The Psychological Record, 2011, 61, 173–198

A RELATIONAL FRAME TRAINING INTERVENTION TO RAISE INTELLIGENCE QUOTIENTS: A PILOT STUDY

Sarah Cassidy and Bryan Roche
National University of Ireland, Maynooth

Steven C. Hayes
University of Nevada, Reno

The current research consisted of 2 studies designed to test the effectiveness of automated multiple-exemplar relational training in raising children's general intellectual skills. In Study 1, 4 participants were exposed to multiple exemplar training in stimulus equivalence and the relational frames of SAME, OPPOSITE, MORE THAN, and LESS THAN across several sessions and weeks. WISC (III-UK) measures were taken at baseline, following stimulus equivalence training, and again following relational frame training. Matched against a no-treatment control group, experimental participants showed significant improvements in full-scale IQ following stimulus equivalence training, and a further significant rise following relational frame training. Study 2 administered an improved multiple-exemplar-based relational frame training intervention to 8 children with a range of educational and behavioral difficulties. In 7 of the 8 cases, full-scale IQ as measured by the WISC (IV-UK) rose by at least 1 SD; the improvement was statistically significant at the group level. These data have important implications for the behavioral analysis of intellectual skills and suggest the basis of an intervention to improve general cognitive functioning.

Key words: Relational Frame Theory, IQ, intelligence, relational training, multiple exemplar training

Measures of intelligence relate strongly to a wide variety of educational (Deary, Strand, Smith, & Fernandes, 2007) and life outcomes (e.g., Schmidt & Hunter, 1998) and tend to be fairly consistent within individuals through the developmental period (Moffitt, Caspi, Harkness, & Silva, 1993). In that context, it is somewhat surprising that there is no well-developed experimental behavior analytic literature showing that behavioral methods can be used to alter these measurable differences in profound ways.

The current research was supported by a Sidney W. and Janet R. Bijou Fellowship award (2006) to the first author from the Society for the Advancement of Behavior Analysis (SABA). The research formed part of the first author's doctoral research program under the supervision of the second author.

Address all correspondence regarding this article to Bryan Roche, Dept. of Psychology, National University of Ireland, Maynooth, Co. Kildare Ireland (e-mail: Bryan.T.Roche@nuim.ie).

Part of the problem may be metatheoretical. Within behavior analysis, the term *intelligence* has often been traditionally treated with caution because it can be used mentalistically, as when behavior is said to be *due to* intelligence (Skinner, 1974). From a behavioral point of view, an interest in intelligent behavior is not controversial, however. In this case, *intelligence* is viewed as a term for a measurable quality of a set of actions.

Behavioral psychologists have been comfortable applying more traditional metrics, such as response fluency (see Binder, 1996) and the learning of three-term contingencies (Williams, Myerson, & Hale, 2008), to an analysis of qualities of complex actions, and interpretive behavioral analyses of intelligence measures have been provided in such terms (e.g., Schlinger, 2003). Unfortunately, these fall short of suggesting immediate research programs that can be employed to improve fluency in the specific behavioral repertoires that relate to measures of intelligent behavior (Williams et al., 2008).

The need for well-crafted learning accounts is underlined by the evidence for modifiability of IQ, such as the fact that traditional forms of schooling have a beneficial impact on intelligence test scores (Ceci, 1991) or that measures of intelligence can be moved by targeting such processes as “working memory” (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). A possible lead worth pursuing is provided by the fact that several recent studies have found that skill in derived relational responding, including stimulus equivalence and derivation of multiple stimulus relations, correlates with performance on intelligence tests. For instance, O’Hora, Pelaez, and Barnes-Holmes (2005) found that participants who successfully completed a complex relational task performed significantly better on the Vocabulary and Arithmetic subtests of the *Wechsler Adult Intelligence Scale* (WAIS-III) as compared to participants who failed to do so. O’Hora et al. (2008) found that accuracy on temporal (before/after) relational responding correlated well with performance on the Block Design subtest of the WAIS-III. Similarly, O’Toole and Barnes-Holmes (2009) found that performance on an Implicit Relational Assessment Procedure (IRAP; D. Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008) designed to assess participants’ fluency in before/after and similar/different relational responding correlated with IQ as measured by the *Kaufman Brief Intelligence Test* (K-BIT).

To apply this knowledge to training, one must be precise about the nature of derived relational responding. Behavior analysts have been interested in derived relational responding since the early work in stimulus equivalence more than 35 years ago (e.g., Sidman, 1971). A relational response is responding to one stimulus in terms of another. In stimulus equivalence, an interlinked set of two trained conditional discriminations (e.g., A → B; B → C) leads to four additional derived (i.e., not specifically trained) relational responses: B → A and C → B (what is termed “symmetry” in the equivalence literature), and A → C (a transitive relation) and C → A (an equivalence relation). As Sidman (2008) noted, however, his approach to stimulus equivalence is “a *limited* theory in that it does not cover other kinds of relations than equivalence, as for example, relational frame theory attempts to do” (p. 331). Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001) expands the analysis of derived relational responses to all types, such as distinction, opposition, comparison, hierarchical, temporal, and the like. This requires an expanded set of terms. In an RFT approach, there are three key features of derived relational responding: *Mutual entailment* refers to situations in which a trained relation between two stimuli (e.g., A →r B, where the →r can refer to *any* type of relation, such as *more than, same as, opposite of*, and so on) leads to a derived mutual relation B →r' A; *combinatorial entailment* refers to the combination of mutually entailed and/or trained relations among three or more stimuli (e.g., A →r B, B →r C entails A →r' C and C →r' A). In RFT these derived relational responses are argued to be controlled by relational contextual cues ("Crel"), not merely the formal properties of the related events, which makes derived relational responses arbitrarily applicable (i.e., they can occur with any set of relata, given the presence of relational cues). The processes giving rise to such arbitrarily applicable derived relational responses are said in RFT initially to be reinforced multiple exemplar training (MET). That is, they are argued to be relational operants. A final defining feature of derived relational responding is the *transformation of function*. That is, given a proper functional context ("Cfunc"), the functions of one relata may alter the functions of others in terms of the derived relation between them. For example if P is the opposite of Q, and Q is a conditioned reinforcer, P may now function as a punisher in a proper relational context that selects consequential functions as relevant (see Barnes, 1994, and D. Barnes-Holmes, Barnes-Holmes, Smeets, Cullinan, & Leader, 2004, for full-length discussions of the differences between stimulus equivalence and RFT).

The possible application of relational responding to intelligence has been outlined in previous theoretical accounts (e.g., D. Barnes-Holmes, Barnes-Holmes, Roche, Healy, et al., 2001; Hayes, Gifford, Townsend, & Barnes-Holmes, 2001; O'Toole, Barnes-Holmes, Murphy, O'Connor, & Barnes-Holmes, 2009) but perhaps most thoroughly in an article by Cassidy, Roche, and O'Hora (2010), which details the variety of derived relations applicable to specific aspects of intellectual behavior. At this point, speculation about the precise relations involved are not as important as realizing the need for breadth and variety. For example, while a few aspects of intellectual behavior (e.g., vocabulary) might be thought of purely in terms of stimulus equivalence, most intellectual skills seem to require fluency in a variety of relational responses. Comprehension of even fairly simple sentences, for instance, requires a number of types of relational responses in order to derive relations between their elements.

Before applying training in derived relational responding to improve IQ, it is important to show that such training is possible. Such data exist. For example, D. Barnes-Holmes, Barnes-Holmes, and Roche (2001; see also a replication by D. Barnes-Holmes, Barnes-Holmes, Roche, & Smeets, 2001) found that explicit multiple-exemplar training was a reliable means by which to facilitate the emergence of generalized mutual entailment where it was found to be absent in a sample of sixteen 4- and 5-year-old children. In a further study, Y. Barnes-Holmes, Barnes-Holmes, Smeets, Strand, and Friman (2004) successfully used interventions suggested by RFT to generate broader repertoires of relational responding, including responding in accordance with MORE THAN and LESS THAN for three children (to avoid confusion about relational terms, from here on they will be in all capitals when we are speaking about relations, if they might otherwise be misunderstood; when using them in normal discourse, they will not be). Y. Barnes-Holmes, Barnes-Holmes, and Smeets (2004; see also an extension by Gomez, Lopez, Martin, Barnes-Holmes, & Barnes-Holmes, 2007) trained children to relate stimuli in accordance with relations of opposition and then to derive novel SAME and OPPOSITE relations across several sets. The relational responding generalized to novel stimulus sets and to a novel experimenter. Berens and Hayes (2007) used multiple-exemplar training to train comparative relations among 4- and 5-year-old children who did not show arbitrary MORE THAN and LESS THAN relations. All participants acquired these relational responses and generalized training to new stimuli and to new relational networks.

