Using computerized procedures
for testing and training abstract comparative relations

Agata Vitale

Health Systems Research Centre
CSI-11
University of Limerick
agata.vitale@ul.ie

Veronica Cullinan

Dept of Psychology
Mary Immaculate College
Limerick
veronica.cullinan@mic.ul.iel

Abstract

Abstract comparative reasoning is involved whenever one arbitrary stimulus (i.e. not defined by its physical properties) is related to another in terms of qualitative or quantitative relations. This kind of reasoning is part of our everyday life, and it is the substratum for other kinds of more complex related reasoning skills, such as hierarchical relations. Previous studies indicated that normally developed adults find difficult to solve some comparative relations, even when these simply involve three elements (e.g. X> U; U<B). The current paper describes the development of automated training procedures for testing and training abstract comparative reasoning. The use of automated procedures was essential to the empirical work reported in this paper. It addressed problems of validity such as the risk that individuals' performances might be affected by the presence of the experimenter. It also improved accuracy in recording individuals' pattern of responses. Furthermore it facilitated the development of more effective training techniques to improve this kind of reasoning. In Experiment 1 participants were exposed to the same comparative relations among three stimuli (e.g. A, B & C) in three subsequent identical phases (Phases1, 2 &3). This was done to verify if they could improve their performance simply by repeated exposures to the same comparative relations. Experiment 2 was identical to Experiment 1, except that in phase 2 individuals were exposed to training (this involved a combination of non arbitrary trials and written feedback). The results indicated that the training was effective, and that individuals in Phase 3 of Experiment 2 performed better than those in Phase 3 of Experiment 1.

Computer programming, more than other fields, requires flexibility of thinking in abstract terms (including abstract comparative reasoning). It is therefore hoped that the automated procedures developed in Experiments 1&2 may provide the basis for similar techniques to test and train comparative relations in individuals who approach this field for the first time (e.g. students).

1. Introduction

Abstract comparative reasoning is involved whenever one *arbitrary* element *(i.e. not defined by its physical properties)* is related to another in terms of qualitative or quantitative relations. This kind of reasoning is part of our everyday life, and provides the basis for other kinds of more complex related reasoning skills, such as hierarchical relations. Previous studies in this field indicated that normally developed adults find it difficult to solve some comparative relations, even when these involve simply three elements (e.g. X> U; U<B), (Vitale et al., 2008; Stenberg 1980; Clark 1994; Clark 1969 a & b). Experimental examination of such relations therefore often involves repeated testing and training sessions on multiple exemplars of the experimental tasks. This can be a laborious and time consuming process using traditional 'paper and pencil' methodologies. The experiments outlined in the current paper therefore explored the use of automated procedures for testing and training different kinds of abstract comparative relations. Given that abstract reasoning plays a large role in the kinds of tasks required of computer programmers, these procedures might be usefully employed to provide testing and training methodologies for disciplines such as programmings.

Abstract comparative reasoning is a specific sub-type of abstract reasoning. Thus, for the purpose of this paper, we will first outline a Behaviour Analytic account of abstract reasoning (including its relevance for computer programmers), followed by an account of abstract comparative reasoning and the rationale for using automated procedures for the empirical work described (i.e. Experiments 1 & 2).

1.1. What is abstract reasoning?

Abstract reasoning refers to the human ability to use symbols instead of concrete objects when processing new information.

A very basic example of abstract reasoning can be the following:

Consider the three shapes below:

'▒', '▓' & ◙

They may meaningless to the reader unless further instructions are given. Thus if one is told that:

◙ is worth more than (>) ▒; & ▒ is worth the same (=) as ▓

→ Then s/he can easily derive that: ◙ is also > than ▓.

The same result can be obtained by using other shapes (e.g. If ▬ > ◙ & ¤ = ◙ → then ▬ >¤.) or by using letters (if A>B; and B=C; → then A>C). In this sense one is able to *abstract the information,* regardless of the manner in which it is presented.

Abstract reasoning involves *arbitrary* stimuli; this means that in deducing specific relations, one does not rely on the physical properties of the material used.

The above example attempts to demonstrate how easily humans are able to think in abstract terms. However, abstract reasoning often involves much more complex tasks such as playing chess or being able to solve a tricky mathematical problem.

1.2. How do we develop abstract reasoning?

Thinking in abstract terms, seems to be a characteristic of the human species. For instance Reese (1968) demonstrated that animals can be trained to select the larger, smaller or dimmer stimulus from an array only when such judgments are based on physical properties of the stimuli (i.e., *non-arbitrary relations)*. Humans however learn how to think in abstract terms during the early stages of childhood development, though experience with concrete objects or aids, allied with reinforcement given by their caregivers (Cullinan &Vitale, 2008; Barnes-Holmes et al., 2001). For instance in a classroom setting, abstract reasoning can be developed in children with the use of materials such as blocks, coloured shapes or beads. A set of blocks may be used to learn both the meaning of numbers and the comparative relations among them (e.g. three blocks are more than two blocks). Furthermore, the teacher typically reinforces correct responding with praise (e.g. 'well done'). Through various examples and reinforcements a child may shift from being able to think in concrete terms to abstract reasoning, and for instance understand the concept of numbers without the presence of physical objects.

1.3. What does this have to do with programming?

Thinking in abstract terms plays a key role in computer science and software engineering (Kramer 2007). Software per se is abstract, and the discipline of producing software requires abstraction skills. Thus for computer programmers having an insight into how humans think in abstract terms, can improve the quality of their products.

