

University of Nevada Reno

Improvements in the Frequency and Accuracy of “I Cannot Know” Type Answers to Three Term Series Word Problems by Multiple Exemplar Training of Relational Skills and Instructional Content.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology

by

Patrick Smith

Dr. Steven C Hayes/Dissertation Advisor

May, 2023

Copyright by Patrick M. Smith 2023

All Rights Reserved



THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

PATRICK SMITH

entitled

**Improvements in the Frequency and Accuracy of “I Cannot Know” Type Answers
to Three Term Series Word Problems by Multiple Exemplar Training of Relational
Skills and Instructional Content.**

be accepted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

Steven C Hayes, Ph.D., *Advisor*

Bethany Contreras, Ph.D., *Committee Member*

Barbara Kohlenberg, Ph.D., *Committee Member*

Matthew Lewon, Ph.D., *Committee Member*

Dave Croasdell, Ph.D., *Graduate School Representative*

Markus Kimmelmeier, Ph.D., *Dean, Graduate School*

May, 2023

Abstract

The ability to accurately use “Not Enough Information” or “I Cannot Know” (ICK) types of responses in everyday life may contribute to the degree to which an individual responds sustainably in situations likely to evoke biased outcomes. This skill may have an analog function previously described but rarely investigated within the Relational Frame Theory (RFT) literature.

This dissertation describes a series of surveys and increasingly elaborate multiple exemplar training (MET) protocol experiments aimed at replicating common ICK inaccuracy in three term series problems and improving that performance via training Known-Unknown (KU) relational responding and introducing an instructional intervention. Accuracy and frequency of ICK responses were analyzed for indications of training effects and evidence of relational responding. Forms of inaccurate use of ICK responses were tested for indications of KU functioning.

Results suggest that (1) US and Irish participants infrequently and inaccurately respond with ICK responses to free response three term series word problems where it is appropriate; (2) KU relational responding can be trained to mastery criterion via MET protocols; (3) KU training is enhanced by building a relational response repertoire of more common functions such as equivalence and comparative relations; (4) trained KU responding can improve the accuracy of ICK responding on three term series word problems; (5) instructions can evoke more frequent, though inaccurate, ICK responding; (6) observed improvements on three term series problems from training KU relations are undetectable against instructional effects; and (7) ICK responses likely serve multiple behavioral functions.

Acknowledgments

A project of this duration and scope required both personal and professional support. Something like this would never have been possible without many people contributing in many different ways.

Personally, Tara, Lucy, and Mavis, this is as much yours as it is mine. Mom, Dad, Sister, Steve, Renee, Kyle, Micah, Anna, and Cam, each of you demonstrated endless support, wisdom, vitality, understanding, persistence, and an ability to steer me away from the most impulsive aspects of me, and by doing that, kept me on track and moving forward time and time again. Emily, you were a steadfast sounding board and crucial guide for how to navigate grad school. John and Laura, at every turn, you unselfishly offered help as if I was your own child. I aspire to live to the fullest good you each see in me and hope in doing so am able to pass along the kindnesses to others.

Professionally, Steve, thank you for building an environment that simultaneously reinforced strong scientific skills and yet allowed me to range far and wide to find what really excited my interest. Allison, you became the accountability buddy I needed to turn wild ideas into a concrete, thought through, and detailed experiment. Neal, thank you for being right there with me on long working weekends in the lab. Stu, your practical experience and dedication to the science generated valuable insight and kept my language sharp. Sara, Jessie, Lauren, Madison, and Annalise, as research assistants, you thought you were being trained by me, but in reality, I was learning just as much by working with each of you. Thank you for volunteering your time and energy toward all of my harebrained ideas.

There are so many who have touched this project. All in a meaningful way. Sufficient acknowledgement of my deep appreciation would outnumber the pages of the manuscript itself. While there is only one name on the byline, I am, and by extension this is, a reflection of my history with each of you. A history I cherish and appreciate making every day. Thank you.

Table of Contents

Introduction	1
Relational Frame Theory	3
Knowing We Don't Know	7
Theoretical Review	8
Figure 1: 10 Stimuli Relational Network Simulation	18
Literature and Pilot Experiments Review	21
Sorting Three Coins Relationally	22
Figure 2: Vitale et al., 2008 non-KU Trial	23
Figure 3: Vitale et al., 2008 KU Trial	23
Table 1: Vitale et al., 2008 48 Permutations of three coins	24
Figure 4: Vitale et al., 2008 modified Trials	25
Forced Wrong Answer for KUs	25
Figure 5: Quinones & Hayes, 2014, Experimental flow chart	26
Figure 6: Quinones & Hayes, 2014, KU Test Trial	27
Table 2: Quinones & Hayes, 2014, Tables 11 and 12 highlighting biasing outcome	27
Pilot 1	28
Table 3: 9-Item 3-Term Series word problems and answers	29
Figure 7: Pilot 1 Trial Screen	32
Pilot 2	33
Figure 8: Pilot 2 & 3 experimental flow diagram	34
Table 4: Three Stimulus networks	35
Figure 9: Pilot 2 & 3 trial logic diagram	35
Table 5: Specific trial training sequences	36
Pilot 3	37
Pilot Experiment Summary	38
Figure 10: Non-KU trials Before and After relative coin size was introduced	39
Figure 11: KU trials after relative coin size was introduced	41
Proposal	43
Planned Experimental Change #1: Extending the Current Training	46
Figure 12: Addition of a mastery criteria and retry contingency	47
Planned Experimental Change #2: Simpler Relational Training	48
Figure 13: Proposed Relational Training Modifications	49
Planned Experimental Change #3: Response Cost	50

Figure 14: Proposed Points Counter	51
Planned Experimental Change #4: Correct Previous Answers Review	51
Figure 15: Proposed Previous Answer	52
Planned Experimental Change #5: Known Word Questions	53
Figure 16: Proposed Known Word Version	54
Planned Experimental Change #6: ICK Evoking	54
Planned Experimental Change #7: Non-arbitrary characteristics.	55
Figure 17: Quinones & Hayes, 2014, Non-Arbitrary Trials	56
Final Study: Post-Training KU Biasing	57
Discussion	58
Experiment 4: Extending the Current Training	59
Method	59
Participants	60
Results	60
Table 6: Experiment 4 General Response Shift Counts For ICK Word Problems	61
Table 7: Experiment 4 Specific Response Shift Counts For ICK Word Problems	62
Discussion	62
Experiment 5: Simpler Relational Training	63
Method	63
Participants	65
Results	65
Discussion	65
Experiment 6: Assessment of the Function of RK Responses	66
Method	66
Table 8: 9-Item 3-Term Series word problems version 2	68
Participants	68
Results	68
Table 9: Experiment 6 Survey Accuracy by Instructional Timing	70
Discussion	72
Experiment 7: ICK Evoking Instructions	77
Method	77
Participants	77
Results	77
Discussion	78

Experiment 8: ICK Evoking Instructions Revisited	81
Method	81
Participants	81
Results	81
Discussion	82
Experiment 4-8 Overall Discussion	85
Next Steps	86
Figure 18: Simplified flow chart of the original bias experiment design	88
Extending Training	89
Less Complex Arbitrary Training	89
ICK Evoking	90
Mastery and Far Transfer	92
Response Cost Contingencies	93
Previous Correct Responses Available	94
Figure 15: Proposed Previous Answer	95
Changing to Known Word Problems	95
Non-Arbitrary Training	96
General Discussion	99
References	102

Introduction

Humans make errors even in simple relational tasks. For example, given two statements “A is smaller than B” and “B is larger than C” and the question “What is the relation between A and C?” more than half of respondents would claim definitively that one is larger or they are the same. The correct answer to this three-term relational task is that we don’t know. There is not enough information provided to derive a definitive relation beyond “unknown.” Since the publication of Relational Frame Theory (Hayes et al., 2001), this high frequency of failure on a relational task has been the subject of limited experiments (Quinones & Hayes, 2014; Vitale et al., 2008, 2012). Where previous research has highlighted specific aspects such as intervention utility and broad theoretical consistency (Vitale et al., 2008, 2012), or interactions with coherence and plausible explanations of “odd forms of thinking” (Quinones & Hayes, 2014), a broader look at the phenomena as understood by the theory illustrates that the scope may have been vastly underestimated. Specifically, such seemingly simple mistakes may be the pathway to a novel and useful understanding of bias.

The term “bias” has definitional ambiguity. Popular usage leans toward "a strong feeling in favour of or against one group of people, or one side in an argument, often not based on fair judgment"(Hornby, 2020) Examples of biased behavior in this usage include racial or ethnic prejudice, political bias, undue favoritism toward particular groups, social stereotypes, irrational preferences, habitually poor judgments in a variety of areas such as relationships or investments, and so on. To avoid ambiguity, this popular usage will be referred to as “unfair judgment.” Within behavior analytic usage, operant bias, as an extension of B. F. Skinner’s response probability metaphor (Skinner, 1965), is the relative strength of any particular behavior occurring as a product of an individual’s history of reinforcement. The term is neutral to the social utility of

the consequent outcomes or their “fairness” in comparison to other possible response options. Operant bias also demonstrates contextual dependence. For example, the probability of words chosen in a conversation change with who is participating in that conversation. Different words are more and less likely when talking to an intimate partner relative to talking to a work supervisor. The context of who we are talking to biases our speaking behavior differentially. All operant behavior is biased.

The overlap between all behavior being biased in an operant sense, and some behavior being biased in a popular sense could be an important one. A basic operant account of the conditions that influence the probability of unfair judgment behaviors could provide specific, actionable, and effective means to intervene on those behaviors. The unfair judgment popular usage itself highlights that addressing such behavior could alleviate pervasive human problems of known social and personal importance. These include behaviors categorized as racist, sexist, classist, or xenophobic to name a few, all of which result in humans responding to each other as relatively lesser beings or otherwise not deserving of equal treatment.

Because operant response bias is systemic across many species, including humans (Brembs, 2003), it is important to differentiate between response bias due to direct reinforcement history and response bias due to generalized derivation of relations as differing sources of unfair judgments. A principle of operant behavioral influence is that contact with a reinforcing outcome will increase the future likelihood of the response that operated on the environment to precipitate said outcome. If an individual engages in a behavior resulting in a reinforcing consequence, that behavior is more likely (has an increased bias) to occur in the future under same or similar conditions (Skinner, 1938). This scenario, where a specific behavior produces a reinforcing consequence, is called direct reinforcement.

In the case of direct reinforcement biasing, it is often inappropriate to assign the popular meaning of bias as unfair judgment to the resultant behavior. This is because, at the time that the bias was strengthened, that response was effective for that individual in that particular moment and it may not have been based on unfair judgments at all. The bias strengthening effect of the consequence of that behavior is evidence of that response's workability. To identify what conditions are most likely to result in an intersection with unfair judgment requires understanding instances of verbal judgment as being based on a special form of operant behavior.

Relational Frame Theory

Distinct from direct reinforcement, much of human behavior is explained through an indirect process. That process is deriving relations among experiences such that future behavior is biased by those relations. The conditions required for behavior to come under the indirect influence of such inferences are described by Relational Frame Theory (RFT) (Hayes et al., 2001). RFT is an operant behavioral account of language and cognition that pivots on three key arguments. First, "language" or "cognition" are behavioral repertoires of responding to relationships amongst stimuli, such as one's worth relative to another (as compared to responding to stimuli directly). Second, many relations amongst stimuli are rooted only in arbitrarily applied social convention, such as value. This is in contrast to non-arbitrary (aka formal) relational responding along physical characteristics of stimuli (e.g., height, weight, shade, surface area, etc) which is common amongst many organisms. For example, turtles (*Terrapene carolina*) will select the darkest or lightest shade of colored paddles for food (Leighty et al., 2013) and New Caledonian Crows (*Corvus moneduloides*) discriminate relative mass, volume, and density to access food (Jelbert et al., 2014). And finally, the conditions for reinforcement of arbitrarily applicable relational responding (AARR) also make reinforcement for derivable relational

responses so frequent that derivation-of-relations per se becomes a widely generalized operant behavior (Healy et al., 2000). That is, rapid and extensive derivation becomes nearly reflexive. For example, someone directly learning that a dime is worth more than a nickel and less than a quarter results in the rapid derivation of more than and less than relations of worth extending across the rest of their known currencies (Berens & Hayes, 2007). Without additional direct reinforcement, they will respond to pennies as less than dimes, dimes less than dollars, and so on. In the case of the simple three-term relational task, the derivation of a relation between A and C occurs almost instantaneously and without notice. This is important because in this special case, we are more likely to derive an inaccurate relation that is only discovered when the unworkable response actually occurs. While that may be abstract in the three-term task, in real life this could be actually committing an unfair judgment, such as a racist, sexist, or xenophobic act, without ever being previously directly reinforced for that unfair judgment behavior.

This learned generalized operant behavior of deriving, technically termed the arbitrarily applicable derived relational response (DRR for short; when referring to non-arbitrary relational responding no acronym will be used), not only accounts for indirect learning but, by extension, the massive generativity of human language (Hayes et al., 2021). It helps explain how humans can learn a small set of rules and symbols and then produce nearly infinite novel meaningful recombinations without having to learn each directly.

Derivation demonstrations are also the evidence required to identify a behavior as being a product of a relational repertoire. To support a claim that an individual is responding to relations amongst stimuli, versus some non-relational characteristic, an expected derived response must regularly occur after one or more relations are trained. When only one relation is trained directly, i.e., a dime is worth more than a nickel, the derivable relationship is the inverse: the nickel

evoking a response reflecting that it is worth less than the dime. This first order of derivation is called Mutual Entailment because one brings about the other in the relational pair: in this case “worth more” entails “worth less.” When more than one relationship is trained, such as in the three-term problem, the combination of two relations involving three stimuli brings with it (i.e., entails) a third relationship. If a quarter is trained as worth more than the dime mentioned above, the relationship between the nickel and quarter can be derived. This second order of derivation is called Combinatorial Entailment, and that relation, derived from combination (whichever direction it is derived [i.e., that is either a nickel is less than a quarter or a quarter is more than a nickel]), also mutually entails the inverse. An experiment training the nickel-dime and dime-quarter relations would only support claims of relational responding and derivation if the participant demonstrated reliable mutual and combinatorial relational responses. Otherwise, it may be reasonably argued that participant behavior was biased by direct reinforcement of non-relational characteristics. This is all to say that an empirical investigation of the intersection of unfair judgment and indirect operant biases must result in specific patterns of indirect behavioral influence.

In the example of coins above, three stimuli connected by two trained relations resulted in four derived relations for a total of six expected responses from just two responses that were directly trained. This six-for-two outcome is the generativity phenomenon of relating and deriving. When this generativity is released from the limits of the physical characteristics (aka forms) of stimuli, such as when the relationship is an arbitrary construct like “worth,” the number of likely derived relations amongst many stimuli becomes exponential. Consider the example of “worth” itself. Once a few exemplars have been established (directly trained), such as the relative worth of a dollar and the relative worth of one’s time, responding to everything

around us in terms of relative worth becomes pervasive. We treat others as worth more or less without considering the dehumanizing effect. Not only do we respond to the relative worth of two physical objects, we also evaluate the relations amongst things as more or less worthy.

Understanding how mass and gravity are related is relatively more worthwhile to most than the relative value of two pieces of sand. This relating of relations further compounds the pervasive, exponential, even fractal or hyperdimensional (Barnes-Holmes et al., 2017) nature of relational behavioral repertoires.

When it comes to unfair judgment, one typical characteristic of such behaviors is that they are inaccurate. The response does not fit the situation. Of the possible behaviors in that moment, the most probable one for that individual has a less than optimal workability, such as when the most qualified candidate for a job is passed over due to reasons unrelated to the effectiveness of the individual in the role (e.g., sexual discrimination [Webb, 1979]). This is the point where indirect operant bias via DRR intersects. Derivation does not have an inbuilt accuracy check. As demonstrated with the three-term problems, derived relations can miss the mark even at low levels of derivation (Vitale et al., 2008, 2012). Thus, as an individual's relational repertoire becomes extensive, millions of combinatorially derived relational responses will bias future behavior. A non-trivial subset of those derivations will be inaccurate and result in the occurrence of unfair judgment with varying degrees of workability. One specific alternative response to inaccurate and unfair judgment is being able to accurately state, and more fairly respond to, when we don't know something.

Knowing We Don't Know

To acknowledge when we don't know something is a powerful action. Consider a time when you noticed that a loved one doesn't know something and how, if they only realized they didn't know, they could take more effective action. Maybe they were struggling with political rhetoric or interpersonal conflict. Either way, a recognition that they are making assumptions and missing key details can highlight nuanced alternatives that are more likely to result in sustainable outcomes. Up until such a revelation of unknowing is made, presenting evidence of these alternative options is unlikely to have the same benefits (Stiernborg et al., 1996). In the worst case, such efforts to help and guide can result in further entrenched biases and assumptions (Howard et al., 2020; Wegner, 1994; Wegner et al., 1987).

In the vastly interconnected environment of online interactions today, navigating hyperpartisan discourse further amplifies the benefits of recognizing when we need more information (Sangin et al., 2011). The ability to quickly and accurately recognize a gap in knowledge can not only improve otherwise discriminatory scenarios (Ben et al., 2017), but also provides opportunities for building interpersonal connection and undermining social polarization (White et al., 2018). In a world of division, the ability to say "I Don't Know" may be more important than ever before. In the practice of science, transforming a sense of unknowing into a signal of discovery shifts the daily experience of the scientist from aversive to appetitive (Schwartz, 2008). From dreading and avoiding each day, to engaging and pursuing each moment of not knowing.

This proposal's goal is to illustrate how training an accurate and timely "I Don't Know" functioning response requires unique methodologies, provides a pathway to accessing

opportunities, and could result in social and self harm reduction. This manuscript will begin by establishing the theoretical foundations for such behavior, review previous empirical studies about closely related behavioral phenomena, present three pilot studies and their implications, and propose a number of next step experimental procedures to further our understanding of the conditions under which “I Don’t Know” responding occurs both accurately and inaccurately.

Theoretical Review

Three foundational claims of this proposal are that “I Don’t Know” functioning responses can be considered (1) operant behavior in general; (2) arbitrarily applicable derived relational operant behavior in particular; and (3) presenting novel utility for scientists and practitioners as a described phenomenon. This theoretical review will walk through each of those claims to establish the proposal scope and requirements for sufficient empirical evidentiary support.

As discussed in the Relational Frame Theory section above, to claim incidences of “I don’t know” as DRR (and thus operant behavior in general) requires demonstrating that the behavior was the result of a relational derivation and not reliant solely on formal characteristics of the procedure. Derived relational responding is characterized by the phenomenon of Transformation of Stimulus Function (ToF) (Dymond & Rehfeldt, 2000; Hayes et al., 2001, pp. 31–33). Definitions of ToF are somewhat ambiguous. The canonical RFT book opens the discussion of ToF with, “When a given stimulus in a relational network has certain psychological functions, the functions of other events in that network may be modified in accordance with the underlying derived relation” (Hayes et al., 2001, p. 31) and later lists the minimum “qualities” of demonstrating relational responding as “mutual entailment, combinatorial entailment, AND [emphasis added] transformation of stimulus function” (2001, p. 105). Though it is never stated

explicitly, the use of “and” in the second passage implies ToF as a third, possibly independent, phenomenon from the two entailment processes requiring additional behavioral demonstration. Work informing the 2001 RFT text suggests that ToF is not independent of entailment processes but the evidence used to demonstrate such processes (M. J. Dougher & Markham, 1996; M. Dougher & Markham, 1994; Hayes, 1991). In other words, ToF is the transformation/modification/change in behavioral function as a result of derivation of relations between psychological events due to processes of either mutual or combinatorial entailment. For this proposal, I suggest the following operational definition of ToF. ToF is the difference of responses evoked in the presence of a stimulating event before and after relational derivation has putatively occurred. Therefore, a response to the same stimulation reflecting relational function(s) of related stimuli not occurring prior to derivation is considered a demonstration of ToF. For example, a child regularly choosing a nickel over a dime may demonstrate ToF after learning about relations of monetary value if they reliably choose the dime after the relational learning. In this case, it would be argued that the controlling stimulus function of the relationship between the dime and nickel has transformed from physical size to worth.

ToF may be due to mutual or combinatorial entailment. In the specific case of “I don’t know,” ToF via combinatorial entailment may be the minimum demonstration of the phenomenon where inaccurate and unfair judgment is likely to be ameliorated. While “I don’t know” can be mutually entailed, the resulting response would always be accurate. For example, an individual trained to respond to A as Greater Than B would demonstrate mutual entailment by responding to B as Less Than A. Other responses inconsistent with training (e.g., B same as A) would not be sufficient to claim mutually entailed DRR occurred. In other words, an adequate demonstration of mutual entailment is mutually exclusive to any response that is inaccurate. In

order for “I don’t know” responses to be qualified as DRR and fall within the scope of unfair judgment, this series of experiments must demonstrate at least combinatorial entailment levels of transformation of stimulus function.

Arbitrary applicability means that the contingency of reinforcement for relational responding has not been determined solely by formal characteristics of the stimulating events, and thus can be modified based on cues established by social whim or convention. For example, the crow and turtle examples given above required specific non-arbitrary characteristics of stimuli involved for an accurate relational response. The available choices had relatively different volume, mass, or color. But in the example of the coin values, no form of the coins (e.g., shape, mass, or diameter) specifically determined that functional relationship. In this case, only the arbitrary coincidence of social reinforcement establishes that these coins have relative value. This arbitrary connection is highlighted when the same coins are equated with vastly differing symbolic and auditory stimuli across social groups and geographies despite relative physical similarity.

To qualify incidences of “I don’t know” as arbitrarily applicable, the experiment must demonstrate that the response occurs reliably independent of characteristics of the evoking stimuli involved, other than the derived relation. Controlling all chances of unintentional formal stimulus influence entirely is unlikely. But systematic alteration of non-arbitrary stimulus characteristics not intended to influence behavior can reduce the probability of behavior coming under the influence of non-arbitrary forms such as position, shape, or orientation of the experimental stimuli. Multiple exemplar training (MET) protocols are commonly used for relational training for exactly this reason (Luciano et al., 2021). MET protocols expose the participant to contingencies of reinforcement where the non-arbitrary characteristics of the

stimulus and response options are varied but the relational characteristics are held consistent. For “I don’t know” training to demonstrate arbitrary applicability, the stimuli being related and the putatively reinforcing response options must involve enough variability in physical form and location while being consistent in relational aspects such that the target relationships of the training come to exert dominant stimulus control over the formal characteristics.

Theoretical implications of the novelty of the “I don’t know” functioning relational response requires reviewing four misconceptions of DRR more broadly. Specifically, the behavior is not strict logic, it is not responding to stimuli per se, it is not constrained to specific forms of behavior, and it does not occur independently of other non-relational operant behaviors. These are particularly important because each can, and has, been a source of significant confusion in the development of this research.

