on their certainty of the network differed significantly depending on which condition they were in. That is, participants in the High Coherence Condition were significantly more certain that KROS meant "least like" than participants in the Low Coherence Condition.

Interestingly, the present findings overlap with the results found by Harte and colleagues (2017), and since the MTS tasks were similar between the studies, some tentative comparisons can be made. The conditions relatively high in derivation, i.e. the Derived Rule Condition in the Harte study, and the two main conditions in the present study, resembled each other quantitatively concerning both rule compliance and contingency sensitivity (the present study was overall high in derivation since nonsense words were used across all conditions to establish the rule network). While highly speculative, this can be said to add further evidence that higher levels of derivation is connected to lower levels of rule persistence (see Barnes-Holmes, Barnes-Holmes, Hussey & Luciano, 2016). Since this study manipulated levels of coherence when derivation could be seen as high, future work could keep the levels of derivation lower, e.g. by only using previous known words.

Due to the exploratory nature of this study, predictions were not made beforehand. Nevertheless, some theoretical assumptions can be made based on the findings. A possible reason the main conditions did not differ with regard to rule-persistancy or contingency sensitivity could be because following the rule actually provided the participants with points during the first 100 trials. This might have served to make the rule network more coherent also in the Low Coherence Condition. More informally, when the rule actually succeeded to provide the participants with points (despite being strange), the participants in the Low Coherence Condition might have forgot about the previous strangeness of the rule network, presented approx. 10-15 minutes earlier. The coherence in this condition can in one way be claimed to be *initially* low, but then to gradually increase when being reinforced for following the rule in the MTS task. Thus, at the time of the switch, it is possible that the main conditions had reached similar levels of coherence, even though this is speculative. One way to control for this possible effect would be to have fewer trials in the MTS task.

Speculative as it is, this alleged increase in coherence in the Low Coherence Condition might be interpreted as a preference for, and strive towards, coherence (Wray et al., 2012; Hayes, Strosahl & Wilson, 1999). In a similar vain, other studies have found consistent rule following to be related to increased levels of confidence, even when performance is poor (Williams, Dunning & Kruger, 2013). Taken together, this could suggest that the incoherence of rules is easily forgotten, speculative as this claim is. Future studies could ask the participants about their certainty of the rule after its been reinforced several times, but before the contingency switch.

Another possible explanation for the similarity between the main conditions could be that the incorrect feedback given in the Low Coherence Condition did not directly challenge the part of the network controlling the rule (A=B=C, i.e. LEAST LIKE=ZID=KROS). Thus, it may be that the feedback did not propagate to the rule word KROS (meaning “least like”). More informally, the participants could be said to have split the network in a coherent and an incoherent part. Future studies could investigate what would happen if the coherence of the rule word was undermined more directly. One potential way to do this would be to tell the participants in the Low Coherence Condition that the A=B=C part of the network was (sometimes) incoherent.

In comparison to research demonstrating that different levels of derivation impact persistent rule-following (Harte et al., 2017), different levels of coherence did not significantly affect either rule compliance or contingency sensitivity measures. Speculative as it is, this might lead to two different conclusions: 1) coherence does not effect RGB to a relevant extent, or 2) the level of coherence has to reach a certain level to have an impact on RGB.

The first conclusion is supported by this study to some extent, which may shed interesting light on RGB, viz. that other variables than coherence may be relevant in explaining RGB. One possible clinical implication could be that the therapist should not focus on establishing a maladaptive rule as incoherent, since that would not have the desired effect (i.e. no effect at all). The therapist could instead work on adding new behaviors or rules to outmaneuver the maladaptive ones. Another somewhat optimistic interpretation might be that clients are prone to follow new rules, even though they seem to be incoherent with their previous knowledge or behavior. This could be said to be in line with findings in social psychology where people can act seemingly odd and immorally when simply instructed to do so (Milgram, 1963; Zimbardo, 2007).

The second tentative conclusion, that the level of coherence has to reach a certain level in order to impact rule-governed behavior, might be a more reasonable conclusion based on the results in this current study. Speculatively, this could be claimed to be in line with newer conceptualizations in RFT regarding how clients can readily shift between different sets, or networks of behaviors (Barnes-Holmes et al., 2018).

