Journal of Behavioral Education, Vol. 12, No. 1, March 2003 (© 2003), pp. 1–21

Effects of Computer-Based Fluency Training on Concept Formation

Eric J. Fox[1,2] **and Patrick M. Ghezzi**[1]

This study provides a preliminary analysis of how the techniques of fluency training can be combined with systematic concept instruction to improve the learning of complex verbal concepts. Fluency techniques, which require the learner to respond accurately at high rates, have typically focused on definition learning when teaching concepts. Instructional psychologists, however, recommend multiple exemplar training for conceptual instruction. To examine this issue, 41 undergraduate students completed a computer-based instructional module on logical fallacies. Participants were assigned to one of four groups, with the modules for each group differing only in the type of practice provided—either fluency or practice with either examples or definitions. Examination of posttest scores revealed significantly higher scores for participants in the examples groups than those in the definitions groups, but low experimental power prevented a clear conclusion to be drawn about differences between the fluency and practice groups. Implications of results and several methodological issues relevant to this area of research are discussed.

KEY WORDS: behavioral fluency; concept formation; precision teaching; computer-based instruction.

An increasing number of researchers, theorists, and educators interested in behavioral education have recently begun to focus their attention on what has come to be known as *behavioral fluency*. The construct of behavioral fluency emerged from the basic research on free-operant conditioning and the related instructional technology of precision teaching (Lindsley, 1971, 1990; Potts, Eshleman, & Cooper, 1993). While traditional measures of learning in the educational setting have focused on percent correct, precision teachers argue that this restricts the assessment of learning, as no additional measurements are possible once response accuracy

[1]Department of Psychology, University of Nevada, Reno, Nevada.
[2]Correspondence should be directed to Eric J. Fox, M. A., Division of Psychology in Education, Arizona State University, Box 870611, Tempe, Arizona 85287-0611; e-mail: eric.fox@asu.edu.

reaches 100%. Instead, precision teaching adopts the dependent measure of operant conditioning laboratories and measures learning with response rate (Lindsley, 1996), working to increase both the accuracy and the speed of academic skills. This emphasis on response rate has also been incorporated into generative instruction (Binder, 1991; Johnson & Layng, 1992, 1994), a technique which combines features of several different behavioral approaches to education.

Behavioral fluency is often described as the combination of accuracy plus speed of responding characteristic of expert or competent performance (Binder, 1988, 1990, 1996). It is similar to what cognitive psychologists call “automaticity” (Laberge & Samuels, 1974), though fluency advocates claim their emphasis on rate of performance rather than mere practice beyond accuracy (or “overlearning”) differentiates fluency from automaticity (Johnson & Layng, 1992). Functionally, fluency is defined as the rate of responding that predicts the retention, endurance, application, and stability of the response (Johnson & Layng, 1992, 1994, 1996). Retention is the accurate and rapid performance of the skill after a significant period of no practice; endurance is the performance of the skill at a high rate for intervals longer than those used during practice; application is the use of the skill as a component of more complex responses; and stability is the accurate and rapid performance of the skill in distracting conditions (Johnson & Layng, 1996). Ideally, these performance standards determine the frequency (or fluency) aims for particular responses, and are often represented by the acronym *REAPS* (Retention, Endurance, Application, Performance standards, and Stability).

The use of frequency-building techniques, either through precision teaching alone or in combination with other instructional strategies, has produced remarkable educational outcomes with both children and adults. In the 1970s, the Precision Teaching Project in Great Falls, Montana revealed that elementary students who received just 30 minutes per day of frequency timings (and performance charting), in addition to their regular curriculum, outperformed other students in the state by 40 percentile points in math and 20 percentile points in reading on standardized tests (Beck & Clement, 1991). Students at Morningside Academy, a private school in Seattle utilizing generative instruction, typically gain between two and three grade levels each year as measured by standardized tests (Johnson & Layng, 1992, 1994). Malcolm X College in Chicago successfully used generative instruction for an adult literacy program (Johnson & Layng, 1992), while Binder and Bloom (1989) achieved remarkable results with fluency-based corporate training programs. The Center for Individualized Instruction (now called Learning Services) at Jacksonville State University has utilized fluency-based computerized instruction to successfully provide academic support to college students for over two decades (McDade & Goggans, 1993).

One potential use of fluency-based instruction that is particularly relevant to teachers of adult learners, but that has not received a lot of attention, is the teaching of complex verbal concepts. Much of adult learning involves abstract verbal concepts, and it would be beneficial to understand how the procedures of fluency-building might be used to enhance the instruction of such material. To date, most of the work on fluency has involved teaching basic academic skills, such as reading and arithmetic, or specific factual knowledge. When fluency techniques have been used to teach conceptual material, it has typically taken the form of training the saying, writing, or identification of the concept's definition to fluency (e.g., McDade, Rubenstein, & Olander, 1983; Polson, 1995). This focus on the definition of the concept is typical of *SAFMEDS* (Say All Fast a Minute Each Day Shuffled; Eshleman, 1985; McGreevy, 1983), a flash card fluency procedure that appears to be the most common way in which fluency procedures are incorporated into college courses (see Korinek & Wolking, 1984; Olander, Collins, McArthur, Watts, & McDade, 1986). Indeed, a review of the sample instructional "decks" distributed with *Think Fast* (Parsons, 2000), a computerized SAFMEDS program, also reveals an emphasis on definition learning.

While a reliance on definition learning is certainly prevalent in both fluency-based and traditional concept instruction, such training is not considered sufficient by most instructional design experts (Engelmann & Carnine, 1991; Gagné, 1970; Gibbons & Fairweather, 1998; Markle, 1969; Merrill, Tennyson, & Posey, 1992). This is because a concept is essentially a class of " . . . objects, events, relations, or other things which vary from one example to another but which are treated as being members of the same group and called by the same name" (Tiemann & Markle, 1990, p. 72), and thus conceptual behavior is the identification of instances of a class. And while a concept's definition may outline the criteria upon which class membership is based, this does not guarantee that the student will accurately discriminate instances based on these criteria. A student may be able to recite verbatim the definition of "positive reinforcement," for instance, but unable to identify an example of the concept in a field setting or in a written passage. Evidence for conceptual understanding includes the correct identification of novel instances as members of the class and the correct rejection of non-instances as not belonging to the class (Tiemann & Markle, 1990); simply reciting or identifying a definition does not provide such evidence.

