The Psychological Record, 2007, 57, 201-218

TEACHING COLLEGE LEVEL CONTENT AND READING COMPREHENSION SKILLS SIMULTANEOUSLY VIA AN ARTIFICIALLY INTELLIGENT ADAPTIVE COMPUTERIZED INSTRUCTIONAL SYSTEM

ROGER D. RAY and NOELLE BELDEN
Rollins College

This paper presents a behavioral model for conceptualizing advanced reading comprehension as a "higher order" behavior class. Also discussed are strategies and tactics utilized by an artificially intelligent adaptive tutoring and testing software system designed to shape such comprehension skills while also teaching subject-specific "content" to college students. The system, called MediaMatrix, offers internet delivery of relatively traditional textbook content using highly individualized and adaptive tutorial and assessment procedures (Ray, 1995a; 1995b, 2004). Extant and new research on the effectiveness of this system is presented, with particular emphasis on a preliminary study of students in two small sections of an introductory psychology course. Students were evaluated during early (pre) and late (post) portions of the semester using two equivalent forms of a specially constructed SAT/GRE type reading comprehension test. A statistically significant 17% gain from pre-to-post reading comprehension scores was observed, suggesting that both the behavioral model and the MediaMatrix strategies and tactics for shaping such higher order behaviors merit further research. Practical implications of teaching both specific course content and higher order behaviors such as reading comprehension without direct teacher contact are especially noted.

Outside of the growing body of literature on various types of relational frames (cf. Hayes, Barnes-Holmes, & Roche, 2001) the behavioral literature is relatively sparse in contributing to our understanding (and engineering) of what is sometimes called "higher order classes of behavior" (Catania, 1998). One of the originating exemplars of higher-order behavior classes is Harlow's learning set phenomenon, which he originally described as "the *learning how to learn efficiently* in the

Portions of the current study were presented in a symposium paper in 2004 at the Boston meetings of the Association for Behavior Analysis. The MediaMatrix adaptive software system described in this paper is owned and copyrighted by $(AI)^2$, Inc., a company partly owned and operated by Roger Ray. Noelle Belden's participation in this study served as a part of her senior undergraduate honors thesis at Rollins College. Correspondence concerning this manuscript should be addressed to Roger D. Ray, Department of Psychology, Rollins College, 1000 Holt Ave., Winter Park, FL 32789. (E-mail: rdray@rollins.edu).

situations the animal frequently encounters." (Harlow, 1949, p. 51). Pérez-González, Spradlin, and Saunders (2000) have demonstrated that such learning set outcomes hold not only for Harlow's monkeys, but also for second-order conditional discriminations in grade-school children.

Other categories of higher order behaviors besides learning sets and relational frames—including of course stimulus equivalences (Sidman, 1994)—have been summarized by Catania (1998). These include identity matching, learned helplessness (Maier, Seligman, & Solomon, 1969), and Estes' (1971) "operation of rules, principles, strategies, and the like" (p. 23). Further, Catania adds some types of attention-getting behaviors, novel behaviors, observational learning, generalized imitation, instruction following, and even some forms of manding and remembering behaviors as classes of higher-order behaviors. But Catania deals only in passing with such abstract classes as those often considered by cognitive psychologists as "meta-skills" (e.g., Karoly, 1993). Especially ignored are those metaskills typically targeted by advanced educational curricula, such as quantitative reasoning, critical thinking and evaluation, and reading comprehension (e.g., Phillips & Bond, 2004).

Perhaps one of the most challenging behaviors confronting behavior analysis is "advanced" reading comprehension: a higher order class of behaviors with significant research from the cognitive perspective (e.g., Kintsch, 1998) but rarely, if ever, discussed in the behavioral analytic literature. One of the earliest attempts to demonstrate how teachers could translate "cognitive" content and "conceptual" learning from a behavioral perspective was offered by Markle and Tiemann (1970). Their model focused on development of the ability to generalize and discriminate between examples, with testing utilizing alternative examples to assess generalization.

Another early first-approximation to comprehension research based on behavioral principles was Miller and Weaver's (1976) investigation of "concept formation" in university students. In a series of experiments these researchers explored the effectiveness of a textbook that incorporated "concept programming." Students were initially heavily prompted within the text to facilitate discrimination of specific concepts described in fictional stories illustrating behavioral concepts taught in each lesson. The concept-programmed textbook resulted in more concept formation than a more traditional textbook. Effective generalization of concepts to novel examples was also demonstrated, and such generalization was present only if concept programming was included in training.

Behaviorally based contributions focusing directly on the development of reading include the Morningside model for teaching reading skills (Johnson & Street, 2004); its on-line spin-off, HeadSprout (cf. www.headsprout.com); and the foundational research conducted within the Morningside/Headsprout context (cf. Layng, Stikeleather, & Twyman, 2004; Layng, Twyman & Stikeleather; 2004). But these programs target only the development of reading fundamentals, mostly in illiterate beginners covering the equivalent of K-2. Nevertheless, Headsprout's incorporation of adaptively interactive computerized delivery makes

it especially powerful, because it allows for automated and internet-distributed shaping of critical foundation reading skills in a massive number of students. Use of its highly engineered instructional techniques has resulted in dramatic outcomes for beginning readers (Layng, Twyman, & Stikeleather, 2003, 2004). But what of the often overlooked college freshman who, when presented with traditional college textbooks, already has a sufficiently developed reading *vocabulary*, but enters higher education with only marginal skills for identifying and organizing salient concepts, specifying relationships among those concepts, and/or elaborating attendant properties of concepts presented in a textbook?