Regarding the measures of mental health, a significant difference was found in the Control Condition only, where contingency insensitivity were related to higher levels of stress, and rule-persistence were related to higher levels of stress and anxiety. The participants in the Control Condition received a rule that was useless in the MTS task. A highly tentative interpretation of this finding is that being shuttlecocked between first following a malfunctioning rule, to subsequently derive a functioning rule that later on ceases to work, but is anyway being stuck to, is associated with increased stress and anxiety. Perhaps, this is because they derived a rule more strongly related to themselves in comparison to the main conditions where a correct rule was given to them.

The current findings could be compared to the result in the study by Harte and colleagues (2017), where participants given a direct rule that was subsequently abandoned when ceasing to work reported significantly higher levels of stress. This was tentatively interpreted as disobeying a clear and well-established rule increases levels of stress. Contrary to the present study, we found no significant increase in stress associated with abandoning a rule when it ceased to work. On the contrary, the association found was between stress and rule compliance, and contingency sensitivity in the Control Condition. This difference between the studies could tentatively be explained by the fact that the control conditions differed across the studies, since in the study by Hart et al., the participants in the control condition did not receive a malfunctioning rule, but only a rule to gain as many points as possible, somewhat more in line with traditional research on RGB where the participants in the control conditions responds by trial and error. The participants in the present study however received identical rules in the different conditions with the only difference being having different meaning associated to the key word in the rule (i.e. KROS). Even though this is a experimental strength, it obstruct comparisons with the Harte study.

Although this research contributed to preliminary insights in the understanding of RGB and coherence, five limitations in the current study should be noted. First, based on the procedural issues mentioned in Harte et al. (2017), related to participants not being able to transfer information from the Derivation Task (i.e. Coherence Task) to the MTS task, some procedural changes were made in attempt to reduce the attrition rate. Procedural instructions were added in this current study, reminding participants that the information in the first part of the study (the Coherence Task) could be useful in the second part of the study (the MTS task). Compared to the study conducted by Harte et al., this resulted in more equal numbers of participants in the main conditions who were able to meet the inclusion criterias (64 % in the High Coherence Condition were excluded, vs. 67 % in the Low Coherence Condition). However, since the exclusion rate was still considerably high, this has to be considered a limitation.

Second, another limitation of the current research is the relatively narrow sample, consisting of university students, from which participants were drawn. This might limit the generalizability of the findings.

Third, a procedural issue in this current study was that accumulating points were not displayed for the participants during the MTS task, which it was in the study by Harte and colleagues (2017). This might have weakened the influence of the feedback, and aggravated their ability to see their behavior in a broader perspective. That is, the direct feedback might not have produced the desired effect and might have lost impact in both reinforcing and punishing functions.

Forth, perhaps the incorrect feedback provided in the Coherence Task in the Low Coherence Condition was not punishing enough to trump the participants sense of being right on their answers. Maybe using a stronger punisher, like losing money, would increase their sense of incoherence. However, the punishment in the Low Coherence Condition (being told they were wrong when they were actually right) was still strong enough to create significant doubt within this condition, in comparison to the other conditions.

Fifth, to follow the procedure in Harte et al. (2017), the DASS-21 measure was taken after participants completed the experimental tasks. Thus, it is possible that the difference in reported levels of stress and anxiety were present before the experiment.

The branch of Relational Frame Theory and RGB has contributed separately to important aspects of psychology and human suffering. However, not until recently did these fields begun to interact research-wise. Putting verbal behavior in the forefront, research is now beginning to understand this previous gap in the literature concerning the core elements in language and cognition. Even though this study is exploratory and many interpretations are speculative, it adds insights to how humans are affected by the rules they (do not) live by and the sense they are (not) making.

References

Alfonsson, S., Wallin, E., & Maathz, P. (2017). Factor structure and validity of the depression, anxiety and stress scale-21 in swedish translation. *Journal of Psychiatric and Mental Health Nursing, 24*(2-3), 154-162. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1111/jpm.12363

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, UK: John Wiley.

Barnes Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From the IRAP and REC model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable relational responding. *Journal of Contextual Behavioral Science, 6*(4), 434-445.

Barnes-Holmes, Y., Boorman, J., Oliver, J. E., Thompson, M., McEnteggart, C., & Coulter, C. (2018). Using conceptual developments in RFT to direct case formulation and clinical intervention: two case summaries. *Journal of Contextual Behavioral Science, 7*, 89-96. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1016/j.jcbs.2017.11.005

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., Lowe, C., & 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.