To promote conceptual understanding by learners, instructional psychologists advise that concept instruction be developed according to guidelines based on the findings of concept formation and discrimination learning research with both humans and non-humans (Clark, 1971; Gagné & Brown, 1961; Johnson & Stratton, 1966; Merrill et al., 1992; Tiemann & Markle, 1990). This literature suggests that learners be exposed to a series of examples and non-examples of the concept, and that these instances be selected with care to prevent classification errors such as overgeneralization, undergeneralization, and misconceptions (Engelmann, 1969; Markle, 1969; Merrill et al., 1992). In selecting a set of instructional and testing examples, it is useful to first conduct a concept analysis, in which the critical or defining attributes and the variable or non-defining attributes of the concept are delineated. After the critical and variable attributes are identified, a *minimum rational set* (MRS; also called the minimum critical subset) of examples and non-examples can be generated that pinpoints the boundary of the concept (Gibbons & Fairweather, 1998; Tiemann & Markle, 1990). These examples and non-examples are then presented to the learner and differential reinforcement is used to increase correct responding and to ensure that "learner performance comes under the fine-grained control of the stimulus features that embody the concept" (Johnson & Layng, 1994, p. 179).

Despite the recommendation of instructional designers to use multiple exemplars when teaching concepts, it has not been demonstrated empirically that this is the best approach to use when employing fluency techniques. It is quite possible that the typical practice of building the recall or recognition of a concept's definition to high rates promotes conceptual understanding just as effectively as training the identification of examples of the concept. Most verbally-sophisticated adults can obviously gain some degree of conceptual understanding simply by learning a concept's definition; indeed, if this were not the case, expanding one's vocabulary would be a painstakingly slow process and dictionaries would be of limited value. Definitions are useful because they can (when well-written) succinctly outline the critical attributes of the concept and specify the manner in which the concept relates to other (hopefully familiar) terms and concepts. Understanding how concepts relate to one another is important, as recent work in the area of derived relational responding suggests that the ability to derive relations among events and stimuli may be the core behavioral process underlying human language and cognition (see Hayes, Barnes-Holmes, & Roche, 2001). Thus, it is feasible that fluency training using definitions may provide the student with sufficient practice and experience to later recall the concept's critical attributes and the ways in which the concept relates to other concepts. In effect, it may allow the student to become fluent at verbally identifying the critical attributes of the concept, which might then allow the student to correctly categorize novel instances of the concept. The danger in relying on definition learning, however, is that students may simply learn the structure of the definition, or memorize it as a string of words, without attempting to understand the meaning of the terms or how they relate to the concept (an effect that may be even more pronounced in fluency training as students may attend more closely to the structural features of the definition in order to achieve a high rate of responding).

The present study was designed to provide a preliminary analysis of these issues. To understand how fluency techniques can best be used to teach verbal concepts, the use of definitions or examples in a fluency-based instructional program was compared. Specifically, this study examined: (a) the effects that fluency training on identifying a concept's definition has on a student's ability to identify novel instances of that concept; (b) the effects that fluency training on identifying examples and non-examples of a concept (based on the MRS) has on a student's ability to identify novel instances of that concept; (c) the differences and similarities in the effects of these two types of training; and (d) whether any positive effects of fluency training can be attributed to mere repetition of the instructional task (practice effects) rather than the achievement of rapid responding. The principal dependent variable was performance on a posttest measuring the student’s ability to correctly identify novel instances of the concept, while the principal independent variables were the type of practice required (either with a high rate requirement or no rate requirement) and the type of practice items used (either examples or definitions). To analyze these issues, college students were exposed to nearly identical computer-based instructional modules about informal logical fallacies, with only the type of practice and the type of practice items varying between groups of students.

METHOD

Participants and Setting

A total of 41 undergraduate college students between the ages of 18 and 24 and enrolled in an introductory psychology course participated. Students were recruited via posted advertisements and classroom announcements, and received course credit for their participation. Prospective participants were given a brief pretest on the logical fallacies used in the experiment (described in detail under the General Procedure section); those scoring at or above 60% correct on the pretest were not allowed to continue with the experiment. The pretest cut-off score of 60% was used because it provided a socially-valid measure of “failing” the test, as most grading systems mark a score below 60% as failing, and because it was low enough to allow measurable improvement on the posttest. Of the 41 students who participated, 2 were ineligible to continue due to their pretest score, and the data for 3 participants was lost due to computer error. The 36 participants (21 female, 15 male) who completed the entire experiment comprise the present data set.

Each participant served in one of four experimental groups: fluency-definitions, fluency-examples, practice-definitions, and practice-examples. The terms used to identify the groups can be understood as follows: the first word (either fluency or practice) indicates the type of practice/training they received—either fluency-building (with a high rate requirement) on the skill, or simple practice on the skill (with no rate requirement); the second word (either definitions or examples) indicates the type of practice items used—either the concepts’ definitions, or examples and non-examples of the concepts.

The first 18 participants were randomly assigned to either the fluency-definitions or the fluency-examples group. Each of the remaining 18 participants were randomly matched to 1 of the 18 participants in the two fluency groups. Random matching was used because all participants in the experiment were already matched on the basis of poor pretest score performance (i.e., those scoring above a criterion level were excluded from the experiment). If a participant was matched to someone in the fluency-definitions group, then he or she was placed in the practice-definitions group; likewise, if they were matched to a participant in the fluency-examples group, then they were placed in the practice-examples group. There were a total of 9 participants in each of the 4 groups.

The purpose of matching participants in the practice groups to participants in the fluency groups was to provide a control for practice effects. That is, to determine whether or not the effects of the fluency training was due to a high rate requirement or simply to repetition of the instructional task, each participant in the two practice groups received the same number of practice trials (without a rate requirement) as was required of their matched counterpart to meet the fluency criteria. For example, if Participant A in the fluency-definitions group required 150 trials to meet the fluency criteria, then his or her matched counterpart in the practice-definitions group was also required to complete a total of 150 practice trials.

Participants sat at a desk in a small room with a computer monitor, keyboard, and mouse in front of them. The room was free from distracting noise and other disruptions. The experimenter sat in an adjacent room, out of view of the participant.