Despite the logical reasoning described in much of the DRR literature, it is a critical mistake to assume that human behavior, particularly relational repertoires, are adherent to strict logical rules. This would be confusing outcomes with process. As with any operant behavior, DRR is a reflection of the individual’s history. If the individual was reinforced for responding logically, the behaviors will result in logically coherent outcomes. But while logical coherence tends to align with natural and social contingencies, the arbitrary applicability of relational responding disconnects what behaviors are reinforced from adherence to strict logic. An individual with a history of reinforcement for contrary, random, or illogical relational responding is predisposed to continue to respond in such a manner (Luciano et al., 2021). In the case of an individual regularly reinforced for taking a specific position on matters they don’t know about, even in conditions where “I Don’t Know” is clearly the most accurate answer, the logical prediction of an “I Don’t Know” response would fail to account for the individual’s history of

reinforcement. By extension, an individual from a context involving harmful bias reinforcement (e.g., participating in extremist groups, growing up in racist/sexist/xenophobic cultures) is likely to derive equivalence between harmed groups and detrimental descriptions without, or in the face of, supporting experience (Belisle et al., 2022; Frederick, 2005; Howard et al., 2020). This explanation of an individual's behavior reflecting their history over logic also supports anecdotal evidence of how evidence-only arguments have limited effects on changing human behavior (Stiernborg et al., 1996). If relational responding was a strictly logical behavior, the world would be a different place.

Relational responding is not responding to stimuli per se. This associative trope of relational responding is a common mistake suggesting that some characteristic inherent in a single stimulus is exerting influence on the individual responding to the experience. If that were the case, relational responding would not be differentiable from non-relational operant behaviors. In that alternative, one would be expected to respond relationally to a single stimulus without ever having a history with other stimuli being related; and with the response having no uniquely identifiable characteristics. This point is easier to make initially with non-arbitrary relational responding. The crow experiment described above is an illustration. In that example, the crow is presented with an array of objects that may be placed in a water filled tube containing a floating treat that is initially out of reach. Some of the objects displace more water than others or vary in density but not size such that they either sink or float. In each trial, the crow will reliably select the object that most quickly raises the water in the tube and brings the treat within reach. This is the case even when object properties are randomized within the array and, in the case of displacement, the one that worked best in a previous trial is the sub-optimal choice in the current trial. If the crow was responding to a specific stimulation per se, they would only use those

objects that previously worked, might randomly choose objects when previously successful objects are unavailable, and not participate in selection at all when object properties exceed specific levels. The fact that none of these latter possibilities occurred supports that the choice behavior was under the influence of relative conditions of stimulation more so than any one stimulus. Similarly in the case of establishing an “I Don’t Know” like response as relational, it requires a reasonable demonstration of such relative stimulus influence. If this is not accomplished, or it becomes clear that a singular stimulus is evoking the response, the claim of relational responding is unsupported.

Relational responding is not constrained to specific forms of behavior. The distinction of the form versus the function of a behavior is essential to operant behavior. The basic premise is that while individual occurrences of behavior may look different, they may still result in the same function, and it is that function, not form, that is influenced by operant principles (Hayes et al., 2001, pp. 40–43). In the example of the coin network above, the function is comparative relational responding (i.e., less than and greater than). That is, in the presence of any of the involved stimuli, the evoked response is likely to serve a comparatively lesser or greater function. Specific to the “I Don’t Know” functioning responses, this form versus function distinction allows for two outcomes. First, behaviors other than the explicit statement of “I Don’t Know” can still result in outcomes distinct from biased or assumptive responses. Second, expressions of “I Don’t Know” do not automatically function in the manner proposed by this discussion. Alternative functions of “I Don’t Know” may be escape or avoidance. An individual may be negatively reinforced by escaping uncertainty or avoiding further inquiry by saying “I Don’t Know” regardless of the statement’s relational functions or accuracy. In this way, any

experimental exploration of such a behavior must provide evidence clarifying the function of the behavior recorded or risk conflating form and function.

Relational responding does not occur independently of non-relational operant behaviors. This is an explicit clarification that operant behavior of all kinds interact, at time in a manner that can make it challenging to parse causal relationships. The interaction of traditionally operant responses with relational training highlights how such interactions are likely. For example, Dougher and colleagues (1994) demonstrated that galvanic skin response, a behavior typically considered non-relational but subject to operant principles, can be impacted by relational training. In their experiment two 4-item arbitrary equivalence (i.e., all items function the same or interchangeably) networks were trained. Each item of that network consisted of a simple yet novel line drawing. One item of one of the equivalence networks was paired with one instance of a mild shock. Further presentation of any objects from the shock equivalence network evoked differentiated galvanic skin responses as if the individual was experiencing the same shock despite only ever experiencing the single instance of shock when that first item was presented. In this example, a learned relational response influenced a traditionally non-relational behavior. In the same way, an individual can come to relate physiological responses as equivalent to appetitive or aversive relational responses and engage in differential relational responding according to non-relational stimuli. This point highlights that procedures purporting to demonstrate relational repertoires must provide particular behavioral evidence, as outlined in the RFT section above, and cannot rely on theoretical reasoning alone.

“I Don’t Know” responses can be characterized as DRR if there is evidence of mutual and combinatorial entailment, if they occur only under specific relational conditions, and if history of reinforcement is correlated with differential response likelihoods. Preliminarily, these

conditions appear to have been satisfied by two prior empirical studies which will be reviewed in depth in following sections. Specifically, Vitali et al (2008) demonstrate that, at the level of the group, the occurrence of an “I Don’t Know” functioning response was differentiated to specific relational conditions and was subject to influence by reinforcement history. Quinones and Hayes (2014) built on these findings when they demonstrated that such responses exhibit transformation of stimulus function supporting mutual and combinatorial entailment as well as could be manipulated (i.e., biased) within individuals via recent reinforcement history to reliably respond inaccurately in particular ways to “I Don’t Know” evoking conditions. I will review these studies later in the proposal.

In addition to the previous empirical evidence, the process of designing and validating the upcoming proposed investigation has uncovered two characteristics of the “I Don’t Know” functioning responses that warrant further elaboration and may be key to their novel utility. The first is the high theoretical likelihood of deriving such relational responses once basic DRR skills have been established. The second is the relative rarity of natural contingencies shaping accurate “I Don’t Know” functioning responses. Arguably, all language capable individuals regularly derive and respond to such relations far more frequently than any other previously relational class currently described in RFT, and very few individuals are ever reinforced for accurately responding to such “I Don’t Know” relationships.

The likelihood of derivation implication is theoretical but grounded in how common the conditions are that result in their derivation. To make this clear, we need to revisit when derivation occurs. In mutual entailment, derivation occurs when an individual is reinforced for responding as if two stimuli are related. The entailed relationship is derived almost automatically as long as the individual has a basic DRR fluency. In mutual entailment, the relational function

of the derived relation may be described as an inverse or equivalent relationship depending on the original relational function. Regardless of the function, the entailed response exhibits that two stimuli are related and how. In combinatorial entailment however, deriving the existence of a relationship does not guarantee that there is enough information to derive what type of relational function is accurate between the two stimuli. In a rough example, to state that John is taller than Joe is to imply that Joe is in a relationship of height with John and the accurate relative height of Joe is “shorter than John.” This is the mutually entailed derivable relation. But if John is also taller than Jane, while there is enough information to accurately derive the Jane:John mutually entailed relationship, there is only enough information to derive that Joe and Jane are related by height (that is, the combinatorially derived relationship exists and could be known if additional stimulus information were provided) but not enough information to resolve if Joe is taller, shorter, or the same height as Jane. In this case, the most accurate relational response would function as “I Don’t Know.” Whenever an individual has a history of three stimuli involved in pairs of relationships (e.g., A:B & A:C), learned generalized derivation will rapidly combine those relationships for a combinatorially entailed derived relational response. This separation of general combinatorial derivation (a relationship exists) and specific combinatorial derivation (what exact relational function is accurate) creates a unique hierarchy. In order to derive an accurate specific relationship, one must have the necessary elements to evoke derivation in the general sense. The opposite is less restricted. In order to derive the existence of a relationship combinatorially, one does not need the necessary elements to evoke derivation in the specific sense. The hierarchy implied here is that the conditions that can evoke a specific combinatorial entailment derivation (e.g., comparative, temporal, hierarchical, deictic, etc) are a subset of the conditions that can evoke a general combinatorial derivation (e.g., “I Don’t Know”). Because the

conditions for an accurate “I Don’t Know” response require fewer, or more ambiguous, experiences to occur than, or are hierarchically inclusive of, all other specific relations, two further conclusions are suggested. First, that for any given situation likely to result in combinatorial derivation, “I Don’t Know” is as or more likely to be derived. Second, if the first claim is true, then relative frequency of “I Don’t Know” combinatorial derivations to all other specific relations is on the order of multiples, or even magnitudes, more frequent. In other words, in any given situation involving DRR, we are more likely to be influenced by “I Don’t Know” like general combinatorial derivations than any other, more specific, relational function.

To claim that RFT research may have not focused much attention on the most common derived relation is a serious statement that requires additional investigation. As a first step, computer models of derivation of comparative and equivalence relational derivations were created (Smith & Hayes, 2022) to simulate various relational networks. The simulation program translated relational derivation processes into a script that took in a set of trained relations and output all mutually and combinatorially derived relationships implied by the trained set. Specific details of the translation and coding are provided in the cited paper. Germaine to this proposal, in test sets of trained relations, “I Don’t Know” like derivations increased as a proportion of output as the number of stimuli in training sets increased and/or the relational conditions included more ambiguous sets. In one example, a training set of eight relations involving ten stimuli resulted in 70 derived relations (8 mutually entailed, 62 combinatorially entailed). Of the combinatorially derived relations, “I Don’t Know” like relations occurred 2.1 times more often than specific relations [Figure 1].

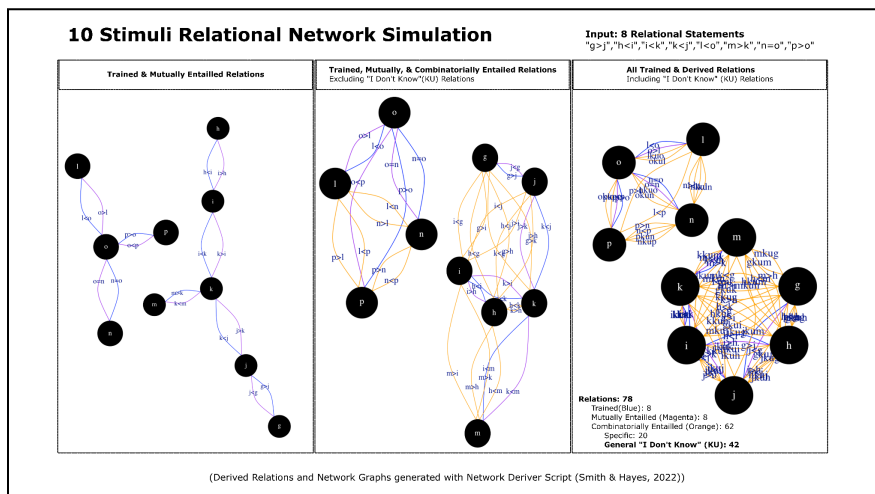


Figure 1:

10 Stimuli Relational Network Simulation. The same network is illustrated with increasing amounts of derivation (Left to Right panels). "I Don't Know" like relations account for 42 of the 62 combinatorially entailed derived relations. Derived relations and network graph visualization generated using the open source R script described in Smith & Hayes (2022).

Simulating the logic of derivation provides some support for the claim that "I Don't Know" like derivations are frequently likely. More so, it highlights that empirical work is needed. To take this simulation data as a substitute for human behavioral data would commit the error of treating logic as a strict process in the human behavior causal chain instead of the current interpretation that logic is a limited byproduct characteristic of DRR behaviors (Luciano et al., 2021; Smith & Hayes, 2022).

The rarity of natural reinforcement of accurate "I Don't Know" relational responding is in contrast with, but not a contradiction of, the basic claim of RFT that relational behaviors are likely to be reinforced relatively regularly just by chance in the natural environment (Hayes et al., 2001, pp. 40–41). That claim is anchored in how common it is as an organism to result in beneficial consequences as a result of responding to the relative characteristics of stimuli. In the crow example above, discriminating tools along physical properties can provide access to lower

competition resources as well as facilitate efficient interaction amongst familial and non-familial social contingencies (Hayes & Sanford, 2014). Contingencies of survival reasonably select for formal relational responding to the largest, smallest, ripest, lightest, or slowest to name a few. In this way, most organisms are likely to have a rudimentary relational responding repertoire just by chance. But formal relational contingencies are nearly devoid of incidences where an “I Don’t Know” functioning response will produce consequences that are materially more beneficial to the organism. In the example of John, Joe, and Jane above, the example works to illustrate the accuracy of the “I Don’t Know” response only as long as the three individuals are otherwise ephemeral concepts in writing. As soon as John, Joe, and Jane are visually available in their physical forms, the relative height becomes readily apparent up to a negligible degree of difference. At that point, an “I Don’t Know” functioning response is likely to be less effective than the alternative more explicit same, taller, or shorter variants. And in all but the rarest cases, a measurable but visually ambiguous height difference has little possibility of requiring discrimination such that responding as if Joe and Jane are effectively the same height is the most likely response to be reinforced. Contrast this with contingencies where the related stimuli are intermediated by written descriptions or formal cue limited stimuli such as internet delivered content. In this context, contingencies where accurate “I Don’t Know” responses may be beneficial are much more likely but reinforcement contingencies may be constrained to the occurrence of searching behavior over the content discovered (Fisher et al., 2015).

Developmental evidence suggesting that accurate “I Don’t Know” responses may be trained out in favor of less accurate responses appear in research with children of varying age probing for awareness of what they don’t know (Mills & Keil, 2004). There, kindergarten age children were significantly more likely to say “I Don’t Know” when asked to explain how a

device (e.g., a stapler or toaster) worked, more than a second or fourth grade child. While the function of this “I Don’t Know” response was not the focus of that experiment, the observed frequency decline at an early age may reflect contingencies selecting against accurate “I Don’t Know” responding. While the arbitrary applicability aspect of DRR and the increasingly common intermediation of access to stimuli by written or spoken descriptions (i.e., via the internet) make “I Don’t Know” responding increasingly useful and relevant, the probability of any one individual experiencing consistent reinforcement for accurate “I Don’t Know” relational responding just by natural chance is low enough to assume that most individuals may be assumed to reliably respond to “I Don’t Know” conditions inaccurately or with chance at best accuracy.

The assumption of chance at best “I Don’t Know” responding may actually severely underestimate the challenge of improving this behavior. A reasonable assumption implied by the kindergartner anecdote above is that by the beginning of compulsory education in the United States, most individuals are being actively taught to not say “I Don’t Know” regardless of accuracy. This proposal is attempting to undermine a behavioral repertoire with a lifetime of history for responding other than the target outcome. If systematic variations of the proposed intervention fail to demonstrate measurable changes in the frequency and accuracy of “I Don’t Know” responding, it is possible that the interventions tested never reached a minimum effective amount given the extensive history of the presenting behavior.

Prior to a detailed review of the empirical literature on this phenomena, the issue of nomenclature of these “I Don’t Know” functioning relational classification should be addressed. “I Don’t Know” is a common phrase with links to a wide range of issues beyond the focus on the present study. Previous literature used monikers such as “unspecified relations” (Vitale et al.,

2008, 2012) and “ambiguous relations” (Quinones & Hayes, 2014) are not sufficiently precise. This proposal adopts the name “Known-Unknown,” abbreviated as KU, to refer to derived relational responding under the stimulus control of a history minimally sufficient to evoke a combinatorially derived relation but insufficient to derive an accurate relational function amongst the psychological events related. The “Known-Unknown” name reflects that behavior under the influence of a KU relation will exhibit a relational function. An untrained individual would behave as if they had sufficient history to combinatorially derive the exact function (i.e., as if the function was known). This response may or may not be accurate by chance. An individual trained to discriminate KU relations may express “I Don’t Know” functioning responses accurately. This response would reflect that they know that they don’t know.

Literature and Pilot Experiments Review

The empirical work most relevant to this proposal includes two previous empirical studies and three recent pilot experiments of our own. Vitale et al (2008) explored KU responding in a sorting task protocol and generated group level baseline measures of KU and non-KU response accuracy and evidence that we may be able to intervene on KU inaccuracy. In a no-right-answer protocol, Quinones and Hayes (2014) targeted the conditions under which such a KU intervention may result in systematically biased inaccurate responses. Our first pilot experiment sought to systematically replicate those previously mentioned group level baseline measures while assessing a potentially more socially valid survey tool. Our second pilot experiment attempted to systematically replicate the MET approach for improving KU used by both preceding studies. And our final pilot integrated probes for evidence of relational responding into the previously piloted MET protocol.

Sorting Three Coins Relationally

The first series of experiments to target KU relational responding (Vitale et al., 2008, 2012) used a coin sorting task where participants were given three same size “coins” (colored circles on a computer screen) and two relational descriptions (aka a 3-Term Series Problem [Vitale et al., 2008, pp. 366–367]) that gave the relative value of each to the others. Using the provided information, participants were tasked to sort the coins into “jars” (four locations on screen) three jars represented most to least value, and the fourth was “I Cannot Know” (ICK). Participants were able to drag and drop one coin to each of the value jars and up to all three coins on the ICK jar. Feedback was not provided during their baseline evaluation experiment but was later provided as the appearance of “Correct” or “Wrong” after each trial in later training experiments that included feedback. In non-KU trials [Figure 2], the relative value statements provided enough information to sort all three coins to all three value jars and none to the ICK jar. In KU trials [Figure 3], one coin could be determined as highest or lowest value and sorted to the appropriate value jar and the other two coins could not be sorted relative to each other and so the most accurate response would be to sort both to the ICK jar.

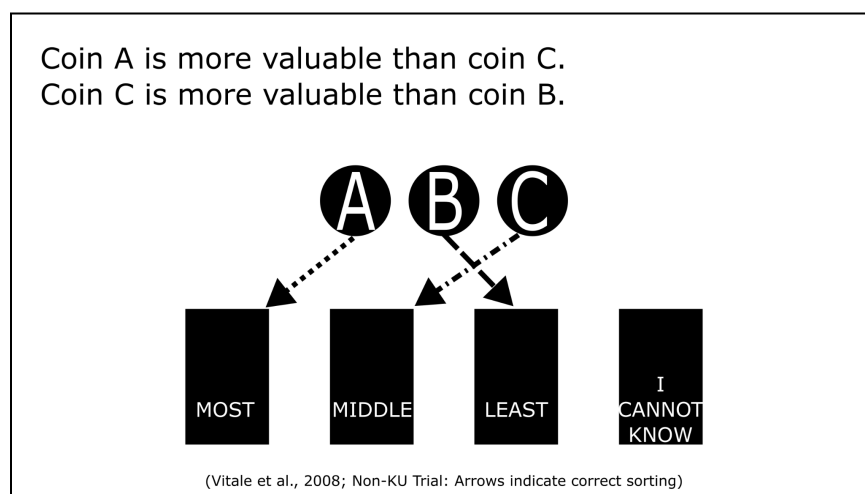
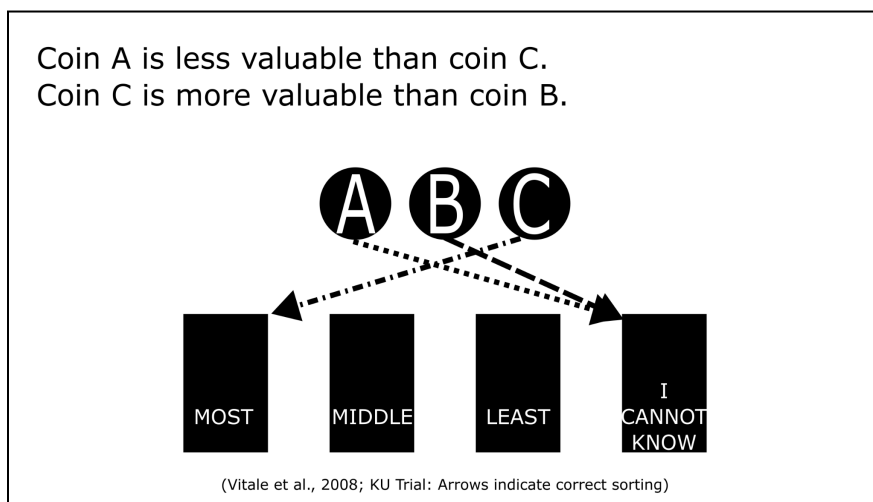


Figure 2:

Vitale et al., 2008 non-KU Trial: Arrows indicate correct coin sorting.

**Figure 3:**

Vitale et al., 2008 KU Trial: Arrows indicate correct coin sorting.

Vitale et al (2008, p. 369) identified 48 permutations [Table 1] of describing the coin relative values such that 24 derive to non-KU relations and the other 24 derive to KU relations.

Specified-Same Relations		Unspecified-Same Relations	
MORE-MORE	LESS-LESS	MORE-MORE	LESS-LESS
A > B; B > C	A < B; B < C	A > B; C > B	A < B; C < B
B > C; A > B	B < C; A < B	B > A; B > C	B < A; B < C
B > A; C > B	B < A; C < B	B > C; B > A	B < C; B < A
C > B; B > A	C < B; B < A	C > B; A > B	C < B; A < B
Specified-Mixed Relations		Unspecified-Mixed Relations	
MORE-LESS	LESS-MORE	MORE-LESS	LESS-MORE
A > B; C < B	A < B; C > B	A > B; B < C	A < B; B > C
B > C; B < A	B < C; B > A	B > C; A < B	B < C; A > B
B > A; B < C	B < A; B > C	B > A; C < B	B < A; C > B
C > B; A < B	C < B; A > B	C > B; B < A	C < B; B > A
Specified-Same Transitive Relations		Unspecified-Mixed Transitive Relations	
MORE-MORE	LESS-LESS	MORE-LESS	LESS-MORE
A > B; C > A	A < B; C < A	A > B; C < A	A < B; C > A
A > C; B > A	C < A; B < A	A > C; B < A	A < C; B > A
C > B; A > C	C < B; A < C	C > B; A < C	C < B; A > C
C > A; B > C	C < A; B < C	C > A; B < C	C < A; B > C

Table 1:

Vitale et al., 2008 48 Permutations of three coins (A, B, & C) in comparative relations producing non-KU (Specified) and KU (Unspecified) combinatorial derivations.

In each version of their experiment, 10 experimentally naive college students were exposed to all of these permutations twice per phase over three phases and their accuracy score from the second attempts of each phase were reported as group averages. Experiment 1 was conducted as an assessment baseline without feedback and results demonstrated that non-KU trial responding was broadly accurate (M=98.8%) while KU trial responding was poor with the most challenging permutations peaking around 65%. Experiment 2 largely repeated the three phase two attempt procedure but during the second phase, each trial attempt was followed by feedback about the participant's accuracy. This training intervention effectively operated as a MET procedure. Phase 1 baseline levels of accuracy mirrored those measured in Experiment 1, and accuracy across phase 2 and 3 almost entirely eliminated the KU/Non-KU accuracy differential previously measured. Mean accuracy of KU trials in phase three reached 96%. Experiments 3-5, as well as their systematic replication series of experiments (Vitale et al., 2012), included a change to using coins of varying size depending on the relational statements provided [Figure 4]. While this adaptation is well justified based on prior relational training literature, in this case it may have unintentionally contaminated their results. This, and the broader implication, only became clear after our third pilot experiment so it will be revisited in detail at that point in this proposal.

