J. EDUCATIONAL COMPUTING RESEARCH, Vol. 37(3) 307-330, 2007

COMPARING STRATEGIES FOR TEACHING ABSTRACT CONCEPTS IN AN ONLINE TUTORIAL

ERIC J. FOX
Western Michigan University

HOWARD J. SULLIVAN
Arizona State University

ABSTRACT

The purpose of this study was to compare traditional classification training for a set of abstract concepts with multiple-relations training consisting of inference practice and the use of a content diagram. To examine this, 200 undergraduate and graduate psychology students completed a Web-based tutorial covering the abstract concepts of a psychological theory of language and cognition. All participants received the same core instructional content and practice activities varied by experimental condition: some participants received classification training, some received multiple-relations training, some received a combination of both, and some received neither. Performance on a posttest with three subsections was evaluated. Participants who received classification training were significantly better at identifying new instances of the concepts than participants who did not. Neither classification training nor multiple-relations training had a significant effect on ability to identify concept definitions or answer application questions. Implications for the development of instruction for abstract concepts are discussed.

Concepts are often considered "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). It is not surprising, then, that most models for the design of concept instruction

doi: 10.2190/EC.37.3.e
http://baywood.com

emphasize the ability to classify instances of the concepts of interest. In particular, these models stress exposing learners to a carefully selected and sequenced series of examples and non-examples to prevent classification errors such as overgeneralization, undergeneralization, and misconceptions (Engelmann, 1969; Gagné, 1965; Markle, 1969). Additional strategies, such as using attention-focusing devices to isolate critical attributes, providing elaborative feedback on interrogatory examples, and focusing on one or two "prototype" or best examples have also been provided (Merrill, Tennyson, & Posey, 1992), but there remains a strong emphasis on teaching and evaluating classification performance.

Recent research and theory suggests that an exclusive focus on categorization in concept instruction may not be sufficient to produce or measure expert knowledge in a domain, however. There are important limitations to relying solely on categories or stimulus classes to account for how a word or concept is used in natural language settings. Research on the cross-classification of concepts (e.g., Ross & Murphy, 1999), the multiple functions of concepts (e.g., Markman, 1981), and derived stimulus relations (e.g., Hayes, Barnes-Holmes, & Roche, 2001) indicate that a broader instructional focus on the multiple functions and relations of concepts may be useful. This may be especially true for complex, abstract concepts (e.g., "justice") because they are defined more by their relations to other concepts than by the physical properties of their instances (Stewart, Barnes-Holmes, Hayes, & Lipkens, 2001).

A number of strategies for concept instruction that go beyond classification training have been offered. Many of these strategies explicitly or implicitly focus on the multiple semantic relations or functions associated with a concept. While multiple-exemplar training is typically still prescribed, techniques such as teaching with analogies, incorporating concepts maps and graphic organizers, and providing use and inference practice are also often recommended (e.g., Tessmer, Wilson, & Driscoll, 1990). Such strategies may be useful because they teach multiple relations between relevant terms and concepts, provide the learner with the opportunity to practice deriving relations among concepts, and enhance the contextual control over these relations.

These newer instructional strategies have received varying degrees of empirical support and attention. Analogies attempt to enhance knowledge of a less familiar domain by comparing it along some dimension or dimensions to a more familiar domain. Researchers have begun examining the impact of analogies on concept learning, particularly in science education (see Dagher, 1995; Duit, 1991 for reviews). Some studies have found analogies to be useful for enhancing conceptual understanding or correcting misconceptions (e.g., Baker & Lawson, 2001; Brown & Clement, 1989; Dupin & Johsua, 1989; Stavy, 1991), but others have failed to detect a beneficial effect (e.g., Gilbert, 1989) or have highlighted the potential dangers and detrimental effects of analogies (e.g., Gentner & Gentner, 1983; Glynn, Britton, Semrud-Clikeman, & Muth, 1989; Venville & Treagust, 1997). Overall, instructional analogies seem to be useful, but it is difficult to draw firm conclusions because the studies in this area often use different ways of employing analogies, different types of analogies, and different evaluation methods (Dagher, 1995).

The research on the instructional use of concept maps, hierarchical displays, semantic networks, tree diagrams, advance organizers, and related techniques can also be difficult to interpret. There are subtle differences between these instructional strategies, but it is useful to group them together because they all share the unifying element of "the visuospatial arrangement of information containing words or statements that are connected graphically to form a meaningful diagram" (Horton, Lovitt, & Bergerud, 1990, p. 12). The literature on these methods can be described as both vast (e.g., Al-Kunifed & Wandersee, 1990, provide 100 references on concept mapping alone) and "conflicting and confusing" (Story, 1998, p. 255), due in large part to the wide variety of techniques used, dependent measures examined, and effects found.

While the results of the research on concept maps and related techniques can be equivocal, some degree of general support for these methods can be found in reviews and meta-analyses of the literature. One meta-analysis of 19 studies (Horton et al., 1993), for example, found that concept mapping raised student achievement and also improved student attitudes toward instruction. Other meta-analyses have found that graphic organizers are generally effective for improving both vocabulary knowledge and comprehension (Moore & Readence, 1984), and that advance organizers have a small facilitative effect on learning and retention (Luiten, Ames, & Ackerson, 1980).

Inference practice has probably received the least empirical attention of the new strategies for concept instruction. Inferential questions are commonly used as a dependent measure, particularly with studies of reading comprehension, but are rarely employed as an independent variable. There has been some analysis of the effects of practice on inferential questions on comprehension ability (e.g., Hansen, 1981), but no studies could be found that directly examine the relationship between inference practice and measures of conceptual understanding such as classification and application skill.

There do not appear to be any systematic examinations of whether combining these newer strategies with traditional classification training improves concept learning. It is also not clear whether these newer strategies alone, without classification training, can positively influence concept learning. Furthermore, it is possible that these newer strategies and classification training influence different measures of concept learning differently. For example, how do these newer strategies alone affect performance on a classification task? How does classification training alone affect the learner's ability to identify concept definitions or answer application questions? Does combining the newer strategies with classification training affect performance on different measures of concept learning differently?