Likewise, the prevalent discipline of understanding existing software (Rajlich 1994) during software maintenance requires abstraction skills. For example, Pennington's theories of Bottom-Up comprehension (Pennington1987) suggest that programmers move from the detail of the source code to abstractions based on chunks of text and based on the domain. Likewise Top-Down theories of software comprehension demand that the programmers relate domain abstractions to the detailed source code (Brooks 1983; Soloway 1984) .

From the other perspective, based on their experience, computer programmers are in an ideal position to help psychologists to develop automated training techniques that improve abstract reasoning.

The empirical work reported in the current paper aims to demonstrate how the interaction between computer programming and the behavior analytic approach to the study of abstract reasoning can provide pragmatic solutions to common problems and can be used to improve a particular sub-type of comparative reasoning: abstract comparative reasoning.

1.4. What is abstract comparative reasoning?

This is involved whenever one event is related to another in terms of qualitative or quantitative relations (e.g., A> B; B>C). This kind of reasoning is part of our daily lives. For instance a child may learn that, although it is physically smaller, a one Euro coin is worth more than a fifty cent coin, therefore, given the choice s/he will choose a one Euro coin.

This comparative reasoning is one of the first kinds of reasoning to appear during cognitive development, and it seems to be the substratum for other complex abstract reasoning skills, such as reasoning about hierarchical relations in Object Oriented programming, designing software systems and deriving domain goals from detailed source code.

Thus identifying methods to enhance the flexibility of comparative reasoning may have a positive effect on other kinds of abstract reasoning tasks. Additionally, Vitale et al. (2008) found that normally developed adults often fail to solve tasks involving abstract comparative reasoning even when this involves only three stimuli (e.g. A, B and C). Thus, given the importance that this specific kind of abstract reasoning has in helping building up other kinds of reasoning, it is important to develop methods that can improve it. The experimental work reported in this paper aims to test and train abstract comparative reasoning among three abstract stimuli through the use of computer based procedures.

1.5. Automated procedures: Why do we use them?

Behaviour Analysis is a psychological discipline interested in language and cognition. It adopts a scientific approach to the study of human behavior and experience. This approach emphasizes the importance of consistency and control in the collection of empirical data. However this is not always possible with some standard ‘paper and pencil’ type experimental procedures which often utilize manual measurement and recording of data.

For example, much contemporary research in Behaviour Analysis has examined the development of derived relational responding (such as abstract comparative reasoning) in both adults and children with and without intellectual disabilities. Such studies typically involve the presentation of discrimination or matching tasks whereby subjects are required to look at stimuli in a display and choose or respond to (typically by pointing or touching) one or more of those stimuli. To date many studies of this kind are conducted using what are often described as non-automated or table-top procedures. These procedures require an experimenter to be physically available to present consecutive experimental tasks and to provide feedback and or programmed consequences. A number of problems with such procedures have been identified in the literature (e.g., Peterson et al. 1982; Saunders and Williams 1998; Dymond et al. 2005) The primary difficulty relates to the physical presence of the experimenter and their role in arranging the experimental setting and managing the sessions. Two issues relating to the physical presence of the experimenter are experimenter cueing effects and experimenter drift. Experimenter cueing effects occur when the experimenter inadvertently ‘cues’ the subject to emit the ‘correct’ response. These cues are typically non-verbal such as glancing at the correct choice or indicating by tone or inflection during verbal instructions what the correct response is. According to Saunders and Williams (1998) "the most important concern is that the experimenter might inadvertently prompt or provide feedback to the subject. Even subjects

with extreme developmental limitations bring to the laboratory a long history of following nonverbal prompts. Moreover, it is surprisingly difficult for many experimenters to suppress inadvertent cues, especially premature motions toward delivering consequences" (p.195).
Experimenter/therapist drift occurs when the experimenter gradually changes the contingencies of the experiment by for example becoming lax with timing of administration of consequences or by modifying the experimental procedure in non-programmed ways. Saunders and Williams (1998) pointed out that one of the main difficulties with the use of non-automated procedures is that "immediate decisions as to whether responses meet the experimental contingencies may be difficult. For example, a subject may barely touch one stimulus and then move quickly to another" (p. 195). The experimenter then has to make a judgement as to which response to record. Unless there are very clear guidelines (and ideally training) for making such judgements then there is a clear possibility that experimenter drift will affect the results of the experiment. As long ago as Peterson et al. (1982) suggested that '..correctly calibrated, mechanical methods of delivering or recording the occurrence of the independent variable which are equally effective as using human therapists and observers, might be preferred because they provide greater accuracy at lower cost.' (p.489)
Other more 'mechanical' aspects of the experimental setting and session management that may affect results were identified by Dymond et al. (2005). For example, the construction of task materials; the format of the individual tasks such as positioning of stimuli and distance from subject; the possibility of subjects being able to see the data sheet thereby identifying criteria for 'correct' and 'incorrect' responding randomization of stimulus positioning across trials; randomization of presentation of trials; contiguity; and consistency of inter-trial intervals.
These and many other considerations have led to the gradual shift away from non automated procedures. The increasing availability and ease of use of computer technology has greatly facilitated this process and thereby facilitated tighter control of variables in Behaviour Analysis research.

2. The Experiments

The empirical work reported in the current article concerns the testing and training of comparative relations among three arbitrary stimuli (A, B and C). As indicated at the beginning of this article, arbitrary elements are elements that are not defined by their physical properties; therefore the only way to determine the relations among them is to follow the instructions. In the current article, these relations were presented in a series of problem solving tasks. Individuals were asked to read the instruction presented on the computer screen and work out the comparative relations among three arbitrary stimuli (e.g. $A>B$; $B<C$). As detailed in the individual method sections below all tasks were presented and responses recorded using automated procedures programmed in Visual Basic.
Two key questions were addressed in this work. First, would individuals improve their performance simply by giving them several exposures to the same comparative tasks (i.e. Experiment 1)? Second, if not, would training with a combination of written feedback and non-arbitrary stimuli improve individuals' performance (i.e., Experiment 2)?
Experiments 1 and 2 were conducted separately using two different participant samples.
Both the experiments consisted of three experimental phases. In Experiment 1, individuals were simply exposed to the comparative relations in three consecutive phases. This was done to verify if their performance in solving the comparative relations was enhanced by the repeated exposures. In Experiment 2 individuals were tested in Phase 1, then trained in Phase 2 and tested again in Phase 3. An overview of the two experiments is given in table 1 below.