Apparatus

The experiment was conducted with a Dell desktop personal computer with a 166 MHz processor, 24 MB of RAM, a 1 GB hard drive, and a 14 in. SVGA color monitor. The experimental program was written in Microsoft Visual Basic (6.0) for Windows.

Stimulus Materials

Four informal logical fallacies were selected to serve as the concepts to be taught during the experimental procedure. They included the bandwagon fallacy, the hasty generalization fallacy, the false cause fallacy, and the appeal to ignorance fallacy. These fallacies were selected because they are all considered common errors in reasoning (Gray, 1991) and because they all have a similar form (i.e., a single premise and a single conclusion). Some fallacies involve multiple premises and are thus considerably more difficult than others, so it was important for the fallacies in the experiment to share a similar form to standardize their complexity.

Concept Analyses

Concept analyses of the fallacies were conducted according to recommended guidelines (Gibbons & Fairweather, 1998; Tiemann & Markle, 1990) and using definitions and information obtained from several books on logic and reasoning (Carney & Scheer, 1980; Copi, 1982; Gray, 1991; Thomas, 1986). A review of these books revealed that each of the four fallacies essentially have two critical attributes: a premise of a specific form, and a conclusion of a specific form. The false cause fallacy, for example, requires a premise of the form "after A happens, B happens" and a conclusion of the form "therefore, A causes B." If either the premise or the conclusion does not take this form (with allowance for the precise wording, of course), it is not classified as a false cause fallacy.

Consultation with a subject matter expert (an instructor of a college-level course on critical thinking) and a previously-published concept analysis of "fallacy" (Tiemann & Markle, 1990, p. 96) revealed that when teaching the form of these fallacies, the primary variable attribute is the apparent truth of the conclusion. This is because a deductively valid argument (a non-fallacy) can result in an apparently false conclusion, and a deductively invalid argument (a fallacy) can result in an apparently true conclusion.[3] For example, an argument with the premise "lots of people think that Elvis Presley is dead" and the conclusion "therefore, Elvis Presley is dead" is a *fallacy* (the bandwagon fallacy, specifically) even though the conclusion seems true. If the apparent truth of the conclusion is not varied, then, students may identify an argument as a fallacy only if it has an apparently false conclusion. Therefore, the concept analyses conducted for the four fallacies considered the apparent truth of the conclusion as the key variable attribute for each, with three dimensions: obviously true (e.g., "dogs have four legs"), obviously false (e.g., "dogs have wings"), and unknown (e.g., "my dog is named Steve"). The precise wording used to state the premises and conclusions varied for all examples and non-examples and was not listed as a variable attribute in the formal concept analysis.

Definitions

The definitions for the fallacies used in this experiment were adapted from those commonly presented in logic textbooks (Carney & Scheer, 1980; Copi, 1982). Further, each definition emphasized the critical attributes of the concept by including the standard form of the premise and conclusion. The definition for the appeal to ignorance fallacy, for example, read:

This fallacy is committed when it is concluded that something is true simply because it has not been proven false. It generally has this form:

Premise: A has never been proven false
Conclusion: Therefore, A is true.

[3]Logicians use the concept of "soundness" to address this issue; deductively valid arguments that have true premises necessarily lead to true conclusions and are deemed "sound" (Lepore, 2000). Instruction in the present study focused on discriminating the form, or validity, of arguments without respect to the soundness.

An inverse version of the appeal to ignorance fallacy (in which something is concluded to be false because it has never been proven true) is also possible, but the present study only addressed the most basic (and most common) version presented above.

Minimum Rational Sets/Fallacy Sets

The concept analyses conducted for the fallacies revealed that the MRS for each would consist of three examples and two non-examples. Three examples were needed so that each level of the variable attribute could be represented (i.e., an example with an obviously true conclusion, an example with an obviously false conclusion, and an example with an unknown conclusion). Two non-examples were needed so that each could have all but one of the two critical attributes (i.e., a non-example with the correct premise form but an incorrect conclusion form, and a non-example with an incorrect premise form but the correct conclusion form). With a total of 5 examples and non-examples needed in the MRS for each concept, and a total of 4 concepts to be taught, the total number of examples and non-examples needed for an entire instructional set (to teach all 4 concepts) was 20. The experimental procedures called for three such instructional sets, hereafter referred to as *Fallacy Sets*.

Example/Non-example Length and Readability

Each example and non-example used in the experiment was matched for word length, with the total number of words (including both the premise and conclusion) for each being no less than 15 and no greater than 20. In addition, examples and non-examples were matched on readability using the Flesch Reading Ease Score, available with Microsoft Word for Windows (9.0). The Flesch Reading Ease Score rates text on a 100-point scale using a formula based on the average number of words per sentence and the average number of syllables per word; the higher the score, the easier it is to understand the text. Each example and non-example used in the experiment had a Flesch Reading Ease Score between 60 and 70 (the range recommended for most standard documents).

Determining Frequency Aims

Some researchers claim that the frequency aims for a specific task should be empirically determined by discovering which rates adequately predict the retention, endurance, application, and stability of the skill (Johnson & Layng, 1996). This process is not without its problems, however. First, the value of using such criteria to describe fluency has received almost no experimental examination (but see Munson, 1998, for one of the few studies addressing this issue). Second, the functional definitions of the REAPS criteria can be ambiguous, making it difficult to develop standardized procedures for determining fluency rates. While proponents of fluency have worked to refine and operationalize these definitions in recent years (e.g., Binder, 1996; Johnson & Layng, 1994), the process of empirically deriving fluency rates from an examination of the REAPS criteria still can be an imprecise, difficult, and often impractical procedure. For these reasons, frequency aims are often set using techniques based on normative criteria (Koorland, Kell, & Ueberhorst, 1990) such as peer comparison (Evans & Evans, 1985) or using the rates of individuals considered to be experts in the target skill (Howell & Lorson-Howell, 1990; Pennypacker & Binder, 1992). These alternative techniques to establishing fluency performance standards are especially useful to individual teachers or researchers focusing on a skill for which there is little or no data on what the appropriate aims might be.