The present study was designed to address such issues. In particular, the study investigated the effects of four different strategies for teaching abstract concepts on three different measures of concept learning. A number of newer strategies for concept instruction require the use of specific software (e.g., Fisher, 1990) or require extensive instructor and/or learner training on how to use the strategy or tool (e.g., Elhelou, 1997). This study focused on strategies that can be readily implemented by most teachers and instructional designers. Specifically, the use of inferential questions and a diagram of the instructional content were employed as practical strategies for emphasizing the relations among the concepts.

These instructional strategies were examined in the context of a Web-based instructional program focused on the abstract technical concepts of a psychological theory of language known as Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001). Abstract verbal concepts are often difficult to teach and seem especially amenable to strategies that emphasize the relations among them. The core concepts of RFT can be particularly challenging, and may benefit from the graphical and animation capabilities of Web-based instruction.

All participants in the study received instruction that included the presentation of the definition and expository examples of the concepts, along with simple practice identifying the definitions. Few instructors would attempt to teach abstract concepts without providing such information, so it was included for all participants to prevent any groups from receiving artificially poor instruction. Treatment groups differed primarily in terms of the kind of practice activities to which participants were exposed.

Participants were randomly assigned to one of four groups in a 2 (Classification Training: Present and Absent) × 2 (Multiple-Relations Training: Present and Absent) factorial design (Campbell & Stanley, 1963). In the *classification group*, participants were exposed to interrogatory examples and non-examples of each concept, providing them with classification training in accord with traditional guidelines for concept instruction (e.g., Merrill et al., 1992). Participants in the *multiple-relations group* were exposed to the newer strategies described above, specifically inference practice (some of which was analogical in nature) and the presentation of a diagram of the instructional content. In the *combined strategies group*, participants were exposed to a combination of all of the instructional elements included in the classification and multiple-relations group. Finally, the *control group* received neither classification training nor multiple-relations training, but were still exposed to definitions, expository examples, and practice identifying concept definitions.

Three different measures of concept learning served as the primary dependent variables in the study. These measures represent common ways of measuring conceptual understanding, and correspond to different types of learning outcomes commonly identified by educational psychologists (e.g., Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Gagné, 1965). They include the classification of novel examples of the concepts, the identification of the concepts' definitions, and answering questions about applying the concepts to relevant domains.

Classification was included as a dependent measure because it is a standard and important measure of conceptual understanding. In Bloom's (1956) taxonomy of learning outcomes, classification skill would go under the category of knowledge, while in Gagné's (1965) taxonomy it would be considered an intellectual skill (specifically, a skill related to defined concepts). Definition identification was included because it is a common measure of conceptual learning in the classroom, and because its validity is supported by research that shows definition learning can facilitate conceptual understanding (e.g., Anderson & Kulhavy, 1972; Tennyson & Park, 1980). Learning definitions would also be categorized as knowledge under Bloom's taxonomy, but would be considered verbal information in Gagné's system. Application questions are included because they may provide a somewhat more socially valid measure of concept learning and also reflect the learner's ability to use and make additional inferences about concepts (Tessmer et al., 1990). The application questions in this study would be categorized as application knowledge in Bloom's taxonomy and considered an intellectual skill involving higher-order rules in Gagné's scheme.

Data on practice performance, attitudes, and time on task were also collected and analyzed for all participants.

The primary research questions for the study were as follows:

- Does instructional strategy (classification training versus multiple-relations training versus a combination of the two versus neither) affect learner ability to identify novel examples of abstract concepts?
- Does type of instructional strategy affect learner ability to identify the definitions of abstract concepts?
- Does type of instructional strategy affect learner ability to answer application questions about abstract concepts?
- Does type of instructional strategy affect practice performance?
- Does type of instructional strategy affect time on task?
- Does type of instructional strategy affect learner attitudes toward instruction on abstract concepts?

METHOD

Participants

A total of 200 psychology students (150 female, 50 male) from several English-speaking countries participated in this study: 175 from the United States, 10 from Canada, 6 from the United Kingdom, 6 from the Netherlands, and 3 from Sweden. Participants were enrolled in 17 different psychology courses (8 undergraduate courses, 8 graduate courses, and 1 combined undergraduate and graduate course) and completed the tutorial as an assignment or extra-credit project for their course. The age of the participants ranged from 18 to 57, with a mean of $M = 25.36$, and 116 of the participants were undergraduates and 84 were graduate students. All participants rated their knowledge of Relational Frame Theory (RFT) with a score of 1 or 2 using a 5-point Likert scale (1 = *not at all knowledgeable*, 5 = *very knowledgeable*). Students who rated their knowledge of RFT with a 3 or higher were excluded from the study.

Materials

All of the materials for this study were developed and incorporated into a Web-based interface using Macromedia® Flash® software. PHP, a scripting language especially suited for Web development, was used to allow the Flash program to interface with a MySQL database on the Web server. The MySQL database was used to store learner performance data and related information.

The study consisted of several core elements. An instructional program, a knowledge quiz, an attitude survey, and an instructor control panel were all embedded within the Flash interface and accessible from a Web site maintained by the author. In addition to these core elements, the Web site for the tutorial included screenshots and additional information about the purpose, nature, and length of the tutorial for students and instructors.

Instructional Program

Four parallel versions of a Web-based instructional program entitled “An Introduction to Relational Frame Theory” were developed for this study using Flash. The program was written for a general audience, with little prerequisite knowledge of psychology required. The instructional program focused on the basic approach of RFT and 12 abstract concepts that are key to understanding it: symbolism of language, generativity of language, functional contextual theories, derived stimulus relations, relational responding, generalized operants, multiple exemplar training, mutual entailment, combinatorial entailment, transformation of stimulus functions, and two forms of contextual control known as C_{rel} and C_{func}. All instructional materials, including assessments and practice activities, were reviewed extensively for scope and accuracy by Dr. Steven C. Hayes of the University of Nevada, Reno and Dr. Dermot Barnes-Holmes of the National University of Ireland, Maynooth. Both Dr. Hayes and Dr. Barnes-Holmes are internationally recognized experts on RFT.