	Experiment 1	**Experiment 2**
Phase 1	Testing	Testing
Phase 2	Testing	*Training*
Phase 3	Testing	Testing

Table 1. An overview of Experiments 1 & 2.

The two experiments were conducted separately with two different group of participants.

2.2. Experiment 1

Method

Participants. Ten naïve individuals participated to Experiment 1. They were undergraduate students and professionals aged between 18 and 24 years recruited through personal contacts of the Experimenter. All subjects participated individually, and they did not receive any remuneration for their participation in the experiment, and all subjects participated individually.

Setting. Experiment 1 was conducted in a room free of distraction. All individuals' participation was conducted by way of interactions with the computer using the same Visual Basic program. During the automated procedure, the participant remained alone in the room, while the Experimenter was waiting outside.

Each participant was exposed to only one experimental phase per day and received exposure to subsequent phases on subsequent days, availability permitting. Individuals, therefore, received two consecutive test exposures each day across three days. Each exposure to a protocol of 48-trials lasted approximately 30 minutes and subjects were given a short break of approximately 5 minutes before the second exposure. During the short break subjects were asked not to leave the experimental room, if possible. When this break was over, subjects received the second exposure immediately.

Apparatus. The apparatus used in Experiment 1 involved an automated procedure that ran on an Apple iBook laptop computer with a Power PC G3 500 MHz processor, a 12.1 inch LCD screen, and standard computer mouse. All stimulus presentations and participant responses within the automated procedure were recorded by a Visual Basic Program (Version 6) presented through Real PC (Version '95). Each automated trial depicted three identically sized circles (referred to as 'coins') that differed only in terms of color -- one red, one blue, and one yellow. For experimental purposes, the three coins were designated as A (red), B (blue), and C (yellow), although participants never saw these labels. Each trial also depicted three brown rectangles (referred to as 'coffee containers'), that differed only in terms of size -- one large (referred to as 'full of coffee'), one medium ('half full of coffee'), and one small ('a little coffee'). During each trial a small black box was presented on the computer screen, and this was referred as 'I cannot know tin'.

Trial-Types. The automated procedure contained a total of 48 trials, divided according to six trial-types that differed according to three basic dimensions: (1) the *target relations* stated among the coins; (2) whether the relations to be derived among the coins were *specified* or *unspecified*; and (3) whether or not the relations to be derived among the coins were *transitive*. The full list of trial-types and their relevant categorizations are presented in Table 2.

<table>
<tr><th colspan="2">Specified-Same Relations</th><th colspan="2">Unspecified-Same Relations</th></tr>
<tr><td>MORE-MORE
A>B; B>C
B>C; A>B
B>A; C>B
C>B; B>A</td><td>LESS-LESS
A<B; B<C
B<C; A<B
B<A; C<B
C<B; B<A</td><td>MORE-MORE
A>B; C>B
B>A; B>C
B>C; B>A
C>B; A>B</td><td>LESS-LESS
A<B; C<B
B<A; B<C
B<C; B<A
C<B; A<B</td></tr>
<tr><th colspan="2">Specified-Mixed Relations</th><th colspan="2">Unspecified-Mixed Relations</th></tr>
<tr><td>MORE-LESS
A>B; C<B
B>C; B<A
B>A; B<C
C>B; A<B</td><td>LESS-MORE
A<B; C>B
B<C; B>A
B<A; B>C
C<B; A>B</td><td>MORE-LESS
A>B; B<C
B>C; A<B
B>A; C<B
C>B; B<A</td><td>LESS-MORE
A<B; B>C
B<C; A>B
B<A; C>B
C<B; B>A</td></tr>
<tr><th colspan="2">Specified-Same Transitive Relations</th><th colspan="2">Unspecified-Mixed Transitive Relations</th></tr>
<tr><td>MORE -|MORE
A>B; C>A
A>C; B>A
C>B; A>C
C>A; B>C</td><td>LESS -LESS
A<B; C<A
A<C; B<A
C<B; A<C
C<A; B<C</td><td>MORE -|LESS
A>B; C<A
A>C; B<A
C>B; A<C
C>A; B<C</td><td>LESS -MORE
A<B; C>A
A<C; B>A
C<B; A>C
C<A; B>C</td></tr>
</table>

Table 2. The 6 Trial-types presented in Experiments 1

(1) Each trial required participants to try to determine the relations among all three coins (based on the instructions). During some trial-types, the relations presented between the coins were the same. For example, participants may have been presented with a task that contained two *more-than* relations: “Coin A is worth more than coin B and coin B is worth more than coin C” (denoted as A>B; B>C). Alternatively, two less-than relations may have been presented (e.g., B<C; A<B). In contrast, some trials presented mixed relations across the three stimuli, in which the first relation between the stimuli contained a more-than relation, while the other contained a less-than relation or vice versa (e.g., B>C; A<B or C<B; B>A). Trials that contained two more-than or two less-than relations were referred to as *same* trials, whereas trials that contained one more-than and one less-than relation were referred to as *mixed* trials.

