School Psychology Review
Volume 17, No. 1, 1988, pp. 81-88

# ASSESSMENT OF ACADEMIC SKILLS OF LEARNING DISABLED STUDENTS WITH CLASSROOM MICROCOMPUTERS

Marley W. Watkins
*Southwest EdPsych Services, Inc.*

Joseph C. Kush
*Deer Valley Unified School District*

ABSTRACT

Computer assisted tests have been recommended as a method to make individualized learning systems more manageable, accurate, and efficient. However, until recently the use of computer-assisted testing has been restricted to settings where sophisticated computing technology and assistance were available. Recent advances in computer technology have made computer-assisted testing possible on microcomputers within typical classroom settings. In the present investigation,a criterion-referenced capitalization test was administered to learning disabled elementary school students in microcomputer and paper-and-pencil versions. Results indicated that similar instructional interventions were generated by both versions of the test, but efficiency and student attitudes favored the computerized test. In addition, students completed the computerized version of the test much quicker than the conventional version, and answers were automatically scored by the computer, subsequently reducing test-scoring time and eliminating the possibility of hand-scoring errors. It was concluded that microcomputer-assisted testing, utilizing a tailored testing or adaptive testing model, holds considerable promise in the assessment of educational skills and the design of instructional programs.

Although generally not a preferred role, school psychologists devote a major portion of their time to student evaluations (Carroll, Bretzing, & Harris, 1981; Lacayo, Morris, & Sherwood, 1981). Many of those evaluations involve assessment of academic skills (Shapiro & Lentz, 1985). Academic assessment for instructional intervention and program planning rather than placement has recently received the greater emphasis (Bardon, 1983; Ysseldyke & Mirkin, 1982). As noted by Bejar (1984), norm referenced tests are of little help in designing instruction and, as a consequence, criterion-referenced academic testing has been recommended for the design of individualized instruction systems (Emrick, 1971; Hambleton, 1982, Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham & Husek, 1969).

Conventional criterion-referenced tests are well suited for diagnostic academic assessment, but measurement and administration problems remain. One major problem inherent in the paper-and-pencil format is that all students, regardless of individual differences, must answer the same test items, which often results in examinees receiving numerous items that may be either too easy or too hard. Additional problems include time limit pressures, test security, scoring difficulty, and violations of standardization procedures (McKinley & Reckase, 1980). The use of computer-assisted tests has been suggested as one means of ameliorating

Requests for reprints should be addressed to Joseph C. Kush, Department of Special Services, Deer Valley Unified School District, 20402 N. 15th Ave., Phoenix, AZ 85027.

The microcomputer software used in the present research was written, copyrighted, and published by the first author. To facilitate research in this field, the author will accept research proposals involving free use of this software. Such proposals should be addressed to the author at P.O. Box 1870, Phoenix, AZ 85001.

the weaknesses of conventional criterion-referenced tests thereby making individualized learning systems more manageable, accurate, and efficient (Dunkleberger & Heikkinen, 1982; Eastmond, 1984; McKinley & Reckase, 1980; Weiss, 1982).

Microcomputer assisted testing in classroom settings has been relatively unexplored (McCullough & Wenck, 1984). Andolina (1982) described a microcomputer-based reading test that produced outcomes comparable to paper-and-pencil testing and resulted in positive student reactions, but failed to provide data and a description of her research design. Hasselbring & Crossland (1981) presented spelling tests to learning disabled students on microcomputers, but did not report what effects the use of a microcomputer had on student performance or attitude. Varnhagen & Gerber (1984) administered The Test of Written Spelling (Larsen & Hammill, 1976) in conventional and computerized forms to underachieving and learning handicapped students. Their results indicated the paper-and-pencil version resulted in fewer errors and required less time than a computerized equivalent. Despite these factors, students preferred the computerized version. In explaining the results, the authors suggested that the increased letter-search time required by students unfamiliar with a computer keyboard, may have served to interfere with the cognitive processes they attempted to assess. That is errors which occurred on the computer version may have been typographical errors rather than spelling errors. As such, the extended search time of students unfamiliar with the keyboard may have reflected an uncontrolled variable which served to create artificial differences between the two mediums of presentation.