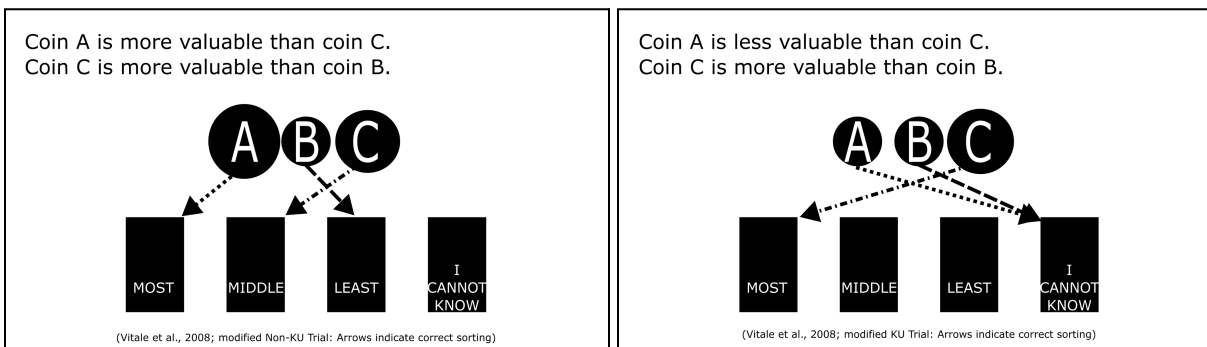


Figure 4:

Vitale et al., 2008 modified Trials: Relative coin size complements relative values. Arrows indicate correct coin sorting.

The results of these experiments highlighted four findings. First, even when participants know “I Cannot Know” is an available response option, their ability to select that option accurately is low. Second, an MET with accuracy feedback can significantly improve KU response accuracy. Third, repeated exposure to KU trials alone does not result in robust improvements in response accuracy. And fourth, the empirical investigation of KU responses is easily unintentionally contaminated.

Forced Wrong Answer for KUs

Investigating the conditions under which cognitive errors occur, Quinones & Hayes (2014) leveraged the poor baseline accuracy of KU trials to explore how recent reinforcement history may contribute to consistencies within those inaccuracies. Participants in this experiment completed a sequence of MET relational trainings [Figure 5] that included a phase where they experienced either training requiring equivalence (SAME) responding or comparative (DIFFERENT, GREATER THAN, and LESS THAN) responding.

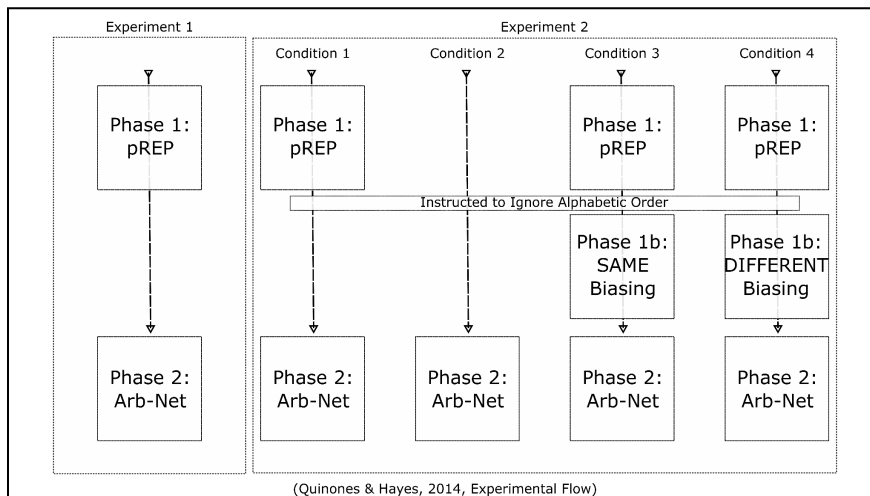


Figure 5:

Quinones & Hayes, 2014, Experimental flow chart. Phase 1 pREP trained the use of arbitrary relational response options. Phase 2 trained and tested two arbitrary networks, one non-KU and one KU. Experiment 2.3 & 2.4 included biasing training in Phase 1b.

After this specific training, participants were then trained on pairs of relations that would derive to a KU network using new to them stimuli and tested on the combinatorially entailed KU responses. In this experiment though, participants were only given the response options of SAME and DIFFERENT, or GREATER THAN and LESS THAN such that they were forced to choose the wrong answer on KU trials [Figure 6].

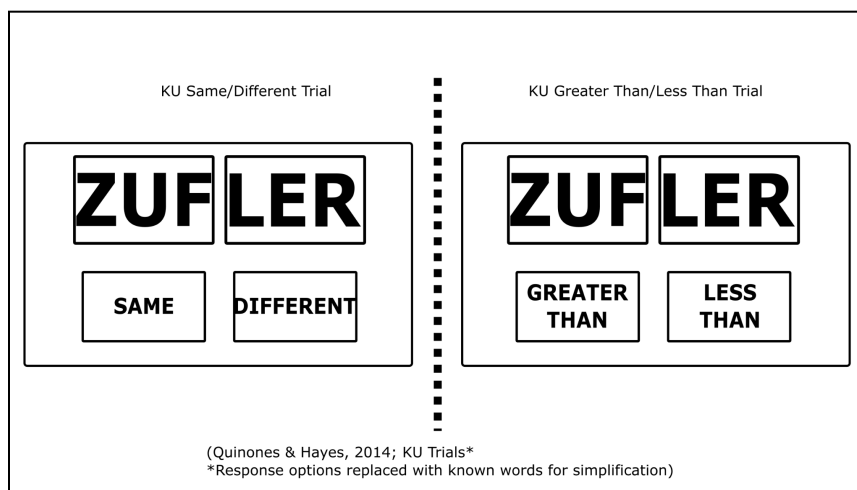


Figure 6:

Quinones & Hayes, 2014, KU Test Trial. Participants could choose between SAME and DIFFERENT (Left) or GREATER THAN and LESS THAN (Right) only. Participants were never provided an ICK option. Response options replaced with known word for simplification. Participants saw arbitrary quadgrams in place of SAME, DIFFERENT, GREATER THAN, and LESS THAN. See “Non-arbitrary Characteristics section of Proposal below for additional information.”

Participant KU response patterns aligned with the manipulated training history such that participants experiencing a history of equivalence responding on unrelated stimuli predominantly chose SAME for KU relations and those who experienced comparative training selected GREATER THAN or LESS THAN for KU relations [Table 2].

Participant	Mutual Entailment		Combinatorial Entailment			No. of probe sets	
	B1 > A1 C1 < A1	B2 < A2 A2 > C2	B1 > C1 C1 < B1	B1 D C1 C1 D B1	B2-C2 as S or D		B2-C2 as > or <
13	100	100	90	95	100 S	IC	5
14	100	100	100	100	100 S	All >	4
15	95	90	100	95	90 S	IC	5
16	85	95	95	95	80 S	IC	5

Note. Greater than = >, less than = <, D = different, S = same, No. = number, IC = inconsistent responding. The number listed in each cell denotes the percentage of trials in which the participant selected the response (averaged from the total number of probes) as indicated in the corresponding column heading.

Participant	Mutual Entailment		Combinatorial Entailment			No. of probe sets	
	B1 > A1 C1 < A1	B2 < A2 A2 > C2	B1 > C1 C1 < B1	B1 D C1 C1 D B1	B2-C2 as S or D		B2-C2 as > or <
17	90	95	95	100	100 D	Inconsistent	5
18	88	92	96	92	92 D	B2 > C2; C2 < B2 (92)	6
19	85	95	100	95	100 D	B2 > C2; C2 < B2 (95)	5
20	100	100	90	90	90 D	B2 < C2; C2 > B2 (85)	5

Note. Greater than = >, less than = <, D = different, S = same, No. = number. Number listed in each cell denotes the percentage of trials in which the participant selected the response indicated in the corresponding column heading. Any relation (e.g., C2 > B2; B2 < C2) indicated will be paired with a number which denotes the percentage of trials that participant selected that relation.

Table 2:

Quinones & Hayes, 2014, Tables 11 and 12 highlighting biasing outcome for each participant during Phase 2 arbitrary network testing. Highlight boxes around KU relational response data.

In addition to this clear demonstration of biasing KU responses, the results highlighted that KU response could be reliably biased to either equivalence or comparative responding, but those who responded to KUs as comparative relations did so idiosyncratically with some relating the B2

stimuli as greater than the C2, others vice versa, and a few responding inconsistently (e.g., both $B2 > C2$ & $B2 < C2$ in repeated trials).

Data provided by these experiments demonstrated three things but also contained a limitation specific to this proposal. First, KU inaccurate responding can be manipulated by an individual's training history that is only related by temporal proximity. Second, the predominant relational function of recent training was predictive of the KU inaccurate response bias. And third, while response bias could be predicted to the level of broad functional class, specific response patterns were idiosyncratic and less predictable. The limitation of these findings specific to this proposal was absence of an ICK functioning response option. While the findings of Vitale et al (2008) suggest that participants wouldn't be impacted by the availability of an ICK option, that is still an empirical question to be clarified.

Taken together with the theoretical implication that KUs may be frequently derived but they are rarely accurate, previous work on KU responses suggests that informative next steps include systematic replication and extensions. These extensions may be aimed at developing socially valid measurement tools, accuracy improving interventions, and reliable methods for testing sources of bias. Towards those ends, three pilot studies have been conducted.

Pilot 1

Pilot study 1 sought to systematically replicate the baseline accuracy measurements observed by Vitale et al (2008) for both KU and non-KU 3-Term Series problems in a socially valid format. To accomplish this, nine word problems [Table 3] were constructed from a sampling of the 48 permutations described in the previous work.

Three Term Series Problem Nine Item Quiz**Structure Key**

- | | |
|---|---|
| 1. $A1 > B1; B1 > C1 \rightarrow A1 > C1$ | 6. $C6 < B6; A6 > C6 \rightarrow A6 \text{ku} B6$ |
| 2. $A2 > B2; C2 > B2 \rightarrow A2 \text{ku} C2$ | 7. $C7 < B7; A7 > C7 \rightarrow A7 \text{ku} B7$ |
| 3. $A3 < B3; C3 > B3 \rightarrow A3 < C3$ | 8. $A8 > B8; C8 > B8 \rightarrow A8 \text{ku} C8$ |
| 4. $C4 > B4; B4 < A4 \rightarrow A4 \text{ku} C4$ | 9. $C9 > B9; B9 < A9 \rightarrow A9 \text{ku} C9$ |
| 5. $A5 > C5; B5 > A5 \rightarrow B5 > C5$ | |

Rare Names:

Alexus, Araceli, Bethzy, Damaris, Itzel, Lizeth, Nevaeh, Meara, Xander, Meara, Amiah, Everleigh, Braylon, Gauge, Hamza, Jett, Monserrat, Semaj, Talon, Yandel, Orion, Orme, Cadogan, Bassel, Harnen, Tancredi, Ryder

Quiz (with solutions):

1. Bethzy is taller than Itzel. Itzel is taller than Nevaeh. What is the relationship between Bethzy and Nevaeh?
 - a. Bethzy is taller than Nevaeh.
2. Gauge is smarter than Talon. Semaj is smarter than Talon. What is the relationship between Gauge and Semaj?
 - a. Not enough information/ I don't know
3. Meara is poorer than Damaris. Alexis is Richer than Damaris. What is the relationship between Meara and Alexis?
 - a. Meara is poorer than Alexis.
4. Hamza is heavier than Lizeth. Lizeth is lighter than Araceli. What is the relationship between Araceli and Hamza?
 - a. Not enough information/ I don't know
5. Braylon is more fashionable than Jett. Monserrat is more fashionable than Braylon. What is the relationship between Monserrat and Jett?
 - a. Monserrat is more fashionable than Jett.
6. Orion is slower than Xander. Yandel is faster than Orion. What is the relationship between Yandel and Xander?
 - a. Not enough information/ I don't know
7. Everleigh is less gracious than Amiah. Meara is more gracious than Everleigh. What is the relationship between Meara and Amiah?
 - a. Not enough information/ I don't know
8. Orme is more diligent than Cadogan. Bassel is more diligent than Cadogan. What is the relationship between Orme and Bassel?
 - a. Not enough information/ I don't know
9. Ryder is more progressive than Tancredi. Tancredi is less progressive than Harnen. What is the relationship between Harnen and Ryder?
 - a. Not enough information/ I don't know

Table 3:

9-Item 3-Term Series word problems and answers including both KU (2,4,6-9) and non-KU (1,3, &5) problems. Solutions represent only one of a number of equivalently functioning acceptable answers. Rare names were used to reduce non-experimental influences and gender matched within questions.

These word problems broke down to three non-KU and six KU type questions. During the development of this questionnaire, it was noted that in most naturalistic settings participants are not made explicitly aware of the availability of a response option functioning as ICK and so the response format deviated from the coin sorting with an explicit ICK to a free response short answer field. Participants are presented with two relational statements about three fictional

people, where Person 1 is compared to Person 2 and then Person 3, and then asked “What is the relation between [Person 2] and [Person 3]?” For example, “Bethzy is taller than Itzel. Itzel is taller than Nevaeh. What is the relationship between Bethzy and Nevaeh?” In this case, this is a non-KU problem and the expected accurate response would be some variant of “Bethzy is taller than Nevaeh.” In KU variants of the problem, a response would be counted as accurate if it was some variant of “Cannot Know.” If the previous data replicated consistently, then overall accuracy scores should approximate 33% with the majority of correct answers occurring on non-KU problems and incorrect answers occurring on KU problems.

It is worth highlighting the difference between KU problems and assessment questions generated based on cognitive theories of bias and dual processing (Kahneman & Frederick, 2002; Tversky & Kahneman, 1974). In the Cognitive Reflection Test (Frederick, 2005) questions, (e.g., “A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?”) each has a specific correct answer other than “I Don’t Know.” While both attempt to evoke a response representative of some ability to respond to ambiguity effectively, the KU problem does not assume that those capable of responding accurately will also experience an initial inaccurate response.

Two cohorts of participants (USA: n=196, Ireland: n=186) were recruited for this initial pilot study as part of a collaboration with Louise McHugh’s lab at the University College of Dublin. Each cohort was approved by their respective institutional review boards and were recruited via electronic means (e.g., Amazon Mechanical Turk, Social Media Advertisements, or Email). Participants were compensated with \$15 as part of a larger pilot study participation (MTurk; US cohort) or completed without compensation as a standalone survey task (Social Ads & Email: IE cohort). Overall group mean accuracies of 36.21% (US) and 34.87% (IE) were

recorded respectively with non-KU problem specific group means (82.82% US; 85.02% IE) falling roughly within the expected range. Possibly due to the lack of an obvious indication that ICK was an acceptable answer, performance on KU problems was significantly lower than the previous work with KU problem specific group means of 13.45% (US) and 9.82% (IE).

Deviation from previous observations may be related to scoring criteria. While there were many accepted forms of ICK (e.g., cannot tell, impossible to know, dunno, unclear, etc...), one common response form excluded from being coded as accurate was a simple restating of the information provided. For example, in one KU problem, Person 1 is faster than Person 2 and Person 3 is slower than Person 1. A typical restating-the-knowns answer was “Person 2 and 3 are slower than Person 1.” While this response is technically accurate, there is ambiguity about the function of this response. Restating-the-knowns may increase contact with available information and improve the likelihood of an accurate outcome or serve to avoid aversive consequences of answering more explicitly incorrectly. If avoidance is the primary function, constructive exploration of what is not known is unlikely and may serve the same function under non-KU conditions rendering it inaccurate for KU relational responding. For this reasoning, restating-the-knowns responses were coded as incorrect for Pilot 1 and the mean accuracies reported above reflect that conservative decision. Considering the relative high frequency of restating-the-knowns responding, future investigations may seek to parse the function of this form of responding.

To investigate the effect of having an explicit ICK response option, 43 participants in the US cohort also completed a 104-trial sequence of baseline no-feedback relational response selections after completing the questionnaire. In this phase, participants were presented with 26 pairs of non-word trigrams (e.g., AHX:GBY) one pair per trial four times in random order and

provided three or four known word relational response options buttons (SAME, GREATER THAN, LESS THAN, and for some I CANNOT KNOW) [Figure 7].

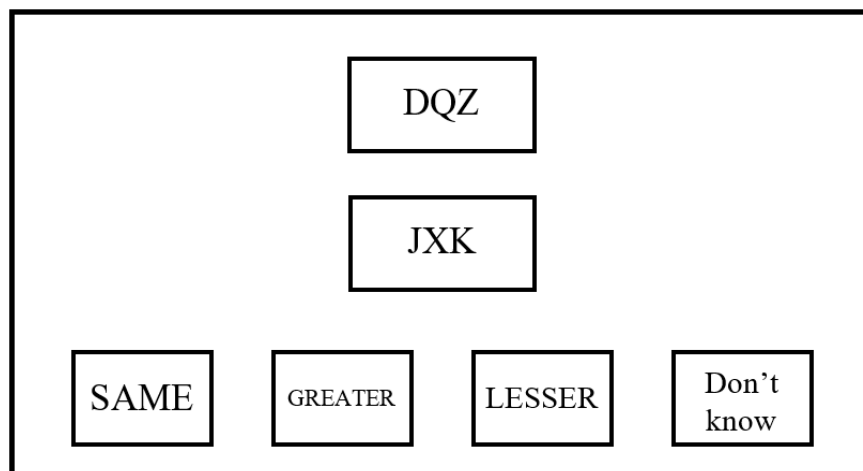


Figure 7:

Pilot 1 Trial Screen: Non-word trigrams arranged vertically with known word relational response options arrayed horizontally along the bottom of the screen.

Participants were allowed 20 seconds to select a response option. No feedback was provided for any trials other than if they ran out of time, in which case the screen showed “TIME EXPIRED” for 3 seconds before beginning the next trial. Response option buttons were arrayed in a row along the bottom of the screen in identical sized rectangles and quasi-randomized such that no button contained the same words more than twice in a row. Five hypothetical response patterns were expected given that the trigrams were putatively novel and would not evoke significant pre-experimental relational history. First, a participant may press one button location regardless of the words it contained. Second, they choose a single relational term regardless of location. Third, they randomly select response options. Fourth, they choose ICK every time it is available because that best reflects their baseline level of history with the stimuli. Or fifth, they respond in a relationally self-coherent manner to the stimulus pairs as they progress through trials. For example, if they consistently choose FGR>GJT and RDU=GJT without any feedback,

they are also choosing $GJT < FGR$, $GJT = RDU$, $FGR > RDU$, and $RDU < FGR$ consistently over repeated trials.

If the presence of an ICK option materially influences choosing other relational response options, the data from this cohort suggested that was not the case. While one participant did consistently select ICK for nearly all trials, the other 42 participants demonstrated near random response selection. There was no indication that the explicit presence of an ICK response exerted reliable and discriminated influence on responding to relational comparisons that participants had no history with.

Pilot 1 data supported that baseline measures and question structures provided by Vitale et al (2008) replicate well. The absence of an explicit ICK response option may have magnified the accuracy difference between KU and non-KU 3-Term Series problems. And the presentation of the 3-Term Series problems as fictitious people in arbitrary relations provided a socially valid questionnaire that is sensitive to the phenomena of interest. This raised the next question of if a MET protocol would replicate previous findings of trainability of accurate KU responding.

Pilot 2

Pilot 2 attempted to extend the previous findings of KU trainability using an arbitrary stimuli MET relational training between pre and post 3-Term series word problems [Figure 8].

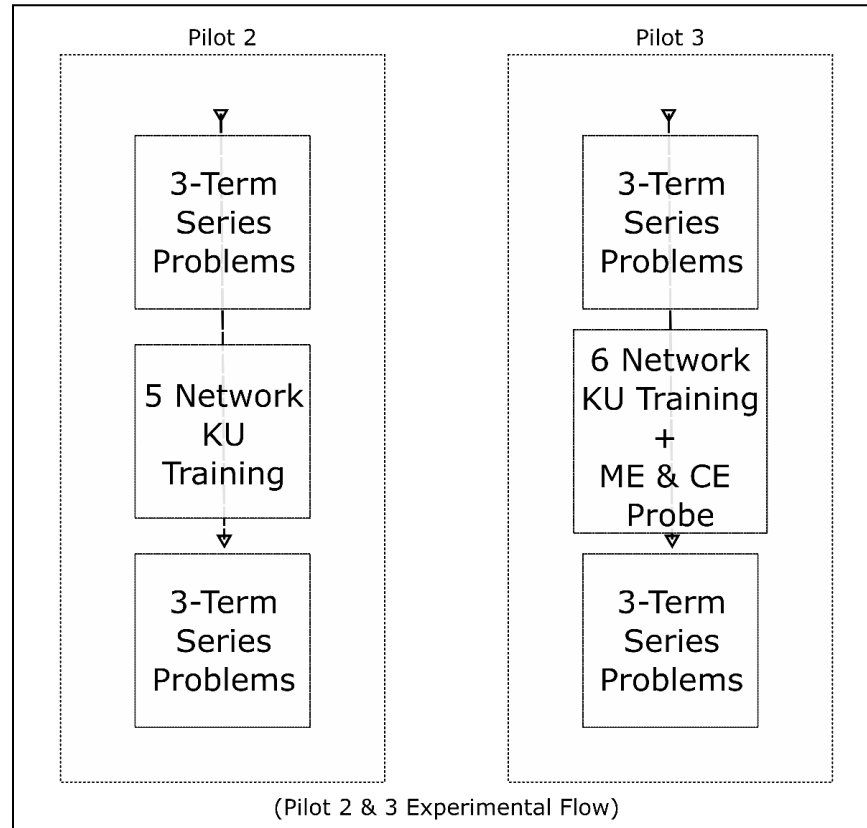


Figure 8:

Pilot 2 & 3 experimental flow diagram. Network training phases were delivered in a Multiple Exemplar Training format.

Instead of presenting known word problems and a sorting task, participants were presented with two non-word trigrams (identical to the later US phase of Pilot 1) and four relational response options. Differing from Pilot 1, participants were given “CORRECT” or “WRONG” feedback contingent on predetermined relations. These relations described five three-stimuli KU networks that were structurally similar to the 3-Term Series problems [Table 4].

Without IDK Networks

- Network 5: $A5 = B5 = C5$
- Network 6: $A6 > B6 > C6$
- Network 12: $A12 < B12 < C12$
- Network 13: $A13 > B13 = C13$
- Network 14: $A14 < B14 = C14$

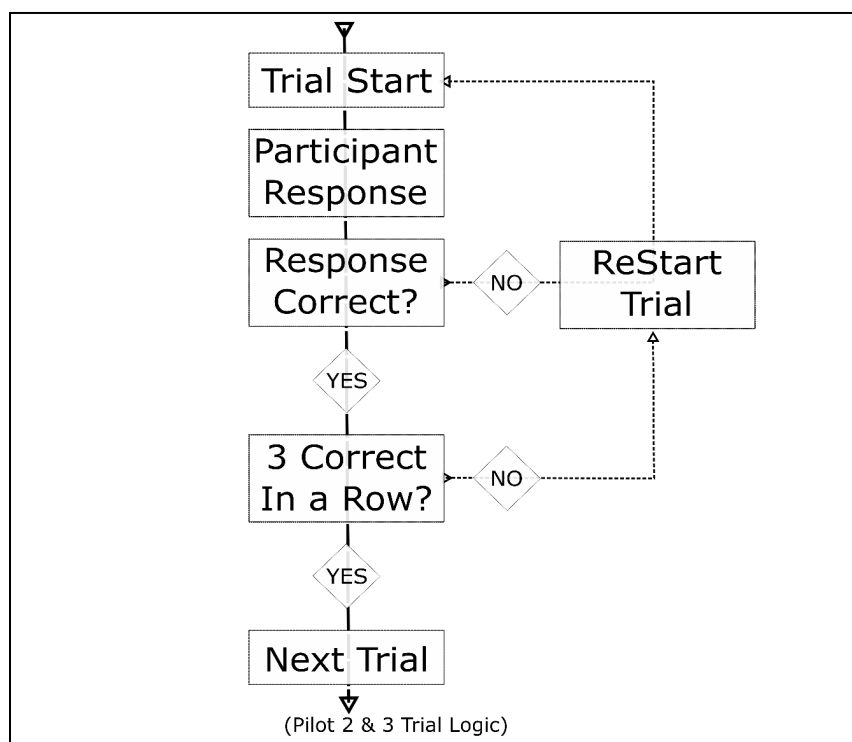
With IDK Networks:

- Network 7: $A7 < B7 > C7$
- Network 8: $A8 > B8 < C8$
- Network 9: $A9 < B9 > C9$
- Network 10: $A10 > B10 < C10$
- Network 11: $A11 > B11 < C11$

Table 4:

Three Stimulus networks of either non-KU (Left) or KU (Right) deriving outcomes. Network numbering sequence (e.g., A5, A6, A12...) is an artifact of the development process. Participants experience non-word trigrams for each stimulus. Both KU and non-KU variants were built but Pilot 2 & 3 focussed on the KU variants.