In summary, the effectiveness of multiple-exemplar training (training participants in a core relational skill across a very large number of exemplars) in improving relational responding is now relatively well established, but no studies have yet applied these methods to increasing intellectual behavior as measured by classical intelligence tests. Sparse empirical evidence exists as of yet to suggest precisely which families of relational frames are involved, but if an RFT account is correct (Cassidy et al., 2010), these should involve much more than mere equivalence. Thus, the first step appears to be to establish fluency in stimulus equivalence, and then fluency in multiple relational responses, and to see if there is any indication that IQ improves with either approach compared to no relational training. That was done in Study 1 with a small group of normally developing children (*N* = 8), half of whom were randomly assigned to receive multiple-exemplar training in stimulus equivalence, and in SAME and OPPOSITE, and MORE THAN and LESS THAN, relational responding, each across multiple stimulus sets. IQ tests were administered at baseline, following the stimulus equivalence training (approximately after 3 months) and again following the completion of the relational frame multiple-exemplar training phases (approximately after 2 years). The remaining participants received only the usual conditional discrimination training and unreinforced testing to criterion for stimulus equivalence (not formal multiple exemplar stimulus equivalence training) and relational testing. The IQs of participants were measured at baseline and at about 3 months (following the multiple exemplar or non-multiple-exemplar-based stimulus equivalence training, depending on condition assignment). A third IQ measure was taken about 2 years after baseline among both control participants and those receiving full relational training.

Study 1

Method

Participants. Eight normally developing 8- to 12-year-old children free from any clinical diagnosis as assessed by or known to their school in the Republic of Ireland, and not presenting with any scholastic difficulties, were recruited through personal contacts for participation as volunteers in Study 1. Six participants were female, and two were male. The mean age was 10 years 3 months (*SD* = 12.8 months; see Table 1).

Table 1
Participants' Sex and Age at the Beginning of Study 1

Participant	Sex	Age at baseline
1 (Exp)	F	12 years 8 months
2 (Exp)	F	10 years 5 months
3 (Exp)	F	10 years 0 months
4 (Exp)	F	10 years 1 month
5 (Cont)	F	8 years 10 months
6 (Cont)	F	10 years 4 months
7 (Cont)	M	10 years 0 months
8 (Cont)	M	10 years 5 months

Note. Exp = experimental participants; Cont = control participants.

Settings and Materials. Each child was administered the *Wechsler Intelligence Scale for Children* (WISC-IIIUK; Wechsler, 1992), an individually administered clinical instrument for assessing children's intellectual ability. It comprises 13 subtests. Twelve of these subtests were administered in this study (i.e., Picture Completion, Information, Coding, Similarities, Picture Arrangement, Arithmetic, Block Design, Vocabulary, Object Assembly, Comprehension, Digit Span, Symbol Search). Three composite scores for performance intelligence quotient (PIQ), verbal intelligence quotient (VIQ), and full-scale intelligence quotient (FSIQ) can be calculated from these subtest scores. Full-scale IQ was considered to be the primary outcome variable, with the subtests used to refine any results found. The Mazes subtest was not administered because it is a seldom-administered supplementary subtest that does not contribute to any of the composite scores (i.e., PIQ, VIQ, or FSIQ) or other calculable indices (e.g., *Freedom from Distractibility*, *Processing Speed*).

All relational training and testing was conducted on a Macintosh™ iBook laptop computer running the experimental generation software PsyScope (Cohen, McWhinney, Flatt, & Provost, 1993). A total of 36 nonsense syllables were employed as stimuli in the conditional discrimination training, symmetry, and transitivity testing. A total of 60 nonsense syllables were employed during MET for SAME relational responding and a total of 132 nonsense stimuli employed during MET for OPPOSITE relational responding. A further 120 nonsense syllables were employed for establishing the relational frames of MORE THAN and LESS THAN. In all cases, stimuli were three-letter nonwords. A list of all stimuli employed can be found online at http://psychology.nuim.ie/Interventions_to_Raise_IQ.shtml. Four contextual cues were also employed. These consisted of typed character strings (i.e., $$$$$$, !!!!!!, %%%%%%, and ******).

General Experimental Sequence

Baseline WISC scores were calculated for all participants before the relational training and testing phases commenced. There were five relational training phases: (1) stimulus equivalence training and testing, (2) multiple-exemplar training for stimulus equivalence, (3) multiple-exemplar training to establish the relational frame of SAME, (4) multiple-exemplar training to establish the relational frame of OPPOSITE, and (5) multiple-exemplar training to establish the relational frames of MORE THAN and LESS THAN. Experimental participants were exposed to all five phases, whereas control participants were exposed only to Phase 1.

All participants began Phase 1 approximately 2 weeks after baseline WISC scores were taken. The five phases were administered to experimental participants across approximately ten 90-minute sessions spanning 5 to 6 weeks. Control participants completed Phase 1 within one session. In all cases, the second WISC was not administered until 12 weeks had passed since baseline IQ measures were taken (to satisfy recommended minimum test–retest interval criteria). A significant interval of time (approximately 18 months) passed before experimental participants were exposed to Phases 3, 4, and 5 (while the relational training procedures were pilot tested with other participants; see Cassidy, 2008, for full details). WISC scores were again calculated for all 8 participants following Phase 5 (approximately 2 years from baseline testing).

IQ assessments were conducted by the main experimenter, who was a trained psychometrician working for the Irish state within the educational system. Thus, the IQ assessor was not blind to treatment assignment. All training and testing was delivered via the laptop computer in a quiet room in the participant's own home. Each session lasted approximately 90 minutes. Standard instructions were delivered on-screen prior to each phase. A digital audio recording of the instructions being slowly read aloud by the female experimenter was also presented simultaneously by the computer software. Full details are available in Cassidy (2008) or by contacting the authors.

Phase 1: Stimulus Equivalence Training and Testing. All participants were exposed to conditional discrimination training to criterion, followed by testing for symmetrical relations. They were then re-exposed to conditional discrimination training to criterion, followed by testing for derived transitive relations. A standard one-to-many matching-to-sample training protocol was used to train the following conditional stimulus relations: A1 → B1 (not B2), A1 → C1 (not C2), A2 → B2 (not B1), and A2 → C2 (not C1), where alphanumerics represent the nonsense syllables randomly assigned to their roles as sample and comparison stimuli.

Participants used the computer mouse to choose one of two on-screen comparisons on each trial. Blocks of 16 training trials (i.e., 4 exposures to each of the four tasks) were administered until 100% correct responding on a single block was observed. Printed on-screen feedback (i.e., the word *correct* or *wrong*) was presented after all responses. The word *correct* was accompanied by a high-pitched beep, and the word *incorrect* was accompanied by a low-pitched beep.

The symmetry test probed for the following relations using a conditional discrimination format: B1 → A1, C1 → A1, B2 → A2, and C2 → A2, while the transitivity test probed for the following relations, also in the absence of feedback: B1 → C1, C1 → B1, B2 → C2, and C2 → B2 (where alphanumerics refer to nonsense syllables). The number of trials, block length, and criterion for testing phases were the same as for training. There was no re-cycling to training phases, irrespective of performance. The entire phase was conducted using a single stimulus set (i.e., six nonsense syllables).

Phase 2: Multiple exemplar training for stimulus equivalence. Only the experimental participants were then exposed to multiple-exemplar training and testing for stimulus equivalence (i.e., symmetry and transitivity). This MET intervention consisted of providing and withdrawing feedback during alternate probe phases until participants could produce symmetry and transitivity with novel stimuli in the absence of feedback (i.e., until stimulus equivalence performance generalized).

Five additional novel stimulus sets were required for the MET intervention. All experimental participants were exposed to all training and testing stages with these stimulus sets, regardless of when generalization emerged. For each stimulus set, experimental participants were exposed to the training and testing cycle once with feedback during testing phases and once without feedback during testing (i.e., two exposures to the entire equivalence training and testing procedure). If a participant failed to pass a symmetry or transitivity test without feedback on the first block, a new training and testing cycle was initiated with a novel stimulus set. However, where feedback was provided during testing, participants were exposed repeatedly to symmetry and transitivity tests until they reached criterion. The final train/test cycle (Stimulus Set 6) did not involve feedback during testing (i.e., MET).

Phase 3: Multiple exemplar training for SAME. Relational pretraining for SAME and OPPOSITE. To establish the contextual functions of SAME and OPPOSITE for two arbitrary stimuli, experimental participants were exposed to a series of contextually controlled conditional discrimination training and testing blocks (see Steele & Hayes, 1991). More specifically, across several stimulus sets participants were required to discriminate between three comparison stimuli related along a physical continuum, given a non–arbitrarily related sample and in the presence of one of the two contextual cues. On a pretraining trial, the SAME (!!!!!!) or OPPOSITE (%%%%%) cue appeared at the top of the computer screen. One second later, a sample stimulus appeared in the middle of the screen, followed 1 s later by three formally related comparison stimuli. One of the comparisons was always the same as the sample, another was different from the sample, and the third was always opposite to, or “most different” from, the sample. As an example, one set of comparisons consisted of three horizontal lines of varying length. In the presence of the SAME cue, given the short line sample, participants were taught to choose the short line comparison, using on-screen corrective feedback. Similarly, in the presence of the OPPOSITE contextual cue, participants were taught to choose the longest horizontal line comparison given the short line sample. There were four pretraining tasks per stimulus set. In a 16-trial block of pretraining, each task was presented four times in a quasi-random order. Blocks were recycled until a participant produced 100% correct responding. If he or she failed to reach criterion within four blocks, the participant was exposed to a new training block with a novel stimulus set. He or she was again exposed to this training block until criterion on a single block of 16 trials was reached, or until four blocks had been administered.