(2) During some trial-types, the information presented across the two premises was sufficient to allow participants to correctly determine the remaining relations among the coins. For example, if presented with A>B; B>C, one could correctly derive the relations A>C and C<A without additional information. In this case, the trials were referred to as *specified*. In contrast, during *unspecified* trials the information presented was not sufficient to allow the participant to derive the target relations. For example, if presented with A>B; B<C, one cannot determine the relations between A and C. Each test exposure contained both specified-same and specified-mixed trial-types.

(3) The trials were also differentiated in terms of whether or not a relation between non-adjacent coins (i.e., a transitive relation) was presented. For example, in the trial C<A; B<A, the first premise identifies a transitive relation between the non-adjacent coins A and C. In contrast, the trial B>C; A<B presents relations between only adjacent coins B and C as well as A and B.

In summary, the presentation of same, mixed, specified, unspecified, and transitive relations generated a total of six basic trial-types that were presented within each test. Specifically, each test contained non-transitive specified-same, unspecified-same, specified-mixed, and unspecified-mixed trial-types, as well as two transitive trial-types -- specified-same transitive and unspecified-mixed transitive (see Table 2).

Procedure. Participants were exposed to the automated procedure twice during each of the three experimental phases (i.e., they each received a total of six exposures), with a 5-min. break between each exposure. At the beginning of the automated procedure, each participant was provided with a printed set of general instructions as follows:

During the experiment, the computer will present you with a number of problems to solve. For each problem, three coins, three coffee jars (with coffee), and a black container will be presented. Each coin is worth the amount of coffee in one of the three containers. Each time the computer will tell you about the relative value of each of the three coins. Your task is to work out which coin is worth which jar of coffee and then drag and drop each coin into the space below the appropriate jar. For some questions it is impossible to know where two of the three coins should be placed. When this is the case, drop these two coins into the space below the black container and drop the remaining coin into the space below the appropriate coffee jar.

The Experimenter left the room shortly after the instructional phase and the automated procedure commenced immediately. Participants simply clicked on an ‘intermediate screen’ to indicate their readiness to proceed.

The automated procedure contained 48 test screens, each of which presented a single trial in which the three coins (red, blue, and yellow), three coffee jars (small, medium, and large), and a black tin (the “I cannot know” tin) were presented. On all trials, the three coins were presented in fixed locations (A/red left, B/blue middle, and C/yellow right), as were the three jars (small/left, medium/middle, and large/right). The basic task required participants to drag and drop each coin under the appropriately sized coffee jar or into the black tin depending upon the relations among the coins stated during the trial.

For illustrative purposes, consider the specified-same trial denoted as A<B; B<C. During this trial, the following instruction appeared on screen: “The red coin (A) is worth less than the blue coin (B); and the blue coin (B) is worth less than the yellow coin (C)” (again participants did not see the alphanumeric labels). In this case, a correct response involved the participant deriving the following information and placing the coins accordingly. The red coin (A) is worth the least (because A is worth less than B and B is worth less than C) and thus it should be placed with the smallest jar. Coin B is in the middle (because B is worth more than A but less than C) and thus the blue coin should be placed with the medium-sized jar. Coin C is the largest (because C is worth more than both A and B) and thus the yellow coin should be placed with the largest jar. Given that the above example depicted a specified trial, no coins would be dragged to the black tin because the three coins could readily be placed beside the three jars.

Now consider the unspecified-mixed trial ‘C>B; B<A’. During this trial, the following instruction appeared on screen: “The yellow coin (C) is worth more than the blue coin (B); and the blue coin (B) is worth less than the red coin (A)”. In this case, a correct response involved the participant deriving the following information and placing the coins accordingly. The blue coin (B) is worth less than the red coin but also worth less than the yellow coin, so B goes beside the smallest jar. However, one cannot determine in this case the precise relationship between A and C and thus the red and yellow coins must be placed in the black tin (because one cannot determine which jars they should go to).

A number of additional minor instructions were presented on each trial to allow participants to interact with the program freely and correctly, and to proceed appropriately through the experiment. For example, an additional button box was located at the bottom of the screen, with the words “Start Again” which returned all of the coins to their original locations, and thus allowed the participant to begin the trial again. Similarly, a second “Finish Trial” button box enabled the participant to proceed immediately to the next screen only after all three coins had been moved either to a jar or to the black tin. On each trial, participants could only place one coin in each jar, but could place any number of coins (including none) in the black tin (alternative responses were not permitted by the program). No feedback was presented after any trial. At the end of the final test trial, the program thanked participants for their time thus far and advised them to contact the Experimenter seated outside the room. They were then given a 5-min. break and were re-exposed to the entire test for a second time. These two consecutive test exposures constituted Phase 1 of the experiment. Phases 2 and 3 were identical to Phase 1, and thus each participant was exposed to a total of six test exposures.

Result and Discussion

Although all participants completed two test exposures in each phase, the data from the three second exposures only were analyzed because pilot work had indicated that responding during the first exposure was often erratic. The accuracies of participants' performances were grouped according to the six relation types and these are presented in