Participants were presented as one relational pair per trial and participants were expected to learn the relations and derive the networks via immediate feedback for selecting a relational response option. A participant was required to select the correct response for a particular pair three trials in a row before advancing to the next pair [Figures 9].

**Figure 9:**

Pilot 2 & 3 trial logic diagram.

Pairs were sequenced so that one network was trained at a time and within that network training, only one stimulus changed between trials if possible following the One-to-Many protocol described by Arntzen, Grondahl, and Eilifsen (2010, pp. 438–439)[Table 5].

Without IDK Condition:		
1. A5:B5	13. A12:B12	25. A14:B14
2. A5:C5	14. A12:C12	26. A14:C14
3. B5:C5	15. B12:C12	27. B14:C14
4. B5:A5	16. B12:A12	28. B14:A14
5. C5:A5	17. C12:A12	29. C14:A14
6. C5:B5	18. C12:B12	30. C14:B14
7. A6:B6	19. A13:B13	31. A15=B15
8. A6:C6	20. A13:C13	32. B15>C15
9. B6:C6	21. B13:C13	33. B15=A15
10. B6:A6	22. B13:A13	34. C15<B15
11. C6:A6	23. C13:A13	35. C15<A15
12. C6:B6	24. C13:B13	36. A15>C15
With IDK Condition:		
1. A7:B7	13. A9:B9	25. A11:B11
2. A7:C7	14. A9:C9	26. A11:C11
3. B7:C7	15. B9:C9	27. B11:C11
4. B7:A7	16. B9:A9	28. B11:A11
5. C7:A7	17. C9:A9	29. C11:A11
6. C7:B7	18. C9:B9	30. C11:B11
7. A8:B8	19. A10:B10	31. A16<B16
8. A8:C8	20. A10:C10	32. B16>C16
9. B8:C8	21. B10:C10	33. B16>A16
10. B8:A8	22. B10:A10	34. C16<B16
11. C8:A8	23. C10:A10	35. C16kuA16
12. C8:B8	24. C10:B10	36. A16kuC16

Table 5:

Specific trial training sequences for non-KU (Top) and KU (Bottom) training versions. Pilot 2 participants experienced only trials 1-30. Pilot 3 participants trials 1-36. Bold trials provided no feedback as mutual and combinatorial entailment probes. Both KU and non-KU variants were built but Pilot 2 & 3 focussed on the KU variants.

For example if the first presentation was stimulus pair A:B, the next pair would be either A:C or C:B (i.e., swapping out one stimulus but the remaining stimulus stays in the same location) but not B:C or C:A (i.e., swapping stimulus and locations). In this pilot training, all relational response selections were consequated and all relations were experienced so a participant completed a minimum of 90 trials because there were 30 unique trials that each required correct responses three times in a row. Mean number of trials to complete this pilot was 198 with a median of 149 trials . Once training was completed, participants had to complete the nine 3-Term Series word problems as their last step.

Based on a power analysis conducted using previous pilot data, a target n of 26 was set for this pilot. Due to a technical issue, 27 new participants based in the US via Amazon Mechanical Turk completed this pilot. Participants were compensated \$15 for their effort. Overall mean accuracy scores on the 3-term series problems replicated pilot 1 during pre-test ($M=37\%$, 95% CI [29%, 44%]) and differed significantly at post-test ($M = 55\%$, 95% CI [42%, 68%], $p<0.005$, $d=0.91$) based on a paired samples t-test ($t=3.078$).

Despite statistically significant pre-post mean accuracies with a large effect size, non-relational explanations were suggested by participant data. Four participants each responded once to a non-KU problem in the posttest with an ICK response. Up until this pilot, false positive occurrences of ICK had not occurred in either pre or post versions of the 3-Term Series word problems. The sudden increase of ICK responses in problem types that did not require then suggested that the observed changes to ICK frequency may be accounted for simply by an increase in base rate ICK responding more so than an increase in accurate ICK responding per se. Thus, while the statistically significant pre-post increase in mean accuracies suggested that an arbitrary trigram MET procedure may have improved accuracy of KU responses, the data also raised the possibility of non-relational sources of this effect. This possibility was explored further in Pilot 3.

Pilot 3

A demonstration of mutual and combinatorial entailment is the accepted minimum sufficient evidence needed to demonstrate arbitrarily applicable derived relational responding. For example, a participant trained on $A<B$ and $B<C$ must respond consistently and accurately to the derivable relations $B>A$, $C>B$, $C>A$, and $A<C$ without feedback. To probe this, six additional

unique trials were added at the end of the MET used in Pilot 2 [Table 5: Trials 31-36]. The first two of these trials trained $A < B$ and $B > C$ with feedback identical to the previous trials. The following four trials presented B:A, C:B, C:A, and A:C stimulus sequences with the four relational response options as usual, but did not provide feedback on participants' selections. Participants were presented each sequence three times in a row for a total of 12 total probe trials. Once probe trials were completed, participants automatically advanced to the final 3-Term series problems as in Pilot 2.

22 new participants, recruited and compensated the same as Pilot 2, completed Pilot 3. Unlike Pilot 2, only two participants showed significant improvement in ICK accuracy from pre to post. Additionally, five false positive ICK responses occurred across three participants in their posttest responses. Mean mutual and combinatorial entailment probe accuracies were moderate (68%) and low (31%) respectively. Individual participant performances on entailment probes had insignificant Pearson's correlation ($r(20) = .28, p = .21$) to their pre-post 3-Term Series change in ICK accuracy. Taken together, this data suggests that the pre-post improvements observed in Pilot 2 were most likely due to a non-relational increase in the frequency of participants' emitting "I Cannot Know."

Pilot Experiment Summary

These pilot data provided key insights into KU responding and highlighted how easy it is to confound the function of an ICK response. In Pilot 1 and the pre-training 3-Term Series responses, participants replicated low accuracy on 3-Term Series KU problems. Human beings appear to be very weak at accurately describing when we don't know, even with seemingly simple logic problems. In pilot 2 and 3, participants demonstrated that exposure to a MET of

ICK responding will increase the frequency of saying “I Cannot Know” but accuracy and relational functions of such a response are not assured.

Pilot 3 suggested that the reference publications needed to be revisited for suggestions on how to train accurate and relational ICK responses more effectively. In Experiment 3 of Vitale et al (2008) investigators incorporated non-arbitrary characteristics into the training procedure. They changed the coins from being the same size to differentially sized in alignment with their relative values in each training problem. For example, if the problem provided $A > B$ and $B > C$, coin A would be the largest diameter, B would be medium diameter, and C would be the smallest diameter [Figure 10].

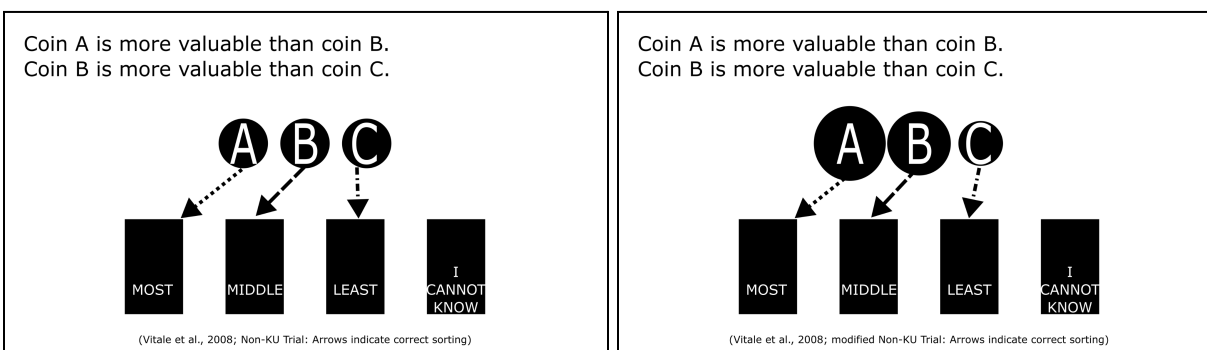


Figure 10:

Non-KU trials Before (Left) and After (Right) relative coin size was introduced to the Vitale et al., 2008 procedure.

The introduction of non-arbitrary (formal) relational characteristics based on the form of related events is an accepted practice within RFT procedures and relational training more broadly (Budziszewska et al., 2022; Dixon et al., 2021) when shaping non-arbitrary relational responding as a precursor to the establishment of arbitrarily applicable derived relational responding. The general explanation for leveraging formal properties is that human children are most likely to learn how to relate stimuli based first on forms before that response comes under the control of arbitrarily social cues. In other words strengthening a non-arbitrary relational

repertoire commonly used to scaffold acquisition of arbitrarily applicable relations (e.g., Berens & Hayes, 2007). Incorporating non-arbitrary relational training may mimic that sequence and enhance acquisition of more challenging arbitrary relations.

In the case of the coin sorting procedure, introducing non-arbitrary elements may have had unintended consequences, however. Specifically, imagine the following presentation for a KU type sorting trial. The provided information is $A > B$ and $A > C$. In form, this would require the A coin to be large and the B and C coin to be smaller than A but the same size relative to each other. If B and C are anything other than the same size relative to each other, their physical forms would no longer compliment the trial's value descriptions. A correct sorting for this trial would require placing the A coin in the "largest" bin and the B and C coins in the ICK bins. And this is what the investigators recorded. Participants in KU trials quickly came to respond accurately when form allowed them to discriminate the one smallest or largest coin from the two of the same size and that accuracy maintained when feedback was withheld but formal characteristics were retained. The issue here is that the ICK sorting bin no longer requires the participant to identify when they cannot know. All the participants need to do to accurately sort two coins to the ICK bin is to notice that the coins are equal in size [Figure 11].

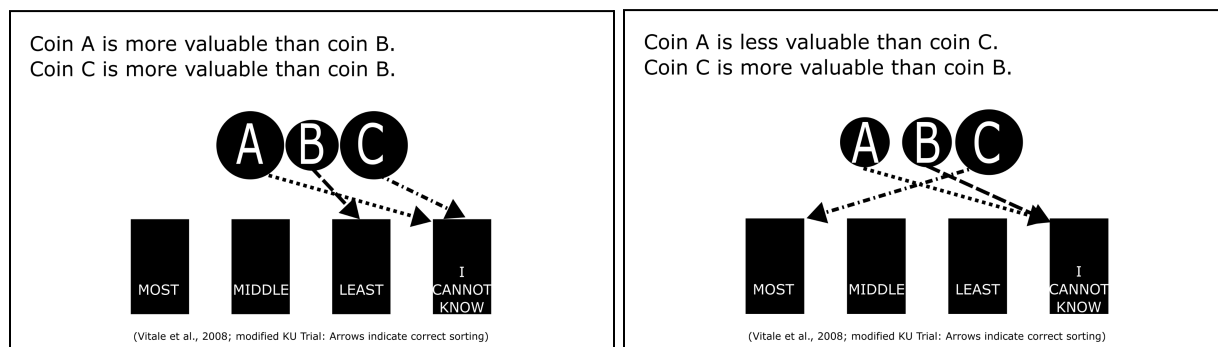


Figure 11:

KU trials after relative coin size was introduced to the Vitale et al., 2008 procedure. Note how same size coins will always be sorted to the “I Cannot Know” bin.

While the form of the response was linked to a label saying “I Cannot Know,” the *function* of the response may simply have been “I know these two are the same.” SAME responding is an equivalence function. As soon as investigators introduced differential sizes into their training procedure, they arguably stopped training and measuring KU accuracy in an unambiguous way, and started training a functional variant to equivalence responses. The accuracy improvements measured after implementing this change were significant, robust, and maintained through a methodological replication published in 2012 (Vitale et al., 2012) but the authors did not appear to be aware of this possible confusion about the actual functional units these performances were measuring.

There are two important limits specific to KU training and responding that derive from this understanding. First, when non-arbitrary forms are incorporated into relational training of KU responses, it is often possible and perhaps even likely that equivalence or difference functions can account for any observed training effect. Take for example a trial where pictures of two people are presented and questions asked about either formal relations that are visually available (e.g. who is older/taller/paler?) or not visually available (e.g., who is wiser/more generous/friendlier?). Once a participant begins to respond to the obvious/non-obvious difference between the two types of questions’ relational criteria, their responding no longer requires KU relating to accurately respond with ICK. They only need to discriminate the presence or absence of the relational cue (e.g., if the question asks for height and relative height is a visually available characteristic, answer based on height, otherwise, if height is not visually available, answer ICK). Similarly, consider trials that use geometric shapes with varying surface areas or perimeter

lengths as stimulus pairs. Participants would be prompted to sort the shapes by surface area or perimeter length, much like the coin sort task, but including extremely fine differences in the task. This relies on an assumption that an accurate ICK response occurs in a discrimination range between clearly different objects and identical objects. This is a flawed assumption. An organism that can detect a difference between two objects enough to evoke a differential response, has discriminated the difference and is responding in accordance. An ICK response resulting from a training using this procedure would not actually be indicative of deriving a relation of unspecified function, but of responding to a specified range of near similarity (or very slight difference). This is functionally equivalent to operant discrimination training (Martínez-Harms et al., 2014; Moll & Nieder, 2014; Sidman & Tailby, 1982; Young & Wasserman, 2001) and does not unambiguously bring the relational functions specific to KU responding into the training context.

The second observation is an extension of the first. The first observation suggests that it is rare, and may be impossible, to train accurate KU responses using formal characteristics of stimulus. Therefore, it is a reasonable assumption that accurate KU responding is unlikely to be learned by chance in naturally occurring contingencies. This is contrasted with the basic RFT assertion that many relational repertoires, both non-arbitrary and arbitrary, can be the result of interactions with naturally occurring and social contingencies (Hayes et al., 2001, pp. 25–28). Despite this contrast, it is not contradictory to RFT. It simply suggests that in the special case of Known-Unknown relational responding, individuals should not be assumed to be able to accurately respond to KU situations in the absence of highly specific relational training or operating in unique contingencies selecting for KU accuracy. This assumption of weak nascent KU responding aligns well with the observed data.

Taken together with the theoretical possibility that KU relations may be derived as, or more, frequently as all other relations, it is likely that most individuals have an extended history of reinforcement for responding inaccurately to their KU relations. This implies that most individuals are starting with a strong inaccuracy bias and will require elaborate, or intensive, training of accurate responding before significant and robust improvements tied to relational responding evidence are observed.

Proposal

The goal of this line of research is to isolate the conditions under which KU biased responding occurs and implement an accuracy training that is significant and robust enough to counteract common biasing conditions. It is not clear if that goal can be met in any given study no matter how well designed and described, since our pilot work has documented how hard it is to refine training procedures that have an unambiguous impact. Previous research and pilot studies have established that KU inaccuracy is persistent and challenging and a viable accuracy training producing relational responding eludes early efforts. The present proposal delineates needed refinements to the piloted MET procedure as a stepping stone to future attempts to examine the impact of KU responding on bias. Instead of a formal experimental proposal, in line with the value of iterative and inductive research in basic behavior analysis, I propose to explore a sequence of iterative steps to see if a robust training procedure can be identified by extending from the current conceptual and empirical base.

I propose a sequence of specific changes that will be pursued in order of implementation complexity to identify such a training approach. They are (1) extending the current training, (2) adding less complex arbitrary relational training, (3) integrating response cost contingencies, (4)

making previous correct responses available immediately, (5) changing the form of training to known word problems, (6) introducing ICK evoking scenarios, and (7) introducing non-arbitrary characteristics. The last of these is proposed not to evoke KU functions, as discussed above, but for complimentary reasons elaborated on below. If these iterations fail, or the results of earlier steps render all planned later steps meaningless, I will defend this set of failures as documentation of how difficult the KU issue is to experimentally analyze at present. Conversely, if some of the preparations lead to major improvements in training, those findings will be the focus of a final study investigating the conditions under which biasing and undermining of biasing of KU responses occur.

In what follows I will describe each of the planned research steps above and defend their relevance. I will then turn to the actual experiments done and the results obtained. First, however, I need briefly to discuss what is meant by improvement of KU responding.

The thresholds for major improvement of KU responses are a combination of within subject performance measures on the entailment probes and post training 3-term series word problems. Successful performance on entailment probes would be demonstrated by high accuracy (i.e., greater than 9 trials (66.6%) correct) on the last complete set of probe trials per participant. Alternatively, a maximum chance level of accuracy could be calculated using the upper bound of the 95% confidence interval for a binomial distribution of 12 trials with 0.25 probability of guessing correctly on any trial (i.e., $X \sim \text{Binomial}[12, 0.25]$). This chance threshold (>5.94 correct) allows for failure of all six combinatorial entailment probe trials. That

would fail to support claims of training effect for combinatorial entailment so the much more conservative threshold will be adopted here.

Successful improvements on the 3-term series word problems can be broken into three measures of overall improvement, improvement in accurate IDK responding, and false positive IDK responding. Overall improvement will be assessed in the same manner as pilot 2 and 3 with a pair samples t-test of pre and post overall mean accuracy scores. During pilots 2 and 3, two participants (~4%) scored perfectly on all pre and post 3-Term Series questions. This potentially diluted the estimates of significance and effect size. The incidence of this occurring is low enough that at this time, all participant data will be included in this analysis. If a significant proportion of participants show similar perfect performance both pre and post, alternative, less conservative, analysis (e.g., percentage of possible improvement scoring) may be explored.

Evaluating accuracy changes of KU and non-KU problems specifically is partially accounted for in the above t-test. But statistical significance does not automatically mean meaningful change. While the 3-term series word problems are free response, we may use the pilot 1 group level accuracy of KU problems (13% in US cohort) as a rough estimate of the probability of a chance correct response for any given trial. Assuming a binomial distribution of chance responding to KU questions (i.e., $X \sim \text{Binomial}[6, 0.13]$), the upper bound of the 95% confidence interval for chance correct ICK responses is 2.39. Within any participant, pre to post improvement on KU questions of three or more is not likely due to chance and may be considered meaningful. For false positive ICK responses, the observed frequency in pilot 1 was just over 1% of all non-KU answers (i.e., 6 of 588 responses in the US cohort). The rarity of this event suggests that any pre to post change to false positive ICK responses can be assumed to be a result of training and undermine any claim of major improvement.

Specific variations discussed below may imply additional process measures of behavior change, but the entailment probe accuracy thresholds and 3-term series problem measures outlined above will be the primary focus of analysis for the proposed sequence below.

Planned Experimental Change #1: Extending the Current Training

In pilot work participants experienced the MET as one time through five three-stimulus networks that are trained and tested through 102 or more individual trials. Including pre and post questionnaires, participation only takes ~15 minutes. Pilot 2 and 3 data suggested that this short training increased the base frequency of ICK responding but the response did not unambiguously demonstrate relational properties. The planned experimental change #1 examined whether applying a mastery criteria to the entailment probe trials and requiring participants to repeat the training trials with novel stimuli until the mastery criterion is achieved would increase accurate KU responding. In different conditions of Quinones & Hayes (2014) minimum mastery criterions of 80% and 100% with similar repetition contingencies were implemented to ensure that participants were performing to an acceptable level. Minimums criteria of 10 out of 12 trials answered correctly (~83%) or better and repetition contingencies can be coded into the training procedure without significant additional time or changes [Figure 12].

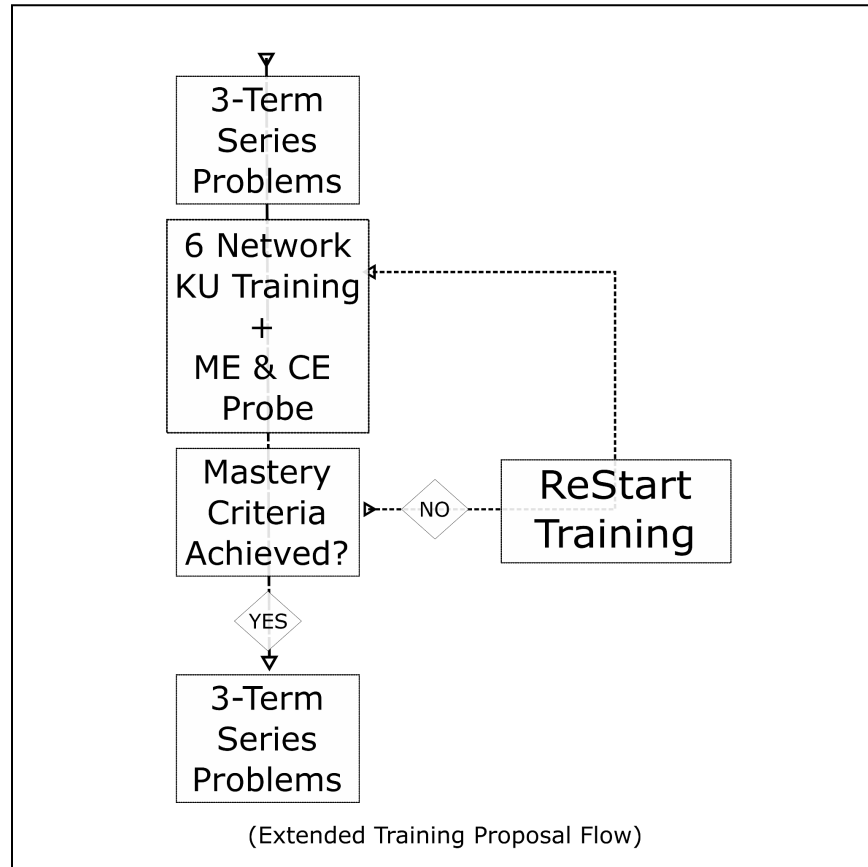


Figure 12:

Addition of a mastery criteria and retry contingency for mutual and combinatorial entailment probes may ensure sufficient training with durations sensitive to each participant's performance.

This modification may provide participants with enough varied examples to bring their, now more frequent, ICK responding under specific control of the KU conditions. If this step is successful, the relational properties of ICK will be seen by participants completing the MET and advancing to the post 3-term series problems. While this may constrain the ability to draw conclusions from the analysis of entailment probe data, the previously discussed three-part analysis of 3-term series problem data may be sufficient to determine training impact on known word problems.