This process continued, using as many novel stimulus sets as necessary until a participant produced 100% correct responding on the first block of training using a novel stimulus set. Once the training criterion was met, participants were exposed to a test for contextual control by the arbitrary cues. The test consisted of the same procedure as for training, with the difference that no corrective feedback was provided following responses. Novel stimuli were also employed during the test.

Multiple exemplar training. A combination of a Relational Evaluation Procedure (see Cullinan, Barnes-Holmes, & Smeets, 2001) and a Yes-No procedure (see Fields, Adams, Verhave, & Newman, 1990) was employed to train three separate two-stimulus arbitrary SAME relations, leading to the emergence of a four-member relation of coordination during testing (i.e., A SAME as B, B SAME as C, and C SAME as D, in a linear training protocol). More specifically, on a given training trial, the two stimuli from a given stimulus pair (e.g., A and B) were presented on-screen, separated by a contextual cue (i.e., in a sentence format reading from left to right). The words *Yes* and *No* were also presented in counterbalanced positions in the bottom left and right corner of the screen. The participant was required to choose "Yes" or "No" by clicking on the relevant word using the mouse and cursor. Choices were guided by corrective feedback following every response. Participants were also trained to respond to the novel stimulus N1 as *not the SAME as N2*. This control task precluded the possibility of direct control over responding by the contextual cue alone. The following training tasks were employed: A SAME B (*Yes*), B SAME C (*Yes*), C SAME D (*Yes*), and N1 SAME N2 (*No*), where the reinforced response is in parentheses (note that the arbitrary contextual cue, and not the actual words *same* or *opposite,* was presented on-screen). The criterion employed to complete training was 100% correct responding across a block of 20 trials (i.e., five exposures to each of the four tasks presented in a quasirandom order).

Testing consisted of probing for the following relations in the absence of feedback: D SAME A (*Yes*), D OPPOSITE A (*No*), C SAME A (*Yes*), and D OPPOSITE A (*No*), where the predicted response is in parentheses. As before, the criterion for passing was 100% correct responding across a block of 20 trials (i.e., five exposures to each of the four tasks). If an experimental participant failed to meet criterion on the first block of a test, he or she was exposed to another training and testing cycle with a novel stimulus set. On this occasion, feedback was provided during testing, which was administered in repeated blocks until 100% correct responding was observed on a single test block. Upon reaching criterion, the participant was once again exposed to a training and testing cycle in which no feedback was presented during the one and only block of testing. This iterative process continued until a participant could produce 100% correct responding during the first exposure to a test employing a novel stimulus set and without feedback (i.e., until SAME relational responding had generalized).

Phase 4: Multiple-exemplar training for OPPOSITE. Contextual control by the arbitrary SAME and OPPOSITE cues had already been established in Phase 3. Following Phase 3, the experimental participants were exposed to an almost identical procedure for establishing the relational frame of OPPOSITE using entirely novel stimulus sets. The following relations were established: A OPPOSITE B (*Yes*), B OPPOSITE C (*Yes*), C OPPOSITE D (*Yes*), and N1 OPPOSITE N2 *(No).* The test blocks probed for the following relations: D OPPOSITE A (*Yes*), D SAME A *(No),* C SAME A *(Yes)*, and C OPPOSITE A *(No).*

Phase 5: Multiple-exemplar training for MORE THAN and LESS THAN. Relational Pretraining for MORE THAN and LESS THAN. This followed the same format as SAME and OPPOSITE relational pretraining and was administered in a single protocol. The cues employed as MORE THAN and LESS THAN cues were $$$$$ and *****, respectively. As an example of one trial, in the presence of the MORE THAN cue, and when presented with an image of two balls as a sample, the participants were trained to choose an image of three balls, rather than an image of two or one from an array.

MET for MORE THAN and LESS THAN relations. MORE THAN and LESS THAN multiple-exemplar training was conducted in a single protocol. The tasks used during training were as follows: A MORE THAN B (*Yes*), B MORE THAN C (*Yes*), and C MORE THAN D (*Yes*). Three additional tasks employed to preclude control by contextual cues or samples alone were as follows: A LESS THAN B (*No*), N1 MORE THAN N2 (*No*), and N1 LESS THAN N2 (*Yes*). The criterion for passing was 100% correct responding across a block of 30 trials (i.e., five exposures to each of the six tasks, presented in a quasirandom order). The testing tasks for MORE THAN/LESS THAN relational responding were as follows: D MORE THAN A (*No*), D LESS THAN A (*Yes*), C MORE THAN A (*No*), C LESS THAN A (*Yes*), A MORE THAN D (*Yes*), and A LESS THAN D (*No*). Once again the criterion for passing was 100% correct responding across a block of 30 trials (i.e., five exposures to each of the six tasks).

The training and testing cycling procedure was identical to that employed for MET for both SAME and OPPOSITE relations, the difference being that tests with feedback (i.e., MET) were administered only once, rather than to criterion. If an experimental participant failed to reach criterion on the first test block without feedback, he or she was exposed to a new training and testing cycle with a novel stimulus set and further feedback during repeated testing, until the individual could pass a single test block (with feedback) with a novel stimulus set. The participant was then re-exposed to a train and test cycle using novel stimuli and with no feedback during testing, and so on, until he or she could pass a test without feedback on the first block.

Results and Discussion

No participants dropped out of the study, and all measures were available for all participants at each measurement occasion. No baseline differences were found on any measure, and no measure violated normality enough to require adjustment. The numbers of training and testing blocks employed across the study are shown in Table 2. Appendix 1 details the total number of training blocks and training trials administered to each participant.

Outcome Analysis Strategy. The small size of this pilot randomized trial makes interpretation of the results of any statistical analysis inherently tentative. That needs to be kept in mind when examining the results, especially when dealing with statistical significance levels or the absolute magnitude of effect sizes. However, it is the pattern of small, medium, and large effect sizes obtained that is a key focus in the present analysis.

Mixed Model Repeated Measures (MMRM) using an unstructured covariance matrix were used to investigate outcomes (Raudenbush & Bryk, 2001). MMRM has advantages in this analytic situation in its use of all data from all participants to model the underlying covariance structure with fewer restrictive assumptions than analysis of variance models.

Effect sizes for *F* values were based on the method suggested for repeated measures and multilevel designs by Rosenthal and Rosnow (1991; see also Verbeke & Molenberghs, 2000). Effect sizes for MMRM contrasts were derived by dividing test estimates by the square root of the sum of the variance estimates at each relevant time point minus 2 times the covariance estimate between the time points; if the contrasts were at the same time point, the square root of the variance estimate at that time point was used as the denominator. All effect sizes were discussed using the cutoffs suggested by Cohen (1988).

Table 2
Total Number of Blocks of Training, Testing, and Novel Stimulus Sets (Phases 3, 4, and 5) Required by Participants in Study 1

Phase 1

Participant	Equivalence training	Symmetry testing	Equivalence retraining	Transitivity testing
1 (Exp)	3	1	1	1
2 (Exp)	10	1	1	1
3 (Exp)	3	1	2	1
4 (Exp)	3	1	1	1
5 (Cont)	2	1	1	1
6 (Cont)	6	5	1	1
7 (Cont)	3	1	1	1
8 (Cont)	1	1	1	1

Phase 2

Participant	Equivalence training	Symmetry testing	Equivalence retraining	Transitivity testing
1	7	6	6	6
2	27	5	5	5
3	13	12	10	8
4	8	7	6	8

Phases 3 and 4

Participant	Pretraining	SAME novel sets	SAME training	SAME testing	SAME MET	OPPOSITE novel sets	OPPOSITE training	OPPOSITE testing	OPPOSITE MET
1	8	3	6	5	0	7	9	4	5
2	13	1	2	1	0	5	7	3	5
3	16	1	2	1	0	5	8	3	5
4	11	5	10	3	4	7	14	4	7

Phase 5

Participant	Pretraining	MORE/LESS novel sets	MORE/LESS training	MORE/LESS testing	MORE/LESS MET
1	6	5	13	3	2
2	8	3	9	2	1
3	4	3	11	2	1
4	6	3	9	2	1

Note. Exp = experimental participants; Cont = control participants.

No attempt was made to adjust for family-wise alpha. Doing so is controversial for many reasons (e.g., see O'Keefe, 2003; Tutzauer, 2003) but seemed particularly unnecessary in the present case, as there was only one primary analysis (Full Scale IQ), and the pattern of supplementary analyses was more important than any particular test, thus reducing the cost of an individual Type I error. However, the reader wishing to apply a Bonferroni-style adjustment (which treats all comparisons as equally important and adjusts individual tests so that across the entire study there is a .05 probability of finding any significant result, assuming only chance is operating) can do so by interpreting only tests with alpha levels of .003 or less.