In conjunction with the paucity of experimental evidence, skepticism concerning the validity and efficiency of microcomputer-based tests is evident among school psychologists. A survey sponsored by National Association of School Psychologists (NASP) discovered that 44% of the respondents believed test administration was a desirable application of computers, but only 14% agreed that tests could be administered more efficiently with a computer (Brantley, Troutman, & Jacob, 1985). The present study was conducted to compare the validity and efficiency of computerized and paper-and-pencil versions of a criterion-referenced capitalization test for learning disabled students in an elementary school setting.

## Method

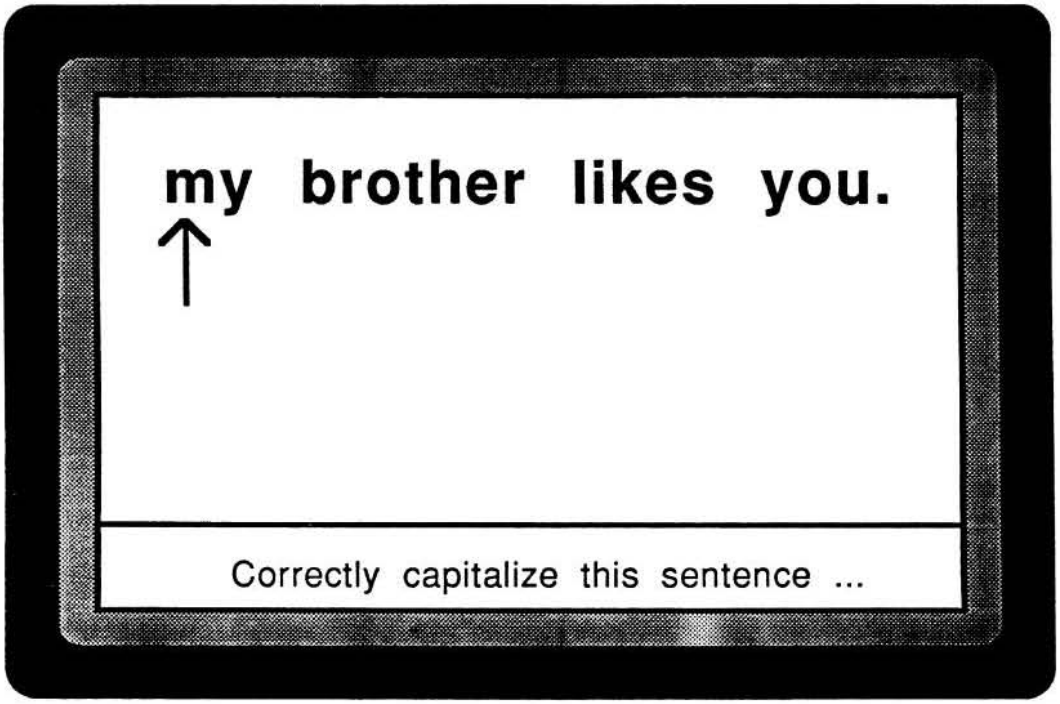
### *Subjects*

All 33 learning disabled students (23 male, 10 female) enrolled in a suburban, southwestern, elementary school served as subjects for this investigation. Average grade placement was 4.5 (range of 1-6) and average full scale WISC-R IQ was 91 ($SD$ = 9.9). Ethnic membership of these students included: 29 Anglo, 3 Mexican-American, and 1 Black. Ten students were members of a self-contained classroom for learning disabled children, whereas the remaining 23 were involved in resource programs. All students had been diagnosed by certified school psychologists as learning disabled based upon a significant discrepancy between ability and achievement.