Planned Experimental Change #2: Simpler Relational Training

In all of the pilot studies, a non-zero number of participants responded to non-KU questions inaccurately. This is unlikely to be due to low-effort-responding of Mechanical Turk workers attempting to game the system by doing the highest volume of low quality work and submitting responses without significant engagement with the questions because the main training sequence includes attentional checks that filter out those who are not attending to a minimum level. Another possible explanation is that some participants lacked even a basic relational repertoire required to participate effectively in the KU training. These participants may need to be excluded from analysis or provided additional relational exemplars that train simple equivalence and comparative networks to build base skills required to benefit from KU accuracy training.

This may take one of three forms [Figure 13].

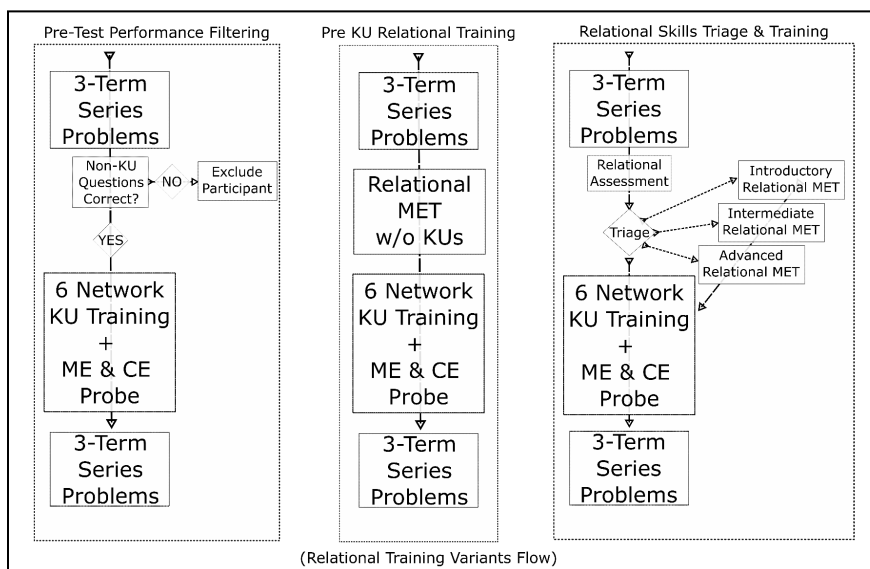


Figure 13:

From simplest to most complex (Left to Right), proposed modifications that may account for many participants requiring some degree of relational training prior to KU training being effective.

The simplest version would be to exclude those individuals who cannot accurately answer the non-KU 3-Term Series word problems in the pre-test. Second, an equivalence and comparative non-KU relational network training sequence may be added at the start of the training and only advance participants to KU training once they meet a minimum mastery level. The most complex version would be to build an entire relational assessment phase and triage participants to specific levels of training based on their assessment outcome. The increasingly complex levels of training based on assessment could be non-arbitrary equivalence, non-arbitrary comparatives, arbitrary equivalence, arbitrary comparatives, and then arbitrary KU training. By the time participants reach the KU training, they have a minimum viable relational repertoire to benefit from the KU accuracy training trials. The first and second versions are relatively quick to implement from the current code base while the most complex version may need to be explored if data strongly indicates this level of refinement is required.

If this step is found to be successful, the relational properties of ICK will be seen by the analysis of entailment probe data and 3-term series problem data as detailed at the top of this proposal section. Additional relational process measures may then be implemented during the non-KU training (either variant 2 or 3) similar to the entailment probes used in pilot 3. This probe data may also be used as mastery criteria as described in the previous section in order to more effectively regulate the progression of each participant.

Planned Experimental Change #3: Response Cost

During all pilots, participants were slightly deceived about the final compensation. They were instructed that compensation was \$10 for any completion and there was a \$5 bonus for those individuals who completed within a specified time limit. In actual implementation, any completion was compensated with \$15. This light deception was intended to function as an enhancement to attend to the feedback provided. Participants were specifically told that they would be able to complete the work fastest if they attended to what gets them “Correct” feedback. For participants working via Amazon Mechanical Turk, there is a reasonable assumption that they are also working on an income maximization goal that further encourages attention to feedback but this may be more reliably addressed by adding more immediate response cost contingencies. For example, Quinones and Hayes (2014) displayed a points counter during all feedback trials that incremented for correct responses and decremented for incorrect responses. Participants were instructed that these points represented their compensation (5 cents per point). This was also a light deception because wrong answers did not actually cost money despite the points counter decreasing. In those experiments, points were not displayed during probe trials but otherwise provided an additional immediate signal of response cost. The proposed modification would implement a similarly contingent points display[Figure 14].

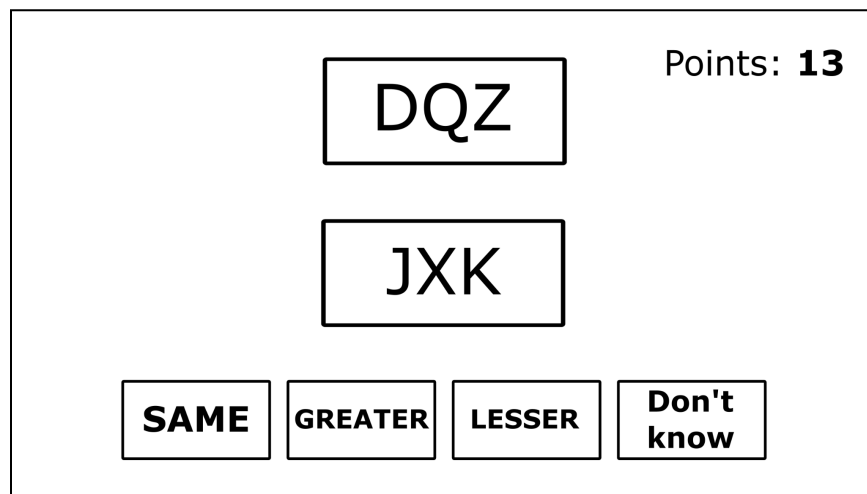


Figure 14:

Proposed trials after response cost points counter (upper right corner) is introduced.

If this step is found to be successful, the relational properties of ICK will be seen by the analysis of entailment probe data and 3-term series problem data as detailed at the top of this proposal section. Additional process measures specific to the response cost contingency described here are not expected. Marginal improvements in the primary outcome measures after implementation of this response cost contingency may suggest additional parametric manipulation of the specific points contingencies involved (e.g. altering the amount of points provided for correct or incorrect answers or modifying the contingency schedule properties).

Planned Experimental Change #4: Correct Previous Answers Review

The relational training methods between Vitale (2008) and Quinones and Hayes (2014) differ in how much information is available to participants during each trial. The latter, as well as the current procedure, follow a standard MET sequence of presenting a sample and then comparison stimulus followed by response options. This is similar to a multiple choice fill in the blank problem of “Sample BLANK Comparison” that relies on a participant responding to recent history. In contrast, during the Vitale (2008) procedures, participants were provided all parts of

the 3-Term Series problem at the same time and asked to fill in the blank given the available information. The proposed modification would build on the Vitale (2008) method by providing participants previously answered correct statements involving the current sample and comparison stimuli on screen alongside the current trial presentation [Figure 15].

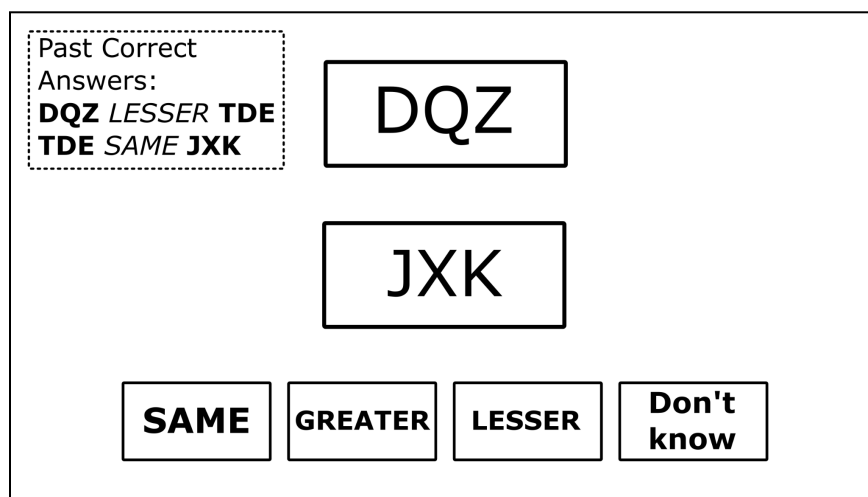


Figure 15:

Proposed trials after correct previous answers list (upper left corner) is introduced.

For example, if a participant is presented with DGV and JKR as sample and comparison and has previously correctly answered “HJT<DGV” and “JKR<HJT,” those two relational statements could be displayed on screen alongside the current sample and comparison stimuli. Displaying the previously correct answers would provide a scenario similar to the 3-Term series training that provides two relational statements and a prompt for the third.

If this step is successful, the relational properties of ICK will be seen by the analysis of entailment probe data and 3-term series problem data as detailed at the top of this proposal section. Additional process measures specific to the availability of previous answer information described here are not expected.

Planned Experimental Change #5: Known Word Questions

The current training presents a sample and comparison stimulus using non-word trigrams. Animation sequences and visual arrangement are used as contextual cues to guide responding within an otherwise novel experience. This follows one of a few standard presentation sequencings for MET within the RFT literature (Delabie et al., 2022; Dixon et al., 2021, p. 202; May et al., 2022; McLoughlin et al., 2020) but may be too abstract for some participants given the otherwise brief experience.

In Vitale et al's (2008) procedures, for example, all sorting training trials were presented with the two known word relational statements. In this way, there were no abstractions that must be learned and responded to at a later time. Instead, the experimental preparation only required the presentation of the two statements and a response opportunity for either KU or non-KU accurate responding. If the planned experimental changes above do not meaningfully impact pre-post changes in KU accuracy and relational responding indicators, this planned change is meant to explore the idea that more direct replication of previous procedures may be required. In this case, training would consist of presenting two known word relational statements, much as in the coin sorting task [Figure 16].

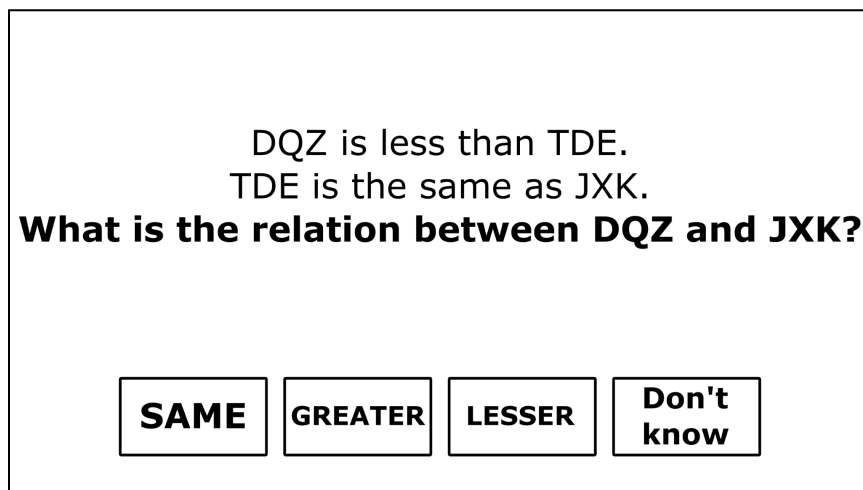


Figure 16:

Proposed trials after sample and comparison are replaced by known word relational statements (3-term series problems) (middle).

If this step is successful, the relational properties of ICK may be limited to combinatorial entailment unless additional probe trials are introduced that describe single relations and test for mutual entailment (i.e., “DQZ is less than TDE. What is the relation between TDE and DQZ?”). If that modification is not made, conclusions from analysis of 3-term series problem performance as detailed at the top of this proposal section will be somewhat more limited.

Planned Experimental Change #6: ICK Evoking

Pilot data highlighted that individuals are unlikely to emit ICK under any conditions. Therefore, increasing the frequency of expressing ICK can be a needed precursor to training accuracy and relational discrimination. One way of accomplishing this may be to begin training with questions so outrageously hard as to evoke an ICK response (e.g., “How many stable solutions are there to the three body problem?” or “What was JRR Tolkien’s intended social message within his *The Lord of the Rings* series of books?” or “What are the differences between isomers of butane?”). In this way, participants would access putative reinforcement for responding ICK correctly, perhaps resulting in an increase in base frequency. Once ICK is a more

common response, questions can transition to increasingly nuanced unknowns such as those used in the 3-Term Series KU word problems. This experimental change might especially make sense if participants continue to demonstrate a persistently low frequency of responding ICK even after training is converted to known word problems.

It is worth noting in passing that when, as in the current case, participants may be able to look up solutions on line, it is increasingly hard distinguish between what people know and what they can look up online quickly (Fisher et al., 2015; Ward, 2021). In part for that reason if this step is successful, additional process measures specific to this change may be important (e.g., number of trials required to consistently evoke ICK responses or correlational data between early ICK frequency and success in the main MET).

Planned Experimental Change #7: Non-arbitrary characteristics.

For all the reasons discussed previously, non-arbitrary characteristics are unlikely to derive to KU responses but may be used similarly to the ICK evoking adaptation to accelerate the learning of the fundamentals required to eventually learn KU relational responding. There is precedent for this within the Quinones and Hayes (2014) procedure. During that procedure they leveraged non-arbitrary characteristics to train participants to select symbol quadgrams (e.g., !!!!, ****, #####, and \$\$\$\$) functioning as Same, Different, Greater Than, and Less Than respectively. For example, a participant presented with a single dot on the left and six dots on the right would be shown “Right!” for selecting \$\$\$\$\$. If the arrangement was reversed (i.e., six left, one right) “Right!” feedback would only occur for selecting ##### [Figure 17].

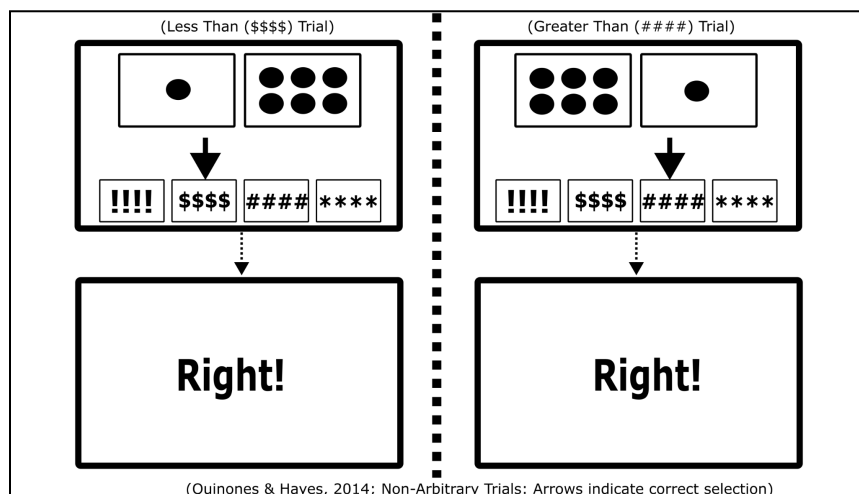


Figure 17:
Quinones & Hayes, 2014, Non-Arbitrary Trials. A Less Than trial (Left) would result in “Right!” feedback for selection of \$\$\$\$ and a Greater Than trial (Right) would result in “Right!” for selection of #####.

In this way participants were provided with basic relational training utilizing non-arbitrary characteristics prior to KU bias training. Participants were subject to a 48-trial probe block at the end of each round of this training that they had to complete error free before continuing on to the main experiment. This ensured that participants in the KU biasing procedure all had a minimum working relational repertoire of equivalence, difference, and comparative relational functions very recently trained to a high level of generalized mastery.

In this final planned experimental change, I plan to implement this same step. If it is successful, the relational properties of ICK will be seen by the analysis of entailment probe data and 3-term series problem data as detailed at the top of this proposal section. Measures of the number of repetitions to mastery and the degree of complexity of non-arbitrary relational trials may serve to further clarify if the observed low ICK accuracy is predicted by an individual’s ability to relationally respond at a much more general and basic level of task.

Final Study: Post-Training KU Biasing

As discussed above, the longer term goal of this project is to isolate the conditions under which KU biased responding occurs and to implement accuracy training or demonstrated impact so as to determine whether it is robust enough to counteract common biasing conditions. That agenda makes sense in the present context only if the planned experimental changes above result in a preparation adequate to address that question. If there are in fact no known robust methods for improving accurate KU responses, then the ultimate target of this investigation will have to await another day.

Because the method in this final study cannot be specified beforehand only a general outline can be provided. The previous literature shows how biasing can be obtained in a MET relational series that has a manipulated mix of correct responses for different relational functions. For example, participants may be exposed to a biasing protocol that is composed of 75% comparative relations (i.e., Greater Than or Less Than) and the other 25% equivalence and/or KU relations. Due to the biasing, these participants would be expected to be more likely to respond to KU relations with inaccurate Greater Than or Less Than responses. That outcome was demonstrated in Quinones and Hayes (2014). The future method would sequence the KU training block developed here prior to such a biasing protocol block such that the individual effects of each were pitted against the other. In order to account for individual pre-experimental histories, novel and arbitrary trigram relational networks could be tested without feedback both before and after the experimental blocks as a baseline and posttest.

Prior to this proposal, a version of this experiment was designed and piloted but a thorough examination of that data made it clear that at that time the KU training block was

unreliable. That conclusion precipitated this proposal. If KU training methods become reliable and robust enough in the future, the larger biasing experiment may further illuminate the social significance of being able to train accurately responding to gaps in our own experience.

Discussion

Suggesting that we respond as if events are related without knowing how they are related is not new or revelatory in itself. Figures throughout history have suggested that to know when we don't know is a powerful tool (Applications, 1981; Dunning, 2011; Frederick, 2005; Luft & Ingham, 1955; Mills & Keil, 2004; "Socrates: I Know That I Know Nothing," 2022; Stebbing, 1939). The phenomena has received relatively little attention within the RFT community, however. It is mentioned only in passing in the canonical book on RFT (Hayes et al., 2001, p. 31) and involves contingencies that are rare in nature and characterized by the absence of a specific history. Because there are less constraints on the conditions for deriving KU relations than any other combinatorially entailed derived relation, the most astounding implication is the possibility that this one functional class may significantly outnumber all other derived relations in relational repertoires as a product of overgeneralization of relational derivation.

Exploring KU responding may provide a new avenue of research and application as well as a direct line into the human phenomena of bias and prejudice, but we cannot know that until we have a better experimental analysis of KU responding itself that then can be used to unpack unfairness. The work on the dissertation has so far documented how hard such an analysis is to construct in an unambiguous fashion.

That leaves this proposal in an unusual situation. Instead of a discrete study or two linked to an experimental hypothesis or question I am proposing a plausible sequence of iterative steps

focused on an inductive experimental analysis. The Known-Unknown relational function is uniquely a product of arbitrarily applicable derived relational responding. It is procedurally devilishly hard to isolate. It may be a doorway to addressing human suffering rigorously and thoroughly in previously underestimated ways, or it may be a kind of behavioral mirage. If it continues to resist experimental analysis it may instead point to a theoretical extension of RFT that has profound implications precisely because it is so hard to establish and thus by extension may be largely functionally absent in everyday human behavior. In an area of ever-expanding knowledge mixing in with ever expanding falsehood, conspiracies, and fake news, we need as human beings to know when we do not know and for that response to be more than lip service. Whether that repertoire exists and can be trained is as yet unanswered.

In what follows I will describe the studies I conducted and in the results they generated in the sequence in which they were done. To steal my own thunder before beginning, I was unable to identify a method to reliably increase the accuracy of KU responding. I did not conduct every originally proposal experimental modification because the obtained data made certain experimental comparisons moot.

Experiment 4: Extending the Current Training

Method

Experiment 4 replicated Pilot 3 methods with training extended by the addition of a mastery criteria contingency where participants must answer at least 9 of the 12 probe trials correctly in order to move from the trigram training trials to the post-word problems. Those participants not achieving the mastery criterion were allowed to retake the trigram training block up to 9 more times. In order to capture data on training failure effects on posttest word problem

responses, those participants who failed all 10 training attempts were allowed to take the posttest word problems.

Participants

26 participants completed the experiment fully. One participant's data was excluded from analysis due to failing all attentional check questions embedded in the trigram training blocks. Two participants' data were excluded from the analysis due to failing to achieve the mastery criteria across all 10 attempts. The remaining 23 participants' data was analyzed for overall differences in word problem answer accuracy, ICK word problems only answer accuracy, and changes in the frequency of false positive "I Cannot Know" responses to non-ICK word problems.

Results

Changes in accuracy, both overall and ICK only, were analyzed using a matched pairs single tailed t-test with a $p < 0.05$ rejection threshold. Effect sizes were calculated using Cohen's d . Improvements in overall ($M_d = 0.28$, $p = 0.00009$, $d = 1.47$) and ICK only ($M_d = 0.44$, $p = 0.00018$, $d = 1.49$) accuracies were both significant with large effect sizes. While both of these results achieved the major improvement criteria detailed in the proposal and are an improvement from the Pilot 3 data, the increase in the frequency of false positives after training (1 pre, 7 post: +6 change) suggested that these improvements may have been at least partially accounted for by base rate increases in indiscriminated ICK responding and fails to meet the zero or negative change criterion of the proposal.

An additional post-hoc analysis of Experiment 4 outcomes was conducted to attempt to provide insight into where improvements in ICK accuracy were occurring. This analysis

compared responses to ICK word problems within each individual pre to post and categorized those responses based on if the response changed and if it did change, what types of responses prior to training were most frequently changing and how. A two by two matrix of right or wrong responses on pre-test to right or wrong responses on posttest produced four general categories of response shifts (Pre & Post Correct, Pre Correct to Post Incorrect, Pre Incorrect to Post Correct, and Pre & Post Incorrect). Counts for each of those categories are shown in Table 6.

		Pre	
		Correct	Incorrect
Post	Correct	21	61
	Incorrect	0	56

Table 6:

Experiment 4 General Response Shift Counts For ICK Word Problems

No participant who accurately responded “I Cannot Know” on the pre-test changed their answer to an incorrect form on the post-test (Table 6: Pre Correct Column). Of those participants that answered incorrectly to start (Table 6: Pre Incorrect Column), just over half of their responses shifted to correct in the post-test (Table 6: Pre Incorrect x Post Correct Cell). This general analysis and the question of the function of the Repeating the Knowns response begged a specific analysis of those Pre Incorrect responses and what specific types were changing, or not, in post. To address this, the Incorrect category was further divided into Repeating the Knowns (RK) and Other Wrong (OW). While the data is coded into six possible wrong answers (Repeat First Relation, Reverse First Relation, Equivalence, Restate Knowns, Other, Incorrect Unknown), the incorrect unknowns (false positive) category does not apply to this analysis of ICK only questions and RK responses were so frequent that further parsing would only create many empty categories. Counts for each of those categories are shown in Table 7.