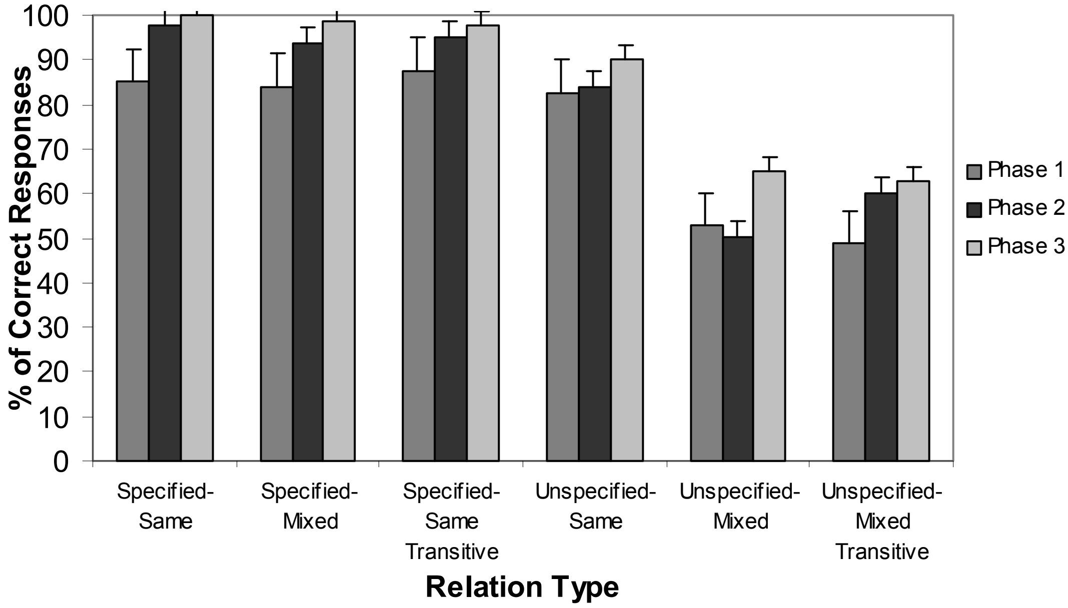

Figure 1. *The percentage of correct responses on the six relation types presented in the second exposure of Phases 1, 2 and 3 of Experiment 1.*

Participants overall performed better on the specified than on the unspecified relations, with the weakest performances recorded on the unspecified-mixed and unspecified-mixed transitive relations. Although initially strong, performances on the three specified relations (specified-same, specified-mixed, and specified-mixed transitive) improved further across phases, with a mean accuracy overall of 98.8% across the three types by Phase 3. However, on the weakest unspecified-mixed and unspecified-mixed transitive relations the mean accuracy across phases changed little, and was not high at any point. Specifically, accuracy on these relations ranged from 52.7% and 48.9% in Phase 1; 50.2% and 60.2% in Phase 2; and 65.1% and 62.7% in Phase 3, respectively. Incidentally, performances on unspecified-same relations were generally more similar to specified than unspecified performances. That is, the mean accuracies recorded for these relations were as follows: 82.6% in Phase 1; 83.9% in Phase 2; and 90% in Phase 3.

A 6 x 3 repeated measures Analysis of Variance (ANOVA) was conducted with both relation type and phase as within participant variables. This analysis revealed significant effects for both variables relation type: [$F(5,45)=14.188$, $p < .0001$, $\eta_p^2 = 0.612$] and phase:[$F(2,18)=4.556$, $p < .0251$, $\eta_p^2 = 0.336$], but no interaction effect [$F(10,90)=0.524$, $p = .8687$, $\eta_p^2 =0.055$]. Post-hoc analyses (Fisher's PLSD) revealed significant superiorities of specified over unspecified responding across all phases particularly when the latter contained mixed and/or transitive relations. Table 3 summarizes these differences. The consistency of responding on the various trial-types across phases indicated that participants' performances on specified relations remained highly accurate, while responding to unspecified-mixed and unspecified-mixed transitive relations remained weak.

Relation Type Comparisons	Phase 1	Phase 2	Phase 3
Specified-same vs. Unspecified-mixed	.0007	< .0001	< .0001
Specified-same vs. Unspecified-mixed transitive	.0002	.0001	< .0001
Specified-mixed vs. Unspecified-mixed	.0011	< .0001	.0001
Specified-mixed vs. Unspecified-mixed transitive	.0003	.0005	< .0001
Specified-mixed transitive vs. Unspecified-mixed	.0003	< .0001	.0002
Specified-mixed transitive vs. Unspecified-mixed transitive	< .0001	.0003	< .0001
Unspecified-same vs. Unspecified-mixed	.0017	.0005	.0036
Unspecified-same vs. Unspecified-mixed transitive	.0005	.0116	.0015

Note. NS= non significant difference

Table 3. Statistical Comparisons of Participant's Performances on Each Relation Type Across All Three Phases Presented in Experiment 1

Participants in Experiment 1 received a total of six complete exposures to the test protocol and from the outset produced significantly better performances on specified versus unspecified relations, particularly when the latter were mixed and/or transitive. Although the strong performances on specified relations improved further, the unspecified-mixed and unspecified-mixed transitive relations remained significantly weaker. This finding suggested that although some improvement resulted from repeated exposure alone, greater improvement especially in the weaker relations would appear to require more explicit forms of intervention. This issue was addressed in Experiment 2.

2.2. Experiment 2

Method

Participants. Ten naïve subjects participated in Experiment.

Once again, they were undergraduate students/professionals aged 18-24, they were recruited through personal contacts, and they did not receive any remuneration for their participation in the experiment.

Setting. The setting for Experiment 2 was identical to that employed in Experiment 1.

Apparatus. The apparatus used in Experiment 2 was identical to the automated procedure employed in Experiments 1, except for modifications to Phase 2. The experimental trials in this phase involved a combination of the written feedback presented with non-arbitrary stimuli (i.e. their physical size was different).

Procedures. The procedure employed in Experiment 1 was identical to Experiment 2, except for Phase 2 (training), which presented non- arbitrary stimuli, and written feedback after each trial-type.