The response rate for the instructional task used in this experiment—responding to computerized multiple-choice questions about logical fallacies—that would reliably predict the REAPS outcomes is not known. Further, the lengthy and complicated process of determining such a rate is beyond the scope of the present study. Fortunately, however, the Center for Individualized Instruction (CII), a multi-disciplinary academic support center at Jacksonville State University, has been using computerized multiple-choice tasks very similar to the one employed in this study for a number of years (McDade & Goggans, 1993; McDade & Olander, 1987). As part of their Computer-Based Precision Learning system, the CII conducts computerized fluency-building exercises with college students for a variety of courses in different disciplines. And while a formal analysis of REAPS outcomes using their system and fluency rates has not been conducted, the CII has achieved impressive results which suggest many of the REAPS outcomes are achieved (McDade & Goggans, 1993). For example, students using their frequency-building program performed better on subsequent essay exams (McDade, Rubenstein, & Olander, 1983) and word problems (measures of application; McDade & Olander, 1987; McDade, Willanzheimer, & Olander, 1981), and retained material longer (Olander et al., 1986) than students who did not use the program.

The CII typically sets the minimum frequency criterion for computerized multiple-choice tasks at anywhere from 20 to 30 correct responses per minute (McDade & Goggans, 1993; McDade & Olander, 1987), with the length of the questions obviously playing a role in the rate (C.E. McDade, personal communication, February 15, 2000). Further, students frequently record rates above 40 correct responses per minute on these tasks (McDade & Goggans, 1993). While such a high rate of responding suggests that readers would need to read at the exceptionally high rate of more than 400 words per minute, in reality students are probably able to respond to questions without reading them entirely by recognizing certain structural features of the question (such as the word or words it begins or ends with, the number of lines in the question, the actual shape of the paragraph, etc.). This is especially likely when it is considered that only a limited number of items typically populate the question pool, allowing the student to see each question many times during their training. As mentioned previously, such discriminative responding based on the structure rather than the meaning of written material is a danger in any form of fluency training that uses textual materials. Nevertheless, the reported success of the CII's practices suggests that this danger may be minimal, and thus their rates served as the basis for the frequency criteria used in the present experiment.

Pilot testing indicated that rates somewhat higher than those typically required by the CII could be obtained in a relatively short period of time. This is probably due to the limited number of items in the item pool used in the present experiment; just 4 items in the definition pool, and just 20 items in the examples pool. Further, higher rates were possible on the definitions task than on the examples task, with the difference again likely due to the greater number of items in the example item pool. Based on the rates used at the CII and on information obtained from the pilot testing, the frequency criterion for the examples identification task was set at 35 correct responses per minute, and the frequency criterion for the definitions identification task was set at 45 correct responses per minute. An accuracy criterion of 90% was also used with these tasks to prevent participants from achieving the frequency criterion simply by clicking on one button at high rates.

Binder (1996) suggests that computer-based fluency programs place a ceiling on response rate if the learner is not fluent at component (or tool) skills, such as typing or using the mouse. To avoid such a problem in the present experiment, participants were exposed to a mouse pre-training phase, a simple identity-matching task designed to promote fast, accurate responding using the mouse. The frequency aim for this pre-training was set at 55 correct responses per minute (with 90% accuracy overall) to ensure that failure to achieve the rates required later (35 and 45 correct responses per minute) could not be attributed to inexperience or dysfluency with computer mouse use.

General Procedure

Participants were exposed to the following experimental phases: Pretest, mouse pre-training, introductory instruction, accuracy training on definitions, probe test, practice, and posttest. All of these phases, with the exception of the practice phase, were identical for all participants. During the practice phase, participants were exposed to different practice conditions according to their experimental group; these differences are outlined in the Practice section below. All phases of the experiment were conducted on the computer and controlled using a custom program written in Visual Basic.

Pretest

Prior to receiving any instruction on the fallacies, participants completed a pretest. The pretest was a multiple-choice task consisting of the presentation of

20 examples and non-examples of the four fallacies. These items were drawn from Fallacy Set 1 and presented in random order. To prevent an ability or inability to recognize premises and conclusions in an argument from affecting their performance, the premise and conclusion for each item was clearly identified on the screen. Upon presentation of the item, the participant was required to identify the type of fallacy being exemplified by clicking on one of five buttons; the first four were labeled with the names of each of the four fallacies, and the fifth was labeled "none of the above." Participants received no programmed feedback on their performance on the pretest, and those answering 60% or more of the questions correctly were not allowed to continue with the experiment.

Mouse Pre-training

Following the pretest, participants scoring below 60% began the mouse pre-training phase of the experiment. As stated in the Determining Frequency Aims section above, mouse pre-training was administered to ensure that dysfluent computer mouse use would not prevent participants from attaining the later fluency aims. Participants were informed to work as fast as they could on the task, and each completed one or more timings. The timings consisted of an identity-matching task in which buttons labeled A, B, C, D and E were present on the lower portion of the screen, while the letter A, B, C, D or E was presented (in random order) on the top portion of the screen. Participants were required to click on the button corresponding to the presented letter, and received feedback in the form of a 22,050-Hz tone presented for 181 ms for incorrect responses. If a participant responded correctly, no tone was presented and the next trial began immediately. After each 1-min timing, the participant received feedback on their performance. This feedback consisted of the participant's actual rate (in responses per minute) and accuracy (in percentage correct), and the rate and accuracy required to continue to the next phase of the experiment. If participants made 55 correct responses per minute while maintaining 90% accuracy during a timing, they began the next phase of the experiment; otherwise, timings were continued until these criteria were met.

Introductory Instruction

After the mouse pre-training, introductory instruction describing the terms fallacy, premise, and conclusion was presented. An example of the three terms was given, and participants were then required to identify the premise and conclusion in a sample argument. Participants answering incorrectly were required to repeat the exercise with new examples until an accurate response was obtained. After answering correctly, participants were informed that the premises and conclusions in the examples used during the experiment would be clearly identified for them. In addition, the convention of using letters to represent statements in arguments was described (e.g., “A is true” where “A” is standing in place of any statement, such as “Johnny is tall”).

Accuracy Training on Definitions

During the beginning of this phase, the definition of each fallacy was presented individually on the screen. The participants used a button labeled *continue* to advance to the next definition. After each definition had been presented once, the participants began accuracy testing on the definitions. The accuracy testing used a multiple-choice format identical to that described in the pretest phase. Rather than presenting examples of the fallacies, however, their definitions were presented in random order. This training continued until the participant completed one full set (all four fallacy definitions) at 100% accuracy.

