

Purpose

Reading achievement has been increasingly politicized in recent years; resulting in intense pressure on schools to improve standardized test scores. High stakes tests are used for accountability purposes, based on the assumption that scores reflect meaningful reading competencies. Children who select a high percentage of correct answers on multiple-choice (MC) reading comprehension tests are judged to be proficient in reading comprehension. MC reading tests are widely used due to their purported reliability and cost-effectiveness. But questions have been raised regarding the validity of these tests in measuring the reading comprehension skills that children use in natural reading activities. To what extent do MC and constructed response (CR) tests measure the same underlying construct, reading comprehension?

A number of studies have examined the relationship between test format and student performance with mixed results (Mehrens, 1991). Substantial differences between MC and CR test scores have been documented in some studies (Shavelson & Pine, 1992; Shepard, 1997). Other studies have provided evidence of near perfect correlations between test formats after scores have been corrected for attenuation (Bennett, Rock, & Wang, 1991; Bridgeman & Lewis, 1994; Bridgeman & Rock, 1993; Lukhele, Thissen, & Wainer, 1994; Traub, 1993). Studies could not be located that examined format differences in the measurement of elementary reading comprehension achievement.

Numerous researchers have pointed out the problems associated with the MC format. Messick (1989) found that children's varying levels of test-taking skills make MC tests susceptible to construct-irrelevant difficulty or easiness, threatening test validity. Shepard (1997) suggested that the distortion of MC scores due to test-wiseness, allows some students to demonstrate knowledge they don't actually have. According to Kintsch (1998) the question-answer format does not tap deep-level understanding of texts. He recommended against the use of questions and answers for scientific purposes in the assessment of comprehension due to the artificial and invalid inferences that may result from such tests. In addition to concerns about validity, a number of studies have found that MC tests provide an advantage to male students (Breland et al., 1994; Lane, Wang, & Magone, 1995; Poplun & Capps, 1999).

Educators agree that reading comprehension proficiency must be measured accurately if educators are to evaluate student progress, determine effective methods of literacy instruction, and hold schools accountable for student learning. The psychometric quality of standardized MC tests is cited as a justification for using these tests for high-stakes purposes. A number of states such as California rely on MC tests as a single measure of reading achievement, providing schools with rewards and sanctions based entirely on students' performance on the SAT 9 test.

The over-reliance on standardized MC test scores highlights the importance of investigating this test format. The purpose of this study was to examine format differences in the measurement of fifth grade students' reading comprehension to determine whether MC tests accurately measure reading comprehension skills. Additionally, the study was designed to explore the association between gender and test format to determine whether MC reading comprehension tests provide an advantage to one gender over another.

Theoretical Rationale

Cognitive theorists, Kintsch and Van Dijk (1978) developed a model for conceptualizing the procedural abilities that underlie reading comprehension. They theorized that comprehension involves a complex series of activities including the organization of text into a coherent whole, the condensation of meaning into a gist and generation of new texts, summaries, based on recall and reconstruction. The process of comprehension is dependent on the identification of key discourse structures known as propositions. According to Kintsch's construction-integration (CI) model, text representations are developed sequentially in a word-by-word, sentence-by-sentence process as meaning is constructed (1998). The CI model provides a framework for understanding how reading comprehension works and how it is measured by MC and CR tests.

Method

Participants and Setting

The study took place in an affluent suburban elementary school located in Northern California. Seventy-two percent of the students are white, 16% are Asian, 5% are Latino. The elementary school SAT 9 reading scores for the prior year (2001) averaged in the 80th percentile. One hundred and twenty-five students in four intact fifth grade classes participated in the study. The participating students included 62 boys and 63 girls.

Research Design

The research design was correlational, designed to explore the association between two sets of reading comprehension tests. The MC tests, developed for a prior study, consisted of four 200-word social studies passages and corresponding 10-item MC item sets. A total of 30 items were designed to measure reading comprehension and 10 items measured vocabulary achievement. The CR tests consisted of two of the same 200-word social studies passages that were found on the MC test that the students were required to read and summarize. Scoring of the CR tests was based on 10 important propositions that the researchers identified in each of the two text passages.

The researchers conducted all assessment activities in the students' regular classrooms. The MC test was administered first. Four weeks later the researchers returned to the classrooms and administered the CR tests. Students were given the first passage and asked to read it. The passages were collected and students were asked to write a detailed summary (narrative or bullet-point format was accepted.) The process was repeated with the second passage.