Almost all students entering college today can read in the literal sense, or as Headsprout's outcome goals would define it. That is, students can respond to sequences of words presented on a printed page or computer monitor by repeating them aloud (or silently) in proper sequences. Students may even be able to demonstrate some of what academics would call "understanding," which is often synonymous with typical definitions of higher order behaviors such as "comprehension." But research on reading comprehension at this advanced student level has focused almost exclusively on finding a metric for measuring comprehension or on using the phenomenon as a primary predictor of college or graduate school success, as exemplified by the proliferation of such standardized tests as the SAT or GRE and their inclusion of reading comprehension assessments.

There is an abundance of preparation strategies commercially offered to improve SAT or GRE scores, including on-line delivery of relatively costly programs that offer guarantees of "satisfaction" with improved scores but little scientific data. Actually, data on the efficacy of such programs vis-à-vis time or monetary investment vs. improvements in test scores is quite sparse. One report based on applications of behavioral strategies in course design demonstrated significant gains in general GRE scores (Miller, Goodyear-Orwat, & Malott, 1996), but this intensive and highly structured self-paced course involved from 66-140 hours of study and targeted only general verbal, quantitative, and combined score improvements. Reading comprehension per se was not isolated for development or measurement.

Even the few studies that have been conducted on direct teaching of advanced reading comprehension skills outside of standardized testing environments typically do not utilize proven behavioral analytic techniques (e.g., Caccamise & Snyder, 2005; Kintsch, 2005) and tend to focus exclusively on tutoring reading skills independently of content-specific knowledge. Behavioral studies, such as Harlow's (1949), suggest that higher order behaviors are likely to be based upon a type of response generalization involving more specific "first-order" skills, such as learning specific discriminations and/or associations. This suggests the possibility that a more efficient approach would be to teach both reading comprehension as a generalized (higher order) skill while also teaching specific content (course-related first-order discriminations

and associations). Thus, Rice (1994) attempted to teach reading comprehension by having participants use text highlighting techniques. However, Rice's primary focus was on whether such procedures applied on paper vs. computer screens made any difference in outcomes, not whether the reading comprehension skill generalized to new content learning. In fact, Rice found that reading comprehension levels, as measured by a text recall (production) test, were the same for students highlighting text on paper vs. highlighting text via a computer screen.

Computer-Based Instructional Designs and Reading Comprehension

Rice's (1994) study is somewhat typical in its focus on computer presented vs. more traditionally presented text. Others have gone further in testing the unique contributions computers can make to the comprehension process. For example, MacArthur and Haynes (1995) conducted a study to compare reading comprehension levels in learning disabled children reading from two versions of computer presented text. Students read two passages of material related to the field of biology in two different modes of presentation. The first passage was presented via a computer in much the same way a textbook would present it. The students were then tested for comprehension of the content. The second passage was presented as an enhanced version that included such features as speech synthesis, highlighting of main ideas and question-to-text linkages. A second comprehension test was then administered. Students attained significantly higher comprehension scores reading the enhanced version and they also stated a preference for that mode over the plain version (MacArthur & Haynes, 1995).

MacArthur and Haynes' (1995) use of enhanced antecedent-stimulus prompting in computerized presentations is not the only study to use this combination of computers and behavioral strategies. We have already mentioned the unique adaptive internet-delivered programs offered in Headsprout's reading development application, as well as its impressive achievements in student outcomes (Layng, Stikeleather, et al., 2004; Layng, Twyman, et al., 2004). An alternative computer-based strategy was detailed by Ray (1995a, 1995b, 2004) wherein he described the philosophy, design strategies, and pragmatics for an artificially intelligent adaptive tutoring and mastery certification system, called MediaMatrix, that was designed from the outset to teach more advanced reading comprehension skills simultaneously with the teaching of subject-specific content. Various versions of the MediaMatrix instructional system have been operational for over a decade (Ray, 1995a, 1995b) but its strategies have always relied upon adaptive presentations of stimulus prompting/fading, as well as the successive approximation techniques of response shaping and the corresponding leaning (i.e., increasing intermittency) of reinforcement/feedback density as a strategy for teaching higher order reading skills.

In a recent summary of this work, Ray (2004) clearly identified the adaptive tutoring and assessment strategies incorporated into MediaMatrix

as applications of *behavioral* principles in artificially intelligent instructional design. Ray noted that such adaptive features were largely prompted by his personal dissatisfaction with the more traditional behaviorally inspired educational content delivery and instructional design technologies created in the mid-twentieth century, including personalized instructional (PSI) course mechanics (Keller, 1968) as well as programmed instructional content presentations based upon "frames," or small units of content and frequent assessment/feedback (e.g., Skinner, 1968; Vargas & Vargas, 1992).

With the exception of advances such as Miller and Weaver's (1976) concept programming, programmed instruction's more traditional focus has been on the teaching of immediate (primary level) discriminative behavior and content while ignoring development of desirable higher order behaviors that might eventually wean the learner from needing such "structured" (i.e., programmed) content presentations. Thus the classical instructional design strategies of programmed instruction inherently emphasized a perpetual reliance upon the programmed instructional model (Ray, 2004). Tiemann and Markle (1990) were among the first to break from this reliance on single-sized units by offering two alternative levels of interactive practice in computer-based tutorials and found that allowing access to "domain-guided" (as opposed to "tutor-guided") interactive practice led to significantly better performance in the use of spreadsheets.

Taking the concept of alternative "levels" even further, MediaMatrix was designed to begin with presentations that are similar to programmed instructional frames, but it gradually and adaptively, based on the dynamic tracking of student performance, increases or fades the obviousness of critical content underlining and other forms of prompting (Ray, 1995a, 1995b). In addition, MediaMatrix's selective presentation of textual units gradually decreases or increases the number of paragraphs (and thus the amount and complexity of content) the student must read and master prior to having mastery assessed and being presented feedback. This contrasts with the consistently small and always obvious framing of text or stimulus units in more traditionally designed "static" (as opposed to adaptive) programmed instruction. As noted, MediaMatrix also includes adaptive strategies that, depending upon student performance, lean the density of reinforcing feedback as better reading/study skills are demonstrated, thereby successively approximating the more traditional use of infrequent testing as assessments of mastery.