All four versions of the instructional program contained the same basic content. For each major concept, a definition, expository examples, and a four-option multiple-choice question about the definition were provided. Expository examples included both verbal examples and, when applicable, graphical and/or animated examples. Interactive elements, such as the use of drag-and-drop activities and hotspot roll-overs to reveal additional information, were also included in the instruction when appropriate. Immediately following the expository examples for a concept, a multiple-choice question about the concept's definition was presented. The core instructional content of the tutorial was common to all treatment groups. This content comprised a total of 478 screens.

Additional practice activities varied according to the experimental group and tutorial version. Each parallel version of the tutorial contained the instruction described above, and three of the four versions included additional materials or activities that reflect the different treatment conditions (the version for the control group contained only the instruction described above). These additional materials or activities are described below.

All of the multiple-choice practice questions (including definition, classification, and inferential questions) used in the various versions of the tutorial shared several features. These included the use of four answer options, the provision of elaborative feedback and knowledge of correct results, and the implementation of a 10-second postfeedback delay for incorrect answers. Elaborative feedback was provided because it is often recommended for the isolation of critical attributes during classification training (Merrill et al., 1992; Tennyson & Cocchiarella, 1986). The postfeedback delay was used because it has been shown to improve performance in computerized programmed instruction (Crosbie & Kelly, 1994; Kelly & Crosbie, 1997; Munson & Crosbie, 1998).

Classification Group

For each of the 12 major concepts, the version of the program for the classification group contained two multiple-choice questions that required the learner to identify examples of the concept. A sample classification question is provided below:

> Which of the following is the best example of the *generativity* of language?
> a) Mr. Tonai wrote a headline that was identical to one he had seen before.
> b) Mr. Tonai had never seen that particular headline before, but he immediately understood it.
> c) Mr. Tonai had seen that headline several times before, and he immediately understood it.
> d) Mr. Tonai wrote a headline that was complete nonsense, even to him.

Although only two questions were provided for each concept, the multiple-choice format exposed the learner to one example and three non-examples per question. Whenever possible, distracters for the classification questions (i.e., the non-examples) were examples of other concepts in the tutorial.

The classification questions were presented immediately following presentation of the concept's definition, expository examples, and definition question. A total of 24 classification questions were presented throughout the tutorial. This version of the tutorial did not include any inferential questions or a diagram of the instructional content.

Multiple-Relations Group

The tutorial version for the multiple-relations group included a total of 24 inferential questions about the concepts, along with a diagram of the instructional content. The inferential questions focused primarily on how the concepts are related to one another in multiple ways. A sample inferential question is provided below:

How are *mutual entailment* and *combinatorial entailment* different?

a) Mutual entailment requires at least three related stimuli or events, whereas combinatorial entailment requires only two.
b) Combinatorial entailment is a property of relational frames, but mutual entailment is not.
c) Combinatorial entailment requires at least three related stimuli or events, whereas mutual entailment requires only two.
d) Mutual entailment is a property of relational frames, but combinatorial entailment is not.

Two of the inferential questions were also analogical in nature, requiring the learner to select the concepts that best complete an analogy in the general form of "concept A is to __________ as concept B is to __________." Because the inferential questions focused on the relations among the concepts, these questions were not evenly distributed throughout the tutorial (early in the tutorial, it is not possible to ask inferential questions about how the current concept relates to concepts that have not been presented yet). For most concepts, one to three inferential questions were presented, and the inferential questions were always presented immediately following presentation of the concept's definition, expository examples, and definition question.

Early in the tutorial, and again at the beginning of each of the 12 major lessons, a diagram of the instructional content was presented to illustrate some of the ways the concepts are related to one another. When first presented, the diagram was gradually revealed from top to bottom, allowing the learner to read each component as it appears. Subsequent presentations of the diagram had the box representing the next lesson or concept highlighted and were accompanied by instructions to review the diagram and notice the highlighted box's relation to the other concepts. Learners were also told to click on the highlighted box to begin learning about that topic. This version of the tutorial did not include any classification questions about the concepts.

Combined-Strategies Group

The tutorial version for the combined-strategies group included a combination of the practice elements from the versions used for the classification and multiple-relations groups. In particular, it included a total of 12 classification questions, 12 inferential questions, and a diagram of the instructional content. The classification and inferential questions were randomly selected from the pool of questions used for the classification and multiple-relations groups. One classification question was provided for each concept, and the inferential questions were presented in the same position and sequence as they were in the version for the multiple-relations group. For each concept, the definition, expository examples, definition question, classification question, and inferential questions (if applicable for that concept) were presented in that order. The diagram of the instructional content was presented early in the tutorial, and again at the beginning of each of the 12 major lessons, just as it was in the version for the multiple-relations group.

Control Group

As mentioned previously, the version of the tutorial for the control group contained only the concept's definition, expository examples, and a definition question. This version of the tutorial did not include any classification questions, inferential questions, or diagrams of the instructional content.

The four treatment conditions described above comprised a 2 (Classification Training: Present and Absent) × 2 (Multiple-Relations Training: Present and Absent) experimental design.

Knowledge Quiz

A 36-item multiple-choice quiz was developed to assess how well participants learned the key concepts of RFT, according to three different measures of conceptual understanding. The quiz consisted of 12 questions about the definitions of the concepts (one question for each major concept), 12 questions requiring the classification of novel examples of the concepts (one question for each major concept), and 12 application questions about how the concepts might be used to interpret natural language interactions (two interaction scenarios with six questions about each). The 24 classification and definition questions were presented in random order first, followed by the 12 application questions.

Participants were able to choose only one answer for each question and were not able to go back and change their answers to previous questions. Feedback on overall quiz performance (total correct) was provided at the end of the quiz. To enhance motivation, at the beginning of the tutorial and again just prior to beginning the quiz, participants were told that they must do well on the quiz in order to receive course credit for completing the tutorial. Internal reliability of the posttest, using Cronbach's alpha, was .87.

Attitude Survey

A 15-item survey was developed to measure participant attitudes toward the instructional program. For 12 of the survey items, respondents used a 5-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*) to rate their attitudes. Three of these items focused on preference for the tutorial in general, three focused on the quality and relevance of the practice questions, three focused on how much was learned about RFT, and the last three focused on miscellaneous issues. One of the miscellaneous Likert-type items varied by treatment condition. Participants in the control and classification groups were given "A diagram of how the concepts are related to one another would have helped me understand the material better," while participants in the multiple-relations and combined-strategies group were given "The diagram of the concepts helped me understand the material." In addition, all participants were asked to rate the length of the program as either too long, too short, or about the right length. The last two items were open-ended questions asking the participants what they liked most and least about the tutorial. The alpha reliability coefficient for the 12 Likert-type items was .94.