Two modifications to the size of the stimuli were necessary in order to transform the trials from arbitrary to non-arbitrary relations, one for specified relations and the other for unspecified relations. Specifically, during all trials involving specified relations the three coins (A, B and C) presented on the screen were all physically different in size -- one was small, one was medium and one was large. The relative sizes of the three coins depended on the trial-type being presented. Consider, for example, a specified-same trial in which the subject was instructed as follows: "The green coin (B) is worth more than the yellow (C) coin and the red (A) coin is worth more than the green coin (B)". During Phase 2, the sizes of the coins matched the instruction. In this example, therefore, the green coin (B) was medium in size, the yellow coin (C) was small and the red coin (A) was large.

When the relations presented in Phase 2 were unspecified, another type of modification to the sizes of the stimuli was required in order to present the trials in non-arbitrary form. During trials containing unspecified relations, a correct response always involved placing two of the coins (whose location in the jars could not be determined) into the "Cannot Know" tin, and placing the third coin into either the large or the small jar (but not in the medium-sized jar). Consider, for example, the following unspecified-same trial in which the participant was instructed as follows: "The green coin (B) is worth less than the yellow (C) coin and the green coin(B) is worth less than the red coin (A)". The correct answer to this trial involved placing the green coin (B) into the small jar (because it is worth less than the other two coins) and placing the remaining coins into the “Cannot Know” tin (because the relation between these coins cannot be determined). In order to present this trial in non-arbitrary form, the two unspecified coins were presented as identical in size and the third coin was different. That is, in the current example, the red (A) and the yellow coins (B) may have been medium in size and the green coin would have been small. As before, however, two coins could not be placed together in any of the coffee jars, only in the “Cannot know container”.

The modifications described previously to the sizes of the coins during specified and unspecified trial-types did not alter the correct responses required during unspecified trials, but were put in place in order to avoid subjects responding to the coins as if they were of a fixed size. All of the 48 trials in each of the two exposures presented during Phase 2 of Experiment 2 were conducted in this way.

Furthermore, after each trial, written feedback appeared in the computer screen.

If a participant responded correctly to a trial, the word “Correct” appeared immediately on the experimental screen. If a participant responded incorrectly to a trial, the word “Wrong” similarly appeared immediately after the response. The correct and wrong feedback labels each lasted seven seconds. During this time the “Start Again” and “Finish Trial” buttons were inactive so that subjects had adequate exposure to the feedback and could not proceed to the next trial during the presentation of feedback.

An example of the training provided in phase 2 of Experiment 2 is given below, in Figure 2.

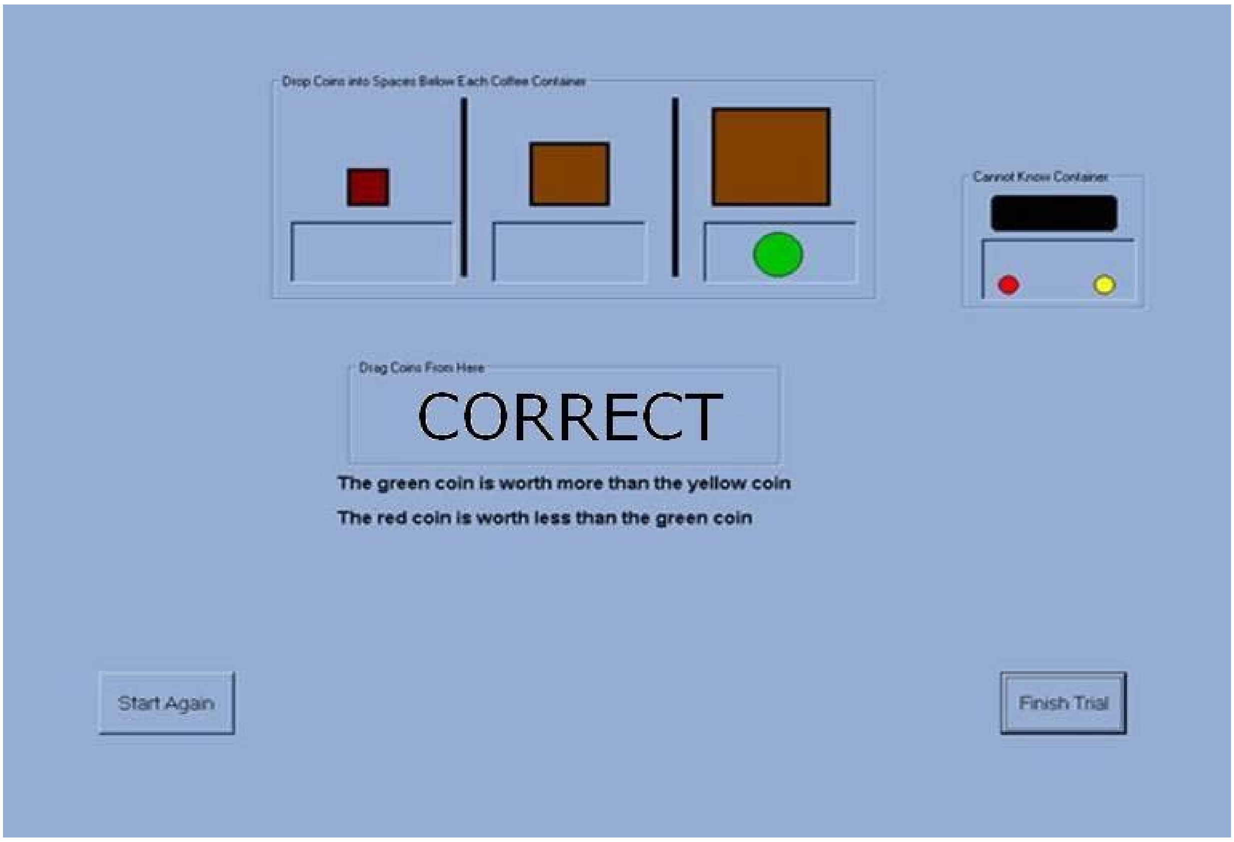

Figure 2. *One of the training screens to which individuals were exposed during the automated procedures employed in Phase 2 of Experiment 2.*

Results and Discussion

Figure 3 presents the percentage of correct responses on the six relation types from the second exposure in each of the three phases. The figure shows the characteristic pattern of differences between specified and unspecified relations, with higher levels of accuracy overall on the specific relations. There were substantive improvements in the three types of unspecified relations in Phase 2 compared to Phase 1 (i.e., 51.4% mean accuracy in Phase 1 compared to 98.4% in Phase 2) and maintenance of these improvements in Phase 3 (99.6% accuracy).