As noted, Ray (2004) also voiced reservations with the traditional use of PSI course formats that relied exclusively upon the student reading and testing, even if peer tutoring was included (cf. Keller, 1968). Although some researchers have explored alterations in the classic design of PSI courses (cf. Conard, Spencer, & Semb, 1978; Miller, Weaver, & Semb, 1974; Spencer & Semb, 1978), Ray has argued that PSI formats more typically favor the well-skilled reader by failing to help students who need to enhance their reading skills. He states that "personalized instruction incorporates peer tutors to help students practice their poor reading skills over and over because such tutors typically are not trained to work on comprehension skill building (p. 150).

Thus Ray's (2004) criticisms both of classical frame-based programmed instruction and traditional PSI formats focused on their typical lack of attention to higher order behaviors that transcend specific content being read and assessed—behaviors that could eventually wean the student from needing either programming or tutoring. Of course the developers of PSI and programmed instruction technologies (e.g., Keller, 1968; Skinner, 1954) would almost certainly have used more "adaptive" or "individualized adjustment" techniques if they had only had access to modern computer technologies to accomplish such adaptations. Adaptive teaching is, after all, the very core of response shaping and errorless discriminative stimulus control (e.g., Terrace, 1963) because each is accomplished via "successive approximation" procedures that shift criteria (i.e., adapt) as respective response classes are reliably produced. But much modern computer-assisted instruction seems to have largely left behind the behavioral technologies that originally inspired machine-based instruction (e.g., Larkin & Chabay, 1992). MediaMatrix's instructional design (Ray, 1995a; 1995b) is a major exception to this generalization, in that it relies quite heavily on the explicitly *behavioral* technologies of prompting/fading, response shaping, and the leaning of feedback density in the development of both immediate content fluency as well as the higher order behavior class of more generalized reading comprehension.

Adaptive Strategies and Tactics for Developing Reading Comprehension

To generate higher order behaviors of text comprehension, MediaMatrix applies strategies that emphasize primary concept terms and their defining and/or discriminative properties when these terms are presented as stimuli within the context of relatively standard textbook narratives. Students are at first prompted by being shown underlined occurrences of all terms that are conceptually "associated," but such prompts are progressively faded as the student demonstrates successful learning of such verbal associates. This more generalized skill of discriminating "primary concept terms" and their appropriately associated "defining property" terms is gradually developed and confirmed by probe assessments which, themselves, fade in their own degree of prompting through the system's changing formats of question presentation (Ray, 2004).

Thus the tactical use of differing question formats in MediaMatrix fades available prompts within questions as well as within the primary text. This is accomplished by gradually moving from the use of selection/recognition question formats to production/recall formats. The various types of question formats begin (in Tutor Level 1) with the use of multiple-choice questions that include a "blank" within the question that is appropriately "filled" by one, and only one, alternative among several terms/phrases. A series of correct answer selections on multiple choice questions causes a shift to questions (in Tutor Level 2) stated in the same form as their multiple choice counterparts, except now questions rely upon a fill-blank answer production format that requires

students to actually type in appropriately associated terms/phrases, rather than merely discriminating/selecting them from a set of accompanying inappropriate alternatives. Successful answering of a series of fill-blank questions brings a subsequent (Tutor Level 3) use of paired associates recognition/selection formats, where a concept term/phrase and property-defining term/phrase are paired and the student subsequently selects "associated/not-associated" as answers. Finally, on a fourth stage of fading prompting stimuli within questions (Tutor Level 4), verbal associate types of questions (Verplanck, 1992) replace the paired associates format. In this verbal associates format, multiple associations must be produced (typed) in response to the presentation of a conceptual term/phrase, such as asking a student to give four distinguishing characteristics for the prompt/question of "cumulative records" (e.g., "Y = cumulative response count," "X = time," "Skinner," and "continuous session").

As such, all strategies and tactics used by MediaMatrix build toward accuracy and fluency in verbal associates test performance (Verplanck, 1992), which is, itself, proposed as being fundamental to advanced reading comprehension skills. Support for this proposition comes not only from Verplanck's behaviorally oriented research program, but also from substantial parallel research literatures in cognitive psychology and constructivist learning theories (Erlmer & Newby, 1993) and their respective uses of "concept mapping" and/or "mind mapping" as a "learn to learn" process (e.g., Jonassen, 1996; Novak & Gowan, 1984). Thus MediaMatrix's approach not only targets the development of specific verbal associates, concept maps, and/or semantic networks (e.g., Sowa, 1984) wherein conceptual terms and their associated property-specific terms/phrases are made specific, it also targets the more generalized class of behaviors underlying the act of discriminating such relevant stimuli when presented within standard textbook narratives.

Research on MediaMatrix as an Adaptive Instruction System

Ray's (2004) recent review of the MediaMatrix adaptive instructional system described an ongoing research program that incorporates, via internet delivery, content as an "electronic textbook" to supplement variously formatted introductory psychology courses at the college/university level. This electronic text for introductory psychology (Kasschau, 2000) includes 17 chapters of relatively traditional text content which was edited specifically to assure inclusion of primary concept terms and their associated (defining or qualifying) properties within each topical narrative unit. The complete electronic system is designed as a replacement for traditional textbooks and is used either as a supplement to a lecture course, as the content for a PSI peer-tutored course, or even as the exclusive source of content in an on-line course with no class meetings or peer tutoring (cf. http://www.ai2inc.com/Instructors/CourseSetup/setup.html).