Instructor Control Panel

To facilitate use of the tutorial as an assignment or extra-credit project for courses, an instructor control panel was developed. Using this control panel, instructors who wished to use the tutorial in their course were able to create an instructor account, register courses, and view which students in their courses had completed the tutorial and their quiz scores. When registering a course, instructors were asked to provide details about the course, such as its title, level, and number of students. A unique Course Access Code was automatically generated for each course, and instructors gave this code to their students to ensure that their tutorial performance was associated with the correct course.

PROCEDURES

Participants were able to access the tutorial's web page from any computer with Web access. They were advised of the approximate length and general procedure for completing the tutorial before beginning. When they first accessed the tutorial, participants were required to register a student account. During this registration process, participants created a unique username and password and provided demographic information about themselves. Participants also indicated their level of training in psychology and general educational background, and rated their knowledge of RFT. They were also required to enter the Course Access Code for their course. Upon completion of the registration process, an email confirmation (containing their username and password) of their registration was automatically emailed to them and the program randomly assigned them to one of the four experimental conditions.

The tutorial was divided into two parts of approximately equal length. Participants were informed that they must complete each part in one sitting, but that they did not need to complete both parts on the same day. Participants were also informed that the second part of the tutorial must be completed within one week of completion of the first part. Upon completion of the first part, participants were given the option to continue on to the second part, or to log out and complete the second part later. If they chose to complete the second part later, the deadline for completing the second part was shown on the screen and emailed to them. They were also informed that if they missed the deadline, they would be required to complete the first part again before beginning the second part. To complete the second part later, participants logged in to the tutorial using the username and password they entered during the registration process.

After completing the instruction for both parts of the tutorial, participants took the knowledge quiz and completed the survey. Upon completion of the survey, a tutorial completion confirmation email containing both the participant's quiz score and a randomly generated eight-digit completion confirmation number was sent to the student. The completion confirmation email served as a safeguard against database errors, as it provided additional evidence of their completion and quiz score. During the tutorial, the program also automatically recorded participant performance on practice questions, overall progress, and time in program to the MySQL database on the Web server.

DESIGN AND DATA ANALYSIS

Student achievement was analyzed using a 2 (Classification Training: Present and Absent) × 2 (Multiple-Relations Training: Present and Absent) analysis of variance (ANOVA) for the mean total posttest scores. Mean scores on the three subsections of the posttest were also analyzed using a 2 (Classification Training) × 2 (Multiple-Relations Training) × 3 (Question Type: Definition, Classification, and Application) multivariate analysis of variance (MANOVA).

Performance on practice items was analyzed using a 2 (Classification Training) × 2 (Multiple-Relations Training) ANOVA for the 12 definition questions common to all four experimental groups. The 12 classification questions common to the classification group and the combined-strategies group were analyzed using an independent-samples *t* test, as were the 12 inferential questions common to the multiple-relations group and the combined-strategies group. Total time in program was analyzed using a 2 (Classification Training) × 2 (Multiple-Relations Training) ANOVA. Responses to each item on the attitude survey were also analyzed using 2 × 2 ANOVAs.

RESULTS

Results are discussed below by achievement, practice performance, time in program, and participant attitudes.

Achievement

The mean scores and standard deviations for the posttest are shown in Table 1. The mean total score was 26.60 (74%) for participants in the control group, 29.28 (81%) for participants in the classification group, 26.58 (74%) for participants in the multiple-relations group, and 27.08 (75%) for participants in the combined-strategies group. A 2 (Classification Training: Present and Absent) × 2 (Multiple-Relations Training: Present and Absent) analysis of variance (ANOVA) of posttest total scores indicated that the interaction between classification training and multiple-relations training was not significant, $F(1, 196) = 1.64$, $p = .20$, partial $\eta^2 = .01$. A significant main effect for classification training was not detected, $F(1, 196) = 3.49$, $p = .06$, partial $\eta^2 = .02$, nor was a significant main effect for multiple-relations training, $F(1, 196) = 1.70$, $p = .19$, partial $\eta^2 = .01$. Table 2 summarizes these findings.

The mean scores for each of the three question types on the posttest are also displayed in Table 1 and illustrated in Figure 1. Correlation coefficients were computed among these scores. Scores on the definition questions were strongly correlated with scores on the classification questions, $r = .71$, $p < .01$, and with scores on the application questions, $r = .59$, $p < .01$. Scores on the classification questions were also strongly correlated with scores on the application questions, $r = .69$, $p < .01$.

Scores on the three subsections of the posttest were analyzed using a 2 (Classification Training) × 2 (Multiple-Relations Training) × 3 (Question Type: Definition, Classification, and Application) multivariate analysis of variance (MANOVA; Table 3). This analysis yielded a significant multivariate classification training by multiple-relations training interaction, Wilk's $\Lambda = .96$, $F(3, 194) = 2.61$, $p = .05$, partial $\eta^2 = .04$. This multivariate interaction reflects the fact that participants who received only classification training generally performed better on all three subsections of the posttest than those who received classification training in combination with multiple-relations training. This effect was accompanied by a significant multivariate main effect for classification training, Wilk's $\Lambda = .94$, $F(3, 194) = 4.52$, $p = .004$, partial $\eta^2 = .07$. The multivariate main effect for multiple-relations training was not significant, Wilk's $\Lambda = .99$, $F(3, 194) = .72$, $p = .54$, partial $\eta^2 = .01$.

As a follow-up to the MANOVA, 2 × 2 univariate ANOVAs were conducted to evaluate the effects of classification training and multiple-relations training on scores on each of the three posttest subsections (Table 3). Each ANOVA was tested at the .05 level of significance.