### *Materials*

*Computer test.* The microcomputer software used to deliver computerized capitalization tests was a modified version of *The Capitalization Machine* (Watkins, 1984). *The Capitalization Machine* is an objective-based, inter-related set of Apple II programs with drill-and-practice, diagnostic assessment, and record keeping functions. For this investigation, the drill-and-practice function was disabled. This package assesses the capitalization domain with 17 discrete objective levels (see Table 1). Presentation consisted of a sentence in 40 column, upper and lower case text (see Figure 1). Students changed letters from upper to lower case, or vice versa, by moving a pointer (via the right and left arrow keys) to the target letter and pressing a key (space bar). Mistakes

FIGURE1
Simulated test item screen from objective level 1 of *The Capitalization Machine*




were corrected by re-pressing the space bar key. Each of the 17 skill levels contained a pool of 10 sentence items tapping that particular capitalization skill (Burns, 1980; Carrico & Welch, 1984; Olson & Kuykendall, 1982). This relatively large number of items per objective was chosen to maximize the accuracy of mastery vs. non-mastery decisions (Berk, 1980; Hambleton, Hutten & Swaminathan, 1976).

There are many approaches to the mechanics of computer-assisted testing (Weiss, 1982). For this investigation, a sequential probability ratio test procedure (Ferguson, 1971) was embedded in the capitalization software. This procedure allowed the a priori specification of mastery and non-mastery criteria for an objective, as well as tolerable Type A and Type B misclassification error rates. Type A error referred to the probability of computer-indicated non-mastery of an objective when, in fact, the student had mastered that particular objective. Conversely, Type B error referred to the probability of computer-indicated mastery of an objective when the student had not, in fact, mastered the objective.

The major advantage of computer assessment which utilizes a sequential probability algorithm is that content domains can be sampled equally or more accurately with smaller numbers of assessment items than if the computerized version were merely a literal translation of the conventional paper-and-pencil version. Specifically, mastery or non-mastery is determined not only by the number of correct or incorrect answers, but equally importantly by the *pattern* of correct and incorrect responses. Mastery criterions and error rates can be altered by the user to establish more or less stringent standards however for the present study, mastery criterion was set at 85% (Millman, 1973), non-mastery level at 40%, allowable Type A error at 20%, and permissible Type B error at 10%.

Given these parameters, in any objective level where a subject missed the first three items in sequence, non-mastery was determined by the computer, and the next

objective level was presented. Similarly, when the subject correctly answered the first five items in a particular level, *in sequence*, mastery was determined and the next objective level was presented. In instances where the subject answered some items correctly but also incorrectly answered other items, assessment continued until the algorithm was "satisfied" that the subject had either "accounted for" the initial errors, in which case mastery was determined, or had some minimal knowledge of the skill objective (or had lucky guesses) in which case non-mastery was determined. Thus, mastery could still be determined even after several errors had been made, however, the subject had to "work harder" to account for his/her mistakes than if s/he had demonstrated mastery initially by correctly answering the first several items in sequence. The exact number of items presented to the subject therefore varied at each objective level, but was determined by the algorithm based on the subject's pattern of correct and incorrect responses.

Initially, the computer randomly selected a capitalization item from a specified objective level and presented it to the student. The student's responses were scored as either correct or incorrect and a proficiency ratio (proportion correct) was calculated by the computer and statistically compared, via the sequential probability ratio test, to the prespecified mastery and non-mastery criteria. If the proficiency ratio was equal to or lower than the non-mastery criterion (taking into account Type A error), a non-mastery decision was made and testing proceeded to the next objective level. If the proficiency ratio was equal to or greater than the mastery criterion (taking into account Type B error), a mastery decision was registered and testing proceeded to the next objective level. Proficiency ratios between mastery and non-mastery criteria resulted in the selection of another item from the same objective level. This iteration continued until a definitive mastery or non-mastery decision was achieved or the pool of sentences for that objective was exhausted. In the later case, a classification of "review" was attached to that objective and the computer proceeded to test the next objective in the sequence. Thus, the sequence of testing was always from objective 1 to objective 17 with a mastery/review/non-mastery classification being registered for each objective.