The researchers rated the summaries based on the list of propositions that had been generated. Forty tests were randomly selected for calibration of scoring. The researchers scored the summaries, compared scores and resolved disagreements by discussion. Inter-rater reliability averaged .85 on the scoring of the two summaries.

Results

Results of the analysis showed an overall correlation of .62 between the MC and CR tests. The correlation between the CR items was .58. The MC mean was 22.69 (30 items on the test) and the standard deviation was 5.59. The CR mean was 9.21 (20 propositions on the test) and the standard deviation was 3.29. Scores were converted into percentages to allow for a meaningful comparison across test formats. Results indicated that student achievement was much higher on the MC test (76%) than on the CR test (46%).

An analysis of histograms for the two tests was particularly interesting. The MC test scores produced highly skewed results (-.95, std error .22), while the CR test produced a slightly skewed pattern more closely resembling normal curve (-.34, std error .22). The histogram for the MC test suggests that the test items may have been too easy for many of the students in this high-performing school, creating a ceiling effect. This pattern is not evident on the CR histogram. The lack of symmetry in the histograms suggests a format-related advantage for high-performing students associated with the MC test.

When data were disaggregated by gender, results indicated that girls outperformed boys on all measures, with significant achievement differences on the CR test. The correlation between the MC and CR test scores was .48 for boys and .72 for girls. (Correlations were significant at the .01 level.) A regression analysis was performed with the MC reading comprehension and vocabulary scores used as predictors of CR test achievement. R^2 was calculated for each gender, indicating that the predictors explained only .26 of the variance of boys' scores compared to .54 of the variance in girls' scores. Results indicated that girls' performance was more consistent across test formats.

Educational Importance

Results of this study suggest that constructing summaries from text propositions was very difficult for the students despite the fact that they had been tested on both text passages four weeks earlier. It seems logical that if children had read and comprehended the texts on the MC test, memory of passage content would have contributed to higher scores on the CR tests. This did not turn out to be the case. Overall scores were much lower on the CR test, suggesting little transfer of information from the MC test that supported the construction of text summaries. What exactly did the MC tests measure?

When children completed items on MC and CR tests distinct differences in test taking behavior were observed. When given the MC test many children skipped reading the text altogether and went directly to the test items. They read the item and then skimmed over the text until a response was located, repeating the process for each of the 30 comprehension items. The behavior observed on the MC test, was a highly fragmented series of activities designed to locate information rather than to construct meaning from a text. The "skim and locate" activities observed in the MC test may have been effective in producing high test scores but reflected little engagement with the text and few of the behaviors have been identified as integral to the reading comprehension process.

The test-taking behaviors observed on the CR test were quite different than those observed on the MC test. When given the text for the CR test nearly all of the children appeared engaged in reading. When they began writing their summaries, children listed propositions in the exact order that they occurred in the text. The children's performance on the CR test supported Kintsch's theory that texts are recalled through a series of propositions and that comprehension is constructed in a sequential manner.

Girls' reading comprehension achievement was substantially more consistent than that of boys across test formats. According to Shepard (1997) variability in test performance results from children's fragile understanding and inability to generalize knowledge from one format to another. To follow Shepard's reasoning, it is possible that the girls in this study had learned to see beyond the differences in test format, discerning the deeper structure in reading comprehension tasks. The less consistent performance of boys may have resulted from greater reliance on a "skim and recognize" strategy for identifying the correct answers on MC test. In short, boys' MC scores may reflect test-taking skills rather than reading comprehension proficiency.

These results suggest the importance of teaching children the cognitive processes that underlie reading. When children learn how to identify propositions, formulate a gist of a text, fill in inconsistencies with inference, and to construct an accurate summary of what has been read they are developing authentic reading comprehension skills that transcend test format. The emphasis on test-taking skills, on the other hand, teaches children to look for correct answers rather than understanding. As Shepard (1998) points out, an emphasis on MC test preparation may reduce children's ability to generalize knowledge in addition to reinforcing ineffective reading skills.

These results highlight a serious deficiency in the way children's reading comprehension achievement is measured. Most high stakes decisions regarding school accountability are based on standardized MC tests. These tests may not be a valid measure of reading comprehension achievement, particularly for girls. MC tests appear to inflate boys' scores substantially more than those of girls, a problem that is accentuated as students grow older. It is likely that MC test are a flawed methods of measuring children's reading comprehension proficiency and should not be used for high stakes accountability purposes.