Bird, A., & Reese, E. (2006). Emotional reminiscing and the development of an autobiographical self. *Developmental Psychology, 42*(4), 613-626. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/0012-1649.42.4.613

Blackledge, J. T., Moran, D. J., & Ellis, A. (2009). Bridging the divide: linking basic science to applied psychotherapeutic interventions—a relational frame theory account of cognitive disputation in rational emotive behavior therapy. *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 27*(4), 232-248.

Bordieri, M., Kellum, J., Wilson, K., & Whiteman, K. (2015). Basic properties of coherence: testing a core assumption of relational frame theory. *The Psychological Record, 66*(1), 83-98.

Borkovec, Robinson, Pruzinsky, & Depree. (1983). Preliminary exploration of worry: some characteristics and processes. *Behavior Research and Therapy, 21*(1), 9-16.

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, NY: Plenum.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*(3), 193-204. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/h0024670

Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization, 126*, 137-154. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1016/j.jebo.2015.10.016

Chomsky, N. (1959). Reviews. Verbal behavior by B. F. Skinner. *Language, 35*, 26-58.

Dymond, S., Bennett, M., Boyle, S., Roche, B., & Schlund, M. (2017). Related to anxiety: arbitrarily applicable relational responding and experimental psychopathology research on fear and avoidance. *The Behavior Analyst,* doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1007/s40614-017-0133-6

Fotopoulou, A. (2008). False selves in neuropsychological rehabilitation: the challenge of confabulation. *Neuropsychological Rehabilitation, 18*(5-6), 541-565. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1080/09602010802083545

Grawe, K. (2007). *Neruopsychotherapy: How the Neurosciences Inform Psychotherapy*. London: Lawrence Erlbaum Associates.

Hammack, P. L. (2011). *Narrative and the politics of identity: The cultural psychology of israeli and palestinian youth.* New York, NY: Oxford University Press.

Harrison, R., & Green, G. (1990). Development of conditional and equivalence relations without differential consequences. *Journal of the Experimental Analysis of Behavior, 54*(3), 225-237.

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:http://dx.doi.org.ezproxy.ub.gu.se/10.1177/0145445517715871

Hayes S. C., Barnes-Holmes, D., & Roche, B. (Eds.) (2001). *Relational Frame Theory: A post-Skinnerian account of human language and cognition*. New York: Kluwer Academic/Plenum Publishers.

Hayes, S., Barnes-Holmes, D., Biglan, A., & Zettle, R. (Eds.) (2016). *The Wiley handbook of contextual behavioral science*. Chichester, West Sussex, UK: Wiley.

Hayes, S. C., & Gregg, J. (2000). Functional contextualism and the self. In C. Muran (Ed.), *Self-relations in the psychotherapy process* (pp. 291-307). Washington, DC: American Psychological Association.

Hayes, S. C., Strosahl, K., & Wilson, K. G. (1999). *Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change*. New York: Guilford Press.

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

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology, 128*(1), 3-31. Retrieved from https://search-proquest-com.ezproxy.ub.gu.se/docview/213790749?accountid=11162

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107-128. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/h0034225

Leonhard, C., & Hayes, S. C. (1991). Prior inconsistent testing affects equivalence responding. Paper presented at the meeting of the Association for Behavior Analysis, Atlanta.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. *Behaviour research and therapy*, *33*(3), 335-343.

Lowe, C., Beasty, A., & Bentall, R. (1983). The role of verbal behavior in human learning: infant performance on fixed-interval schedules. *Journal of the Experimental Analysis of Behavior, 39*(1), 157-164.

Luciano, C., Rodrigquez, M., Manas, I., & Ruiz, F. (2009). Acquiring the earliest relational operants: coordination, difference, opposition, comparison, and hierarchy. In R. A. Rehfeldt & Y. Barnes-Holmes (Eds.), *Derived relational responding: Applications for learners with autism and other developmental disabilities* (pp. 149–172). Oakland, CA: New Harbinger.

McAdams, D., Reynolds, J., Lewis, M., Patten, A., & Bowman, P. (2001). When bad things turn good and good things turn bad: sequences of redemption and contamination in life narrative and their relation to psychosocial adaptation in midlife adults and in students. *Personality and Social Psychology Bulletin, 27*(4), 474-485.