*Conventional test.* The conventional paper-and-pencil test was composed of the same capitalization sentences found in the computerized test. Using the same group of 10 sentence items per objective level the paper-and-pencil capitalization test contained 170 sentence items, randomly sequenced by objective level and presented 20 per page in 14 point upper and lower case type. Students changed letters from upper to lower case, or vice versa, by writing that letter onto the test protocol atop the incorrect letter case.

### *Procedure*

Students were randomly divided at mid-year into two groups for initial testing. One-half of the students received computerized assessment of their capitalization skills while the other students took the equivalent paper-and-pencil capitalization test. Following an 8 week interim period, all students were again tested on capitalization skills, but with opposite versions of the test. Thus, all students were administered both versions of the capitalization test in a counterbalanced design.

*Computer test.* Testing was conducted on Apple //e microcomputers equipped with monochrome monitors. Testing was supervised by the student's special education teacher, and began with scripted instructions and two practice sentences. Teacher assistance was available during the practice items, however no assistance was provided after testing began. Students were encouraged to do their best and to guess if they were uncertain of an answer.

*Conventional test.* Conventional testing utilized the paper-and-pencil version. Testing was supervised by their special education teacher who presented instructions similar to those used for computer testing. As with the computer testing no

assistance was provided following completion of practice items and all but two students completed the entire 170 item test. These two students were in first-grade and were unfamiliar with higher level skill objectives. For these students non-mastery was registered for each incomplete level.

## Results

Student performance on each of the 17 capitalization objectives was characterized as mastery, review or non-mastery. These classifications were performed automatically by the computer, whereas the conventional tests were scored by hand and classified into the three ordinal categories utilizing the same mastery and non-mastery criteria applied by the computer (i.e., >84%, 41%-84%, and <41%, respectively). Classifications were then transformed to numerical values, where 3 was equal to mastery, 2 equated to review, and 1 equaled non-mastery. Finally, scores were summed across all 17 objectives to produce two total capitalization test scores: one for the computerized version and one for the conventional version.

The computerized test resulted in a mean of 25.36 ($Md$ = 25) and standard deviation of 8.23. The conventional test version had a mean of 27.67 ($Md$ = 26) and standard deviation of 8.83. The difference between test means was significant ($t$ = 2.47, $p < .05$), while test variances were homogeneous ($t$ = .65, $p > .10$). The Pearson product-moment correlation between scores on computerized and conventional tests was .806 ($p < .001$).

From a diagnostic assessment perspective, it is more meaningful to examine test results in terms of resultant instructional interventions than to make a normative comparison of sample means. Consequently, student performance on each objective was characterized as instructional (performance at the non-mastery or review level suggested instruction was needed) or non-instructional (classification at the mastery level indicated no further instruction was required) and submitted to a series of 2 (mastery vs. non-mastery) by 2 (computer vs. paper-and-pencil) chi-square analyses. Table 1 indicates that computerized and paper-and-pencil test versions did not significantly differ ($p > .10$) in their assignment of students to instructional interventions.

Student attitudes about testing medium were assessed on a 5-point pictorial scale (happy/sad faces) two weeks subsequent to the completion of academic testing. The difference between student attitudes toward computerized and conventional capitalization assessment was significant ($t$ = 7.43, $p < .001$), with the computerized test (Mean = 4.61; $SD$ = .86) being perceived in a more favorable light than the conventional paper-and-pencil test (Mean = 2.67; $SD$ = 1.43).

## Discussion

The present investigation found small, but statistically significant, mean differences between criterion-referenced academic tests administered in microcomputer versus conventional paper-and-pencil versions. Substantial correlations between test methods were obtained, as were homogeneous test variances. In reviewing the testing literature Green (1984) predicted that response differences would cause means to differ when computerized and conventional tests were compared. He suggested further, as the present study verified, that correlations between versions would be high and that variability would be similar, and concluded that "a mean shift of this sort is not a serious psychometric problem; a new set of norms is definitely required, but they can be obtained by a simple adjustment of the present norms" (Green, 1984, p. 6). In this regard, mean differences in the present study could easily have been modified by relaxing the prespecified Type B error rate employed in the program's statistical decision rule.