Probe Test

After accuracy training, participants received a probe test to assess their ability to identify examples and non-examples of the fallacies. Since instructional design experts claim that learning a definition is usually not sufficient to produce true conceptual understanding (Tiemann & Markle, 1990), this test was conducted to determine whether participants would be able to correctly identify novel instances of the fallacies after learning only their definitions. The format and procedure of this test was identical to the pretest, and also used the items from Fallacy Set 1. Participants received no programmed feedback on their performance during this phase. After completing the 20-item probe test, participants progressed to the Practice phase of the experiment.

Practice

During this phase, participants received practice either at identifying examples of the fallacies or at identifying the definitions of the fallacies. Exact practice conditions and contingencies depended upon the experimental group to which the participant was assigned.

Fluency-Definitions Group. Participants in the fluency-definitions group received fluency-building practice on definition identification. Before beginning, participants were informed that they were going to practice identifying the definitions of the fallacies, that they would be completing 1-minute timings on the task, that they would hear a beep immediately after each incorrect answer, that they would receive feedback on their overall accuracy and speed after each timing, and that they should work as fast and accurately as possible. The practice task was identical to the accuracy testing involved in the accuracy training on definitions phase described above. Practice consisted of 1-min timings with rate and accuracy feedback provided after each timing. Feedback consisted of the participant's actual rate (in responses per minute) and accuracy (in percentage correct), and the rate and accuracy required to continue to the next phase of the experiment. If participants made 45 correct responses per minute while maintaining 90% accuracy during a timing, they began the next phase of the experiment; otherwise, timings were continued until these criteria were met.

Practice-Definitions Group. Participants in the practice-definitions group received practice on definition identification without a rate requirement. Before proceeding, participants were informed that they were going to practice identifying the definitions of the fallacies, that they would hear a beep immediately after each incorrect answer, and that they should work as fast and accurately as possible. The practice task was identical to the accuracy testing involved in the accuracy training on definitions phase described above. Practice sessions were not divided into 1-min timings; rather, participants continued with the task until they had completed the same total number of trials as their matched counterparts in the fluency-definitions group. For example, if Participant A in the fluency-definitions group needed 150 trials to meet the rate and accuracy criteria, his or her matched counterpart in the practice-definitions group would also be required to complete a total of 150 trials during the practice phase. The principal difference between the matched participants was that the member of the practice-definitions group was not required to meet a rate or accuracy requirement, and did not conduct practice in 1-min timings.

Fluency-Examples Group. Participants in the fluency-examples group received fluency-building practice on example identification. The first part of this phase involved accuracy training on the examples. The format and procedure of the accuracy training on the examples was similar to the pretest and the probe test, except that novel examples and non-examples from Fallacy Set 2 were used instead of Fallacy Set 1. Participants also received feedback after each response in the form of the words “Correct!” or “Incorrect!” appearing on the screen and, in the case of incorrect responses, the presentation of a brief (181 ms) 22,050-Hz tone and the correct answer.

After participants had responded to the entire set of 20 items in Fallacy Set 2, the accuracy of their responses was checked by the computer program. If they did not answer 90% or more of the items correctly, the entire set of items was re-presented (in a different random order). This process was repeated until the participant achieved 90% accuracy on one full set of items.

Once the accuracy criterion was met, participants were informed that they were now going to practice identifying examples of the fallacies, that they would be completing 1-minute timings on the task, that they would hear a beep immediately after each incorrect answer, that they would receive feedback on their overall accuracy and speed after each timing, and that they should work as fast and accurately as possible. The task remained the same as during the accuracy training on examples, except that practice now consisted of 1-min timings with rate and accuracy feedback provided after each timing. Feedback consisted of the participant's actual rate (in responses per minute) and accuracy (in percentage correct), and the rate and accuracy required to continue to the next phase of the experiment. Also, the words "Correct!" or "Incorrect!" and the correct answer no longer appeared after each response, but a brief tone was still presented for incorrect responses. Examples and non-examples from Fallacy Set 2 continued to be used during rate-building. If participants made 35 correct responses per minute while maintaining 90% accuracy during a timing, they began the next phase of the experiment; otherwise, timings were continued until these criteria were met.

Practice-Examples Group. Participants in the practice-examples group received practice on example identification without a rate requirement. They first received accuracy training on the examples identical to that received by participants in the fluency-examples group at the beginning of their Practice phase. Once participants met the accuracy criterion, they were informed that they were going to practice identifying examples of the fallacies, that they would hear a beep immediately after each incorrect answer, and that they should work as fast and accurately as possible. The task remained the same as during the accuracy training on examples, except that the words "Correct!" or "Incorrect!" and the correct answer no longer appeared after each response. A brief tone was presented immediately after each incorrect response, and the items from Fallacy Set 2 continued to be used. Practice sessions were not divided into 1-min timings; rather, participants continued with the task until they had completed the same total number of trials as their matched counterparts in the fluency-examples group. Only trials completed after the accuracy training counted toward this total.

Posttest

After meeting the requirements of the practice phase, all participants were then given the posttest. Participants were informed that they would not receive any feedback on their performance during this part of the experiment, and were asked to answer as accurately as they could. The posttest was identical to the pretest and probe test in format and procedure, but used entirely novel examples and non-examples from Fallacy Set 3. Participants received no programmed feedback on their posttest performance.

RESULTS

Table I provides descriptive statistics (including the mean, standard deviation, skewness, standard error of skewness, kurtosis, and standard error of kurtosis) for the pretest, probe test, and posttest scores for each experimental group. An alpha level of .05 was used for all of the statistical tests described below.

Table I. Descriptive Statistics for Pretest, Probe Test, and Posttest Scores[a] for the Experimental Groups

Group[b]	Test	Mean	*SD*	Skewness	SE of skewness	Kurtosis	SE of kurtosis
Fluency-Examples	Pretest	7.22	2.73	.27	.72	−1.38	1.4
	Probe Test	9.11	1.90	−.20	.72	−4.9	1.4
	Posttest	13.56	1.24	−.93	.72	1.37	1.4
Practice-Examples	Pretest	6.00	2.78	.54	.72	−.12	1.4
	Probe Test	10.56	2.88	1.38	.72	3.00	1.4
	Posttest	12.67	3.97	−.72	.72	.50	1.4
Fluency-Definitions	Pretest	5.67	2.06	.25	.72	−.75	1.4
	Probe Test	10.56	3.00	.73	.72	−.38	1.4
	Posttest	8.78	3.03	1.07	.72	.87	1.4
Practice-Definitions	Pretest	6.56	2.40	−.09	.72	−1.37	1.4
	Probe Test	10.00	2.55	−1.98	.72	3.98	1.4
	Posttst	11.11	3.22	.57	.72	.76	1.4

[a] Scores are number correct out of 20.
[b] $n = 9$ for each group.