		Pre		
		Correct	Repeat Knowns	Non-RK Wrong
Post	Correct	21.00	58.00	3.00
	Repeat Knowns	0.00	41.00	5.00
	Non-RK Wrong	0.00	3.00	7.00

Table 7:

Experiment 4 Specific Response Shift Counts For ICK Word Problems

Discussion

Nearly all of the responses that changed to ICK (58 of 61) came from RK pre-test responses (Pre Repeat Knowns x Post Correct Cell). At the same time, a minority of non-RK pre-test responses changed to accurate ICK responses (3 of 15). This contrast suggests that either the RK response may be a functional analog to ICK and thus any claim to training effect may just be a topographic shift without change in function, or those individuals already responding in this way are more likely to be more receptive to the training procedure. This second hypothesis would align with the discussion during the Simpler Relations proposal section. Some participants may not have the foundational relational repertoire to respond to the training procedure and those that do are also more likely to present the RK response. If that is the case, training focusing on simpler relations may be helpful.

Results from Experiment 4 thus suggested two distinct next steps. One was partially anticipated in this proposal: exposing participants to simpler relational training prior to the ICK training block. A second step emerged in the committee discussion at the initial proposal meeting, namely, to explore the function of the Repeating the Knowns response through the use of repeated assessment and instructions. Experiment 5, below, examines the impact of adding less complex arbitrary relational training, augmented by other small training modifications

including one identified in the dissertation proposal meeting discussion. Experiment 6, further, describes the RK response assessment analysis.

Experiment 5: Simpler Relational Training

Method

Experiment 5 extends the method of Experiment 4 and adds three additional changes. The major planned change was a basic relational MET block prior to the ICK MET block. This basic block followed the same procedure as ICK but only featured three response options (SAME, GREATER THAN, and LESS THAN). All relations trained and tested in this block were fully derivable to those three options. It was hypothesized that providing this additional training may better prepare participants who lack basic relational skills for the ICK training. It was also hypothesized that such training could inoculate against the high frequency of false positive responses in the post-test word problems.

The second and third changes introduced to the procedure were the ability to alter the number of sequential correct answers required to move forward (figure 9: above) and a visible animation of the correct response button in each training trial. The Pilot 2 choice to require participants to extend their training by repeating the same correct answer three times in a row was somewhat arbitrary. Adding another 36 novel training trials with the new relational MET block implied that participants could expect a doubling of participation time required compared to Experiment 4. With MET emphasizing training effects due to participants experiencing many novel exemplar responses, it was hypothesized that participants could experience more novel exemplars in the same relative amount of time if the repetition contingency was reduced or eliminated. The contingency, which had previously been hard coded in the program, was

converted to an adjustable parameter and a small pilot (n=5) was run to specifically test if the repetition contingency could be eliminated to balance out the time required for the additional MET block. Where participants averaged 36 minutes in Experiment 4 with the repetition contingency in place and only needing to pass one MET block, these participants that completed this pilot averaged 105 minutes. With a nearly 3x increase in time required, the data suggested that the repetition contingency had some additive effect on success of the training. Two more small pilot groups (n= 6 & 7) were then run with the repetition parameter set to twice and three times respectively. The average time to completion for each of these were 48 and 62 minutes respectively. There were no significant differences in training outcomes between these latter two groups and the ratio of participants who completed versus abandoned favored the faster procedure, so the rest of Experiment 5 was run with the repetition parameter set to two correct answers in a row on the same trial in order to advance to a novel trial.

The animation introduced in Experiment 5 serves as a visual cue for flawless responding during training trials. Participants see the correct response option button vertically bounce every two seconds. This change was implemented based on correspondence with recent participants who described adhering to ineffective verbal strategies despite the programmed feedback in combination with Dr. Contreras' suggestion during the proposal about instituting a flawless training mechanism. With this animation in place, all training trials are clearly indicated for the correct answer while all probe sessions have no animations or correct/incorrect feedback. Participants were still required to respond to the relational patterns with a high level of accuracy in order to advance to the second training block and post-test.

Participants

Excluding the parametric pilots described above, 30 new participants completed the experiment fully. Five participants' data were excluded from the analysis due to failing to achieve the mastery criteria across all 10 attempts of the final ICK block. 4 of these participants also failed to achieve mastery during the simple relations block. The remaining 25 participants' data was analyzed for overall differences in word problem answer accuracy, ICK word problems only answer accuracy, and changes in the frequency of false positive "I Cannot Know" responses to non-ICK word problems.

Results

Changes in accuracy, both overall and ICK only, were analyzed using a matched pairs single tailed t-test with a $p < 0.05$ rejection threshold. Effect sizes were calculated using Cohen's d . Improvements in overall ($M_d = 0.18$, $p = 0.00109$, $d = 0.86$) and ICK only ($M_d = 0.22$, $p = 0.0052$, $d = 0.78$) accuracies were both significant with large and medium effect sizes. The frequency of false positive ICK responses reduced to just one instance for one participant in the posttest-word problems.

Discussion

The reduction in effect size observed, combined with the near elimination of false positive ICK responses, suggests the procedure was beginning to produce a reasonably discriminated KU relational response. This conclusion is supported by the demonstration of mastery in mutual and combinatorial entailment probes during the MET blocks. This conclusion was tempered by the fact that only four participants demonstrated perfect responses in the post-test word problems, one only missed one question, and two showed some improvement. The

remaining 18 participants who did not fail the training probes broke down into 15 who showed no training effect and three who got perfect responses on both the pre and post word problems. One conclusion from these results is that the training either works or it doesn't for any given participant. With the available data, it is not clear why only about 20% of the participants that passed the mastery criteria demonstrated a far transfer to the word problems. Results from Experiment 6 (below) indicated a possible next step that may broaden training efficacy however – namely focusing on the kind of errors made.

Experiment 6: Assessment of the Function of RK Responses

Method

Experiment 6 extends the 3-Term Series problem large group survey method of Pilot 1 and added nine additional problems (Table 8) as well within and between group manipulations of normalizing instructions for ICK responding. The within subject manipulation was that one cohort of participants completed the first (Table 3) set of 9 questions and then were provided the set of 9 additional questions with the further instruction “For some of these problems, "Not enough information" or "I don't know" are the correct answers.” The between group manipulation was that the other cohort of participants were provided that same instruction at the beginning of all 18 questions.

The reasoning behind the instructional manipulation was that if the Repeat the Knowns (RK) responses were functioning as KUs, as was suggested by the Experiment 4 analysis, then two outcomes may be expected. First, those who got the instruction upfront would be more likely to accurately respond with an ICK response where others commonly repeat the known information. And second, those who get the instructions between the two structure matched sets

of questions are more likely to shift their response from repeating the knowns in the first half, to ICK in the second half.

Three Term Series Problem Nine Item Quiz Version 2

Structure Key

- | | |
|---------------------------|---------------------------|
| 1. A1>B1; B1>C1 -> A1>C1 | 6. C6<B6; A6>C6 -> A6kuB6 |
| 2. A2>B2; C2>B2 -> A2kuC2 | 7. C7<B7; A7>C7 -> A7kuB7 |
| 3. A3<B3; C3>B3 -> A3<C3 | 8. A8>B8; C8>B8 -> A8kuC8 |
| 4. C4>B4; B4<A4 -> A4kuC4 | 9. C9>B9; B9<A9 -> A9kuC9 |
| 5. A5>C5; B5>A5 -> B5>C5 | |

Rare Names:

Vinnyla, Starlette, Sianna, Charmay, Antwoh, Wiatt, Tenysi, Kairo, Kanaan, Brayan, Abrielle, Caoimhe, Damita, Eydie, Hima, Myaree, Paignton, Rumer, Tuiren, Winry, Calix, Exton, Ido, Llyr, Ohene, Tripp, Yarden

Quiz (with solutions):

1. Vinnyla is louder than Starlette. Starlette is louder than Sianna. What is the relationship between Vinnyla and Sianna?
 - a. Vinnyla is louder than Sianna.
2. Calix is nearer than Exton. Ido is nearer than Exton. What is the relationship between Calix and Ido?
 - a. Not enough information/ I don't know
3. Charmay is smaller than Antwoh. Wiatt is bigger than Antwoh. What is the relationship between Charmay and Wiatt?
 - a. Charmay is smaller than Wiatt.
4. Llyr is heavier than Ohene. Ohene is lighter than Tripp. What is the relationship between Tripp and Llyr?
 - a. Not enough information/ I don't know
5. Yarden is more serious than Rumer. Winry is more serious than Yarden. What is the relationship between Winry and Rumer?
 - a. Winry is more serious than Rumer.
6. Tenysi is sadder than Kairo. Brayan is happier than Tenysi. What is the relationship between Brayan and Kairo?
 - a. Not enough information/ I don't know
7. Kanaan is less hungry than Abrielle. Caoimhe is more hungry than Kanaan. What is the relationship between Caoimhe and Abrielle?
 - a. Not enough information/ I don't know
8. Damita is more friendly than Eydie. Hima is more friendly than Eydie. What is the relationship between Damita and Hima?
 - a. Not enough information/ I don't know
9. Myaree is more awake than Paignton. Paignton is less awake than Tuiren. What is the relationship between Tuiren and Myaree?
 - a. Not enough information/ I don't know

Table 8:

9-Item 3-Term Series word problems version 2 and answers including both KU (2,4,6-9) and non-KU (1,3, &5) problems. Solutions represent only one of a number of equivalently functioning acceptable answers. Rare names were used to reduce non-experimental influences and gender matched within questions.

Participants

418 new participants completed Experiment 6 fully. Because previous pilots and experiments illustrated that responses to these questions can vary widely, no participants were excluded from the data for wrong answers as long as the answers demonstrated some reasonable response to the questions. 14 submissions were excluded due to either duplicate, incomplete, or blatantly off topic submissions. One example of an off-topic submission that was excluded was a participant that pasted the text of the first google result from searching the first name in each question. 404 submissions distributed across two cohorts (n=200 & 205) were analyzed for within subject changes in ICK responding, within subject accuracy on both ICK and non-ICK problems, and between group comparisons of frequency of types of responses. All responses were categorically coded as one of seven response types (Correct, Repeat First Relation, Reverse First Relation, Equivalence, Restate Knowns, Other, Incorrect Unknown) in the same manner as previous pilots and experiments.

Results

Cohort 1 (n=200) read the additional instruction at the beginning of the 18 questions. Cohort 2 (n=204) read the additional instructions after the ninth question. Responses were coded using the previously described seven possible outcomes. General accuracy was scored within cohorts across surveys. Because the form of questions changed between part 1 and part 2 of each cohort, questions were matched by underlying structure using the following question number

pairings 1:10, 2:18, 3:11, 4:16, 5:12, 6:13, 7:15, 8:17, and 9:14. For clarity, question numbers below 10 are referred to as Survey 1 and above as Survey 2. Response outcomes were further coded within subjects across surveys based on changes. Three general change analyses within each cohort were conducted using McNemar's chi-squared method. Those analyses were the changes between correct and incorrect responses across surveys, the changes to and from each specific incorrect response type, and the changes between correct and repeating the knowns responses types. This resulted in eight chi-squared tests conducted on each cohort data set and as such planned $p < 0.05$ was reduced using the Bonferroni method (planned p / number of tests) to a threshold $p < 0.00625$ for these series of change analysis.

Cohort overall accuracy across surveys [Table 9] was not significantly different for Cohort 1 and were significantly different for Cohort 2 ($M_{diff} = 11.11\%$, $SD = 33\%$, $t(406) = 3.4454$, $p < 0.001$).

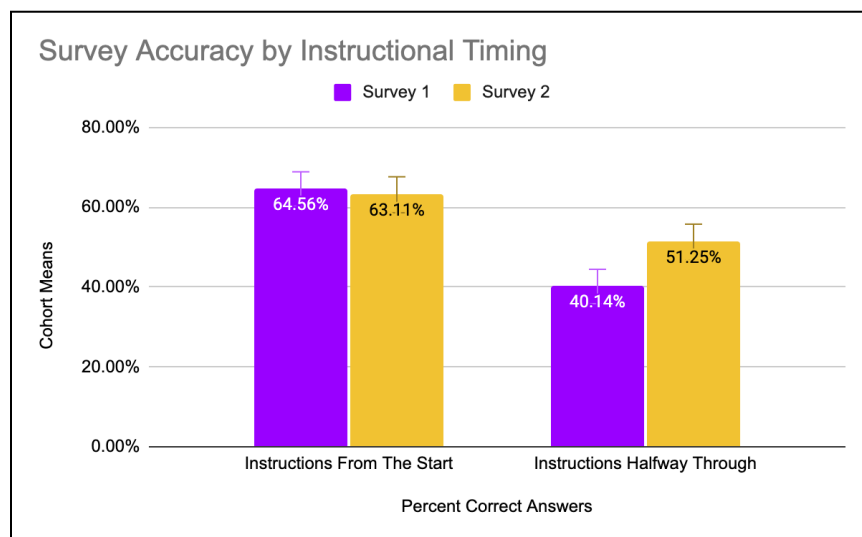


Table 9:

Experiment 6 Survey Accuracy by Instructional Timing. Cohort 1 (Instructions from the Start) had no significant change in accuracy. Cohort 2 (Instructions Halfway Through) significantly improved their accuracy after experiencing instructions.

Changes between correct and incorrect responses across surveys were not significant for Cohort 1 ($X^2(1, N = 1800) = 2.77, p = 0.096$), but were significant for Cohort 2 ($X^2(1, N = 1806) = 111.98, p < 0.0001$). Approximately 13% and 20% of Cohort 1 and 2 participants respectively changed their responses on the structurally matched questions of the second survey. Cohort 2 participants who changed were 3.48 times more likely to change to a correct answer if they got the first wrong than to change to a wrong answer if they got the first right after experiencing the instructions.

Each wrong answer subtype was analyzed for changes to and from that type. The wrong answer of repeating whatever first relation was presented in the 3-term series problem was not significant for either cohort.

Changes in reversing the first relation presented in the 3-term series problem were not significant for Cohort 1 and were significant for Cohort 2 ($X^2(1, N = 160) = 9.42, p = 0.0021$). Approximately 9% of all Cohort 2 responses involved this wrong answer type and of those about 82% changed. Cohort 2 participants who changed were 1.77 times more likely to change this wrong answer to some other answer on Survey 2 than change another answer type to this after experiencing the instructions.

Changes in answering as if the two individuals being equivalent were not significant for Cohort 1 and were significant for Cohort 2 ($X^2(1, N = 94) = 11.76, p < 0.0001$). Approximately 5% of all Cohort 2 responses involved this wrong answer type and of those about 66% changed. Cohort 2 participants who changed were 2.65 times more likely to change this wrong answer to

some other answer on Survey 2 than change another answer type to this after experiencing the instructions.

Changes in answering by repeating the known information were not significant for Cohort 1 and were significant for Cohort 2 ($X^2(1, N = 538) = 42.10, p < 0.0001$). Approximately 30% of all Cohort 2 responses involved this wrong answer type and of those about 47% changed. Cohort 2 participants who changed were 2.41 times more likely to change this wrong answer to some other answer on Survey 2 than change another answer type to this after experiencing the instructions.

Changes in answering incorrectly in an otherwise not categorized wrong way (Other) were significant for Cohort 1 ($X^2(1, N = 163) = 8.35, p = 0.0039$) and Cohort 2 ($X^2(1, N = 270) = 68.89, p < 0.0001$). Approximately 9% and 15% of all Cohort 1 and 2 responses respectively involved this wrong answer type and of those about 50% and 37% changed respectively. Cohort 1 participants who changed were 2.00 times more likely to change a different answer to this wrong answer on Survey 2. Cohort 2 participants who changed were 11.5 times more likely to change this wrong answer to some other answer on Survey 2 than change another answer type to this after experiencing the instructions.

Changes in answering by stating “I Don’t Know” or “Not Enough Information” incorrectly were not significant for Cohort 1 and were significant for Cohort 2 ($X^2(1, N = 124) = 60.49, p < 0.0001$). Approximately 7% of all Cohort 2 responses involved this wrong answer type and of those about 65% changed. Cohort 2 participants who changed were 15.2 times more likely to change to this wrong answer from some other answer on Survey 2 than change this wrong answer to another answer type after experiencing the instructions.

Repeating the Known information additionally analyzed for changes within cohorts to and from Correct responses. This was done to explore the hypothesis that this sort of wrong answer may function as an alternative topography for stating “I Cannot Know” or “Not Enough Information.” Changes from Correct to Repeating the Knowns and vice versa for Cohort 1 were not significantly different. Cohort 2 relative changes were significantly different ($X^2(1, N = 1083) = 106.54, p < 0.0001$) and largely in the direction of repeating the knowns responses on Survey 1 being changed to correct answers on Survey 2 within this cohort. Approximately 8% of all Cohort 2 responses involved a Repeating the Knowns answer changing to a Correct answer or vice versa. Cohort 2 participants who changed were 14.8 times more likely to change this wrong answer to a correct answer on Survey 2 than change a correct answer to this wrong answer after experiencing the instructions. This specific variant of changing to or from a Repeating the Knowns wrong answer accounted for 26.4% of all response change pairs to or from Repeating the Knowns.

Discussion

Experiment 6 provided insight into three key questions. Does a topographically different but structurally matched questionnaire (Survey 2) evoke similar responses to the original 9 questions (Survey 1)? Does an instructional intervention as simple as telling participants that “I Cannot Know or Not Enough Information may be a valid option” impact KU behavior? And does the common response of Repeating the Knowns function as an alternate topography of ICK responding?

Analysis of correct and incorrect responses under consistent conditions for Cohort 1 highlighted that there was no difference in responding between both surveys. This supports that

Survey 1 and 2 may function similarly. It further suggests that the structure utilized in their design and construction may have additive utility and be used in the future with additional subject and relational terms to reliably similar effect.

Analysis of correct and incorrect responses across the instructional intervention timing with Cohort 2 highlighted that the instruction was likely to have been the key evoking stimulus accounting for the significantly increased accuracy on Survey 2. Within this question there is a secondary question of if participants were able to use the ICK response accurately or if the gains were due to increases in indiscriminated ICK responding. The critical data that suggests that accuracy gains were due to a general increase in ICK responding is the analysis of ICK wrong answer change rates. Cohort 2 participants who changed answer involved incorrect ICK responding were 15.2 times more likely to change any other answer to an incorrect ICK response than change an incorrect ICK to any other answer. In other words, after the instructions participants just responded with ICK much more often in general but did not demonstrate any ability to apply the ICK response only when appropriate. Returning to the original question of effects of the instructional intervention, the instructions make ICK responding much more common but even when participants know it may be a correct answer, their correct usage is more by chance than skill. In other words, just being aware of the response option does not seem to be enough for participants to reliably use “I Cannot Know.”

Analysis of Cohort 2 response changes involving Repeating the Knowns illustrated that this is likely an area ripe for additional study. Of all the response types, this was the most common and between Survey 1 and 2 for Cohort 2 the changed pairs involving this response type alone accounted for 14% (252) of all response pairs. If this was an alternate topography of ICK responding, much of the data from previous experiments (including the hopeful results of

Experiment 5) have to be reanalyzed with that in mind. While the general change analysis highlighted that Repeating the Knowns much more frequently changed to a different answer than vice versa, the specific analysis of changes involving correct responses as well showed that about 75% of those changes from Repeating the Known went to Correct ICK responses but those responses only accounted for about 52% of all changed responses involving Repeating the Knowns. The higher percentage continues to give credence to the possibility that some people may be using Repeating the Knowns as an alternative for ICK, but the second percentage suggests that ICK isn't a universal function of this response topography. In other words, for some people, some of the time, a response of Repeating the Knowns may reflect KU functions, but not for everyone all the time.

Instructions increasing the frequency of an undiscriminated ICK response has implications for the current line of research. Experiment 5 made clear that only a minority of participants demonstrate a far transfer of discriminated relational training effects for KU responding. Additional manipulations that increase the frequency of undiscriminated ICK responses may set the stage for more participants exhibiting a discriminated KU response after training. There are two immediately obvious risks to implementing this instructional manipulation. The first is a decrease in sensitivity of the measurement tool. The second is research drift.

Measurement sensitivity in this case is the remaining range on our measurement tool after accounting for the likely response level of participants. Experiment 6 Table 9 Cohort 2 pre instructions (6.2 pt1) indicated that we can expect untrained participants to average 3-4 correct responses on the nine question word problems and nearly zero incorrect ICK responses (i.e., false positive) on a non-KU question. This provides the opportunity for the five remaining

questions to change and detect a training effect as well as two non-KU questions to detect increases in false positive responses. In other words, 55% and 100% of the measurement scales can be expected to be useful for/sensitive to detecting training effects. Introducing the ICK normalizing instruction reduces those ranges.

Experiment 6 Table 9 Cohort 1 first nine questions (6.1 pt1) indicated that untrained participants are likely to average 5-6 correct responses and at least 1 false positive. This reduces the sensitivity of the scales to 33% and 66% respectively. This constraint increases the likelihood that a training effect may not register on the measurement scale because participants benefitting from training may still vary enough within the limited sensitivity range to produce an overlapping confidence interval to untrained participants. This sensitivity risk may be mitigated by also introducing the second version of 3 Term Series problems created for Experiment 6 part 2. All Experiment 6 participants in both cohorts responded to the second version questions after the instructional intervention which provides a significant validation data set for future uses.

Research drift in this case is a question of whether the experiments are moving away from the focus of the research. The central proximal question for this research was whether discriminated relational KU responses can be trained and can that training demonstrate far transfer to questions replicating the naturalistic free response contingencies, so that the question of resistance to bias could then be examined. Note, however, that in naturalistic settings, individuals are not provided a list of response options. Providing a prompt, such as the ICK normalizing instruction, begins to pull the experiments away from a naturalistic model of KU behavior and toward a more contrived learning scenario.

This trajectory is not entirely negative. Through the previous experiments, the low frequency of a discriminated KU relational response in US based participants has been made very clear. Experiment 5 has begun to demonstrate that this response is likely to qualify as relational in nature because it is a discriminated response and requires demonstration of mutual and combinatorial entailment as well as benefits from training on the relational response functions that support the KU function. Additionally, Experiment 6 highlights that the form of saying “I Don’t Know,” “Not Enough Information,” or other similar phrases can indicate a function distinct from seemingly useful alternatives such as repeating the information provided by the question.

While a number of basic questions are still to be addressed, the constraint of replicating naturalistic conditions in this particular way may be reaching a momentarily useful limit. Most participants failed to demonstrate far transfer of KU training. If a less naturalistic instruction can materially change the far transfer outcome, then most participants can realize a more socially significant benefit.

Considering the above discussion, Experiment 7 incorporated the instructional intervention from the beginning and added the second version of 3 Term Series problems to both pre and post tests. In terms of the original proposal, the instructional intervention can likely be considered an ICK evoking stimulus. The standard of Major Improvement described earlier continued to be in effect. The focus was on a clear demonstration of a well discriminated KU response within subjects. Broadly that means increases in overall accuracy, accurate ICK responding, and elimination of false positive ICK responses.

Experiment 7: ICK Evoking Instructions

Method

Experiment 7 extended the final method of Experiment 5 with alterations to the pre and post test 3 Term Series Problems. Those alterations were the inclusion of the ICK normalizing instruction “For some of these problems, "Not enough information" or "I don't know" are the correct answers” prior to the pre and post test 3-Term Series problems and the addition of the nine new 3-Term Series problems from the version 2 created for Experiment 6 (Table 8) to both the pre and post test word problems as questions 10 through 18.