IQ Results. Results for Full Scale IQ and the Verbal IQ and Performance IQ subscales are shown graphically in Figure 1 and in more detail in Table 3. Individual performances for Full Scale IQ are shown in Figure 2 (individual data for all measures can be found in Cassidy, 2008).

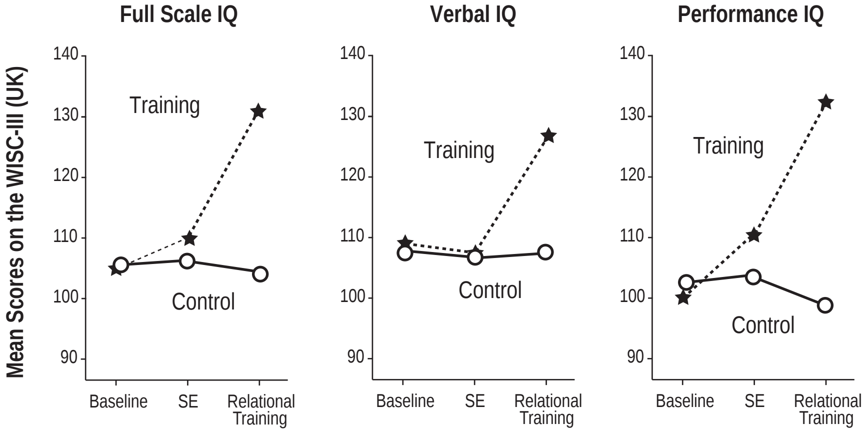

Figure 1. Mean Full Scale, Verbal, and Performance IQs for those in the relational training and control conditions at each measurement period in Study 1.

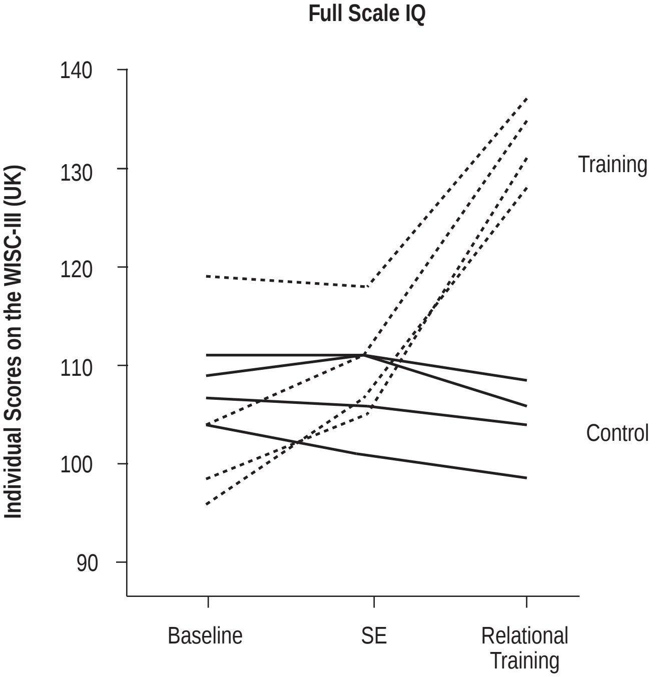

Figure 2. Individual relational training and control participants' Full Scale IQ scores at each measurement period in Study 1.

Table 3
Means, Standard Deviations, and Ranges for Control and Relational Training Participants in Each Phase of Study 1

		Control			Relational training		
		M	*SD*	Range	*M*	*SD*	Range
Full Scale	Baseline	106.50	3.32	104–111	105.50	10.66	96–119
	SE training	107.25	4.79	101–111	110.25	5.74	105–118
	Relational training	104.25	3.86	99–108	132.75	4.03	128–137
Verbal	Baseline	108.25	4.86	101–111	109.25	8.88	101–120
	SE training	107.50	6.66	98–113	107.75	9.03	100–120
	Relational training	108.50	8.85	99–117	127.00	12.99	111–139
Performance	Baseline	102.75	6.13	94–107	100.25	11.24	91–115
	SE training	105.00	7.07	96–113	111.50	3.32	107–115
	Relational training	98.75	8.58	88–109	132.75	2.99	130–137

Full Scale IQ. Results for the MMRM revealed a significant effect for phase, *F*(2, 6) = 67.55, *p* = .000; condition, *F*(1, 6) = 7.16, *p* = .037; and their interaction, *F*(2, 6) = 113.30, *p* = .000, effect size = 8.69 (a large effect). The interaction was explained by the fact that the two conditions did not differ at baseline (*p* = .86, effect size = .13) or after stimulus equivalence training (*p* = .45, effect size = .57, a medium effect), but they did after a more complete set of relational training, Mdiff estimate = –28.5, *SE* = 2.79, *t*(6) = –10.21, *p* = .000, effect size = 7.22 (a large effect). The medium between group effect size difference after stimulus equivalence training emerged because although there was no improvement merely with relational and IQ testing in the control group (Mdiff estimate = .75, *p* = .77, effect size = .15), there was a marginally significant improvement from baseline for those actually receiving stimulus equivalence training (Mdiff estimate = 4.75, *SE* = 2.43, *t*(6) = 1.96, *p* = .098, effect size = .98, a large effect). After full relational training, the control group deteriorated slightly but nonsignificantly with testing alone as compared to baseline levels (Mdiff estimate = –2.25, *p* = .43, effect size = .42, a small effect), while the relational training group increased their scores markedly from baseline levels (Mdiff estimate = 27.25, *SE* = 2.66, *t*(6) = 10.25, *p* = .000, effect size = 5.13, a large effect) and from the levels reached after stimulus equivalence training alone (Mdiff estimate = 22.50, *SE* = 1.21, *t*(6) = 18.63, *p* = .000, effect size = 9.31, a large effect).

Subscale IQ Results. This pattern of results was similar for verbal and performance subscales. On Verbal IQ, results for the MMRM revealed a significant effect for phase, *F*(2, 6) = 13.91, *p* = .006; no significant difference for condition, *F*(1, 6) = 1.32, *p* = .295; and a significant interaction between them, *F*(2, 6) = 11.60, *p* = .009, effect size = 2.78 (a large effect). The interaction was explained by the fact that the two conditions did not differ at baseline (*p* = .85, effect size = .14) or after stimulus equivalence training (*p* = .97, effect size = .03), but there was a marginally significant difference after a more complete set of relational training, Mdiff estimate = –18.5, *SE* = 7.86, *t*(6) = –2.35, *p* = .057, effect size = 1.66 (a large effect). The change from baseline to the final phase in the relational training group was both large and statistically significant, Mdiff estimate = 17.75, *SE* = 3.56, *y*(6) = 4.99, *p* = .002, effect size = 2.11.

The effects on Performance IQ were somewhat more marked. Results for the MMRM revealed a significant effect for phase, *F*(2, 6) = 40.58, *p* = .000; condition, *F*(1, 6) = 12.26, *p* = .037; and their interaction, *F*(2, 6) = 159.46, *p* = .000, effect size = 10.31 (a large effect). The interaction was explained by the fact that the two conditions did not differ at baseline (*p* = .71, effect size = .28) or after stimulus equivalence training (*p* = .147, effect size = 1.18, a large effect), but they did after a more complete set of relational training, Mdiff estimate = –34.0, SE = 4.78, *t*(6) = –7.49, *p* = .000, effect size = 5.29 (a large effect). The large effect size difference after stimulus equivalence training emerged because while there was no improvement merely with relational and IQ testing in the control group (Mdiff estimate = 2.25, *p* = .66, effect size = .24), there was a marginally significant improvement for those actually receiving stimulus equivalence training, Mdiff estimate = 11.25, *SE* = 4.78, *t*(6) = 2.35, *p* = .057, effect size = 1.18 (a large effect). After full relational training the control group deteriorated slightly but nonsignificantly with testing alone at compared to baseline levels (Mdiff estimate = –4.0, *p* = .51, effect size = .53, a medium effect), while the relational training group increased their scores markedly both from baseline levels, Mdiff estimate = 32.5, *SE* = 5.64, *t*(6) = 5.76, *p* = .001, effect size = 2.88 (a large effect) and from the slightly improved levels reached after stimulus equivalence training alone, Mdiff estimate = 21.25, SE = 1.28, *t*(6) = 16.56, *p* = .000, effect size = 8.28 (a large effect).

Another way to consider these changes that is less dependent on inferential parametric statistics is to look at the number of participants who showed reliable changes in IQ as measured by a change of 1 *SD* or more (using standard deviations derived from baseline scores). Table 4 shows the percentage of participants showing deterioration or improvement using that criterion. The conditions did not appear to be different following stimulus equivalence training but showed large differences after full relational training. At the final measurement period, only 1 of 12 scores was significantly improved (changes of 1 *SD* or more) in the control condition, while 2 were deteriorated and 9 unchanged. In the relational training condition, 12 of 12 participants improved at least 1 *SD*, 11 of 12 were improved more than 2 *SD*, and 7 of 12 scores were improved more than 3 *SD*. A Fisher's exact test shows that these differences were all highly statistically significant, regardless of whether changes were considered at 1 (*p* < .0001), 2 (*p* < .0001), or 3 (*p* < .0046) *SD* of improvement. This adds to the earlier analysis by showing that without parametric assumptions, and focusing only on notable within-participant changes, the two conditions differed statistically.