Table 1. Mean Posttest Scores

	Treatment group			
Measure	Control (*n* = 50)	Classification (*n* = 50)	Multiple relations (*n* = 50)	Combined (*n* = 50)
Posttest total				
M	26.60	29.28	26.58	27.08
SD	(5.50)	(5.35)	(6.63)	(6.51)
Definition				
M	10.24	10.34	9.90	10.06
SD	(2.18)	(2.07)	(2.23)	(2.48)
Classification				
M	8.74	10.20	8.96	9.34
SD	(2.10)	(2.05)	(2.62)	(2.32)
Application				
M	7.62	8.74	7.72	7.68
SD	(2.36)	(2.13)	(2.38)	(2.39)

Note: The maximum possible score for the posttest total was 36 and the maximum possible score for each question type was 12.

Table 2. Two-Way Analysis of Variance for Posttest Total Scores

Source	*df*	*SS*	*MS*	*F*
Classification training (C)	1	126.41	126.41	3.49
Multiple-relations training (M)	1	61.61	61.61	1.70
C × M	1	59.41	59.41	1.64
Error	196	7109.94	36.28	

For the classification questions, the mean scores were 8.74 (73%) for the control group, 10.20 (85%) for the classification group, 8.96 (75%) for the multiple-relations group, and 9.34 (78%) for the combined-strategies group. The classification training by multiple-relations training interaction was not significant, $F(1, 196) = 2.80$, $p = .10$, partial $\eta^2 = .01$, but a significant main effect for classification training was found, $F(1, 196) = 8.13$, $p = .005$, partial $\eta^2 = .04$.

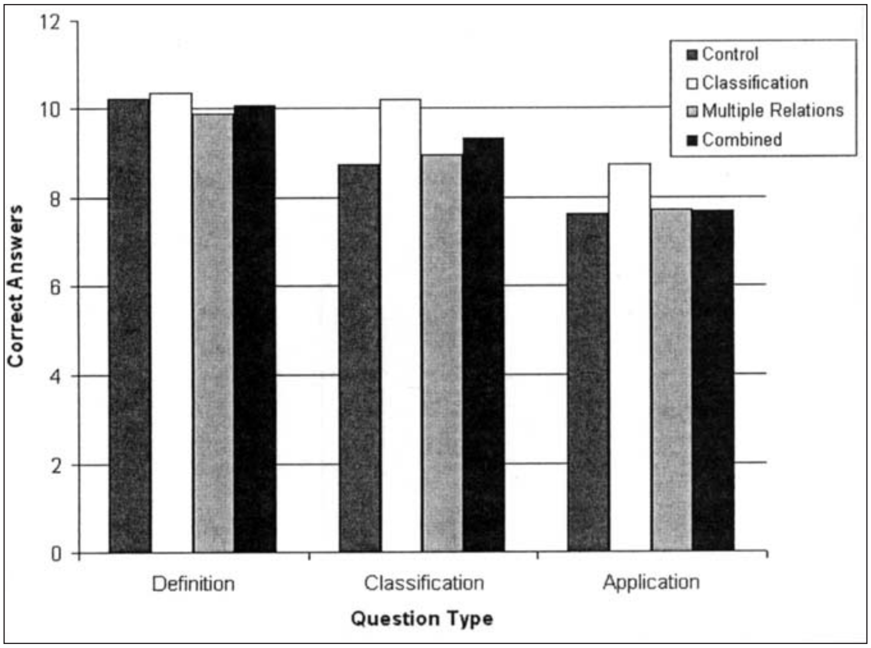

Figure 1. Performance on posttest questions by treatment condition.

The main effect for multiple-relations training was not significant, *F*(1, 196) = .98, *p* = .32, partial η^2 = .01. The classification training main effect indicated that students who were exposed to classification training, either alone or in combination with multiple-relations training, answered more classification questions correctly on the posttest than students who were not exposed to classification training.

For the definition questions, the mean scores were 10.24 (85%) for the control group, 10.34 (86%) for the classification group, 9.90 (83%) for the multiple-relations group, and 10.06 (84%) for the combined-strategies group. The classification training by multiple-relations training interaction was not significant, *F*(1, 196) = .01, *p* = .93, partial η^2 = .0001. The main effect for classification training was also not significant, *F*(1, 196) = .17, *p* = .68, partial η^2 = .001, nor was the main effect for multiple-relations training, *F*(1, 196) = .96, *p* = .33, partial η^2 = .005.

For the application questions, the mean scores were 7.62 (64%) for the control group, 8.74 (73%) for the classification group, 7.72 (64%) for the multiple-relations group, and 7.68 (64%) for the combined-strategies group. The classification training by multiple-relations training interaction was not significant,

Table 3. Two-Way Multivariate and Univariate Analyses of Variance *F* Ratios for Posttest Subsection Scores

		ANOVA		
Variable	MANOVA *F*(3, 194)	Definition questions *F*(1, 196)	Classification questions *F*(1, 196)	Application questions *F*(1, 196)
Classification training (C)	4.52**	0.17	8.13**	2.72
Multiple-relations training (M)	0.72	0.96	0.98	2.15
C × M	2.61*	0.01	2.80	3.13

Note: Multivariate *F* ratios were generated from Wilks's Lambda. MANOVA = multivariate analysis of variance; ANOVA = univariate analysis of variance.
$*p < .05$. $**p < .01$.

$F(1, 196) = 3.13$, $p = .08$, partial $\eta^2 = .02$. The main effect for classification training was also not significant, $F(1, 196) = 2.72$, $p = .10$, partial $\eta^2 = .01$, nor was the main effect for multiple-relations training, $F(1, 196) = 2.15$, $p = .15$, partial $\eta^2 = .01$.

Practice

Two different types of en route practice items were presented during the tutorial. Participants in all treatment conditions including the control group were given the same 12 definition questions, but the other practice questions varied according to treatment condition. Participants in the classification group were given 24 classification practice questions, participants in the multiple-relations group were given 24 inferential practice questions, participants in the combined-strategies group were given 12 classification and 12 inferential questions, and participants in the control group were given no practice questions other than the definition questions. Responses to these two different types of practice items (definition questions and treatment questions) were analyzed separately to determine whether differences occurred in participant performance. Table 4 shows the means and standard deviations for practice performance.