.

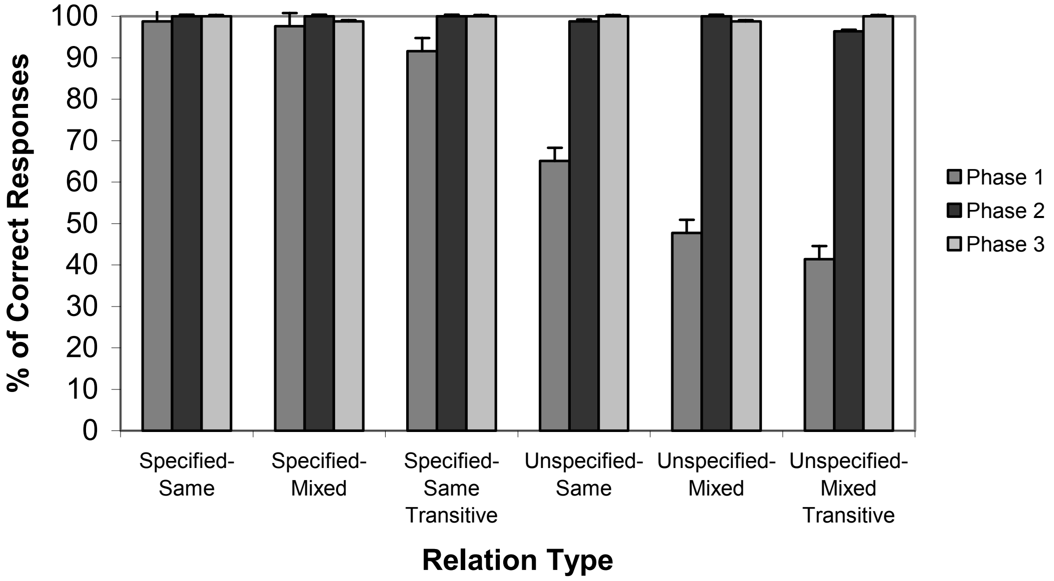

Figure 3. *Percentage of correct responses to the 6 relation types in Phases 1, 2 & 3 of Experiment 2.*

A 6 x 3 repeated measures ANOVA was conducted on the data and yielded significant main effects for both relation type [F(5,45)=36.514, p< .0001, □p2 = 0.802] and phase [F(2,18)=310.902, p< .0001, □p2 = 0.972], as well as a significant interaction effect between the two variables [F(10,90)=34.229, p< .0001, □p2 = 0.792]. Fisher's post-hoc analyses indicated that the majority of significant differences were recorded between specified and unspecified relations in Phase 1 only (see Table 4). Only three significant differences remained in Phase 2, all of which included unspecified-mixed transitive relations. This pattern of results suggests significant improvements in unspecified relations in Phase 2 that were maintained in Phase 3. Indeed, performances on unspecified relations in Phases 2 and 3 reached the same levels of accuracy as on the other relations, thus highlighting the efficacy of combining written feedback and non-arbitrary trials.

Relation Type Comparisons	Phase 1	Phase 2	Phase 3
Specified-same vs. Unspecified-same	< .0001	NS	NS
Specified-same vs. Unspecified-mixed	< .0001	NS	NS
Specified-same vs. Unspecified-mixed transitive	< .0001	.0077	NS
Specified-mixed vs. Unspecified-same	< .0001	NS	NS
Specified-mixed vs. Unspecified-mixed	< .0001	NS	NS
Specified-mixed vs. Unspecified-mixed transitive	< .0001	.0077	NS
Specified-mixed transitive vs. Unspecified-same	< .0001	NS	NS
Specified-mixed transitive vs. Unspecified-mixed	< .0001	NS	NS
Specified-mixed transitive vs. Unspecified-mixed transitive	< .0001	.0077	NS
Unspecified-same vs. Unspecified-mixed	.0054	NS	NS
Unspecified-same vs. Unspecified-mixed transitive	.0002	NS	NS

Note. NS= non significant difference

Table 4. *Statistical Comparisons of Participant's Performances on Each Relation Type Across All Three Phases Presented in Experiment 2.*

3 General Discussion

The experimental work reported in the current paper aimed to test and train comparative relations among three arbitrary stimuli by using automated procedures.

The outcome of both experiments demonstrated that individuals perform better on the specified than the unspecified relations, especially when these latter are mixed and transitive. Experiment 1 demonstrated that repeated exposures to the trials types are not effective in improving individuals' performance on the unspecified relations, especially when these are mixed or mixed transitive. According to Vitale et al. (2008) abstract unspecified comparative relations are not part of our everyday life, and this might explain why individuals found it difficult to solve them. Experiment 2 demonstrated that the training techniques of written feedback and non-arbitrary trials had a significant effect on individuals' ability to solve both specified and unspecified comparative relations.