In examining patterns of effect sizes, those in the control condition showed small or nonexistent effect sizes for improvements and deterioration seen throughout. For the experimental participants, stimulus equivalence training alone produced improvements in Performance IQ and Full Scale IQ. Although not significantly different from the control condition in this very small trial, these changes reached marginal levels of significance for within-condition comparisons. In the final phase, the broader relational training protocol produced additional large improvements in all IQ measures. These changes were sufficiently large that they did reach conventional levels of statistical significance within and between groups, even when focusing nonparametrically only on notable levels of change.

Table 4
Percentages of Participants Showing Significant Improvement or Deterioration in Full Scale, Verbal, and Performance IQ in Study 1

	Baseline to post-stimulus equivalence					
	Control participants			Experimental participants		
	Full	Verbal	Performance	Full	Verbal	Performance
1 *SD* Deterioration	0%	25%	0%	25%	0%	0%
1 *SD* Improvement	0%	25%	25%	0%	0%	50%
	Baseline to post-relational frame training					
	Control participants			Experimental participants		
	Full	Verbal	Performance	Full	Verbal	Performance
1 *SD* Deterioration	0%	25%	25%	0%	0%	0%
1 *SD* Improvement	0%	25%	0%	100%	100%	100%
2 *SD* Improvement	0%	0%	0%	100%	75%	100%
3 *SD* Improvement	0%	0%	0%	75%	25%	75%

Despite the randomized nature of Study 1, the very large increases in the final IQ test cannot be laid entirely at the feet of the full relational training. Several months passed after multiple-exemplar stimulus equivalence training and the beginning of full relational training. An additional IQ test was not taken immediately before the final training phase. Thus, the large rise in IQ perhaps was due in part to delayed effects of multiple-exemplar stimulus equivalence training, which may have also made normal educational exposure more potent in the interim. Further research will be needed to untangle these possibilities.

In Study 1, only four participants received full relational training. A larger sample needs to be examined, even in a pilot project such as this one. As an applied matter, it seemed important to replicate these effects using a single coordinated relational curriculum, and focusing on its effects on the IQs of those who struggle with educational achievement. These issues were addressed in Study 2.

Study 2

In Study 2, eight 11- and 12-year-old schoolchildren from a school in the Republic of Ireland who had been identified by their teachers as having educational difficulties were exposed to an intensive battery of SAME, MORE THAN, LESS THAN, and OPPOSITE multiple-exemplar training across 6 to 14 weeks of training (administered across approximately 9 calendar months). IQ was assessed before and after the intervention. A Relational Abilities Index (RAI) was also devised and administered at baseline and following the intervention to ensure that relational skill repertoires were indeed changing as a result of the MET intervention.

The order in which these four relations were trained was altered from that employed in Study 1 because concurrent research conducted by the first author (see Cassidy, 2008) found that a different sequence of relational training was more effective at establishing a full relational repertoire. Relational training phases were also improved by the use of additional control tasks, to further preclude the possibility of extraneous sources of control over responding. In addition, following the suggestion of Berens and Hayes (2007), remedial training protocols were employed to accelerate the generalization of the relational operants when they were slow to emerge. Finally, to better control for the impact of testing and focus retraining on the underlying relational operant, all tests with feedback for SAME, OPPOSITE, MORE THAN, and LESS THAN were administered only once, rather than to criterion, followed by additional training and testing blocks, always with novel stimulus sets. This procedure required that a greater number of stimulus sets be employed during training and testing, thereby ensuring a greater level of generalization of the relational operants than was established in Study 1.

Method

Participants. Three male and five female participants (S9–S16) aged between 11 years 6 months and 12 years 11 months (*M* = 12 years, *SD* = 5.6 months) were identified by a local school principal and resource teacher as experiencing ongoing educational difficulties. Seven of the eight participants (all but S12) had received generic learning support at some time during their academic careers as a result of scoring below the 10th percentile on school-based standardized tests for reading or mathematics. S10, S11, S14, and S15 had recently been, or were currently, under the care of a psychologist due to behavioral difficulties, intellectual difficulties, or both. S14 and S15 also had regular contact with psychiatrists and had both been diagnosed with ADHD. S10, S14, and S15 had been diagnosed with specific learning difficulties in reading, mathematics, or both. S11 had been diagnosed with mild general learning disability as well as mild-moderate expressive and receptive language delay. Thus, the sample had a range of known educational and behavioral problems.

Setting and Materials. The measurement of Full Scale IQ was taken using the WISC-IVUK (Wechsler, 2004). Like its predecessor, this is an individually administered, comprehensive clinical instrument for assessing children's intelligence. It provides composite scores that represent intellectual functioning in specified cognitive domains, (i.e., Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index, and Processing Speed Index), as well as providing a composite score that represents a child's general intellectual ability, or full scale IQ. IQ measures were taken individually in a private room in the school setting. The Relational Abilities Index (RAI) and the MET interventions were administered in a quiet room in which all participants were exposed to the intervention simultaneously in approximately 90-minute sessions. Participants were seated approximately 4 feet apart, each facing an Apple iMac (G3: 300 MHZ) computer with a 15-inch monitor. Each participant wore headphones so that auditory feedback provided by the computer software was not audible by other participants. Participants could not see each other's computer screens and were unaware of the phase to which other participants were being exposed. The same computer software and sets of nonsense syllables as employed in Study 1 were again employed. However, 120 additional nonsense syllable stimuli were required for the RAI. A list of all stimuli employed can be found on-line at http://psychology.nuim.ie/Interventions_to_Raise_IQ.shtml.

General Experimental Sequence

Participants were first administered the WISC-IVUK IQ test to assess baseline levels of IQ. IQ assessments were conducted by the main experimenter, who was a trained psychometrician working for the Irish state within the educational system. Thus, the IQ assessor was not blind to treatment assignment.

The study was conducted in, typically, twice-weekly 90-minute sessions when access to the children was possible (i.e., excluding term breaks, school outings or other activities, family constraints, illness, etc.). Thus, the actual session time required was spread across approximately 9 months for most participants. In a session subsequent to the baseline IQ assessment, participants were administered the specially designed RAI to assess baseline levels of SAME, MORE THAN, LESS THAN, and OPPOSITE relational responding. Participants were then exposed to the necessary pretraining and MET relational training for SAME, MORE THAN, LESS THAN, and OPPOSITE relational frames, in that order. The training and testing cycling procedure and all other features of the MET protocol were identical to that employed in Study 1, except for (a) the addition of two further control tasks during training for SAME and OPPOSITE relations; (b) the presentation of SAME and OPPOSITE testing blocks (without feedback) only once, rather than to criterion; and (c) the use of a remedial training protocol where generalized relational responding was slow to emerge during MET phases. Once all relational responding was at criterion levels, the RAI test and the WISC-IVUK were re-administered as follow-up measures. In all cases, the second WISC was not administered until approximately 9 months had passed since baseline IQ measures were taken.

Standard instructions were delivered onscreen prior to each phase. A digital audio recording of the instructions being slowly read aloud by the female experimenter was also presented simultaneously by the computer software. Full details are available in Cassidy (2008) or by contacting the authors.

Relational Abilities Index. The RAI consisted of three stages of successive blocks of onscreen statements and questions to assess the fluency of SAME, MORE THAN, LESS THAN, and OPPOSITE relational frames. Each relation type was assessed across 60 test trials (20 trials per stage). There were no criteria for passing, and each test block was administered only once, providing a score out of 20 for each of the three stages and an overall composite RAI score out of 60.

On every trial, a statement such as “A is the SAME AS B” was presented onscreen. (The alphanumerics A and B represent nonsense syllables chosen randomly.) One second later, a second statement such as “B is the SAME AS C” appeared on the screen underneath the first statement. After a further 1-s interval, a question such as “Is A OPPOSITE to C?” appeared on the screen underneath the previous two statements, along with the words *Yes* and *No* in the bottom right and left of the computer screen (the positions of which were counterbalanced across trials). No feedback was provided following the participant’s response. None of the stimuli were employed across more than one trial. SAME and OPPOSITE tests blocks were administered separately, while MORE THAN and LESS THAN relational abilities were assessed simultaneously in a single block; the order was SAME, mixed MORE THAN/LESS THAN, OPPOSITE.

In RAI Stage 1 and Stage 2, all stimuli remained onscreen until the participant responded by clicking on *Yes* or *No* using the computer mouse.

Stage 2 replicated Stage 1, but the order of the first two statements was reversed (e.g., in the trial described above, the statement “B is the same as C” was followed by “A is the same as B”). Stage 3 of the RAI was identical to Stage 2 except that it was timed: If the participant failed to respond within 5 s, the nonresponse was recorded as an incorrect response, the screen was cleared, and the next trial was presented.