Participants

30 new participants completed the experiment fully. 16 participants’ data were excluded from the analysis due to failing to achieve the mastery criteria across all 10 attempts of both or either relational training MET block (simple or ICK). 4 of these participants failed to achieve mastery during both blocks. 11 failed to achieve mastery during the simple relational training block but passed the ICK block. 1 passed the simple block but failed the ICK block. The remaining 14 participants’ data was analyzed for overall differences in word problem answer accuracy, ICK word problems only answer accuracy, and changes in the frequency of false positive “I Cannot Know” responses to non-ICK word problems.

Results

Changes in accuracy, both overall and ICK only, were analyzed using a matched pairs single tailed t-test with a $p < 0.05$ rejection threshold. Effect sizes were calculated using Cohen’s d . Where Experiments three through five aggregate results were based on nine questions overall,

six ICK questions, and three non-ICK that may result in false positive ICK responding, the addition of the second version of the 3-Term Series questions doubled all of those values respectively. Improvements in overall ($M_d=0.14$, $p=0.05218$, $d=0.61$) accuracy were not significant at the planned p value. Improvements on ICK only ($M_d=0.24$, $p=0.04043$, $d=0.81$) accuracies were significant with large effect sizes. The frequency of false positive ICK responses increased pre to post with 12 incorrect ICK responses across 6 participants during pre-test to 30 incorrect ICK responses across 7 participants during post-test.

Discussion

Two empirical areas suggested that something beyond the experiment may have occurred to systematically shift the data so drastically compared to Experiment 5: a much higher rate of failure of one or more MET blocks and the similarly increased average time to complete. While these two are logically correlated, the only material change between this experiment and Experiment 5 was the addition of 18 more word problems (9 pre & 9 post). Completion of the same 18-word problems in Experiment 6.1 and 6.2 averaged 22 minutes so it was hypothesized that the current Experiment's overall average time to completion would be around the sum of Experiment 5 (58 min) and 6 (22 min) average times. In other words, it was expected that these participants would average roughly 80 minutes to complete all word problems and training blocks. This was not the case. Even excluding those who failed one or more training blocks, the average time of the 14 participants who achieved mastery on both training blocks was 116 minutes or nearly 50% longer than expected. If this were attributable to the instructional change or the additional questions, the time difference would have been expected to also show up in Experiment 6.1 and 6.2 where both elements were present.

Considering the above, it is hard to draw conclusions from the results of Experiment 7 as compared to the previous experiments. Additionally, the significance of the difference in ICK accuracy fell far short of the pre-planned major improvement threshold. Doubling the power of the experiment by increasing the number of ICK questions per participant from six to twelve may have increased sensitivity to undiscriminated increases in ICK responding. Considered in the light of the major increase in incorrect ICK responses, the simplest conclusion is that participants increased undiscriminated ICK responses much like in Pilot 3 and Experiment 4. This also aligns with conclusions from Experiment 6 where the instructional intervention increased undiscriminated ICK responding. Overall, the results of Experiment 7 did not provide clear evidence of a training effect or discriminated KU responding.

Investigation into influences beyond the experiment revealed concerning developments with the M-Turk platform. One of the benefits of M-Turk is access to a global participant pool. For researchers who only want participants from a specific geographic region or language fluency, there are optional filters provided for both. These filters are not foolproof. Most researchers only request North American English speakers. As a participant outside North America, if you have some English language skill, there is a strong incentive to get around the geographic restrictions. This has been a known issue for quite some time and various methods have been deployed by researchers to detect bad actors and prevent their responses from contaminating research data. For the experiments so far discussed here, a combination of other available filter options were employed as well as the geographic filter in order to minimize this influence. When the Experiment 7 data suggested that those efforts may have been completely subverted, more organized efforts were discovered. Specifically, in M-Turk user groups across the internet, North American English speaking participant account holders were being offered

money for their account login credentials. The offers were very frequent, sometimes daily, and coming from accounts seemingly originating from countries that were former UK colonies where English is a common second language and exchange rates amplify the monetary incentives. From what could be gathered, it appears that hundreds, maybe thousands, of account credentials were being “rented” such that a small organization can farm revenue off cycling through these accounts systematically and providing minimal effort responses that get funded. If such farming efforts are distributed across enough accounts and enough different M-Turk research projects, the low effort responses are mostly treated as a nuisance.

If such farming efforts ended up concentrated on a single experiment, frequent artifacts due to limited English fluency would be expected in the data. This is a plausible explanation for the extra time to completion and high frequency of failure on MET blocks in Experiment 7. It is an unfortunate issue because it also suggests that further use of the M-Turk platform is not advisable. If no additional bad actor detection methods are implemented, continued use of the platform risks extra-experimentally contaminated data. Alternatively, developing and deploying detection methods is costly, distracts from the research effort, and introduces additional sources of potential confounding. One way to test if Experiment 7 data was a result of farming efforts is to repeat the experiment using an alternative platform that has already implemented more advanced bad actor filtering methods. If the data matches Experiment 7, bad actors and farming could be reasonably ruled out and between experiment conclusions may be more valid. Alternatively, if the data align more closely with expectations established by Experiments 5 and 6, the organized farming hypothesis is more supported and Experiment 7 data may be reasonably set aside.

Experiment 8: ICK Evoking Instructions Revisited

Method

Experiment 8 repeated the conditions of Experiment 7 on the Prolific.co platform in order to avoid the possible problems that we encountered on M-Turk. In order to avoid accidentally recruiting individuals who may have previously participated via Amazon Mechanical Turk, eligibility included the criteria that the participants reported not working for any other online work platform.

Participants

28 new participants completed the Experiment fully. 5 participants' data were excluded from the analysis due to failing to achieve the mastery criteria across all 10 attempts of both or either relational training MET block (simple or ICK). 2 failed to achieve mastery during the simple relational training block but passed the ICK block. 3 passed the simple block but failed the ICK block. The remaining 23 participants' data was analyzed for overall differences in word problem answer accuracy, ICK word problems only answer accuracy, and changes in the frequency of false positive "I Cannot Know" responses to non-ICK word problems.

Results

Changes in accuracy, both overall and ICK only, were analyzed using a matched pairs single tailed t-test with a $p < 0.05$ rejection threshold. Effect sizes were calculated using Cohen's d . Improvements in overall ($M_d = 0.01$, $p = 0.40852$, $d = 0.05$) accuracy were not significant at the planned p value. Improvements on ICK only ($M_d = 0.03$, $p = 0.35061$, $d = 0.09$) accuracies were not significant at the planned p value. The frequency of false positive ICK responses decreased pre

to post with 15 incorrect ICK responses across 6 participants during pre-test to 13 incorrect ICK responses across 4 participants during post-test but this difference was not statistically significant.

It should be noted in passing that Experiment 8 had the lowest ratio of attempts to completions of all the MET experiments so far with 77 starts and 28 completions. In other words, slightly more than one participant in every three starts finished the experimental protocol. Because there was no formal method of capturing feedback from abandoned participants, abandonment may not be a good proxy of how challenging the protocol was (digital workers may abandon a task merely because the pay does not seem adequate, for example) but it does seem worth mentioning. Some of these participants that did not complete the experiment reached out via email to report technical difficulties or frustrations and so anecdotal evidence suggests that at least for a subset of abandonments, aversiveness of the protocol contributed to non-completion.

Discussion

Unlike Experiment 7, mean time to completion and mean attempts required to pass each block (60 minutes and 4 for both blocks respectively) aligned with extrapolations from the Experiment 5 and 6 data. This provided evidence that the unexpected data from Experiment 7 data was likely due to the inclusion of a different population or otherwise impacted by extra-experimental conditions that undermined the consistency of the M-Turk platform. In that basis Experiment 7 should be disregarded for the purpose of this investigation.

Experiment 7 was an attempt to merge the increase in undiscriminated ICK responding observed in Experiment 6 and the apparent discriminated ICK training effect observed in

Experiment 5 to more reliably produce a discriminated response in participants. Planned analysis of results from this experiment suggests that this attempt failed. One possible explanation for a failure to detect a training effect is the increase in indiscriminated ICK responding. Prior to Experiment 6, it was very rare for a participant to respond ICK during pre-test and those participants that demonstrated a training effect tended to shift from getting all KU questions wrong to all, or nearly all, correct. With KU questions making up 67% of the 3-Term Series questions, even when only a few participants showed a training effect, it surfaced as statistically significant. Once the instructional intervention was introduced in this experiment, participants were much more likely to respond ICK to any question which subsequently decreased the pre-post contrast for those participants demonstrating a training effect.

An alternative analysis that may be more sensitive to change in this lower contrast condition is the usage of an agreement calculation that allows for chance and inaccurate changes in response. Cohen's Kappa (Cohen, 1960) is commonly used in behavior analytic methods when multiple raters are coding behavior to quantify to what degree their ratings match each other. Kappa is calculated on an agree/disagree binary scoring transformation that accommodates various base rates of occurrence for the underlying outcome being coded as well as the chance occurrence of agreements and disagreements between raters. Kappa values are constrained to a negative one to positive one range with one representing perfect agreement and negative one perfect disagreement. In this alternative application of Kappa, the pre and posttest responses can be treated as separate coders and agreement can be calculated across time points. The assumption here would be that an individual unaffected by training would not change their responses beyond chance variation, which would translate into a near perfect agreement value. As training impacts responding, changes beyond chance variation would reduce the agreement value. A hypothetical

alternative outcome that could produce the same low agreement would be all random variation between pre and post. In this case, agreement would be low regardless of the presence of an intervention. While this is a possible outcome, it would violate both theory and previously observed data. The theoretical violation would be that of the assumption of coherence as a learned generalized reinforcer. This specific position within RFT claims that an individual's history of regular reinforcement for behavior consistent with their past behavior (coherence) results in continued coherence taking on reinforcement properties itself (Hayes et al., 2001, pp. 70 & 199; Healy et al., 2000). In other words, if an individual is reinforced for being predictable consistently enough, predictability becomes reinforcing on its own. If this holds true in the current case of these participants, then barring stronger stimulus influence, once an individual has responded a certain way once (e.g., on a pre-test), they are predisposed to repeat that behavior given same or similar conditions (e.g., on a post-test). The previously observed from Experiment 6 data somewhat supports this assumption of consistency as well. Participants in the 6.1 and 6.2 cohorts (n=200 & 204 respectively) were internally consistent to structurally matched questions to a group $k=0.67$ and 0.73 respectively. The Experiment 6 data is limited because the comparison is between topologically different but structurally matched questions such that the participants' antecedent stimulation is not identical but the structural matching still provides some insight into likely values of participant behavioral variation as imputed by Kappa. One specific case where such randomness may occur would be a failure to discriminate between all of the questions at all. Under these conditions, one may respond with some degree of randomness, but for both of the above reasons it is more likely that they will respond the same way to many of the questions (coherence) or the same to some (limited discrimination).

For this analysis, responses are categorized as correct or incorrect in both pre and post and a 2X2 (Pre Correct, Pre Incorrect X Post Correct, Post Incorrect) contingency table is generated representing counts of questions that were answered correctly both before and after training, answered correctly before and incorrectly after, incorrectly before and correctly after, and incorrectly both before and after. Individual cell values as well as row and column sums provide the input to the Kappa calculation and the agreement value is output. For the 23 included participants of Experiment 8, their pre-post agreement Kappa value was $k=0.59$. This value is considered moderate agreement and suggests that post training responding did not significantly exceed chance variability and does not reflect a training effect.

Post-hoc Kappa analysis of Pilot 3 through Experiment 5 data ($k=0.45, 0.36, \& 0.64$ respectively) highlights that this was the second to least behavioral variability detected within this line of investigation. The previously observed low rate of ICK responding in both pre and post questionnaires would translate to higher individual agreement amongst participants who showed no training effect which would raise the group computed value. With the addition of the instructional intervention, any effect evoked by the training is pitted against a normally low probability response that has been temporarily brought to higher probability. Under those conditions, it is reasonable to hypothesize that any possible effect due to training failed to bring the target discriminated accurate response to high enough probability to overcome the temporarily higher probability response due to the instructions in the posttest questions.

Experiment 4-8 Overall Discussion

This proposal suggested that additional training, mastery criteria, simple relational training, ICK evoking, non-arbitrary relational training, formal changes to the training trials, and

informative additions to training trials could each provide incremental improvements to a training effectiveness for building accurate and discriminate KU relational repertoires. Over these five experiments, the first four changes have been explored to mixed results. As of Experiment 5, discriminated KU responding can be trained with multiple exemplar methods via a sequence of basic relational training followed by KU training with repetition contingencies requiring mastery of one type of relation before moving on to the next. Experiment 6 indicated that a very simple instructional intervention that normalizes the usage of ICK responses can evoke an increased frequency of ICK responses but those responses are undiscriminated and frequently wrong. Additionally, Experiment 6 indicated that a common form of wrong answer, repeating the known information of the question, may function distinctly from KU relational responding for most people. Experiment 7 attempted to leverage the evoked response of the instructional intervention to amplify the training effect observed in Experiment 5. While the data may support realizing the evoked response but not the training effect, there is reason to doubt any conclusions from Experiment 7 due to possible confounding by population changes in M-Turk. Experiment 8 repeated the conditions of Experiment 7 using an alternative online work platform and confirmed that confounding hypothesis as well as highlighted that any training effect identified in Experiment 5 is not strong enough to reliably overcome the evocative effects of the instructional intervention demonstrated in Experiment 6.

Next Steps

Before reviewing each of the proposed steps, it may be worthwhile to circle back and review the basic premise for this line of investigation as well as the larger experiment that these

incremental steps are intended to inform. The premise for this line of research is that Known-Unknown (KU) responding is a trainable relational behavior exhibiting unique characteristics that may provide insight into broader human phenomena including biased responding. The general goal of this investigation is (1) to identify the conditions under which KU behavior can be brought to reliable and accurate occurrence within individuals (2) validate if such behavior generalizes into performance on logical tests and (3) experimentally demonstrate how likely sources of bias interact systematically with KU repertoires.

The final biasing experiment that was originally designed to address all three goals and was made up of four distinct elements: pre-test of novel stimuli, training of a robust KU repertoire, systematic biasing of that KU repertoire while training relational networks made up of the pre-test stimuli, and then testing the relational networks just trained for predicted biased outcomes. [Figure 18: Biasing Experiment Flow Chart]

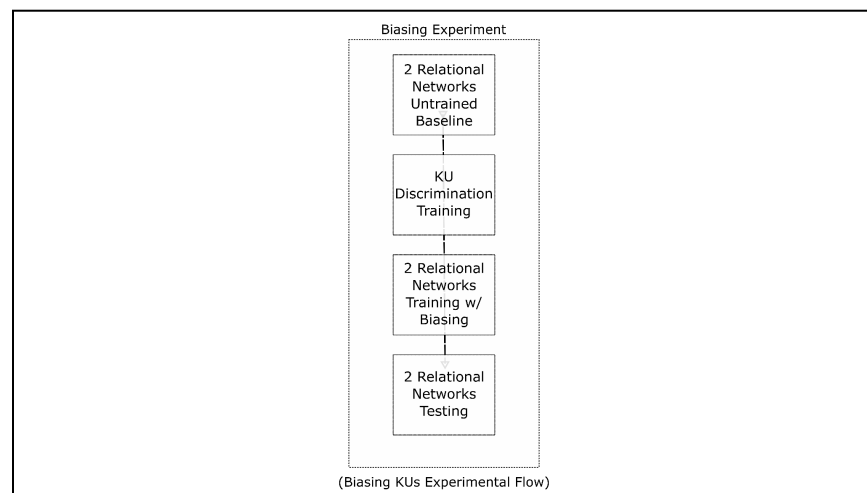


Figure 18:

Flow chart of the original bias experiment design simplified to the four main elements. (Top) Pre testing 2 relational trigram networks as a baseline measure of responding to novel stimuli. (Second from Top) Training a reliable KU repertoire in participants. (Second from bottom) Training the previously tested 2 relational trigram networks in a manner that should systematically bias KU responding. (Bottom) Post test of those 2 now trained relational networks.

As mentioned in the proposal, prior to Pilot 1, this experiment was built and conducted and the results emphasized that the training method for KU responding was unreliable. That unreliability meant that all data from the biasing and post-test blocks was unreliable if it was predicated on consistent outcomes from the KU block. This precipitated the current line of investigation under the general direction that the larger experiment could not be rerun until we were confident that the KU training block was establishing a reliable relational response exhibiting KU characteristics.

In the studies conducted I have generated a multiple exemplar training (MET) and tests for near (entailment probes) and far transfer (3-Term Series word problems) demonstrations of KU responding. The training methods deployed produced measurable changes on near and far tests but it was not clear that the observed behavior was relational in nature nor that the effect was reliable enough to fold back into the original biasing experiment.

While the committee approved the strategy of seven experimental changes to the MET protocol meant to potentially enable the biasing experiment, the unexpected results of the experiments done argue against continuing with all seven studies. The planned seven steps, in original order of proposal, were (1) extending the current training, (2) adding less complex arbitrary relational training, (3) integrating response cost contingencies, (4) making previous correct responses available immediately, (5) changing the form of training to known word problems, (6) introducing ICK evoking scenarios, and (7) introducing non-arbitrary

characteristics. At this point in time, steps 1, 2, and 6 have been implemented and additional experiments have been added to address emerging findings. The originally planned steps 3, 4, 5, and 7 have not been conducted because the results of the present series of studies arguably do not support them. In this section, I'll walk through each step, briefly summarize what was done and what was found, and how the results reflect on that change or argue against a step that was not implemented. I will begin with the conditions implemented: (1) extending the current training, (2) adding less complex arbitrary relational training, and (6) introducing ICK evoking scenarios,

Extending Training

Extending training was accomplished by adding a mastery criteria of 9 out of 12 trials correct on no feedback probe trials during the MET. Participants who failed to meet this criteria were cycled back to the beginning of the training block and experienced novel trigram sets under the same training conditions until they achieved mastery or had failed 10 attempts at training. This was implemented in Experiment 4 and results supported their efficacy with large effect sizes for increased accuracy on both overall and KU specific questions. While these results were promising, the frequency of incorrect ICK responses after training was unacceptably high suggesting that the extension of training increased the base rate frequency of an undiscriminated ICK response without improving discrimination of a KU relational response.

Less Complex Arbitrary Training

The main thesis of this line of investigation is that KU responses are a product of relational repertoires which implies that an individual with a weak basic arbitrary relational repertoire will be less likely to benefit from training of KU relations. Less complex arbitrary relational MET as a precursor to the KU MET was integrated into the protocol in Experiment 5.

This training block was nearly identical to the KU block but only trained mutual and combinatorial entailment of the relational functions SAME, GREATER THAN, and LESS THAN. Participants in Experiment 5 demonstrated a more discriminated KU response compared to Experiment 4, mostly attributed to the reduction of incorrect ICK responses. This may have been a direct result of the new training block producing competing responses or an indirect result of it scaffolding a more discriminated KU repertoire.

Either way, Experiment 5 stands out as a demonstration that KU relational responding can be trained via MET to an acceptable level on both near (entailment) and far (word problems) transfer tests. While the effect was not reliable enough at this stage to fold into the biasing experiment, this is, to the best of our knowledge, the first known demonstration of KU responding from an arbitrary stimulus MET transferring to known word problem behavior.

ICK Evoking

Throughout early stages of this investigation, one commonality that stood out was the very low rate of participants responding to any question with an ICK like response. In natural context, this limits any individual's chance of ever being reinforced for such a response regardless of accuracy. In this investigation, it was noted that any intervention that may evoke ICK responses may improve the protocol by creating an opportunity to consequence such behavior. A version of this was introduced in Experiment 5 and another in Experiment 6.

Experiment 5 introduced a visual animation of the correct response option button during feedback trials. In specific trials, this had an ICK evoking effect when participants would select "I Cannot Know" and experience "Correct!" feedback during the appropriate moments of the KU

MET block. Because Experiment 5 also introduced the simple relational training MET block as well, it is unclear if the animated button had an additive effect on training outcomes.

Experiment 6 introduced a much less subtle ICK evoking intervention in the form of instructions during the 3-Term Series problems stating that “Unknown or Not Enough Information may be the correct answer for some of these questions.” Compared to participants who did not get this instruction for some of the questions, those that did were much more likely to respond with an ICK response but that response was not reliably accurate. Participants were just as likely to respond with ICK to a question where it was not the correct answer as they were to one where it was.

The use of this ICK evoking instruction in Experiment 6 also allowed exploration of a long standing question about the function of the form of wrong answer categorized as Repeating the Knowns. Repeating the Knowns responses were when a participant restated the premise for any particular 3-Term Series question. For example, when told Joe is taller than Jane and Jane is shorter than John, and asked the relationship between Joe and John, participants would frequently state that Joe and John are both taller than Jane. While technically accurate, this response was coded as wrong since the target response was to identify that not enough information was given to answer the direct relation between Joe and John. Because this topography of response was so frequent relative to other wrong responses, there was a legitimate question of if it may be an alternative topography for a KU functioning response. If this had been the case, data up to Experiment 6 would need to be recategorized and reinterpreted.

By including the ICK evoking instruction in a B-B / A-B design survey, data could be analyzed not only on the evocative effect of the instruction, but also on the function of the

Repeating the Knowns response. If participants in the A-B version reliably changed Repeating the Known responses from A to correct ICK responses in B, that would support the hypothesis that these two topographies had shared function. In actuality, while Repeating the Knowns occurred less frequently during the B condition, the Repeating the Knowns responses during the A condition did not reliably predict correct ICK responses during the B condition on matched questions. With this, the hypothesis that Repeating the Knowns functioned differently from KU could not be rejected.

The additional instruction in Experiment 6 did significantly increase the undiscriminated ICK response and so was incorporated into Experiment 8 as a possible method of enhancing the training effect observed in Experiment 5. The undiscriminated ICK response was again observed in pre-test questions and the improvements in discrimination observed in Experiment 5 were not detected in Experiment 8's posttest questions leaving doubts about the reliability and strength of any training effect detected in Experiment 5.

I will now turn to the conditions not implemented: (3) integrating response cost contingencies, (4) making previous correct responses available immediately, (5) changing the form of training to known word problems, and (7) introducing non-arbitrary characteristics. Before doing so, however, I will briefly detour into a discussion of mastery criteria since that issue will appear in the discussion of these proposed studies.

Mastery and Far Transfer

With respect to MET protocols in general, mastery criteria are the field standard for demonstrating efficacy. An individual who satisfies the mastery criterion of a target behavior is considered to, as the term suggests, mastered the behavior to the planned standard. In the case of

this line of investigation, demonstrating mutual and combinatorial entailment to a high level of reliability in previously untrained pairs of trigrams is accepted as a common demonstration of mastery. This is what was tested during the no feedback probe trials at the end of each training block. In this discussion this is referred to as a near transfer demonstration and is the first level of evidence for claiming a behavior has been trained.

In the context of a concern over relational reasoning, however, it needs to be said that an individual choosing an appropriate relation button in the presence of arbitrary stimuli without feedback is a far cry from translating that behavior to more naturalistic outcomes. Because the originally planned final biasing experiment necessitated confidence in the outcomes of the KU training block, the more naturalistic far transfer test of the 3-Term Series problems was included in this series of studies so as to provide convincing evidence of the claimed effects of KU training. That way, even though far transfer per se was not included in the originally designed final biasing experiment, the methods used to establish KU responding could be vetted for their impact on more commonly encountered or “real world” stimuli and reasoning tasks.