The overall mean score across all participants on the 12 definition practice questions was 9.59 (80%). The mean score on the definition practice questions was 9.36 (78%) for the control group, 9.72 (81%) for the classification group, 9.74 (81%) for the multiple-relations group, and 9.54 (80%) for the

Table 4. Mean Number of Practice Questions Answered Correctly

	Treatment group			
Question type	Control (*n* = 50)	Classification (*n* = 50)	Multiple relations (*n* = 50)	Combined (*n* = 50)
Definition				
M	9.36[a]	9.72[a]	9.74[a]	9.54[a]
SD	(2.06)	(2.23)	(1.72)	(2.23)
Classification				
M	—	18.32[b]	—	7.54[a]
SD	—	(3.10)	—	(2.04)
Inferential				
M	—	—	14.90[b]	7.70[a]
SD	—	—	(4.40)	(2.58)

[a]Maximum possible score was 12 items correct. [b]Maximum possible score was 24 items correct.

combined-strategies group. A 2 × 2 ANOVA revealed that the classification training by multiple-relations training interaction was not significant, $F(1, 196) = .1.07$, $p = .30$, partial $\eta^2 = .005$. There was also no significant differences for classification training, $F(1, 196) = .09$, $p = .77$, partial $\eta^2 = .0001$, or for multiple-relations training, $F(1, 196) = .136$, $p = .71$, partial $\eta^2 = .001$.

Classification practice questions were analyzed by comparing the performance of participants in the classification group to the performance of participants in the combined-strategies group. Participants in the combined-strategies group were given 12 classification questions randomly selected from the pool of 24 questions used in the classification group. Performance on the 12 classification questions common to both groups was compared. Participants in the classification group on the average answered more classification practice questions correctly ($M = 9.14$, $SD = 1.74$) than those in the combined-strategies group ($M = 7.54$, $SD = 2.04$). An independent samples *t* test revealed that this difference was statistically significant, $t(98) = 4.22$, $p < .001$.

Inferential practice questions were analyzed by comparing the performance of participants in the multiple-relations group to the performance of participants in the combined-strategies group. Participants in the combined-strategies group were given 12 inferential questions randomly selected from the pool of 24 questions used in the multiple-relations group. Performance on the 12 inferential questions common to both groups was compared. An independent-samples *t* test revealed that the mean number of correct answers by participants in the multiple-relations group (M = 7.76, SD = 2.04) was not significantly different from the mean number of correct answers by those in the combined-strategies group (M = 7.70, SD = 2.58), $t(98) = .13$, $p = .90$.

Time in Program

The overall mean number of minutes spent completing the tutorial across all participants was 176.88 min, SD = 72.94. Participants in the control group spent an average of 167.76 min, SD = 74.29; participants in the classification group spent an average of 171.41 min, SD = 78.27; participants in the multiple-relations group spent an average of 182.43 min, SD = 64.42; and participants in the combined-strategies group spent an average of 185.93 min, SD = 74.69. A 2 × 2 ANOVA did not reveal a significant classification training by multiple-relations training interaction, $F(1, 196) = .00$, $p = .99$, partial $\eta^2 < .001$. The main effect for classification training was also nonsignificant, $F(1, 196) = .12$, $p = .73$, partial $\eta^2 = .001$, as was the main effect for multiple-relations training, $F(1, 196) = 1.99$, $p = .16$, partial $\eta^2 = .01$.

Participant Attitudes

Responses to the 12 Likert-type items on the attitude survey were scored on a 5-point scale, with a 1 for the most negative response ("strongly disagree") and a 5 for the most positive response ("strongly agree"). The overall mean score across the 12 items was 3.90, a favorable rating indicating agreement with positive statements about the instructional program. The three highest-rated statements on the survey were "The questions I answered during the tutorial were relevant to the instruction" (M = 4.17), "There were enough examples for me to understand the concepts" (M = 4.14), and "The questions I answered during the tutorial helped me understand the material" (M = 4.08). The three lowest-rated statements were "I liked this tutorial" (M = 3.56), "I would recommend this tutorial to other people" (M = 3.60), and "I would rather learn from a tutorial like this than from a traditional textbook" (M = 3.70).

The survey was intended to have three clusters of items representing attitudes toward three distinct topics: three items on preference for the tutorial in general, three items on the quality and relevance of the practice questions, and three items on how much was learned about RFT. However, *all* of the survey items were strongly correlated with one another, as suggested by the very high alpha reliability coefficient of .94 for the entire survey, and thus were not analyzed in clusters. Using an alpha level of .01, 2 × 2 ANOVAs conducted on the responses to each of the survey statements revealed a significant main effect for multiple-relations training for one statement: "The questions I answered during the tutorial were relevant to the instruction," $F(1, 196) = 7.66$, $p = .006$, partial $\eta^2 = .04$. Participants who were exposed to multiple-relations training responded more negatively to this question than those who were not, regardless of whether they also received classification training. For this question, the classification group had a mean of $M = 4.42$, the multiple-relations group had a mean of $M = 3.90$, the combined-strategies group had a mean of $M = 4.12$, and the control group had a mean of $M = 4.24$. No other statistically significant differences were found for the survey items.

An analysis of the responses to the open-ended questions on the attitude survey indicated that what participants liked most about the tutorial was the number, quality, and/or relevance of the examples given for the concepts, a response given by a total of 63 of the 200 participants (32%) across all four groups. Other common responses to what users liked most included the humor and lighthearted tone of the tutorial (56 participants, 28%), the graphics and animations (55 participants, 28%), and the interactivity (30 participants, 15%). When asked what they liked least about the tutorial, 93 participants (47%) reported the length or time required to complete it, 34 (17%) indicated the practice questions and/or the quiz, 28 (14%) listed the lack of learner control over pacing and/or presentation order, and 23 (12%) replied the redundancy or repetitiveness of the material. There was no noticeable difference in the pattern of student comments across treatment groups.

DISCUSSION

The primary purpose of this study was to compare traditional classification training for concepts to newer, non-classification strategies on several measures of concept learning. While most instructional design models focus on classification or multiple-exemplar training for concepts, some theorists argue that newer strategies such as inference practice and the use of graphic organizers may be particularly useful for teaching abstract verbal concepts. The results of this study suggest that classification training is important for helping learners identify new instances of abstract concepts, but neither classification practice nor multiple-relations training were found to have a significant impact on learners' ability to identify concept definitions or answer application questions. Although a significant multivariate classification training by multiple-relations training interaction was detected, this interaction was not found to be significant in follow-up univariate analyses of the posttest subsections.