Multiple-exemplar training for SAME relations. Pretraining for SAME and OPPOSITE contextual control was identical to that employed in Study 1. SAME relational training was also identical to that employed in Study 1, except for the addition of two further control tasks. In effect, six relations in total were established during this phase: A1 SAME B1 (*Yes*), B1 SAME C1 (*Yes*), C1 SAME D1 (*Yes*), A1 OPPOSITE B1 (*No*), N1 SAME N2 (*No*), and N1 OPPOSITE N2 (*Yes*), where the correct response is indicated in parentheses. Participants were required to reach a criterion of 100% correct responding across a block of 30 trials (i.e., five exposures to each of the six tasks). Training was repeated until criterion was met.

The SAME relational testing procedure was identical to that employed in Study 1, with the difference that tests with feedback (i.e., MET) were administered only once, rather than to criterion. Specifically, if a participant failed to reach criterion on the first test block without feedback, he or she was exposed to a new training and testing cycle with a novel stimulus set and further feedback during repeated testing, until the participant could pass a single test block (with feedback) with a novel stimulus set. He or she was then re-exposed to a train and test cycle using novel stimuli and with no feedback during testing, and so on until the individual could pass a test without feedback on the first block.

Remedial training. If participants did not pass the SAME relational testing phase within seven cycles of training, testing, and (interspersed) MET testing (i.e., seven stimulus sets), they were exposed to remedial training and testing. This was identical to standard training except that non–arbitrarily related stimulus sets were employed in the place of the nonsense syllables (e.g., lines, circles, boxes, etc., as employed for relational pretraining but novel in each iteration of the remedial train–test cycle). Once participants reached criterion (i.e., 100% correct performance on a test without feedback and using a novel stimulus set), they were returned to relational training for SAME relations. The iteration between standard and remedial training and testing was cycled until performances reached criterion during relational testing. Remedial training for OPPOSITE, and MORE THAN/LESS THAN relations followed the same strategy.

Multiple-exemplar training for MORE THAN, LESS THAN, and OPPOSITE. MORE THAN/LESS THAN relational pretraining, and relational training and testing, were identical to those employed in Study 1. OPPOSITE relational pretraining had already been administered as part of the pretraining for SAME relations. OPPOSITE relational training was identical to that described in Study 1.

Results

No participants dropped out of the study, and all measures were available for all participants at each measurement occasion. No baseline differences existed on any measure, and no measure violated normality sufficiently to require adjustment. The numbers of training and testing blocks across the study are shown in Table 5. Individual participant data can be found in Cassidy (2008). Appendix 2 details the total number of training blocks and training trials administered to each participant.

Table 5
Total Number of Blocks of Pretraining, Relational Training, Relational Testing, Novel Stimulus Sets, and Remedial Training and Testing Required by Participants in Study 2

Participant	SAME relational pretraining	Novel sets	SAME relational training	SAME relational testing	SAME MET	SAME remedial training	SAME remedial testing
9	5	1	11	1	0	0	0
10	6	3	10	2	1	0	0
11	15	1	18	1	0	0	0
12	13	1	6	1	0	0	0
13	11	3	13	2	1	0	0
14	7	3	39	2	1	0	0
15	12	1	10	1	0	0	0
16	22	1	13	1	0	0	0
Participant	MORE/LESS relational pretraining	Novel sets	MORE/LESS relational training	MORE/LESS relational testing	MORE/ LESS MET	MORE/LESS remedial training	MORE/LESS remedial testing
9	16	3	52	1	0	0	0
10	9	18	43	10	8	8	6
11	8	10	48	6	4	2	2
12	9	14	36	8	6	7	4
13	6	18	36	10	9	8	6
14	10	5	15	3	2	0	0
15	6	7	49	4	3	0	0
16	5	14	41	8	6	10	4
Participant	OPPOSITE relational pretraining	Novel sets	OPPOSITE relational training	OPPOSITE relational testing	OPPOSITE MET	OPPOSITE remedial training	OPPOSITE remedial testing
9	6	1	5	1	0	0	0
10	5	10	24	6	4	4	2
11	5	16	49	9	7	6	5
12	6	10	19	6	4	6	2
13	4	7	12	4	3	0	0
14	9	12	30	7	5	5	3
15	17	14	53	8	6	5	4
16	8	16	52	9	7	8	5

Outcome Analysis Strategy. Given that all data were available, paired associate *t* tests were used to assess whether posttests differed from pretests. However, given the small number of participants, significance levels for *t* tests were determined using bootstrapping. Bootstrapping is a nonparametric procedure in which samples of the same size as the original data set are created, drawing individual scores from the existing data with replacement after each selection. In the present case 1,000 datasets were generated and *t* tests were calculated on each. Ninety-five percent confidence intervals were derived using bias-corrected and accelerated values, which are similar to percentile values for the obtained distribution of test scores but with *z*-score-like corrections. Bootstrapping prevents significant findings from emerging from a few participants in small data sets. Said another way, for bootstrapped values to be significant, very consistent effects need to be seen throughout the data set.

As in Study 1, there was no attempt to adjust for family-wise error, because there are only two primary tests (overall improvement in relational ability and full scale IQ). However, readers wishing apply a Bonferroni-style adjustment to multiple comparisons can do so by interpreting tests with $p < .005$.

Improvements in Relational Ability. Training was recursive and linked to a criterion, and thus the best measure of improvements in relational ability was not the training data but the RAI. At pretesting, the three stages of the RAI correlated significantly with each other (Stage 1 with Stage 2 = .66, Stage 1 with Stage 3 = .67, both *p*s = .05, one tailed), and thus we will address only the RAI values summed across stages of testing here.

The training protocol led to a significant increase in overall relational performance as assessed by the RAI. Means, standard deviations, and ranges at pre- and posttesting are presented in Table 6. At pretesting, participants responded at near chance levels, with a mean of 11.69 correct items per 20-trial block (*SD* = 2.48), or 58.5% correct (50% correct is chance level on this test). After the intervention, participants had 18.47 (*SD* = 1.08) trials correct per 20-trial block, or 92.4% correct. This difference was significant summing across all four relations: Bias-corrected mean difference estimate = 27.22, *SE* = 3.76, $t(7) = 7.14$, $p = .003$, 95% CI: 19.33, 37.26, effect size = 2.52 (a large effect). It was also true for each of the specific relational performances that made up this total score, including SAME, $t(7) = 11.24$, $p = .001$, effect size = 3.97 (a large effect); MORE THAN, $t(7) = 3.01$, $p = .048$, effect size = 1.06 (a large effect); LESS THAN, $t(7) = 3.14$, $p = .034$, effect size = 1.12 (a large effect); and for OPPOSITE, $t(7) = 5.65$, $p = .004$, effect size = 2.00 (a large effect). Focusing on change scores of more or less than 1 (or 2) *SD* (using the standard deviation from the baseline scores), all participants improved at least 1 *SD* and 5 of 8 (62.5%) improved 2 *SD* on the RAI.

Improvements in IQ. Although retesting IQ alone can create improvements (Wechsler, 1992), such effects are not expected when IQ tests are widely spaced. On average, the baseline-to-follow-up IQ tests were separated by 9 months in Study 2. Thus, significant within-condition improvements were taken to reflect possible effects for the relational training protocol and whatever academic experiences the child was being exposed to in school, as opposed to repeated testing.

Individual results for Full Scale IQ are shown graphically in Figure 3 (individual data for all measures can be found in Cassidy, 2008). Means, standard deviations, and ranges for Full Scale IQ and the subscales for pre- and postassessment are shown in Table 6.

Table 6
Means, Standard Deviations, and Ranges in Full Scale IQ, IQ Subtests, and Mean Relational Abilities Indices at Baseline and Following Relational Training in Study 2

	Baseline			Post-Intervention		
	M	*SD*	Range	*M*	*SD*	Range
Full Scale IQ	82.88	8.29	70–92	95.88	10.62	76–111
Verbal Comprehension	82.25	7.32	73–93	92.38	9.20	83–110
Perceptual Reasoning	82.13	10.25	65–96	94.50	6.65	84–106
Working Memory	94.88	16.56	59–116	97.50	12.29	77–116
Processing Speed	91.00	9.84	83–109	107.00	15.64	78–121
Total RAI	46.75	9.90	29–60.7	73.88	4.31	65.3–78
SAME RAI	9.37	2.88	6–15.3	19.46	.47	18.7–20
More RAI	11.50	5.46	2–18.3	17.75	2.73	11.3–19.7
Less RAI	12.58	4.96	6–18.3	17.58	1.75	15–19.7
Opposite RAI	13.29	2.85	9.7–17.3	19.08	1.21	16.3–20

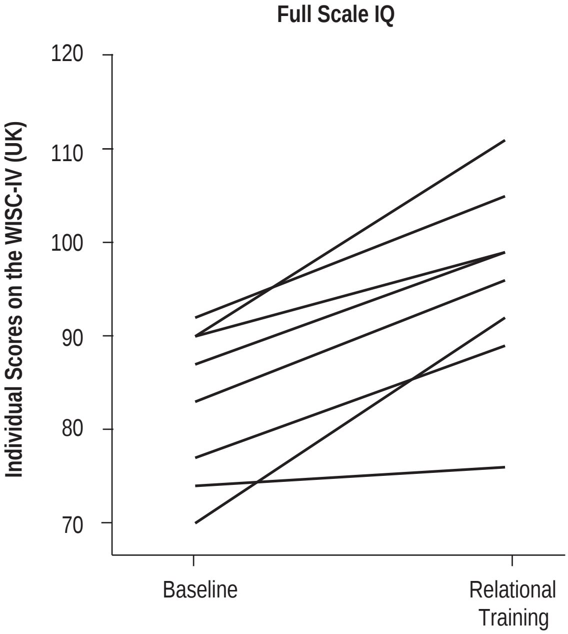

Figure 3. Individual participants' Full Scale IQ scores pre and post relational training in Study 2.