Response Cost Contingencies

Response Cost was originally proposed as a method of making the training contingencies more salient to participants. Salience in this case means that participants are demonstrating consequential stimulus control during the MET blocks via achievement of near transfer mastery criterion in a reasonable amount of time. In the proposal, it was suggested that the appearance of “Correct” or “Wrong” consequential feedback alone may not shape the target behavior in a reasonable amount of time because the experience is relatively ephemeral. To address this, it was

proposed that a constant score counter that increments and decrements along with the feedback could amplify the consequential stimulus control of the training.

This study is not needed because in Experiments 4, 5, and 8, participants were consistently able to achieve the mastery criterion and reached that level of mastery in less than half of the allowed attempts. The distribution of the number of attempts to achieve mastery was consistently right skewed with both median and modal counts lower than the mean count. There is no indication that there was a lack of consequential impact and thus the obtained data does not support implementation of response cost contingencies.

With the primary decision data not supporting implementation, it may still be worth walking through hypothetical outcomes and impacts on far transfer performance of response costs. The assumption that response cost would enhance the efficacy of feedback suggests that participants would complete the MET block faster. That means they would experience fewer training exemplars of both simple and KU relations. Experiment 4 specifically explored the impact of adding exemplars and highlighted the increased impact on both near and far transfer demonstrations. Based on that data, it may be extrapolated that implementing response cost may have a negative impact on far transfer. That is, participants may reach mastery sooner but show even less of a discriminated KU response on the 3-Term Series problems. Such an outcome would be counterproductive toward the long term bias experiment goal.

Previous Correct Responses Available

This step was proposed as a way of reducing possible issues with the ephemeral nature of each trial. The focus of the planned implementation was on near transfer outcomes. The proposed change was to present previous correct answers that were relevant to the current trial

pair somewhere on screen so a participant could refer to those as they work through the current trial [Figure 15: reprinted below].

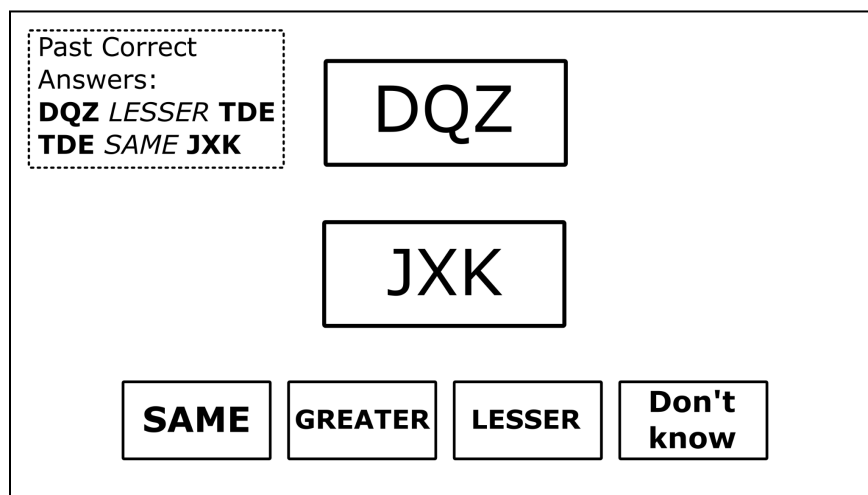


Figure 15:

Proposed trials after correct previous answers list (upper left corner) is introduced.

The primary goal of this change was to reduce the total number of block attempts to reach mastery. Without the primary near transfer data suggesting the need for such a change, this was not implemented. The presence of the previously correct stimuli more closely resembled the 3-Term Series problems which might enhance far transfer but the lack of far transfer was not the barrier preventing the conduct of the final biasing experiment. Said in another way, this planned implementation is interesting but does not address the immediate experimental problem.

Changing to Known Word Problems

In the event that all previous steps failed to produce a discriminated KU relational response, changing the MET protocol to known word problems was proposed as a step toward a more direct replication of the previously published research. It was included as a planned implementation because if all the previously approved steps had been implemented to no avail,

there would be serious doubts about the Vitale et al (2008) Experiment 1 and 2 findings and an attempt at direct replication would have been appropriate. The data in the present investigation did not fail to replicate Vitale et al however. For example, Experiment 5 produced discriminated KU relational responses in some participants, albeit not robustly. In addition to that consideration, the final biasing experiment was based on arbitrary matching to sample, and it would not be possible to implement the change to known word problems there. For that reason, this step was not implemented: it was designed to address a problem that did not emerge and could not be used directly to solve the problems that did.

Non-Arbitrary Training

Non-Arbitrary relational training was proposed as a last step in MET block additions in order to provide a foundation for arbitrary relational training efficacy. For individuals who struggle to learn arbitrary relational responding, trials that use forms such as count, brightness, shape, and size can provide the requisite component skills as a skill foundation for building arbitrary relational skills. In this case, participants would have experienced geometric shapes in both the sample and comparison boxes, instead of trigram pairs, during each trial that varied on either size or number of shapes within each box. Trials would provide feedback for choosing SAME, LESS THAN, or GREATER THAN in coordination with relative counts or sizes of the sample and comparison shape sets.

This modification proved to be unnecessary because it targets a problem we did not encounter. In Experiment 5 mastery of both simple and KU MET blocks was regularly achieved in less than half of the available attempts. As I have already discussed earlier there is no direct non-arbitrary “I don’t know” relation so the non-arbitrary training originally envisioned would

be targeting foundational skills participants already adequately displayed. Thus, the obtained data did not support creation and implementation of a non-arbitrary training block.

There is an alternative hypothetical argument that could be made for introducing this training block related to motivational operations. During all versions of the MET blocks, participants regularly reported frustration working through them and they created many unsuccessful strategies for getting the most correct answers. It is possible that providing a training block that is both easier to complete successfully and is more familiar to participants, may (1) facilitate quick completion of the more unfamiliar MET blocks resulting in the participant being in a favorable motivational state during the post-test and (2) constrain the self-generation of unsuccessful strategies.

There is a set counter argument to this alternative, however. First, implementing a third MET block would have increased the minimum training time required by 50%. For a participant to realize material time savings, they would have to reach mastery on the new block in a single trial and reduce their total trial attempts across all three MET blocks by two. This is not likely. In Experiment 8, the mean combined trials to completion for participants who reached mastery criteria was 6.39 with a median of 4 and mode of 2. This right skewed distribution of attempts to mastery means that most participants are already near the minimum possible completion time and either could not speed up or would have to achieve perfect completion across MET blocks. Thus, even if a marginal gain in acquisition is realized, time to completion will stay the same or increase and if there is a motivational factor tied to time to completion, this change is more likely to harm than help such a factor.

Second, if loss of motivation due to time spent on a task is the issue, time to completion should significantly and negatively correlate with improvements in accuracy on the 3-Term Series problem. Post hoc analysis of the Experiment 8 data shows a weak non-significant correlation between the combined number of attempts to reach mastery across both MET blocks and changes to overall accuracy on 3-Term Series problems. When including all 23 participants, the Pearson's correlation coefficient was $r(21) = -.32, p = .139$. There is a chance that value could have been impacted by the 9 participants who achieved perfect scores on both pre and post word problems resulting in no change to accuracy. When only including those 14 participants who had some opportunity to improve pre to post, the Pearson's correlation coefficient was nearly identical $r(12) = -.37, p = .206$. These data suggest there is essentially no correlation between time to completion and pre-post changes in accuracy on the 3-Term Series word problems.

Finally, there is no data to suggest that non-arbitrary training would reduce the self-generated rules that interfere with timely completion of the MET blocks. Self-generated rules can interfere with relational learning tasks (Hayes et al., 1986; Rosenfarb et al., 1992, 1993) but it is particularly common for participants to resort to using formal properties of stimuli in arbitrary matching to sample tasks. For example, in this series of studies participants commonly described an unsuccessful strategy of trying to assign numerical value to letters in trigrams in order to determine their relative value. Training using non-arbitrary relations seems likely to increase, not decrease, this rule-based imposition of form into arbitrary relational learning tasks.

In summary, there does not seem to be a good theoretical reason to use non-arbitrary training to increase motivation or to decrease self-rule generation, and the data obtained do not

support the use of the originally planned study to help participants master relational learning tasks. This originally planned condition is no longer needed in the project.

General Discussion

While the proposal called into question certain methods of Vitale et al (2008, 2012), the core findings that participants struggle to be accurate on ICK questions and an MET protocol can improve discriminated KU relational responding accuracy have been systematically replicated as well as extended in these current experiments. Known-Unknown relational responding has been demonstrated in some participants. This finding was empirically supported via mastery demonstrations of combinatorial entailment. This form of relational responding was trained using standard multiple exemplar methods with arbitrary trigram stimulus and strengthening a more basic relational repertoire was shown to complement the efficacy of KU training. And finally, improvements in accuracy for responding to 3-Term Series logical problems that are topographically different from the MET protocol suggests some far transfer of these nascent non-KU and KU repertoires to more naturalistic conditions.

All of these findings support the idea that KU responding can be analyzed as an aspect of relational learning. This is important because it means this socially important and frequent form of relational responding is susceptible to experimental analysis using the conceptual and empirical tools of RFT.

That being said, the biasing findings of Quinones and Hayes (2014) could not be systematically replicated in the context of a KU training program because none of the planned implementations were robust enough empirically or conceptually for that purpose. For example, while most participants were able to demonstrate mutual and combinatorial entailment on probes

during MET blocks, far transfer of training the core relational skills was not reliable enough to test for susceptibility to biasing conditions. This point is particularly highlighted in the data from Experiment 8 where a very simple instructional intervention overwhelmed the previously measured training effect.

Until basic questions of how to generate a robust and reliable discriminated KU repertoire are addressed, the compound question of KU biasing will have to wait. Some possible changes seem impractical. For example, the mastery criterion implemented from Experiment 4 onward required participants to achieve at least 9 of 12 probe trials correct in order to move on. This 75% threshold was on the lower end of the field's accepted range and in Quinones & Hayes (2014) -- they employed a varying mastery criteria for different training blocks up to perfect performance (i.e., 100% or 12 out of 12 in this case). An obvious step would be to increase the mastery criterion but had such a higher criterion had been implemented in Experiment 8, 8 (35%) of the remaining 23 included participants would have been required to reattempt one or more training block and at least one would have failed their 10th attempt and been excluded from the final analysis. In light of the attrition mentioned earlier (nearly 2/3 of the participants abandoned Experiment 8), such a change would very likely negatively impact the attempts to completion ratio in a manner that limits the scalability and generalizability of any data that comes from the work. In other words, if the training effect becomes more robust at the cost of materially increasing attrition, any conclusions drawn about how the training works may become unrepresentative.

Combinations of the planned implementations present themselves as possible training avenues. For example, a combination of a known word training that left previous correct training available might be able to establish training effects that could then be transferred to arbitrary

stimuli. Guided practice at the beginning of the MET blocks that walked a participant through the procedure could include instructions such as, “Treat the letter combinations as names of things you’ve never heard of before. Just like how a Giraffe is taller than an Ant and a Crocodile is scarier than a Puppy, in the following sequences apply what you have learned to these arbitrary letter sequences. Note that their appearance or the alphabetic order does not affect the relationship between the things shown.”

Despite the empirical difficulties the present research program seems worthy of further exploration because knowing when you do not know is a socially important form of knowledge in dire need of improvement. The very fact that it is devilishly hard to change in a reliable and robust way underlines why this question needs experimental and conceptual attention. If nothing else the present series of studies documents that the KU problem is a problem that is hard to solve. The data so far supports that MET protocols and simple instructional interventions alter the usage of “I Cannot Know” or “Not Enough Information” in ways that set the stage for more accurate KU responding. Together that means that the tools of behavior analysis are relevant to a major social problem of known importance, but that concentrated work will be needed fully to solve that problem. In the history of behavior analysis, that combination has led to progress when the sustained attention of the field was engaged. It is my hope that these studies will set the stage for such an outcome.

References

- Applications, U. S. C. H. C. on S. and T. S. on S. S. and. (1981). *NASA Program Management and Procurement Procedures and Practices: Hearings Before the Subcommittee on Space Science and Applications of the Committee on Science and Technology, U.S. House of Representatives, Ninety-seventh Congress, First Session, June 24, 25, 1981*. U.S. Government Printing Office.
- Arntzen, E., Grondahl, T., & Eilifsen, C. (2010). The Effects of Different Training Structures in the Establishment of Conditional Discriminations and Subsequent Performance on Tests for Stimulus Equivalence. *The Psychological Record*, *60*(3), 437–461.
<https://doi.org/10.1007/BF03395720>
- Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEntegart, 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. <https://doi.org/10.1016/j.jcbs.2017.08.001>
- Belisle, J., Payne, A. N., & Paliliunas, D. (2022). *A Socio-Behavioral Model of Racism Against the Black Community and Avenues for Anti-Racism Research* [Preprint]. PsyArXiv.
<https://doi.org/10.31234/osf.io/akq6h>
- Ben, J., Cormack, D., Harris, R., & Paradies, Y. (2017). Racism and health service utilisation: A systematic review and meta-analysis. *PLOS ONE*, *12*(12), e0189900.
<https://doi.org/10.1371/journal.pone.0189900>
- Berens, N. M., & Hayes, S. C. (2007). Arbitrarily Applicable Comparative Relations: Experimental Evidence for a Relational Operant. *Journal of Applied Behavior Analysis*, *40*(1), 45–71. <https://doi.org/10.1901/jaba.2007.7-06>
- Brembs, B. (2003). Operant conditioning in invertebrates. *Current Opinion in Neurobiology*,

13(6), 710–717. <https://doi.org/10.1016/j.conb.2003.10.002>

Budziszewska, L., Villarroel Carrasco, J., & Gil, E. (2022). Hierarchical Classification from Relational Frame Theory: A Review. *International Journal of Psychology & Psychological Therapy*, 22(2), 143–162.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>

Delabie, M., Cummins, J., Finn, M., & De Houwer, J. (2022). Differential Crel and Cfunc Acquisition through Stimulus Pairing. *Journal of Contextual Behavioral Science*. <https://doi.org/10.1016/j.jcbs.2022.03.012>

Dixon, M. R., Paliliunas, D., Barron, B. F., Schmick, A. M., & Stanley, C. R. (2021). Randomized Controlled Trial Evaluation of ABA Content on IQ Gains in Children with Autism. *Journal of Behavioral Education*, 30(3), 455–477. <https://doi.org/10.1007/s10864-019-09344-7>

Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior*, 62(3), 331–351. <https://doi.org/10.1901/jeab.1994.62-331>

Dougher, M. J., & Markham, M. R. (1996). Stimulus classes and the untrained acquisition of stimulus functions. In T. R. Zentall & P. M. Smeets (Eds.), *Advances in Psychology* (Vol. 117, pp. 137–152). North-Holland. [https://doi.org/10.1016/S0166-4115\(06\)80107-X](https://doi.org/10.1016/S0166-4115(06)80107-X)

Dougher, M., & Markham, M. (1994). Stimulus equivalence, functional equivalence and the transfer of function. In S. C. Hayes, L. J. Hayes, M. Sato, & K. Ono (Eds.), *Behavior Analysis of Language and Cognition* (pp. 71–90). Context Press.

- Dunning, D. (2011). The Dunning–Kruger Effect. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 247–296). Elsevier.
<https://doi.org/10.1016/B978-0-12-385522-0.00005-6>
- Dymond, S., & Rehfeldt, R. A. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst, 23*(2), 239–254.
- Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144*(3), 674–687. <https://doi.org/10.1037/xge0000070>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Hayes, S. C. (1991). A relational control theory of stimulus equivalence. In L. J. Hayes & P. N. Chase (Eds.), *Dialogues on verbal behavior: The first international institute on verbal relations* (pp. 19–40). Context Press.
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition*. Springer Science & Business Media.
- Hayes, S. C., Brownstein, A. J., Zettle, R. D., Rosenfarb, I., & Korn, Z. (1986). Rule-governed behavior and sensitivity to changing consequences of responding. *Journal of the Experimental Analysis of Behavior, 45*(3), 237–256.
<https://doi.org/10.1901/jeab.1986.45-237>
- Hayes, S. C., Law, S., Assemi, K., Falletta-Cowden, N., Shamblin, M., Burleigh, K., Olla, R., Forman, M., & Smith, P. (2021). Relating is an Operant: A Fly Over of 35 Years of RFT Research. *Perspectivas em Análise do Comportamento*.

<https://doi.org/10.18761/PAC.2021.v12.RFT.02>

Hayes, S. C., & Sanford, B. T. (2014). Cooperation came first: Evolution and human cognition.

Journal of the Experimental Analysis of Behavior, *101*(1), 112–129.

<https://doi.org/10.1002/jeab.64>

Healy, O., Barnes-Holmes, D., & Smeets, P. M. (2000). Derived Relational Responding as

Generalized Operant Behavior. *Journal of the Experimental Analysis of Behavior*, *74*(2),

207–227. <https://doi.org/10.1901/jeab.2000.74-207>

Hornby, A. Sydney., Lea, Diana. ., Bradbery, Jennifer. ., (2020). *Oxford advanced learner's*

dictionary of current English. Oxford University Press; /z-wcorg/.

Howard, S., Kennedy, K., & Tejeda, F. (2020). Social Media Posts About Racism Leads to

Evaluative Backlash for Black Job Applicants. *Social Media + Society*, *6*(4),

2056305120978369. <https://doi.org/10.1177/2056305120978369>

Jelbert, S. A., Taylor, A. H., Cheke, L. G., Clayton, N. S., & Gray, R. D. (2014). Using the

Aesop's Fable Paradigm to Investigate Causal Understanding of Water Displacement by

New Caledonian Crows. *PLOS ONE*, *9*(3), e92895.

<https://doi.org/10.1371/journal.pone.0092895>

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in

Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and*

Biases: The Psychology of Intuitive Judgment (pp. 49–82). Cambridge University Press.

Leighty, K. A., Grand, A. P., Pittman Courte, V. L., Maloney, M. A., & Bettinger, T. L. (2013).

Relational responding by eastern box turtles (*Terrapene carolina*) in a series of color

discrimination tasks. *Journal of Comparative Psychology*, *127*(3), 256–264.

<https://doi.org/10.1037/a0030942>

- Luciano, C., Torneke, N., & Ruiz, F. J. (2021). Clinical behavior analysis and RFT: Conceptualizing psychopathology and its treatment. In M. P. Twohig, M. E. Levin, & J. M. Petersen (Eds.), *The Oxford handbook of acceptance and commitment therapy*. Oxford University Press.
- Luft, J., & Ingham, H. (1955). The Johari window, a graphic model of interpersonal awareness. *Proceedings of the Western Training Laboratory in Group Development*, 246, 2014–2003.
- Martínez-Harms, J., Márquez, N., Menzel, R., & Vorobyev, M. (2014). Visual generalization in honeybees: Evidence of peak shift in color discrimination. *Journal of Comparative Physiology A*, 200(4), 317–325. <https://doi.org/10.1007/s00359-014-0887-1>
- May, R. J., Tyndall, I., McTiernan, A., Roderique-Davies, G., & McLoughlin, S. (2022). The impact of the SMART program on cognitive and academic skills: A systematic review and meta-analysis. *British Journal of Educational Technology*, 00. <https://doi.org/10.1111/bjet.13192>
- McLoughlin, S., Tyndall, I., Pereira, A., & Mulhern, T. (2020). Non-verbal IQ Gains from Relational Operant Training Explain Variance in Educational Attainment: An Active-Controlled Feasibility Study. *Journal of Cognitive Enhancement*. <https://doi.org/10.1007/s41465-020-00187-z>
- Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, 87(1), 1–32. <https://doi.org/10.1016/j.jecp.2003.09.003>
- Moll, F. W., & Nieder, A. (2014). The long and the short of it: Rule-based relative length discrimination in carrion crows, *Corvus corone*. *Behavioural Processes*, 107, 142–149.

<https://doi.org/10.1016/j.beproc.2014.08.009>

Quinones, J. L., & Hayes, S. C. (2014). Relational coherence in ambiguous and unambiguous relational networks. *Journal of the Experimental Analysis of Behavior*, *101*(1), 76–93.

<https://doi.org/10.1002/jeab.67>

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. <https://doi.org/10.1037/0021-843X.102.4.642>

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

<https://doi.org/10.1901/jeab.1992.58-107>

Sangin, M., Molinari, G., Nüssli, M.-A., & Dillenbourg, P. (2011). Facilitating peer knowledge modeling: Effects of a knowledge awareness tool on collaborative learning outcomes and processes. *Computers in Human Behavior*, *27*(3), 1059–1067.

<https://doi.org/10.1016/j.chb.2010.05.032>

Schwartz, M. A. (2008). The importance of stupidity in scientific research. *Journal of Cell Science*, *121*(11), 1771–1771. <https://doi.org/10.1242/jcs.033340>

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. <https://doi.org/10.1901/jeab.1982.37-5>

Skinner, B. F. (1938). *The behavior of organisms; an experimental analysis*,. D.

Appleton-Century Company, Incorporated; /z-wcorg/.

Skinner, B. F. (1965). *Science and human behavior*. Simon and Schuster.

- Smith, P., & Hayes, S. C. (2022). An Open-Source Relational Network Derivation Script in R for Modeling and Visualizing Complex Behavior for Scientists and Practitioners. *Frontiers in Psychology, 13*. <https://www.frontiersin.org/article/10.3389/fpsyg.2022.914485>
- Socrates: I know that I know nothing. (2022). In *Wikipedia*.
https://en.wikipedia.org/w/index.php?title=I_know_that_I_know_nothing&oldid=1070607033
- Stebbing, S. L. (1939). *Thinking to Some Purpose*. Penguin Books.
- Stiernborg, M., Zaldivar, S. B., & Santiago, E. G. (1996). Effect of didactic teaching and experiential learning on nursing students' AIDS-related knowledge and attitudes. *AIDS Care, 8*(5), 601–608. <https://doi.org/10.1080/09540129650125551>
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Vitale, A., Barnes-Holmes, Y., Barnes-Holmes, D., & Campbell, C. (2008). Facilitating Responding in Accordance with the Relational Frame of Comparison: Systematic Empirical Analyses. *The Psychological Record, 58*(3), 365–390.
<http://dx.doi.org/10.1007/BF03395624>
- Vitale, A., Campbell, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2012). Facilitating Responding in Accordance with The Relational Frame of Comparison II: Methodological Analyses. *The Psychological Record, 62*(4), 663–676.
<https://doi.org/10.1007/BF03395827>
- Ward, A. F. (2021). People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences, 118*(43), e2105061118.
<https://doi.org/10.1073/pnas.2105061118>

Webb, J. (1979, November 1). Women Can't Fight. *Washingtonian*.

<https://www.washingtonian.com/1979/11/01/jim-webb-women-cant-fight/>

Wegner, D. M. (1994). *Ironic Processes of Mental Control*. 19.

Wegner, D. M., Schneider, D. J., III, S. R. C., & White, T. L. (1987). Paradoxical Effects of Thought Suppression. *Journal of Personality and Social Psychology*, 53(1), 5–13.

White, A. A., Logghe, H. J., Goodenough, D. A., Barnes, L. L., Hallward, A., Allen, I. M., Green, D. W., Krupat, E., & Llerena-Quinn, R. (2018). Self-Awareness and Cultural Identity as an Effort to Reduce Bias in Medicine. *Journal of Racial and Ethnic Health Disparities*, 5(1), 34–49. <https://doi.org/10.1007/s40615-017-0340-6>

Young, M. E., & Wasserman, E. A. (2001). Entropy and variability discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 278–293. <https://doi.org/10.1037/0278-7393.27.1.278>