Ray (2004) also summarized some preliminary research that resulted in statistically reliable improvements in content mastery and certification

accuracy (and corresponding course grades) as a function of increasing time that students were in contact with MediaMatrix's adaptive tutoring services. One study reviewed was a conference presentation by Belden, Miraglia, and Ray (2003) that investigated different instructor-established contingencies (via internal settings available within MediaMatrix) for bringing students into actual contact with the (user-optional) adaptive tutoring services within MediaMatrix. These researchers investigated both (a) the use of bonus points for tutoring and (b) the application of a feature wherein tutoring was required following limited numbers of less-than-satisfactory performances on mastery certification tests administered within MediaMatrix. Five instructors used different combinations of bonus and required tutoring supports as independent or combined contingency settings, thus allowing for a qualitatively ranked ordering of different degrees of "contingency stringency." Students in courses with the highest degrees of contingency stringency had four times the mean contact time with tutoring than students in lowest stringency courses, and a systematic increase in time of contact occurred with increasing stringencies. But even more relevant was the finding that, although average maximum certification mastery test scores (out of variously allowed numbers of retakes using similar tests) across all of these same instructors were approximately the same (i.e., between 80-83% accuracy), the more stringent criteria were associated with achieving this maximum score within an average of two attempts vs. four to five attempts for the least stringently managed course.

Belden et al. (2003) also ranked students by quartiles in their distribution of total tutoring time across all instructors. Students within the lowest quartile of tutoring time had test scores that averaged 78% accuracy. Test scores increased quartile by quartile up to the highest quartile of tutoring time resulting in 84% accuracy. The authors note that the practical implications of this finding include a difference between a high C letter grade and mid-B grades for testing performances when lowest quartile was compared to highest quartile of tutoring time.

Ray (2004) also summarized a presentation by Butterfield and Houmanfar (2003) wherein the use of the MediaMatrix system with Kasschau's (2000) incorporated text was compared to use of a commercially popular traditional textbook (Gray, 2002) within the context of alternative sections of a PSI-formatted university course in introductory psychology. Adaptive mastery certification testing within the MediaMatrix sections of the course was time limited for each question, but not supervised. Certification testing for the printed-text sections of the course was administered via a WebCT-based computerized system and was supervised by teaching assistants/tutors. Students in both sections were assessed with the same pretests and posttests constructed to sample content common to both textbooks. Butterfield and Houmanfar generated samples of 41 students each to represent the corresponding research conditions in the large introductory sections (200-400 students per section) across two semesters. The Fall semester samples revealed

nearly twice as many adaptive instruction students being in the A and the B grade ranges for the final exam compared to the WebCT students. Although not as dramatic in magnitude, a similar difference was reported for the Spring samples as well.

A Preliminary Investigation of Reading Comprehension

Importantly, while Ray's (2004) review of available research on MediaMatrix suggested a positive outcome for the system's ability to impact *content-based* performance, his review included *no* evidence that the system actually delivers on the other primary purpose for which it was designed: the improvement of more generalized (higher order) reading comprehension skills. This is because no research had yet been conducted on MediaMatrix and generalized reading comprehension. Thus we pursued a preliminary investigation to evaluate whether improved reading comprehension scores might be obtained simultaneously with traditional content-oriented certifications of student progress (grades based on test performances) in the context of a lecture-and-discussion-based offering of an introductory psychology course. To this end, general reading comprehension performance was probed early and late in the semester, with special interest focused both on statistically reliable and on practically useful improvements on measures representative of those used for predicting future collegiate success and/or student selection.

Method and Procedures

Passages and corresponding questions used to evaluate reading comprehension were obtained from one SAT test preparation book (Robinson & Katzman, 2001) and two GRE test preparation books (Alexander-Travis et al., 2002; Martinson, 2003). Two equivalent forms (A and B) of reading comprehension tests were constructed so students could be exposed to different content on pretests and posttests. Each test included two single and one dual passage on nonpsychological content, a total of 20 multiple choice questions, and was administered in class with an imposed 30-min time limit.

The study was conducted within the context of a course that met twice per week for 75-min class periods throughout a 16-week semester. In-class activities focused on a stated goal of improving audio-visual (A/V) comprehension skills (not measured) as well as relatively standard goals focusing on knowing content (vocabulary, concepts, methods, and apparatus used in studies, comparative theories, etc.). In-class activities were paralleled with assigned out-of-class adaptive on-line text and tutoring services in MediaMatrix to substitute for a more traditional textbook. However, no attempts to teach reading comprehension per se were implemented during class periods. Typical in-class activities included lectures, commentaries, and discussions based on presentations from the 27 volumes of the Zimbardo-hosted *Discovering Psychology* video series (Yourgrau, 1990). Typical in-class

use of these videos included showing 2-5 min segments at a time, with pauses to highlight via lecture, commentaries, question/answer, and so forth, the use of settings, apparatus, selections of participants and other relevant variables and properties of experiments and/or processes being illustrated by the current segment—an *A/V* process designed to parallel the process implemented within MediaMatrix for developing *text* comprehension (cf. Ray, 2000). In-class testing was all associate-type questions wherein some questions presented concept terms/phrases to prompt associate recall/production (verbal prompt for verbal associates) and other questions presented representative "single frame-grabs" of video scenes shown in class as prompts for productions of verbal associate "answers" (A/V prompt for verbal associates).