Significant improvements were seen in Full Scale IQ scores, with a bias-corrected mean improvement estimate of 13.10 points, *SE* = 2.15, *t*(7) = 5.77, *p* = .002, 95% CI: 9.88, 16.25, effect size = 2.04 (a large effect). Significant improvements were also seen for three of the four IQ subscales in the WISC-IV(UK): Verbal Comprehension showed a bias-corrected estimated mean improvement of 10.17 points, SE = 1.88, *t*(7) = 4.99, *p* = .006, 95% CI: 6.50, 13.75, effect size = 1.76 (a large effect); Perceptual Reasoning showed a bias-corrected estimated mean improvement of 12.38 points, SE = 3.05, *t*(7) = 3.73, *p* = .006, 95% CI: 7.00, 18.38, effect size = 1.32 (a large effect); and Processing Speed showed a bias-corrected estimated mean improvement of 16.17 points, SE = 3.73, *t*(7) = 3.98, *p* = .008, 95% CI: 7.64, 23.13, effect size = 1.41 (a large effect). There was no significant change in Working Memory, which showed a bias-corrected estimated mean improvement of 2.82 points, *SE* = 6.51, *t*(7) = .39, *p* = .75, 95% CI: –7.50, 14.45, effect size = .14 (no effect).

Focusing on change scores of more or less than 1 (or 2) *SD* (using the standard deviation from the baseline scores), the percentages of participants who improved significantly were as follows: Full Scale IQ: 88% (25%); Verbal Comprehension: 63% (38%); Perceptual Reasoning: 50% (25%); Processing Speed: 75% (38%); and Working Memory: 13% (13%). No participants deteriorated significantly on any measure, and no participants improved 3 *SD* or more except 1 participant (13%) in Processing Speed.

Explaining IQ Improvements. Improvements in Full Scale IQ from baseline to follow-up were not predicted by Full Scale IQ at baseline, *r* = .03, *p* = .94. End-stage relational learning was made more similar across participants by the design of the intervention, since training persisted until criteria were reached, which undermines the usefulness of this metric in a small dataset for attempting to explain IQ changes. The two correlated *r* = .50, *p* = .21, which is a large effect but not significant in this small study. Instead, the fluency of relational learning was examined, defined by the total relational score at follow up (summing tested performances in SAME, MORE, LESS, and OPPOSITE) divided by the number of training blocks required to complete the relational learning program (similar to the concept of "celeration" in a precision teaching approach). Fluency of relational learning did significantly predict changes in Full Scale IQ, *r* = .86, *p* = .006, which is a significantly greater correlation than that of IQ changes with baseline IQ, *z* = –1.99, p = .047 (two tailed). The fluency of relational learning also did not relate to baseline levels of IQ, *r* = .01, *p* = .99. Taken together, this pattern of results suggests that it was those who learned derived relational responding efficiently and effectively who improved in their Full Scale IQ scores.

General Discussion

The current data lend further support to the RFT idea that fluency in derived relational responding is related to intelligence quotients. Across two small studies, this relationship was shown to be likely a functional one, since full relational training appeared to have led to rises in IQ. In Study 1, increases in IQ were observed following extensive stimulus equivalence training and testing, but only if a multiple-exemplar-based training protocol was employed. A subsequent and still larger increase in IQ was observed only for those participants exposed to the multiple-exemplar-based intervention for the relational frames of SAME, OPPOSITE, MORE THAN, and LESS THAN. Of course, the long time delay and sequential nature of the design cannot distinguish delayed effects of multiple-exemplar-based stimulus equivalence training from effects of the full relational training per se. For example, it may be the case that the stimulus equivalence training provided in Study 1 actually facilitated the acquisition of the subsequent relational frames and, thereby, the concurrent large rises in IQ recorded. In effect, given the current experimental design, it is difficult to assess the relative impact of stimulus equivalence and relational frame training on the observed increases in FSIQ.

Study 2 found that the relational training intervention was at least sufficient (if not necessary) to produce significant increases (more than 1 *SD* improvement) in IQ for 7 of the 8 educationally disadvantaged participants. At baseline, Full Scale IQ scores ranged from 70 to 92, with half of the children falling below 85. Following the intervention, scores ranged from 76 to 111, with only one child still below an IQ of 85. All participants also showed notable increases in relational ability as assessed by the RAI. Relational ability, and in particular the fluency of relational learning, was correlated with these rises in IQ. Taken together, these findings provide preliminary evidence that an RFT-based intervention may be effective in raising the fluency of cognitive skills for both normally developing and developmentally delayed populations.

It seems unlikely that the current findings resulted from practice effects across subsequent exposures to the IQ tests. In Study 1, both experimental and control participants were exposed to IQ tests at equal intervals. No practice effect was observed for the control participants in that study. It also seems unlikely that the current IQ increases can be explained in terms of normal educational, maturational, or other developmental processes, both because the size of the increases outstrip previous evidence on such effects and because of the controlled nature of Study 1.

One possibility that cannot be fully ruled out in the area of IQ is testing bias, because the assessor was not blind to treatment assignment, and no reliability checks were taken on the IQ tests employed. While IQ tests are fairly structured and quantitative, there is room for unconscious bias in areas like tone, facial expressions, presentational style, inconsistencies, and scoring errors. It is somewhat reassuring that the RAI was fully computerized and RAI fluency correlated very highly with the IQ changes seen—for bias alone to explain these patterns precisely would be difficult. Nevertheless, due to both the lack of blinding and the very small size of both studies, the present findings should be viewed as tentative and preliminary. Given the findings, we would argue that an RFT-based intervention for intellectual deficit merits a larger randomized trial with educationally challenged and typical children alike, but would not yet conclude that the impact of relational training on IQ is established.

That caution should not take away the possibilities that RFT affords for progress in this area. RFT offers the advantage of a well-considered technical nomenclature for examining the types of skill sets required for good performance on an IQ test (see Cassidy et al., 2010, for extensive examples; see also Y. Barnes-Holmes, Barnes-Holmes, Roche, et al., 2001). Consequently, appropriate MET interventions can be administered to target specific relational deficits. By applying a thoughtful taxonomy of relational skills, the RFT approach should allow researchers and therapists to more accurately identify which aspects of an IQ test pose a problem for a particular person.

Effective use of the RFT approach in applied settings will require further research that will identify the nature and number of multiple exemplars that are needed to establish particular repertoires of relational responding. This research will need to functionally map the development of specific repertoires of relational skills in terms of their impact on specific aspects of cognitive abilities. Such an endeavor would allow behavioral psychologists to speak more directly than ever before to the concept of intelligence as interpreted and measured by widely employed psychometric tests.

Some readers may find remarkable the idea that IQ can be raised substantially with a computer-based relational training intervention. Intelligence is widely viewed as an invariant trait that is normally distributed across the population (but see Gardner, 1993). The evidence suggests otherwise. Raw IQ scores *do* typically rise by a considerable amount across a lifetime, and measurably so from year to year. This effect is called "IQ drift" (Flynn, 1998), and psychometricians compensate for its disruptive effect on the presumed stability and distribution of IQ scores by adjusting for chronological age in calculating IQ scores, and revising IQ tests every decade or so. In addition, it is known that normal educational and programmed interventions can have an impact on IQ (Ceci, 1991; Jaeggi et al., 2008). Thus, the large improvements in raw IQ scores (e.g., 1 *SD* or more) are not completely unexpected, particularly if RFT is correct and IQ tests in part assess core relational skills.

Previous behavior analytic studies have occasionally included IQ test measures as part of interventions for severe disability. For example, Lovaas (1987) reported IQ gains as large as 30 points from the outset of a 3-year intensive applied behavior analysis intervention for autism. However, other authors have raised serious methodological concerns (Connor, 1998; Gresham & MacMillan, 1997; Magiati & Howlin, 2001), and the effectiveness of the Lovaas intervention for improvements in IQ is still in doubt (Reed, Osbourne, & Corness, 2005). Nevertheless, Sallows and Graupner (2005) also recorded significant IQ rises in a recent replication of the Lovaas (1987) study, as did Smith, Eikeseth, Klevstrand, and Lovaas (1997) following an applied behavior analysis intervention to improve expressive speech and adaptive behavior among children with severe mental retardation and autistic features. These studies were concerned with IQ as one part of a larger range of dependent measures in wide–ranging and multifaceted studies, however. In contrast, the current approach directly targeted core relational skills in both typical and educationally disadvantaged children.