As previously noted, mean comparisons reflect a normative test orientation which is of little help in designing instruction (Bejar, 1984). When compared within a diagnostic assessment framework both

TABLE 1
Classification Differences Between Computerized and Conventional Capitalization Objectives

| Objective | Percent Requiring Instruction | | Percent Requiring Non-Instruction | | Chi-Squares |
|---|---|---|---|---|---|
| | Computer | Conventional | Computer | Conventional | |
| First word in sentence | 39 | 55 | 61 | 45 | .97 |
| Proper nouns — person's name | 55 | 64 | 45 | 36 | .25 |
| Proper nouns— geog. names | 70 | 88 | 30 | 12 | 2.27 |
| Proper nouns — awards, buildings | 76 | 91 | 24 | 9 | 1.75 |
| Proper nouns — Businesses | 91 | 94 | 9 | 6 | 0.00 |
| Proper nouns — school classes | 79 | 91 | 21 | 9 | 1.06 |
| Proper nouns — compass points | 73 | 91 | 27 | 9 | 2.55 |
| Proper nouns — race religion | 76 | 73 | 24 | 27 | 0.00 |
| Pronoun "I" | 48 | 48 | 52 | 42 | .24 |
| Calendar days, months etc. | 88 | 97 | 12 | 3 | .87 |
| Abbreviations — Mr., Mrs., etc. | 61 | 52 | 39 | 48 | .25 |
| Abbreviations - others | 91 | 88 | 9 | 12 | 0.00 |

test versions produced decisions which would have resulted in similar patterns of instructional interventions.

Student attitudes were significantly more favorable toward the microcomputer administered capitalization test than toward the traditional paper-and-pencil version of the test. Anecdotically, teachers reported that, on average, students required 75 minutes to complete the conventional capitalization test, but only 30 minutes to take the computerized test. In addition, the computerized test automatically scored student responses and thereby eliminated hand marking of test protocols.

This reduction in both administrator time and potential scoring errors are obvious advantages of computerized testing. Such positive results are contrary to those reported by Varnhagen and Gerber (1984) for a normative spelling test and suggest academic testing on classroom microcomputers should emulate the methods proven effective on large computers within defense and university settings, rather than attempting to literally translate a conventional test to computerized form. That is, an algorithm for selecting test items, evaluating responses, and making mastery decisions for individual students must be an integral component of any microcomputer-based test. This type of system has been labeled adaptive or tailored testing (McKinley & Reckase, 1980; Vale, 1981; Weiss, 1982) and takes advantage of the microcomputer's strengths. In contrast, translating conven-

tional tests to a computer format simply duplicates the weaknesses of conventional testing methods while confounding them with new sources of error (e.g., keyboard-unfamiliarity).

Results from the present study suggest several advantages of computerized academic assessment over a traditional paper-and-pencil version. While the computerized version was more time-efficient and was preferred by students, accuracy was not sacrificed as substantial correlations between versions were demonstrated, as were similar (mastery vs. nonmastery) patterns for subsequent educational instruction. Although the present investigation represents a starting point for the generation of future research, additional data which examines test-retest reliability and internal consistency must be collected before equivalence between the two testing modes can be determined. In this regard, potential ethical and technical problems exist if computerized psychodiagnostic assessment is not evaluated with the same critical rigor as are more traditional methods of assessment (American Psychological Association, 1985; Jacob & Brantley, 1986; Thomas, 1984; Walker & Myrick, 1985). One microcomputer system which offers promise in this direction has been developed by Nitko and Hsu (1984). This general purpose, tailored-testing system allows the user the opportunity to modify test items to pre-established criteria, throughout the construction of the test. Once psychometric equivalence is established however, computerized academic assessment may become preferred to more traditional methods of paper-and-pencil assessment if as the present study shows, computerized assessment remains superior in terms of efficiency, student-preference and accuracy.

## References

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Andolina, M. (1982). Reading tests: Traditional vs. computerized. *Classroom Computer News, 2,* 39-40.