Probe test scores, obtained after the participants were trained to accurately identify the fallacy definitions, were higher than the pretest scores for all but 4 of the 36 participants. Since the probe test was administered before manipulation of independent variables, all participants were treated as a single group in a one-factor repeated measures analysis of variance (ANOVA) conducted on the pretest and probe test scores. The probe test scores were found to be significantly higher than the pretest scores, $F(1, 35) = 63.31$, $p < .001$ with an effect size (η^2) of .64 and observed power of 1.00.

The posttest scores were evaluated using a two-factor mixed design ANOVA, with training (fluency or practice) serving as the repeated-measures factor and items (examples or definitions) serving as the independent-measures factor. Although a repeated-measures factor typically involves several measures being obtained from the same subject, in this analysis the scores for the repeated-measures factor were obtained from a pair of matched subjects (see the Participants and Setting section above). That is, rather than using a score for fluency training and practice training from the same participant, these scores were obtained from two different participants (one in a fluency condition and the other in a practice condition) who had been matched based on total number of practice trials completed. While this is uncommon, it does not violate the assumptions of the model (Howell, 1997, p. 494).

The results of the ANOVA revealed a significant effect for the type of items used (examples vs. definitions), $F(1, 16) = 7.84$, $p = .01$($\eta^2 = .33$, observed power = .75). No significant effect for the type of training (fluency vs. practice) was found, $F(1, 16) = .68$, $p = .42$($\eta^2 = .04$, observed power = .12). These two effects are illustrated in Fig. 1. The interaction between the two factors was not significant, $F(1, 16) = 3.38$, $p = .08$($\eta^2 = .18$, observed power = .41). It should be

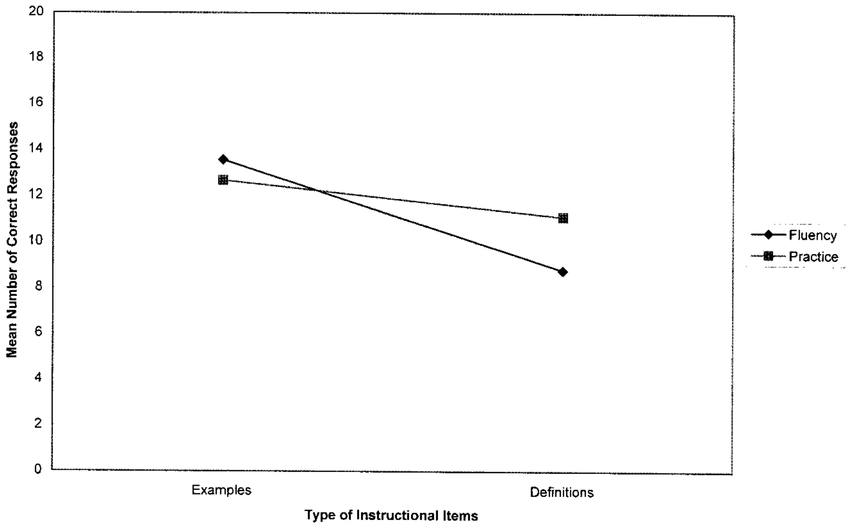

Fig. 1. Mean posttest scores.

noted with regard to the posttest scores that all 9 of the participants in the fluency-examples group and 6 of 9 participants in the practice-examples group achieved higher scores on the posttest than the probe test, while this was accomplished by only 4 of 9 participants in the practice-definitions and none of the participants in the fluency-definitions group. A related t-test conducted on the probe test and posttest scores of participants in the fluency-definitions group revealed that the posttest scores were significantly lower than the probe test scores, $t(8) = 2.87$, $p = .02$.

DISCUSSION

The results of the present study support the recommendation of instructional designers that conceptual material should be taught using multiple exemplars, but is less clear with regard to the comparison between fluency training and simple overtraining. These findings may prove useful for both educational researchers and educators who wish to improve their teaching of conceptual material or use of fluency techniques. In addition, this study illuminates several of the methodological challenges associated with conducting research on fluency-based concept instruction that should be addressed in future research in this area.

With a great deal of concept instruction emphasizing definition learning, it is perhaps heartening to learn that scores on the probe test were significantly higher than scores on the pretest. This implies that definition learning alone had a positive effect on the participants' ability to identify novel examples of the fallacies. The fact that the overall mean percent correct for all participants on the probe test was a meager 50%, however, suggests serious limitations for definition learning in promoting conceptual understanding.

An analysis of the posttest scores, which revealed that participants who received practice on identifying instances of the concepts performed significantly better than those who received practice on identifying the definitions of the concepts, suggests conceptual understanding can be enhanced with multiple exemplar training. This finding is not surprising given the solid empirical foundation on which the recommendation to use multiple exemplars rests (see Clark, 1971; Gagne & Brown, 1961) and the success of instructional systems that use such training (e.g., Becker & Carnine, 1980; Horner & Albin, 1986). Nevertheless, the finding is revealing, given the heavy emphasis placed on definitions in most instructional materials (including textbooks, flash cards, etc.). In essence, this finding supports the notion that there should be an alignment between instructional objectives, the type of activities on which students receive instruction and practice, and assessment instruments. When one's instructional objectives include being able to identify instances of a concept, then students should be provided the opportunity to practice such behavior.