Students could retest on chapter-wide mastery certification tests (adaptively constructed by MediaMatrix to be unique on each offering) for each text chapter up to posted deadlines that varied chapter by chapter. Only the maximum score for each chapter counted towards a student's testing scores in course grading, and final grades for the course were determined exclusively by (a) these (unsupervised) on-line chapter-wide certification testing scores for 17 chapters of the electronic textbook, plus (b) four in-class exams, and (c) one in-class final that was weighted as two in-class tests. In-class exams covered materials from both textbook readings and video-based lectures. On-line chapter testing scores accounted for 50% of the total course grade while all combined in-class testing accounted for the other 50%. We began the course with no required use of tutoring services in MediaMatrix, regardless of certification testing scores on assigned chapters.

We administered the first (pretest) 30-min assessment of reading comprehension during the 2nd week of class, which was after deadlines had expired for two on-line chapter mastery certifications. After the first in-class course exam was administered in the 4th week of the course, we implemented more stringent contingencies (automatically administered by MediaMatrix) for use of the MediaMatrix certification and tutoring system. As noted earlier, this system allows an instructor to set: (a) a specified criterion for minimally accepted accuracy in chapter-level mastery certification testing and (b) the number of retests allowed to reach this criterion. The contingencies used in this study required a student to attain a 90% or greater accuracy score within three attempts on a given chapter's certification testing. If the student failed to reach the 90% criterion in the three opportunities allowed, that student was then required to use the adaptive tutoring services of MediaMatrix to tutor on all topics within a chapter that were prescribed by the MediaMatrix system based on that individual's performance. Completion of 100% accurate tutoring of all specified topics allowed another set of three chapter-certification attempts to reach 90% on equivalent test forms (and this cycle repeats until criterion is reached or the student accepts a lower than 90% testing score). Thus students had the option of accepting a score less than the assigned criterion and could choose not to utilize tutoring for retesting. New chapters began the cycle anew.

Just before the last week of classes, which was before students took their last in-class test and final exam (both supervised), we administered a second assessment of reading comprehension under conditions matching those of the first assessment, except administrations were of the alternative form of the reading comprehension test taken previously.

Results

The original design of this study intended to make comparisons of two similar sections (8AM and 2PM) of the same introductory psychology course (offered by the same instructor using the same syllabus and assignment schedule), with the order of AB vs. BA forms of reading comprehension assessment being the primary difference between the two sections. Although the study began with 38 participants, drop/add changes and class absences during the reading comprehension assessment days resulted in complete comparison data for 24 original enrollees. Pre/post comparison scores for each participant on all evaluations existed for 10 participants for the 8AM section and 14 participants for the 2PM section.

Unfortunately the 8AM section's posttest data were seriously compromised by an administration error involving the distribution of reading comprehension tests to students too near the end of the class period to allow total required time (30 min) for completion. Many students were not able to complete the test, while others reported being rushed by the shortness of the time allowed. Thus we did not consider the 8AM section's posttest data reliable for comparison purposes.

However we deemed both sections' pretest administrations successful and thus evaluated the equivalence of Form A vs. Form B using an independent group means comparison. The mean pretest raw score on Form A for the 8AM class was 10.10, and the mean pretest score on Form B for the 2PM class was 10.14. A *t* test for independent group means reveals an observed $t(22) = .02$ (alpha .05, nondirectional critical value = 2.07, n.s.).

The remaining viable comparison was between the 2PM section's pretest (Form B) vs. posttest (Form A) reading comprehension test scores. The Form B raw-score mean (pretest) was 10.14 and the Form A (posttest) was 11.86. These equate to a mean percentage correct for the 2PM section on the pretest of 50.7% and a mean posttest of 59.3% correct, or a 17% improvement. The calculated value for differences from pretest to posttest means using a correlated group $t = -2.22$ and is thus statistically reliable at an alpha level of .05, $t(13) = 2.16$ (nondirectional).

Pretest and posttest measures were also considered from a single-subject perspective, even though multiple baseline and multiple change measures were not practically feasible in this study. Of special interest was whether students with the lowest pretest comprehension scores might account for the greatest amount of change in posttest measures. Figure 1 presents an individual-by-individual comparison of pre/post numbers of questions answered correctly, ordered according to pretest scores. Three participants (individuals 4, 9 and 11) had a decline in number of correct

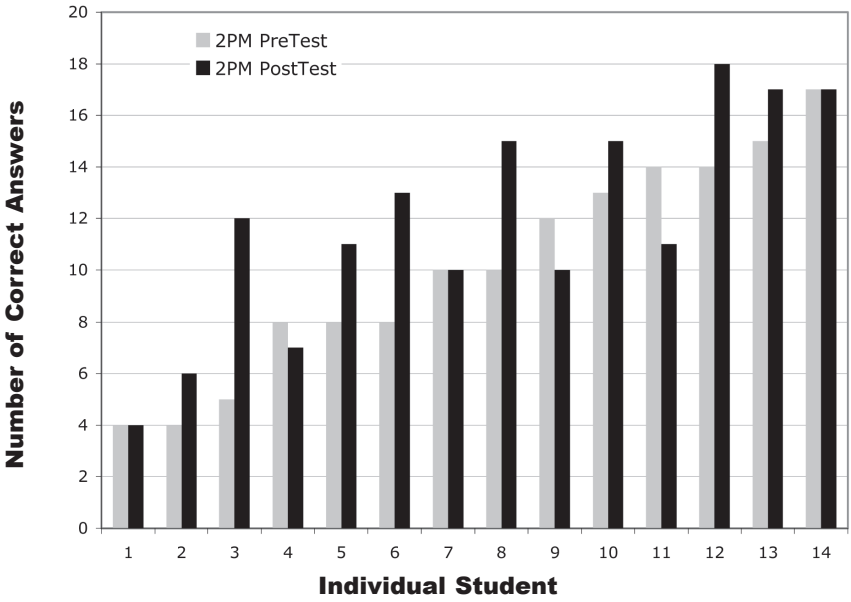

Figure 1. Number of correctly answered questions on pretest and posttest measures of reading comprehension for each individual student presented in ranked order of lowest-to-highest pretest scores.