The finding of a significant main effect for classification training on the classification questions of the posttest supports the emphasis placed on such training by most researchers and instructional design experts (e.g, Merrill et al., 1992; Tennyson & Cocchiarella, 1986). It is noteworthy that this effect was detected even though the differences between the treatment conditions were rather small, as discussed below. Due to the extensive amount of core instruction included, less than 10% of the entire instructional program differed between treatment conditions. Despite the relatively minor difference in practice activities, participants who received classification training performed significantly better on the classification posttest questions and tended to score higher on all three subsections of the posttest. This suggests that classification training plays a vital role in the understanding of abstract concepts.

While multiple-relations training did not seem to have a positive impact on any of the three measures of concept learning, this result should be interpreted cautiously. It is possible that such training would be beneficial only after sufficient classification training has been provided. The value of teaching multiple relations among concepts may not be realized until students are able to correctly discriminate the defining attributes of the concepts, an outcome most directly engendered by systematic classification training. Participants in the combined-strategies group in the present study received only half as many classification practice questions as participants in the classification group, and this may have been insufficient to produce accurate generalization and discrimination. Additional research in which the amount of classification training is varied before adding multiple-relations training may help answer these questions.

This study also failed to find significant differences between treatment conditions on the definition and application subsections of the posttest. Results of the current study suggest that neither classification training nor multiple-relations training have a significant effect on these measures of concept learning. Participants in the control group—who received only the core instruction of definitions, expository examples, and definition questions—performed comparably on these types of questions to participants in the other groups. There may have a ceiling effect for the definition subsection of the posttest, however, as participants in the control group answered 85% of the definition questions correctly.

Analyses of en route performance did not reveal any significant differences between treatment conditions on the definition practice questions. When the 12 classification questions common to both the classification group and the combined-strategies group were analyzed, however, it was discovered that participants in the classification group answered significantly more of these questions correctly than participants in the other group. This is a somewhat curious effect, particularly since a similar analysis of the inferential questions answered by participants in the multiple-relations group and the combined-strategies group did not reveal any significant differences between the groups. This effect may be because performance on classification questions improves with increased practice (the classification group received twice as many classification questions as the combined-strategies group), whereas performance on inferential questions does not. Alternatively, it could be that answering inferential questions after classification questions, as was typically the case in this instructional program, has a retroactive inhibitory effect on classification learning.

As expected, participants in the control condition, who did not receive the additional 24 practice questions required of participants in the other three conditions, had the lowest mean time in the program. While this mean time was more than 18 minutes shorter than the mean time of the combined-strategies group, no significant differences in overall time spent completing the tutorial were detected. This lack of a significant difference is partly due to the considerable variability in time required to complete the tutorial within each treatment condition: the standard deviation for each group was over 60 minutes. Such a high degree of within-group variability makes it difficult to detect between-group differences in statistical analyses.

Overall, student attitudes toward the instructional program were not strongly affected by the instructional strategy to which they were exposed. The only item for which a statistically significant difference was detected was "The questions I answered during the tutorial were relevant to the instruction." Participants who were exposed to multiple-relations training responded more negatively to this statement than did those who did not receive multiple-relations training. They may have perceived the inferential practice questions to be less relevant because of their less direct relation to the content of the program and/or the posttest, especially since inferential questions were the only type of practice question not directly represented on the posttest.

Implications for Instruction

This study has a number of implications for the design and development of instruction for abstract concepts. In particular, the results suggest that the strong emphasis traditionally placed on classification training for concepts is justified for abstract concepts, at least for some types of learning outcomes. Classification practice seems to improve the learner's ability to identify new instances of a concept, whereas strategies such as inference practice and content diagrams did not have a significant effect on this skill in this study.

Neither classification training nor the newer strategies had a significant impact on definition learning or application skill in the present study. The posttest measure may have been insufficient for detecting differences on the definition subsection, however, given that the control group achieved a mean score of 85% correct on the definition questions. It is also not clear whether classification training or multiple-relations training would have more of an effect on ability to answer definition and application questions if additional practice were provided. Nevertheless, the results indicate that neither classification practice nor inference practice, in the amount and form presented in this study, is very effective for enhancing student ability to learn definitions or apply concepts.

Implications for Research

Several features of the present study limit the ability to generalize the results to other settings. This study focused on practice activities and assessments that are relatively easy to implement and evaluate in online learning environments. Multiple-choice questions were used for both practice and assessment, and the content diagram was simply presented on the screen, rather than created by the learners. Other strategies or measures of concept learning that are somewhat more difficult to implement in an online format, but perhaps more feasible in face-to-face training environments, may be more useful or reveal different effects. These strategies or measures might include constructed-response questions that focus on recall rather than recognition, essay and short answer questions, and guided discussion. The present study also did not allow an analysis of the relative value of inference practice and the content diagram, as they were always presented together in the multiple-relations training conditions.

Another feature that undoubtedly affected the results of this study is that all versions of the tutorial included an extensive amount of core instruction and the treatments comprised a very small proportion of the overall instructional program. The core content comprised a total of 478 screens, while the treatments consisted of only 24 screens of practice questions and, for the multiple-relations and combined-strategies groups, 12 additional screens displaying the diagram of instructional content. Thus, less than 10% of the entire instructional program differed between treatment conditions. Such a small difference between treatments may explain why stronger effects were not obtained for the experimental conditions. Furthermore, the relatively high posttest total score of 74% obtained by the control group suggests that the core content alone resulted in considerable learning gains that may have masked the potential effects of the different practice activities in an environment with considerably less core instruction.

Finally, the tutorial was not completed by participants in a controlled or standardized setting, and this likely increased within-treatment variability and made it more difficult to detect between-treatment differences. This increased variability, of course, can be a challenge for any researcher seeking to examine performance differences in relation to online training in a real-world setting. This study attempted to control for some aspects of the participants' experiences (e.g., by using a login system, requiring each part to be completed in one sitting, including the tutorial as a course assignment, etc.), but a great deal of variability likely remained in how users interacted with the tutorial and the settings in which they did so. Reviews of participant responses to the open-ended questions on the survey revealed, for example, that some participants took copious notes during the tutorial while others completed it in a noisy and distracting computer lab.