Future research in the laboratory should consider the role of the sequence of relational frames trained in interventions such as the current one, and to relate this training to other general factors, such as attentional skills, discrimination abilities, or a host of other educational, social, biological, and psychological variables. Moreover, a wider range of further relational skills sets, such as the frames of hierarchy and deictic (perspective) and temporal (before-after) relations, could have been trained, and should be in future interventions.

There is a great deal yet to be learned about the process and outcome of relational multiple-exemplar training, but enough studies exist to demonstrate that it can develop relational skills and as a result impact important psychological processes (e.g., Y. Barnes-Holmes, Barnes-Holmes, Roche, Smeets, 2001; Y. Barnes-Holmes et al., 2001; Berens & Hayes, 2007; for a book length treatment see Rehfeldt & Barnes-Holmes, 2009). The current findings suggest that such training can foster broadly assessed intellectual skills, but they need to be replicated and extended before that conclusion is firm. Nevertheless the present study at least provides hope that RFT may help behavior analysts to develop practical interventions to increase intellectual skills substantially in both typical and educationally deficient populations.

References

BARNES, D. (1994). Stimulus equivalence and Relational Frame Theory. *The Psychological Record, 44*, 91–124.

BARNES-HOLMES, D., BARNES-HOLMES, Y., SMEETS, P. M., CULLINAN, V., & LEADER, G. (2004). Relational frame theory and stimulus equivalence: Conceptual and procedural issues. *International Journal of Psychology and Psychological Therapy, 4*, 181–214.

BARNES-HOLMES, D., HAYDEN, E., BARNES-HOLMES, Y., & STEWART, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations. *The Psychological Record, 58*, 497–516.

BARNES-HOLMES, Y., BARNES-HOLMES D., & ROCHE, B. (2001). Exemplar training and a derived transformation of function in accordance with symmetry. *The Psychological Record, 51*, 287–308.

BARNES-HOLMES, Y., BARNES-HOLMES, D., ROCHE, B., & SMEETS, P. (2001). Exemplar training and a derived transformation of function in accordance with symmetry II*. The Psychological Record, 51*, 589–603.

BARNES-HOLMES, Y., BARNES-HOLMES, D., ROCHE, B., HEALY, O., LYDDY, F., CULLINAN, V., & HAYES, S. C. (2001). Psychological development. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 157–180). New York: Plenum Press.

BARNES-HOLMES, Y., BARNES-HOLMES, D., SMEETS, P., STRAND, P., & FRIMAN, P. (2004). Establishing relational responding in accordance with more-than and less-than as generalized operant behavior in young children*. International Journal of Psychology and Psychological Therapy, 4*, 531–558.

BARNES-HOLMES, Y., BARNES-HOLMES, D., & SMEETS, P. (2004). Establishing relational responding in accordance with opposite as generalized operant behavior in young children. *International Journal of Psychology and Psychological Therapy, 4*, 559–586.

BERENS, N. M., & HAYES, S. C. (2007). Arbitrarily applicable comparative relations: Experimental evidence for a relational operant. *Journal of Applied Behavior Analysis, 40*, 45–71.

BINDER, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19*, 163–197.

CASSIDY, S. (2008). *Relational Frame Theory and human intelligence: A conceptual and empirical analysis*. Unpublished doctoral thesis, National University of Ireland, Maynooth.

CASSIDY, S., ROCHE, B., & O'HORA, D. (2010). Relational Frame Theory and human intelligence. *European Journal of Behavior Analysis, 11*, 37–51.

CECI, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27*, 703–722.

COHEN, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

COHEN J. D., MACWHINNEY B., FLATT M., & PROVOST J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments and Computers, 25*, 257–271.

CONNOR (1998). A review of behavioural early intervention programmes for children with autism. *Educational Psychology in Practice, 14*, 109–117.

CULLINAN, V., BARNES-HOLMES, D., & SMEETS, P. M. (2001). A precursor to the relational evaluation procedure: The search for the contextual cues that control equivalence responding. *Journal of the Experimental Analysis of Behavior, 76*, 339–349.

DEARY, I. J., STRAND, S., SMITH, P., & FERNANDES, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13–21.

FIELDS, L., ADAMS, B. J., VERHAVE, T., & NEWMAN, S. (1990). The effects of nodality on the formation of equivalence classes. *Journal of the Experimental Analysis of Behavior, 53*, 345–358.

FLYNN, J. R. (1998). WAIS III and WISC III IQ gains in the United States from 1972–1995: How to compensate for obsolete norms. *Perceptual and Motor Skills, 86*, 1231–1239.

GARDNER, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic Books.

GOMEZ, S., LOPEZ, F., MARTIN, C. B., BARNES-HOLMES, Y., & BARNES-HOLMES, D. (2007). Exemplar training and a derived transformation of function in accordance with symmetry and equivalence. *The Psychological Record, 57*, 273–294.

GRESHAM, F. M., & MACMILLAN, D. L. (1997). Autistic recovery? An analysis and critique of the empirical evidence on the Early Intervention Project. *Behavioral Disorders, 22*, 185–201.

HAYES, S. C., BARNES-HOLMES, D., & ROCHE, B. (2001). *Relational Frame Theory: A post-Skinnerian account of human language and cognition*. New York: Plenum.

HAYES, S. C., GIFFORD, E. V., TOWNSEND, R. C., JR., & BARNES-HOLMES, D. (2001). Thinking, problem-solving, and pragmatic verbal analysis. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational Frame Theory: A post-Skinnerian account of human language and cognition* (pp. 87–101). New York: Plenum Press.

JAEGGI, S. M., BUSCHKUEHL, M., JONIDES, J., & PERRIG, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Science, 105*, 6829–6833.

LOVAAS, O. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting Clinical Psychology, 55*, 3–9.

MAGIATI, I. & HOWLIN, P. A. (2001). Monitoring the progress of preschool children with autism enrolled in early intervention programmes: Problems in cognitive assessment. *Autism, 5*, 399–406.

MOFFITT, T. E., CASPI, A., HARKNESS, R., & SILVA, P. A. (1993). The natural history of change in intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry, 34*, 455–506.

O'HORA, D., PELAEZ, M., & BARNES-HOLMES, D. (2005). Derived relational responding and performance on verbal sub-tests of the WAIS-III. *The Psychological Record, 55*, 155–175.

O'HORA, D., PELAEZ, M., BARNES-HOLMES, D., RAE, G., ROBINSON, K., & CHAUDHARY, T. (2008). Temporal relations and intelligence: Correlating relational performance with performance on the WAIS-III. *The Psychological Record, 58*, 569–584.

O'KEEFE, D. J. (2003). Against familywise alpha adjustment. *Human Communication Research, 29*, 431–447.

O'TOOLE, C., & BARNES-HOLMES, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligence test: The importance of relational flexibility. *The Psychological Record, 59*, 119–132.

O'TOOLE, C., BARNES-HOLMES, D., MURPHY, C., O'CONNOR, J., & BARNES-HOLMES, Y. (2009). Relational flexibility and human intelligence: Extending the remit of Skinner's *Verbal Behavior. International Journal of Psychology and Psychological Therapy, 9*, 1–17.

RAUDENBUSH, S., & BRYK, A. (2001). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

REED, P., OSBORNE, L., & CORNESS, M., (2005). *The effectiveness of early intervention programmes for autistic spectrum disorders*. Unpublished report.

REHFELDT, R. A., & BARNES-HOLMES, Y. (Eds.) (2009). *Derived relational responding: Applications for learners with autism and other developmental disabilities*. Oakland, CA: New Harbinger.

ROSENTHAL, R., & ROSNOW, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

SALLOWS, G. O., & GRAUPNER, T. D. (2005). Intensive behavioural treatment for children with autism. *American Journal on Mental Retardation 110*, 417-438.

SCHLINGER, H. D. (2003). The myth of intelligence. *The Psychological Record, 53*, 15–32.

SCHMIDT, F. L., & HUNTER, J. E. (1998). The validity and utility of selection methods in psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

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

SIDMAN, M. (2008). Symmetry and equivalence relations in behavior. *Cognitive Studies, 15,* 322–332.

SKINNER, B. F. (1974). *About behaviorism*. New York: Random House Inc.

STEELE, D. L., & HAYES, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior, 56*, 519–555.

SMITH, T., EIKESETH, S., KLEVSTRAND, M., & LOVAAS, O. (1997). Intensive behavioral treatment for preschoolers with severe mental retardation and pervasive developmental disorder. *American Journal on Mental Retardation, 102*, 238–249.

TUTZAUER, F. (2003). On the sensible application of familywise alpha adjustment. *Human Communication Research, 29*, 455–463.

VERBEKE, G., & MOLENBERGHS, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

WECHSLER, D. (1992). *Wechsler intelligence scale for children–Third edition–UK Manual.* London: Psychological Corp, Europe.

WECHSLER, D. (2004). *WISC-IVUK administration and scoring manual.* London: Harcourt Assessment.

WILLIAMS, B., MYERSON, J., & HALE, S. (2008). Individual differences, intelligence, and behavior analysis. *Journal of the Experimental Analysis of Behavior, 90*, 219–231.