answers from pretest to posttest measures, ranging from three fewer correct to one fewer. Three participants (individuals 1, 7, and 14) showed no change between pretest and posttest measures. The remaining 8 of our 14 participants showed improvement, ranging from 2 additional questions correct (individual 2) to 7 additional correct (individual 3) out of the 20 questions. There appears to be no systematic relation between the pretest scores and the "improvement" reflected in posttest scores. Some of the highest pretest scores are associated with negative change, some with no change and some with relatively large positive changes. Likewise for the low-scoring students with respect to pretest levels. Of course the more standard descriptive statistic for this comparison is a correlation coefficient, which was determined to be -.31 between pretest scores and change scores.

Discussion

The loss of the 8AM class pre-post comparisons was unfortunate, in that this negated the intended assessment, through counter balancing, for "order of presentation" effects in reading comprehension for test forms A and B. Nevertheless, there is minimal likelihood of an order effect in this study given that our independent group comparison of pretest means for the two forms demonstrates form equivalence. Thus we are reasonably confident in focusing on the primary question concerning pre-post course comparisons of reading comprehension, which was the only statistically reliable outcome found.

Because this study was conducted as an integral part of an ongoing course, it clearly qualifies at best as incorporating a quasi-experimental design (Campbell & Stanley, 1963), and thus includes potentials for various confounds. It was especially unfortunate that no other introductory psychology sections were available that semester to offer a control group for comparison with a more traditional teaching method and use of a standard textbook. Further, the study did not use randomly selected participants, and the relatively high loss rate of participants who started in this course might argue for the possibility of selective retention. We have no counter argument for this, other than to reflect on the reported individual-by-individual pretest reading comprehension characteristics. Pretest comprehension scores for students actually contributing data to the outcome comparison ranged from 4 of 20 (20%) questions correct to 17 of 20 (85%) questions correct (as reported, the mean was 50.7 % for the entire class). At a minimum, this wide range of student pretest reading comprehension levels argues for a fairly robust effect across a diversely prepared sample. In addition, our single subject analysis, as well as our assessment of the correlation between pretest measures and posttest changes in these measures, show that the majority of students made a positive change, but that the degree of change is not reliably predicted by their beginning reading comprehension scores.

The lack of a control section for in-class activities as they relate to potential for impact on our reading comprehension measures is perhaps even more serious a limitation. One might suggest that it was the instructor and/or the A/V comprehension activities in class that were most responsible for changes in student reading comprehension skills. There was neither in-class reading practice nor any direct instructional focus on reading skill development, but the Media-Based Introductory Psychology course upon which this study was conducted exposed students to lectures, video clips, and discussions. A relatively clear argument might thus be made supporting the possibility of an instructor-based enhancement of A/V comprehension, which was a stated learning goal. But is it plausible that improvement in reading comprehension was developed by focusing on A/V skills? There is no literature of which we are aware to suggest such a transfer effect, much less any evidence that A/V skills were themselves enhanced during this study. We thus feel our results are more likely attributable to MediaMatrix and its adaptive tutoring system, in that MediaMatrix was actually designed to improve the higher order behavior of reading comprehension. Regardless, it seems unlikely that most traditionally taught introductory psychology courses would result in the change in generalized reading comprehension skills observed in this study.

A similar argument might be made regarding potential confounds from student experiences in other courses as possibly accounting for the reading comprehension improvements. However, these students shared no common activity as a group aside from the introductory psychology course, and no courses of which we are aware within the college's curriculum target reading comprehension skill development.

Conclusion

Further study is obviously needed to clarify the potential alternative interpretations in this preliminary study on reading comprehension. But the data available certainly suggest the worthiness of investing time and effort in the conduct of a larger study—a study designed to measure both A/V comprehension changes and reading comprehension changes in sections taught by different instructors and with the added value of a control section representing a more traditionally taught introductory psychology course using a traditional textbook. If enhanced A/V comprehension skills can also be documented, this would suggest incorporating adaptive A/V instruction as a natural next step in MediaMatrix's adaptive instructional design. That is, the in-class activities of the instructor are literally an attempted "personal" management of much the same "adaptive processes" of prompting/fading, shaping, and feedback approximations as those automated within MediaMatrix—except within the class they are focused upon video-based content presentations as opposed to text-based content presentations. As such, projected research should focus on the efficacy of this teacher-based adaptive A/V procedure prior to investing extremely expensive development resources into video content and software that might automate the process and thus potentially accomplish the same outcomes.

It is also worth noting the practical, as opposed to statistical, significance in the reading comprehension findings just reported. Students in this course demonstrated a nearly 17% (i.e., the posttest score was 117% of the pretest score) mean improvement in reading comprehension while taking an Introduction to Psychology course—whether or not this resulted, as suggested, from using the MediaMatrix system. If this same improvement occurred in the context of graded activities in a course (such as test scores), it would represent a practical grade level change. For example, this *could* represent a shift from, say, a test grade of 69 (typically a borderline upper-D grade) to a test grade of 80 (typically a borderline lower-B). Most students would consider this quite a practical (significant) difference in grades.

But observed improvements have implications that extend beyond the Introduction to Psychology course these students completed. Students who decide to continue their education by enrolling in graduate studies might also have gained a presumed advantage on the comprehension portion of typical standardized tests, such as the GRE. A 17% improvement in reading comprehension could easily elevate a student on the corresponding sections of this "qualifications" exam. It was for this potential practical value that one verbal component of GRE practice tests was used in our outcome measures for this study. If implemented at the secondary school level, a similar achievement in improved reading comprehension on the SAT could have potential significance for college placement as well.