Two additional points should be made about the higher posttest scores achieved by the participants exposed to examples during training. First, while most of the participants (15 of 18) in the examples groups recorded a higher score on the posttest than the pretest, few of the participants (4 of 18) in the definitions groups did the same. Therefore, not only did the posttest scores for the definitions participants not improve *significantly*, for most of these participants they did not improve *at all*. In fact, participants in the fluency-definitions group had posttest scores that were significantly lower than their probe test scores, $t(8) = 2.87$, $p = .02$. This suggests that "overtraining" on definitions (either via fluency building or simple repeated practice) has little effect—and possibly even a deleterious effect—on the learner's understanding of the concept (especially when only a limited number of definitions are being taught at once; see below). Such a finding directly challenges the utility of definition-based fluency procedures, such as SAFMEDS, for promoting conceptual understanding. Of course, if students are tested only on their knowledge of the concept's definition, this negligible or negative effect would be rarely detected. Further investigation into the effects of fluency-based definition learning procedures on concept formation is clearly warranted.

Second, while the scores for the participants in the examples group are significantly higher than the scores for the participants in the definitions groups, they are still not large. The mean posttest score for the participants in the examples group was only 13.11, or 66% correct, obviously leaving much room for improvement. This lends credence to the following warning by Tiemann and Markle (1990): "The phrase 'Minimum Rational Set' does NOT mean 'a set adequate for instruction.' To teach a complex concept, you may need far more cases to bring students to mastery than those specified in the minimum rational sets" (p. 120). The present experiment used only as many examples and non-examples as prescribed by the MRS, and it was clearly insufficient for producing complete mastery of the concepts.

The posttest scores for those participants who received fluency training were not found to be significantly different from the scores of their matched counterparts in the practice groups, but this comparison is best deemed inconclusive. This is because the power of the experimental design and data analytic plan employed was very low (observed power estimate = .12), greatly reducing our ability to detect any true differences between the fluency and practice groups. Further, the use of randomly matched pairs of participants to control for practice effects is probably not optimal, as there is typically considerable individual variation in the number of trials required to reach fluency (from 152 to 689 for participants in the Fluency-Definitions group in the present study, for example). A more powerful design might be to use a within-subjects control in which each participant learns two separate sets of concepts (of relatively comparable complexity), one with a rate requirement (fluency training) and one without (matched practice). Controlling for response opportunities (practice) in research on behavioral fluency is important, as recent evidence suggests that rate-building procedures, while perhaps providing a more efficient means of practice, do not necessarily enhance learning beyond the effects of practice alone (Haag, O'Shields, & Chase, 2002).

Future studies in this area should also give careful consideration to the number of items in the instructional pool used during fluency training. As mentioned previously, students are more likely to attend to purely structural features of questions during fluency training in order to respond at the high rates required. Increasing the number of items in the instructional pool can reduce the student's ability to memorize structural features of the questions, perhaps increasing discrimination based on the critical, non-structural features. In the present study, for example, participants in the examples groups may have performed better on the posttest partly because the larger number of items in their instructional pool (20, as compared to 4 for the participants in the definitions groups) forced them to attend more to the meaning—rather than the structure—of the questions. The negative effect of fluency training with definitions found in the present study might also be ameliorated if a larger number of terms and definitions were used. Parametric analyses of such variables may provide educators and instructional designers with a better sense of the minimum number of instructional items that should be used during fluency-based instruction.

In addition to the issues addressed above, several other methodological considerations are highlighted by this study. Massed versus distributed practice, while receiving considerable attention in the traditional educational literature, should be more closely examined in relation to fluency training. Participants in the present study completed all instructional activities in one day, while most fluency programs distribute practice over several days, weeks, months, or even years. The process of determining fluency rates should also be modified, when possible, to meet the functional definition of fluency by incorporating measurement of the REAPS outcomes. Although this can be a lengthy and complicated process, it may improve the construct validity of future experimental analyses. Further, terminal response rates from practice conditions should be measured to allow researchers to determine whether terminal rate of responding is, in fact, an important variable in predicting learning outcomes. The present study did not collect such data, and it is possible that participants in the practice conditions were responding at rates comparable to those of participants in the fluency conditions, despite the lack of programmed contingencies to do so.

The current study takes an important step toward understanding how the techniques of fluency training and systematic concept instruction can be combined to improve the teaching of complex verbal concepts. A great deal of education, both formal and informal, involves the learning of such concepts, so it is valuable to understand how this instruction can become more effective and efficient. Although this study is but a preliminary analysis of these issues, the results indicate that the most common way in which fluency techniques are currently used to teach verbal concepts—requiring the identification or recitation of definitions at a high rate—may have minimal impact on the learner's understanding of the concept. It seems evident that further experimental examination of fluency training, concept instruction, and the combination of the two can yield information important to the improvement of instructional practices.

ACKNOWLEDGMENTS

This study formed part of a thesis submitted by the first author in partial fulfillment of the requirements for a master's degree at the University of Nevada, Reno. We thank Elizabeth Athens for her help in conducting the research and the members of the thesis committee, Thomas E. Boyce and Melissa Olive, for their advice and support. The helpful comments of several anonymous reviewers and Carl Binder are also greatly appreciated. A complete script of the instructional materials used in this study can be obtained by contacting the first author.

REFERENCES

Beck, R., & Clement, R. (1991). The Great Falls Precision Teaching Project: An historical examination. *Journal of Precision Teaching, 8*(2), 8–12.

Becker, W. C., & Carnine, D. W. (1980). Direct Instruction: An effective approach to educational intervention with the disadvantaged and low performers. In B. B. Lahey & A. E. Kazdin, (Eds.), *Advances in Clinical Child Psychology* (Vol. 3, pp. 429–473). New York: Plenum Publishing.

Binder, C. (1988). Precision teaching: Measuring and attaining exemplary academic achievement. *Youth Policy*, *10*(7), 12–15.

Binder, C. (1990, September). Closing the confidence gap. *Training*, pp. 49–56.