The automated procedures employed in Experiments 1 and 2 support the trend of considering computers as the new cognitive technologies for improving aspects of abstract reasoning (Pea 1985). It is hoped that the procedures reported here may have pragmatic implications for testing and training abstract comparative reasoning in other populations, including individuals who may be approaching computer programming for the first time (e.g. BA students). According to Sitaraman et al. (2000) in fact it is impossible to understand the behaviour of software objects by appealing to physical analogies since software objects do not correspond one-to-one to physical objects. Thus flexibility in thinking in abstract terms, including for comparative reasoning, is an essential cognitive skill for programmers.

It is hoped therefore that the automated procedures reported here may provide the foundation for the development of similar procedures designed for pragmatic rather than experimental applications. For example, such procedures could form the basis for aptitude tests for prospective computer programmers, and/or for the development of training techniques for improving abstract reasoning skills in novice programmers. These procedures can best be generated through interactions between psychologists, who have an understanding of abstract thinking and how it can be improved, and computer programmers, who have the tools to generate software for improving cognitive skills.

4. References

Barnes-Holmes, Y., Barnes-Holmes, D., Roche, B., Healy, O., Lyddy, F., Cullinan, B. & 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. New York: Plenum.

Barnes-Holmes, Y., Barnes-Holmes, D. & Cullinan, V. (2001). Education. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), Relational Frame Theory: A post-Skinnerian account of human language and cognition. New York: Plenum

Byrne, R. M. J. & Handley, S. J. (1992). Reasoning strategies. *Irish Journal of Psychology, 13,* 111-124.

Byrne, R. M. J. & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language, 28*, 564-575.

Brooks, R., (1983). Towards a Theory of the Comprehension of Computer Programs. *International Journal of Man-Machine Studies*, 18, pp 543-554.

Clark, H. H. (1994). Discourse in production. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics.* London: Academic Press.

Clark, H. H. (1969a). Linguistic processes in deductive reasoning. *Psychological Review*, *76*, 387-404.

Clark, H. H. (1969b). The influence of language in solving three-term series problems. *Journal of Experimental Psychology, 82*, 205-213.

Cullinan, V. & Vitale, A., (2008). The contribution of Relational Frame Theory to the development of interventions for impairments of language and cognition. *The Journal of Speech & Language Pathology - Applied Behavior Analysis* 2.4-3.1, 122-135

Dymond, S., Rehfeld, R. A. & Schenk J. (2005) Non automated procedures in derived stimulus relations research: a methodological note. *The Psychological Record,* 55, 461-481.

De Soto, C., London, M., & Handel, S. (1965). Social reasoning and spatial para logic. *Journal of Personality and Social Psychology, 2,* 513-521.

Ehrlich, K. & Johnson-Laird, P. N. (1982). Spatial description and referential continuity. *Journal of Verbal Learning and Verbal Behavior, 21*, 296-306. *Psychology*, *48,* 286-298.

Hayes, S., Fox, E., Gifford, E., Wilson, K., Barnes-Holmes, D., & Healy, O. (2001). Deriver relational responding as learned behaviour. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), Relational Frame Theory: A post-Skinnerian account of human language and cognition. New York: Plenum.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review, 75,* 550-560.

Johnson-Laird P. N. & Byrne, R. M. J. (1990). Metalogical reasoning: Knights, knaves and Rips. *Cognition, 36,* 69-84.

Johnson-Laird, P. N. & Byrne, R. M. J. (1989). Only reasoning. *Journal of Memory and Language, 28,* 313-330.

Kramer, J.(2007). Is abstraction the key to computing? *Communication of the ACM,* 50, 37-42.

Mani, K. & Johnson-Laird, P. N. (1982). The Mental representation of spatial description, *Memory and Cognition,10,* 181-187.

Pea, R., (1985). Beyond amplifications: using the computer to reorganise mental functioning. *Educational Psychologist*, 20 (4), pp167-182.

Pennington, N., (1987). Stimulus Structures and Mental Representations in Expert Comprehension of Computer Programs, *Cognitive Psychology,* Vol. 19, pp 295-341

Peterson, L., Homer, A.L., & Wonderlich, S.A., (1982). The Integrity of independent variables in Behavior Analysis. *Journal of Applied Behavior Analysis, 15,* 477-492, 4.

Rajlich V., (1994), Program Reading and Comprehension. *Proceedings of Summer School on Engineering of Existing Software*, pp 161-178

Reese, H. W. (1968). *The perception of stimulus relations: Discrimination learning and transposition.* New York: Academic Press.

Russell, J., McCormack, T., Robinson, J., & Lillis, G. (1996). Logical (versus associative) performance on transitive reasoning tasks by children: Implications for the status of animals' performance. *The Quarterly Journal of Experimental Psychology, 49b,* 231-244.

Saunders, K.J., & Williams, D.C. (1998). Stimulus control procedures. In K.A. Lattal & M. Perone (Eds), *Handbook of Research Methods in Human Operant Behaviour.* (pp. 193-228). New York: Plenum.

Sitaraman, M., Atkinson, S., Kulczycky, G., Weide, B.W., Long, T., Bucci, P., Heym, W., Pike, S. M., & Hollingsworth, J. (2000). Reasoning about software-components behaviour. In William B. Frankes Eds. Software Reuse: Advance is Software Reusability.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General, 109,* 119-159.

Sternberg, R. J. (1981). Reasoning with determinate and indeterminate linear syllogisms. *The British Journal of Psychology, 72,* 407-420.

Soloway, E., Ehrlich, K., (1984), Empirical Studies of Programming Knowledge. *IEEE Transactions on Software Engineering,* Vol. SE-10, 595-609

Vitale A., Barnes-Holmes Y., Barnes-Holmes D., & Campbell C. (2008). Facilitating Responding in Accordance with the Relational Frame of Comparison: Systematic Empirical Analyses I. *The Psychological Record, 2008, 58, p. 365-390.*