The MediaMatrix system was designed to accomplish two goals: to teach students specific content and to shape advanced higher order

behaviors. Although the system does not directly instruct students on how to improve their reading comprehension, which is the typical approach of advanced reading comprehension courses, it does directly incorporate stimulus prompting, behavioral shaping, and declining densities of feedback as instructional design features that target these specific higher order behaviors. As such, the system offers significant promise because all of its activities are fully automated and it allows for distributive (internet-based) development on a highly individualized basis. This stands in stark contrast to the more typical and extremely expensive "student resource center" approach for improving reading comprehension. Further, it is worth noting that we are aware of virtually no centers for improving A/V comprehension skills, despite the fact that an argument might be made that, in addition to video, chalk/marker-board illustrated lectures are ubiquitous forms of the A/V communication medium applied to teaching! It is thus truly ironic that so little research on A/V comprehension as a higher order behavior exists.

Further research also needs to be conducted in a more controlled environment to assess how pretest scores might relate parametrically to posttest scores, including how total time individually spent doing adaptive tutoring impacts the percentage of change, and so forth. But MediaMatrix is designed to gradually fade its services in an adaptive fashion as students improve their accuracy and fluency with respect to various types of question format as described earlier. This makes studying individual trajectories quite complicated. Is it when highest tutoring (probe only) levels are reached that one should assess reading comprehension, rather than at the end of the entire course? Probably not, because more complex materials might be encountered subsequently by a student that would result in the adaptive system lowering tutoring levels, thus offering more support. Even such global measures as total time spent tutoring become quite variable and highly individualized based on moment-to-moment student performance fluctuation with respect to different types (levels) of dynamically changing tutoring services.

Perhaps the most stable phenomenon in need of elaboration beyond this paper is the implicit assumption made in MediaMatrix's design that verbal associate test accuracy/fluency is the most salient correlate with the popular academic notion of reading comprehension. Novak and Gowin (1984) present arguments and data that certainly suggest a heuristic relation between associative verbal networks (e.g., their "concept maps") and what most teachers recognize as gains in "knowledge." Research demonstrates that concept mapping enhances comprehension (Chang, Sung, & Chen, 2001, 2002). But behavioral analysis faces a significant challenge in effectively operationalizing such abstractions as "higher order behaviors." The challenge becomes even more daunting when audio *and* visual presentations are proposed as stimuli controlling higher order behaviors similar to those controlled by text in "comprehension."

Nevertheless, when instructors define what they really want to teach in higher education, almost all would target "process" over "content" as

the more lasting contribution a teacher can make to students. We all are likely to include "creative problem solving," "critical thinking," and similar "skills" as being our most desired educational outcomes. Computer-based artificially intelligent and adaptive tutoring systems may finally bring those phenomena within the reach of a broad audience, but research on how these outcomes can be procedurally developed and effectively assessed is a challenge that hopefully will realize a greater future than is offered by its meager past and fledgling present research base.

References

ALEXANDER-TRAVIS, P., BELL, D., DALEY, J. W., DAVIS, A. P., DIBENEDETTO, M., FREEMAN, L. M., et al. (2002). *REA's testbuster for the GRE (Test Preps)*. Piscataway, NJ: Research & Education Association.

BELDEN, N., MIRAGLIA, K., & RAY, R. D. (2003). *Getting students to use adaptive tutorial services: Strategies, issues and outcomes.* Paper presented at the meeting of the Association for Behavior Analysis, San Francisco, CA.

BUTTERFIELD, S., & HOUMANFAR, R. (2003). *Self-paced interactive system of instruction (SPIN) & Psych-Al adaptive instruction: A systematic comparison.* Paper presented at the meeting of the Association for Behavior Analysis, San Francisco, CA.

CACCAMISE, D., & SNYDER, L. (2005). Theory and pedagogical practices of text comprehension. *Topics in Language Disorders,* 25(1), 5-20.

CAMPBELL, D. T., & STANLEY, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

CATANIA, A. C. (1998). *Learning* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

CHANG, K. E., SUNG, Y. T., & CHEN, S. F. (2001). Learning through computer-based concept mapping with scaffolding aid. *Journal of Computer Assisted Learning,* 17, 21-33.

CHANG, K., SUNG, Y., & CHEN, I. (2002). The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education,* 71(1), 5-24.

CONARD, C. J., SPENCER, R. E., & SEMB, G. (1978). An analysis of student self-grading versus proctor grading in a personalized university course. *Journal of Personalized Instruction,* 3(1), 23-28.

ERLMER, P. A., & NEWBY, T. J. (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly,* 6(4), 50-72.

ESTES, W. K. (1971). Reward in human learning: Theoretical issues and strategic choice points. In R. Glaser (Ed.), *The nature of reinforcement* (pp. 16-36). New York: Academic Press.

GRAY, P. (2002). *Psychology* (4th ed.). Gordonsville, VA: Worth Publishers.

HARLOW, H. F. (1949). The formation of learning sets. *Psychological Review,* 56, 51-65.

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.

JOHNSON, K., & STREET, E. M. (2004). *The Morningside Model of generative instruction*. Concord, MA: Cambridge Center for Behavioral Studies.

JONASSEN, D. H. (1996). *Computers in the classroom: Mindtools for critical thinking*. Eaglewoods, NJ: Merrill/Prentice Hall.

KAROLY, P. (1993). Mechanisms of self-regulation: A systems view. *Annual Review of Psychology,* 44, 23-47.