Binder, C. (1991). Morningside Academy: A private sector laboratory for effective instruction. *Future Choices*, *3*(2), 61–66.
Binder, C. (1993). Behavioral fluency: A new paradigm. *Educational Technology, 33*(10), 8–14.
Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19,* 163–197.
Binder, C. & Bloom, C. (1989, February). Fluent product knowledge: Application in the financial services industry. *Performance and Instruction*, pp. 17–21.
Carney, J. D., & Scheer, R. K. (1980). *Fundamentals of logic* (3rd ed.). New York: Macmillan Publishing Co., Inc.
Clark, D. C. (1971). Teaching concepts in the classroom: A set of teaching prescriptions derived from experimental research. *Journal of Educational Psychology, 62,* 253–278.
Copi, I. M. (1982). *Introduction to logic* (6th ed.). New York: Macmillan Publishing Co., Inc.
Engelmann, S. (1969). *Conceptual learning.* Belmont, CA: Fearon Publishers.
Engelmann, S., & Carnine, D. (1991). *Theory of instruction: Principles and applications* (Revised ed.). Eugene, OR: ADI Press.
Eshleman, J. W. (1985). Improvement pictures with low celerations: An early foray into the use of SAFMEDS. *Journal of Precision Teaching*, *6*(3), 54–63.
Evans, S. S., & Evans, W. H. (1985). Frequencies that ensure skill competency. *Journal of Precision Teaching*, *6*(2), 25–30.
Gagné, R. M. (1970). *The conditions of learning.* (2nd ed.). New York: Holt, Rinehart & Winston.
Gagné, R. M., & Brown, L. T. (1961). Some factors in the programming of conceptual learning. *Journal of Experimental Psychology, 62,* 313–321.
Gibbons, A. S., & Fairweather, P. G. (1998). *Computer-based instruction: Design and development.* Englewood Cliffs, NJ: Educational Technology Publications.
Gray, W. D. (1991). *Thinking critically about New Age ideas*. Belmont, CA: Wadsworth Publishing Company.
Haag, S. S., O'Shields, E., & Chase, P. N. (2002). *A literature review of empirical studies of rate-building and practice*. Paper presented at the annual meeting of the Association for Behavior Analysis, Toronto.
Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Kluwer Academic/Plenum Publishers.
Horner, R. H., & Albin, R. W. (1986). Generalization with precision: The role of negative teaching examples in the instruction of generalized grocery item selection. *Journal of the Association for Persons with Severe Handicaps*, *11*(4), 300–308.
Howell, D. C. (1997). *Statistical methods for psychology*. (4th ed.). Belmont, CA: Duxbury Press.
Howell, K. W., & Lorson-Howell, K. A. (1990). What's the hurry?: Fluency in the classroom. *Teaching Exceptional Children*, *22*(3), 20–23.
Johnson, K. R., & Layng. T. V. (1992). Breaking the structuralist barrier: Literacy and numeracy with fluency. Special Issue: Reflections on B. F. Skinner and psychology. *American Psychologist*, *47*(11), 1475–1490.
Johnson, K. R., & Layng, T. V. (1994). The Morningside model of generative instruction. In R. Gardner, D. M. Sainato, J. O. Cooper, T. E. Heron, W. L. Heward, J. W. Eshleman, and T. A. Grossi (Eds.), *Behavior analysis in education: Focus on measurably superior instruction* (pp. 173–197). Belmont, CA: Brooks/Cole.
Johnson, K. R., & Layng, T. V. (1996). On terms and procedures: Fluency. *The Behavior Analyst*, *19,* 281–288.
Johnson, D. M., & Stratton, R. P. (1966). Evaluation of five methods of teaching concepts. *Journal of Educational Psychology*, *57,* 48–53.
Koorland, M. A., Keel, M. C., & Ueberhorst, P. (1990). Setting aims for precision learning. *Teaching Exceptional Children*, *22*(3), 64–66.
Korinek, L., & Wolking, B. (1984). Study methods in graduate school. *Journal of Precision Teaching*, *5*(3), 64–66.
Laberge, D., & Samuels, J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6,* 293–323.

Lepore, E. (2000). *Meaning and argument: An introduction to logic through language.* Malden, MA: Blackwell Publishers Inc.
Lindsley, O. R. (1971). From Skinner to precision teaching: The child knows best. In J. B. Jordan & L. S. Robbins (Eds.), *Let's Try Doing Something Else Kind of Thing* (pp. 1–11). Arlington, VA: The Council for Exceptional Children.
Lindsley, (1990). Precision teaching: Discoveries and effects. *Journal of Applied Behavior Analysis, 25*, 51–57.
Lindsley, O. R. (1996). The four free-operant freedoms. *The Behavior Analyst, 19,* 199–210.
Markle, S. M. (1969). *Good frames and bad: A grammar of frame writing* (2nd ed.). New York: John Wiley & Sons, Inc.
McDade, C. E., & Goggans, L. A. (1993). Computer-based precision learning: Achieving fluency with college students. *Education and Treatment of Children, 16*(3), 290–305.
McDade, C. E., & Olander, C. P. (1987, March). Precision management of instructional technology: A program update. *Educational Technology*, 44–46.
McDade, C. E., Rubenstein, S. B., & Olander, C. P. (1983). Parallel between frequency testing and performance on essay questions in a theories of personality course. *Journal of Precision Teaching, 4*(1), 1–5.
McDade, C. E., Willanzheimer, L. A., & Olander, C. P. (1981). *Precision teaching in a psychological statistics course: Parallels between frequency testing and performance on word problems*. Invited paper presented at the meeting of the Association for Behavior Analysis, Milwaukee, WI.
McGreevy, P. (1983). *Teaching and learning in plain English* (2nd ed.). Kansas City, MO: Plain English Publications.
Merrill, M. D., Tennyson, R. D., & Posey, L. O. (1992). *Teaching concepts: An instructional design guide* (2nd ed.). Englewood Cliffs, NJ: Educational Technology Publications.
Munson, K. (1998). *Effects of celeration rate on behavioral fluency*. Unpublished doctoral dissertation, West Virginia University.
Olander, C. P., Collins, D. L., McArthur, B. L., Watts, R. O., & McDade, C. E. (1986). Retention among college students: A comparison of traditional versus precision teaching. *Journal of Precision Teaching*, (4), 80–82.
Parsons, J. A. (2000). Think fast [Computer software]. Victoria, BC: Author.
Pennypacker, H. S., & Binder, C. (1992, January). Triage for American education. *Administrative radiology*, 18–25.
Polson, D. A. (1995). Fostering multiple repertoires in undergraduate behavior analysis students. *The Behavior Analyst, 18*(2), 293–299.
Potts, L., Eshleman, J. W., & Cooper, J. O. (1993). Ogden R. Lindsley and the historical development of precision teaching. *The Behavior Analyst, 16,* 177–189.
Thomas, S. N. (1986). *Practical reasoning in natural language* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
Tiemann, P. W., & Markle, S. M. (1990). *Analyzing instructional content: A guide to instruction and evaluation*. Champaign, IL: Stipes Publishing Company.