Bardon, J. I. (1983). Psychology applied to education: A specialty in search of an identity. *American Psychologist, 38,* 185-196.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21,* 175-189.

Berk, R. A. (1980). Practical guidelines for determining the length of objectives-based, criterion-referenced tests. *Educational Technology, 20,* 36-41.

Brantley, J., Troutman, L., & Jacob, S. (1985). A survey of computer use, attitudes, and ethical concerns among school psychologists. *CTASP Newsletter,* Summer, 7-10.

Burns, P. C. (1980). *Assessment and correction of language arts difficulties.* New York: Charles Merrill.

Carrico, J. P., & Welch, P. (1984). *English writing and language skills.* New York: Harcourt.

Carroll, J. L., Bretzing, B. H., & Harris, J. D. (1981). Psychologists in secondary schools: Training and present patterns of service. *Journal of School Psychology, 19,* 267-273.

Dunkleberger, G. E., & Heikkinen, H. (1982). A review of computer-generated repeatable testing. *AEDS Journal, 15,* 219-225.

Eastmond, D. V. (1984). Math assessment videodisc. *The Computing Teacher, 12,* 57-59.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement, 8,* 321-326.

Ferguson, R. L. (1971). *Computer assistance for individualized measurement* (Report No. 71/8). Pittsburgh: University of Pittsburgh, Learning Research and Development Center.

Green, B. F. (1984, August). *Computer-based ability testing.* Paper presented at the meeting of the American Psychological Association, Toronto.

Hambleton, R. K. (1982). Advances in criterion-referenced testing technology. In C. Reynolds & T. Gutkin (Eds.), *The handbook of school psychology.* New York: John Wiley.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. (1976). A comparison of several methods for assessing student mastery in objectives-based instructional programs. *Journal of Experimental Education, 45,* 57-64.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1-47.

Hasselbring. T. S., & Crossland, C. L. (1981). Using microcomputers for diagnosing spelling problems in learning-handicapped children. *Educational Technology, 21*, 37-39.

Jacob, S., & Brantley, J. C. (1986). Computer applications in school psychology: Proposed guidelines for good practice. *CTASP Newsletter*, Winter.

Lacayo, N., Morris, J., & Sherwood, G. (1981). Daily activities of school psychologists: A national survey. *Psychology in the Schools, 18*, 184-190.

Larsen, S., & Hamill, D. (1976). *Test of written spelling*. Austin: ProEd.

McCullough, C. S., & Wenck, L. S. (1984). Current microcomputer applications in school psychology. *School Psychology Review, 13*, 429-439.

McKinley, R. L., & Reckase, M. D. (1980). Computer applications to ability testing. *AEDS Journal, 13*, 193-203.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research, 43*, 205-216.

Nitko, A. J., & Hsu, T. (1984). A comprehensive microcomputer system for classroom testing. *Journal of Educational Measurement, 21*, 377-390.

Olson, M. C., & Kuykendall, C. B. (1982). *The writing process: Composition and applied grammar*. Boston: Allyn & Bacon.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

Shapiro, E. S., & Lentz, F. E. (1985). Assessing academic behaviors: A behavioral approach. *School Psychology Review, 14*, 325-338.

Thomas, A. (1984). Issues and concerns for microcomputers in school psychology. *School Psychology Review, 13*, 469-472.

Vale, C. D. (1981). Design and implementation of a microcomputer-based adaptive testing system. *Behavior Research Methods & Instrumentation, 13*, 399-406.

Varnhagen, S., & Gerber, M. M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly, 7*, 266-270.

Walker, N. W., & Myrick, C. C. (1985). Ethical considerations in the use of computers in psychological testing and assessment. *Journal of School Psychology, 23*, 51-57.

Watkins, M. W. (1984). *The capitalization machine* (Computer program). Phoenix, AZ: Southwest EdPsych Services, Inc.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Ysseldyke, J. E., & Mirkin, P. K. (1982). The use of assessment information to plan instructional interventions: A review of the research. In C. Reynolds & T. Gutkin (Eds.), *The handbook of school psychology*. New York: John Wiley.