Knowledge of complex, abstract concepts is important in many areas of education and training. Further research is clearly needed to help us develop more effective strategies for fostering such knowledge. The present study suggests that classification training is an important component of such instruction, but perhaps not adequate for all measures of conceptual knowledge. Future research could minimize the limitations of the present study by systematically manipulating the amount of practice participants receive, examining other strategies and measures of concept learning, and training and testing in more controlled conditions.

Basic research on human categorization and relational responding may hold additional implications for instruction that should also be explored. Whether recent cognitive and behavioral theories on concept learning will lead to effective new instructional methods remains to be seen, but they may provide a useful framework for investigating this important type of human learning.

REFERENCES

Al-Kunifed, A., & Wandersee, J. H. (1990). One hundred references related to concept mapping. *Journal of Research in Science Teaching, 27* (10), 1069-1075.

Anderson, R. C., & Kulhavy, R. W. (1972). Learning concepts from definitions. *American Educational Research Journal, 9*(3), 385-390.

Baker, W. P., & Lawson, A. E. (2001). Complex instructional analogies and theoretical concept acquisition in college genetics. *Science Education, 85*(6), 665-683.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: Cognitive domain*. New York: McKay.

Brown, D. E., & Clement, J. (1989). Overcoming misconceptions via analogical reasoning: Abstract transfer versus explanatory model construction. *Instructional Science, 18*, 237-261.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Crosbie, J., & Kelly, G. (1994). Effects of imposed postfeedback delays in programmed instruction. *Journal of Applied Behavior Analysis, 27*(3), 483-491.

Dagher, Z. R. (1995). Review of studies on the effectiveness of instructional analogies in science education. *Science Education, 79*(3), 295-312.

Duit, R. (1991). On the role of analogies and metaphors in learning science. *Science Education, 75*(6), 649-672.

Dupin, J., & Johsua, S. (1989). Analogies and "modeling analogies" in teaching some examples in basic electricity. *Science Education, 73*, 207-224.

Elhelou, M. A. (1997). The use of concept mapping in learning science subjects by Arab students. *Educational Research, 39*(3), 311-317.

Engelmann, S. (1969). *Conceptual learning.* Belmont, CA: Fearon Publishers.

Fisher, K. M. (1990). Semantic networking: The new kid on the block. *Journal of Research in Science Teaching, 27*(10), 1001-1018.

Gagné, R. M. (1965). *The conditions of learning.* New York: Holt, Rinehart & Winston.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teaming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99-129). Hillsdale, NJ: Lawrence Erlbaum.

Gilbert, S. (1989). An evaluation of the use of analogy, simile, and metaphor in science texts. *Journal of Research in Science Teaching, 26*, 315-327.

Glynn, S. M., Britton, B. K., Semrud-Clikeman, M., & Muth, K. D. (1989). Analogical reasoning and problem solving in science textbooks. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *A handbook of creativity: Assessment, research and theory*. New York: Plenum.

Hansen, J. (1981). The effects of inference training and practice on young children's reading comprehension. *Reading Research Quarterly, 3*, 391-417.

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

Horton, S. V., Lovitt, T. C., & Bergerud, D. (1990). The effectiveness of graphic organizers for three classifications of secondary students in content area classes. *Journal of Learning Disabilities, 23*(1), 12-22.

Horton, P. B., McConney, A. A., Gallo, M., Woods, A. L., Senn, G. J., & Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. *Science Education, 77,* 95-111.

Kelly, G., & Crosbie, J. (1997). Immediate and delayed effects of imposed postfeedback delays in computerized programmed instruction. *The Psychological Record, 47*, 687-698.

Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal, 17*, 211-218.

Markle, S. M. (1969). *Good frames and bad: A grammar of frame writing* (2nd ed.). New York: John Wiley & Sons, Inc.

Markman, E. M. (1981). Two different principles of conceptual organization. In M. E. Lamb & A. L. Brown (Eds.), *Advances in developmental psychology* (pp. 199-236). Hillsdale, NJ: Erlbaum.

Merrill, M. D., Tennyson, R. D., & Posey, L. O. (1992). *Teaching concepts: An instructional design guide* (2nd ed.). Englewood Cliffs, NJ: Educational Technology Publications.

Moore, D. W., & Readence, J. E. (1984). A quantitative and qualitative review of graphic organizer research. *Journal of Educational Research, 78*(1), 11-17.

Munson, K. J., & Crosbie, J. (1998). Effects of response cost in computerized programmed instruction. *The Psychological Record, 48*, 233-250.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38*, 495-553.

Stavy, R. (1991). Using analogies to overcome misconceptions about conservation of matter. *Journal of Research in Science Teaching, 28*, 305-313.

Stewart, I., Barnes-Holmes, D., Hayes, S. C., & Lipkens, R. (2001). Relations among relations: Analogies, metaphors, and stories. In S. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 73-86). New York: Kluwer Academic/Plenum Publishers.

Story, C. M. (1998). What instructional designers need to know about advance organizers. *International Journal of Instructional Media, 25*(3), 253-261.

Tennyson, R. D., & Cocchiarella, M. J. (1986). An empirically based instructional design theory for teaching concepts. *Review of Educational Research, 56*(1), 40-71.

Tennyson, R. D., & Park, O. (1980). The teaching of concepts: A review of instructional design literature. *Review of Educational Research, 50*, 55-70.

Tessmer, M., Wilson, B., & Driscoll, M. (1990). A new model of concept teaching and learning. *Educational Technology Research & Development, 38*(1), 45-53.

Tiemann, P. W., & Markle, S. M. (1990). *Analyzing instructional content: A guide to instruction and evaluation.* Champaign, IL: Stipes Publishing Company.

Venville, G. J., & Treagust, D. F. (1997). Analogies in biology education: A contentious issue. *The American Biology Teacher, 59*(5), 282-287.

Direct reprint request to:

Dr. Eric J. Fox
Department of Psychology
Western Michigan University
Kalamazoo, MI 49008-5439
e-mail: eric.fox@wmich.edu