KASSCHAU, R. A. (2000). *Psychology: Exploring behavior.* Winter Park, FL: $(Al)^2$, Inc.

KELLER, F. S. (1968). "Goodbye, teacher…". *Journal of Applied Behavior Analysis,* 1(1), 79-89.

KINTSCH, W. (1998) *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

KINTSCH, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders,* 25(1)*,* 51-64.

LARKIN, J. H., & CHABAY, R. W. (1992). *Computer-assisted instruction and intelligent tutoring systems: Shared goals and complementary approaches.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

LAYNG, T. V. J., STIKELEATHER, G., & TWYMAN, J. S. (2004). Scientific formative evaluation: The role of individual learners in generating and predicting successful educational outcomes. In R. Subotnick & H. Walberg (Eds.), *The scientific basis of educational productivity.* Washington, DC: American Psychological Association.

LAYNG, T. V. J., TWYMAN, J. S., & STIKELEATHER, G. (2003). *Headsprout Early Reading*™: Reliably teaches children to read. *Behavioral Technology Today,* 3, 7-20.

LAYNG, T. V. J., TWYMAN, J. S., & STIKELEATHER, G. (2004). Engineering discovery learning: The contingency adduction of some precursors of textual responding in a beginning reading program. *The Analysis of Verbal Behavior,* 20, 99-109.

MACARTHUR, C. A., & HAYNES, J. B. (1995). Student assistant for learning from text (SALT): A hypermedia reading aid. *Journal of Learning Disabilities*, 28, 150-159.

MAIER, S. E., SELIGMAN, M. E. P., & SOLOMON, R. L. (1969). Pavlovian fear conditioning and learned helplessness: Effects on escape and avoidance behavior of (a) the CS-US contingency and (b) the independence of the US and voluntary responding. In B. A. Campbell & R. M Church (Eds.), *Punishment and aversive behavior* (pp. 299-342). New York: Appleton-Century-Crofts.

MARKLE, S. M., & TIEMANN, P. W. (1970). "Behavioral" analysis of "cognitive" content. *Educational Technology*, 10(1), 41-45.

MARTINSON, T. H. (2003). *Master the GRE CAT 2004 (Academic Test Preparation Series)*. Lawrenceville, NJ: ARCO, Thomson Learning, Inc.

MILLER, J. M., GOODYEAR-ORWAT, A., & MALOTT, R. W. (1996). The effects of intensive, extensive, structured study on GRE scores. *Journal of Behavioral Education,* 6(4)*,* 369-379.

MILLER, L. K., & WEAVER, F. H. (1976). A behavioral technology for producing concept formation in university students. *Journal of Applied Behavior Analysis,* 9(3), 289-300.

MILLER, L. K., WEAVER, F. H., & SEMB, G. (1974). A procedure for maintaining student progress in a personalized university course. *Journal of Applied Behavior Analysis,* 7(1), 87-91.

NOVAK, J. D., & GOWAN, D. B. (1984). *Learning how to learn*. New York: Cambridge University.

PÉREZ-GONZÁLEZ, L. A., SPRADLIN, J. E., & SAUNDERS, K. J. (2000). Learning-set outcome in second-order conditional discriminations. *The Psychological Record,* 50(3), 429-442.

PHILLIPS, V., & BOND, C. (2004). Undergraduates' experiences of critical thinking. *Higher Education Research & Development,* 23, (3), 277-294.

RAY, R. D. (1995a). MediaMatrix: An authoring system for adaptive hypermedia teaching-learning resource libraries. *Journal of Computing in Higher Education,* 7(1) 44-68.

RAY, R. D. (1995b). A behavioral systems approach to adaptive computerized instructional design. *Behavior Research Methods, Instruments, & Computers,* 27(2), 293-296.

RAY, R. D. (2000). Multimodality concept maps and video documentary reconstructions: New uses for adaptive multimedia in learning. In L. Lloyd (Ed.), *Teaching with technology: Rethinking tradition* (pp. 347-359). Medford, NJ: Information Today.

RAY, R. D. (2004). Adaptive computerized educational systems: A case study. In D. Moran & R. Mallott (Eds.), *Evidence-based educational methods* (pp. 143-170). San Diego, CA: Elsevier, Academic Press.

RICE, G. E. (1994). Examining constructs in reading comprehension using two presentation modes: Paper vs. computer. *Journal of Educational Computing Research*, 11, 153-178.

ROBINSON, A., & KATZMAN, J. (2001). *Cracking the SAT* (2002 Ed). New York: The Princeton Review, Random House Inc.

SIDMAN, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA: Authors Cooperative.

SKINNER, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review,* 24(2), 86-97.

SKINNER, B. F. (1968). *The technology of teaching*. New York: Macmillan.

SOWA, J. F. (1984). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley.

SPENCER, R. E., & SEMB, G. (1978). Giving students the opportunity to increase unit size: A performance-based system for personalized instruction. *Journal of Personalized Instruction,* 3(2), 76-80.

TERRACE, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior,* 6, 1-27.

TIEMANN, P. W., & MARKLE, S. M. (1990). Effects of varying interactive strategies provided by computer-based tutorials for a software application program. *Performance Improvement Quarterly,* 3(2), 48-64.

VARGAS, E. A., & VARGAS, J. S. (1992). Programmed instruction and teaching machines. In R. P. West & L. Hamerlynch (Eds.), *Designs for excellence in education: The legacy of B. F. Skinner* (pp. 33-69). Longmont, CO: Sopris.

VERPLANCK, W. S. (1992). A brief introduction to the Word Associate Test. *The Analysis of Verbal Behavior,* 10, 97-123.

YOURGRAU, T. (Sr. Producer) (1990). *Discovering psychology.* WGBH Boston & American Psychological Association.