McAuliffe, D., Hughes, S., & Barnes-Holmes, D. (2014). The dark-side of rule governed behavior. *Behavior Modification, 38*(4), 587-613.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, *67*(4), 371.

Mineka, S., & Hendersen, R. (1985). Controllability and predictability in acquired motivation. *Annual Review of Psychology, 36*, 495-529.

O'Connor, M., Farrell, L., Munnelly, A., & McHugh, L. (2017). Citation analysis of relational frame theory: 2009–2016. *Journal of Contextual Behavioral Science, 6*(2), 152-158.

Peterson, C., & Seligman, M. E. (1984). Causal explanations as a risk factor for depression: theory and evidence. *Psychological Review, 91*(3), 347-374. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/0033-295X.91.3.347

Pilgrim, C., & Galizio, M. (1995). Reversal of baseline relations and stimulus equivalence: I. adults. *Journal of the Experimental Analysis of Behavior, 63*, 225–238.

Quinones, J. L., & Hayes, S. C. (2014). Relational coherence in ambiguous and unambiguous relational networks. *Journal of the Experimental Analysis of Behavior, 101,* 76–93.

Roche, B., Barnes-Holmes, D., Barnes-Holmes, Y., & Hayes, S. C. (2001). Social processes. In S. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 197–210). New York, NY: Kluwer Academic/Plenum Publishers.

Rosenfarb, I. S., Burker, E. J., Morris, S. A., & Cush, D. T. (1993). Effects of changing contingencies on the behavior of depressed and nondepressed individuals. *Journal of Abnormal Psychology, 102*(4), 642-646. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/0021-843X.102.4.642

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

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

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

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: from consistency theories to constraint satisfaction. *Personality and Social Psychology Review, 6*(4), 283-294. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1207/S15327957PSPR0604_03

Skinner, B.F. (1936). The verbal summator and a method for the study of latent speech. *The Journal of Psychology, 2*(1), 71-107.

Skinner, B.F. (1966). An operant analysis of problem solving. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, teaching* (p. 225-257). New York: Wiley.

Starr, B. J., & Katkin, E. S. (1969). The clinician as an aberrant actuary: illusory correlation and the incomplete sentences blank. *Journal of Abnormal Psychology, 74*(6), 670-675. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/h0028466

Stewart, C., Stewart, I., & Hughes, S. (2016). A contextual behavioral approach to the study of (persecutory) delusions. *Journal of Contextual Behavioral Science, 5*(4), 235-246.

Törneke, N. (2014). *Relationsinramningsteori - RFT : Teori och klinisk tillämpning* (2nd ed.). Lund: Studentlitteratur.

Törneke, N., Luciano, C., & Salas, S. (2008). Rule-governed behavior and psychological problems. *International Journal of Psychology and Psychological Therapy, 8*(2), 141-156.

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, NY: Plenum.

Vaughan, M. E., & Michael, J. L. (1982). Automatic reinforcement: an important but ignored concept. *Behaviorism, 10*(2), 217-227. Retrieved from https://search-proquest-com.ezproxy.ub.gu.se/docview/616838992?accountid=11162

Villatte, M., Villatte, J. L. & Hayes, S. C. (2016). *Mastering the clinical conversation: Language as intervention*. New York: The Guilford Press.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin, 134*(2), 163-206. doi:http://dx.doi.org.ezproxy.ub.gu.se/10.1037/0033-2909.134.2.163

Williams, E., Dunning, D., Kruger, J., & Smith, Eliot R. (2013). The hobgoblin of consistency: algorithmic judgment strategies underlie inflated self-assessments of performance. *Journal of Personality and Social Psychology, 104*(6), 976-994.

Wilson, K. G., & Hayes, S. C. (1996). Resurgence of derived stimulus relations. *Journal of the Experimental Analysis of Behavior, 66*, 267–281.

Wray, A. M., Dougher, M. J., Hamilton, D. A., & Guinther, P. M. (2012). Examining the reinforcing properties of making sense: a preliminary investigation. *The Psychological Record, 62*(4), 599-622. Retrieved from https://search-proquest-com.ezproxy.ub.gu.se/docview/1269432429?accountid=11162

Zimbardo, P. (2007). *The lucifer effect: Understanding how good people turn evil.* New York, NY: Random House.