DOCUMENT RESUME

ED 392 015 CS 012 338

AUTHOR Greer, Eunice Ann
TITLE Examining the Validity of a New Large-Scale Reading Assessment Instrument from Two Perspectives. Technical Report No. 623.
INSTITUTION Center for the Study of Reading, Urbana, IL.
PUB DATE Nov 95
NOTE 54p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Evaluation Methods; Factor Analysis; Grade 3; Primary Education; *Reading Achievement; Reading Comprehension; Reading Research; *Reading Tests; *Test Validity
IDENTIFIERS *Illinois Goal Assessment Program; *Large Scale Assessment

ABSTRACT

A study evaluated the validity of the Illinois Goal Assessment Program (IGAP) grade-3 reading assessment from 2 perspectives: its relationship to 21 other measures of reading via factor analyses, and the sensitivity of the IGAP to instruction in 3 schools via multiple regression. Data were collected in a longitudinal study involving 350 students. Factor analyses indicated the tests load on 2 factors, a comprehensive factor and a factor contrasting reading rate/accuracy with narrative comprehension. Two elements of the IGAP failed to load on either factor: the literacy survey or the metacognitive measure. The three regressions models were consistent across sites. After controlling for entering ability, home influences, and grade-1 and grade-2 teachers, instructional activities in grade 3 that focus students' attention on comprehension are associated with higher IGAP scores. Instructional activities that focus students' attention on decoding are associated with lower scores. (Contains 95 references, and 8 tables and 1 figure of data.) (Author/RS)

Technical Report No. 623

EXAMINING THE VALIDITY OF A NEW LARGE-SCALE READING ASSESSMENT INSTRUMENT FROM TWO PERSPECTIVES

Eunice Ann Greer
University of Illinois at Urbana-Champaign

November 1995

Center for the Study of Reading

TECHNICAL REPORTS

College of Education
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

CENTER FOR THE STUDY OF READING

Technical Report No. 623

EXAMINING THE VALIDITY OF A NEW LARGE-SCALE READING ASSESSMENT INSTRUMENT FROM TWO PERSPECTIVES

Eunice Ann Greer
University of Illinois at Urbana-Champaign

November 1995

College of Education
University of Illinois at Urbana-Champaign
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

Abstract

A study was conducted to evaluate the validity of the Illinois Goal Assessment Program [IGAP] grade-3 reading assessment from two perspectives: (a) its relationship to 21 other measures of reading via factor analyses, and (b) the sensitivity of the IGAP to instruction in three schools via multiple regression. Data were collected in a longitudinal study involving 350 students. Factor analyses indicate the tests load on two factors, a comprehension factor and a factor contrasting reading rate/accuracy with narrative comprehension. Two elements of the IGAP failed to load on either factor: the literacy survey and the metacognitive measure. The three regressions models were consistent across sites. After controlling for entering ability, home influences, and grade-1 and grade-2 teachers, instructional activities in grade 3 that focus students' attention on comprehension are associated with higher IGAP scores. Instructional activities that focus students' attention on decoding are associated with lower scores.

EXAMINING THE VALIDITY OF A NEW LARGE-SCALE READING ASSESSMENT INSTRUMENT FROM TWO PERSPECTIVES

In the 1980s, reading instruction and instructional materials began to reflect research from the previous decade that defined reading as a dynamic interaction of a text, a reader, and a context (Michigan Reading Association, 1987; Wixson & Peters, 1984). This definition stressed the importance of *active, thinking* readers: readers who use their knowledge of topic, context, and text structure to strategically construct their own models of meaning from the text. The extent to which readers are successful in this effort is not influenced by cognitive factors alone. Affective factors, including literacy habits and attitudes, have an equally important influence on a reader's ability to construct meaning. In 1986, reading researchers surveyed the large-scale, reading assessment instruments available to schools and warned consumers and educators alike that reading assessment had not kept pace with the advancements in reading research and instruction (Valencia & Pearson, 1987; Wixson & Peters, 1987).

This dynamic definition of reading more closely mirrors the way children learn (Eva Baker, personal communication, April 1989). It also has implications for the way teachers teach reading. Among other things, it suggests that it is important to help students learn to use what they know to make sense of what they read, to strategically monitor their attention as well as their success in interpreting the author's message, to make timely and parsimonious use of fix-up strategies when necessary, and to discriminate between important and unimportant information.

These changes in the way reading is taught lead to changes in teachers' goals for their students, the way teachers and schools defined gains and successes, and in some of the abilities teachers looked for when they attempted to diagnose students' strengths and weaknesses. However, until very recently, these widespread curricular advancements have not been reflected in large-scale assessments. In 1987, Valencia and Pearson warned that when this mismatch exists between what is taught and what is measured, teachers are not rewarded sufficiently for good teaching. "High" scores on the assessments are assumed to mean that things are going well when, in fact, they may not be. Also, teachers have to take time out from teaching to prepare students for tests that do not measure what the teachers teach for the other 8 months of the school year.

This "mismatch" can be manifested in any or all of the following ways:

- Research has led both basal authors and teachers to realize and address the role of prior knowledge in reading comprehension, but older assessments attempt to vitiate the influence of prior knowledge on comprehension by using a number of short passages on various topics;

- Research has led both basal authors and teachers to teach students to recognize and use common text structure and organizations to support comprehension, but older assessments make use of short, contrived "test text" that does not approximate the complexity or structural integrity of texts that teachers use for reading instruction;

- Recent research has led both basal authors and teachers to place more emphasis on the role that inferencing within and across text(s) plays in successful comprehension, but because of the lack of length and complexity of passages found on older tests of reading, the questions that accompany the test passages are predominantly literal;

- The use of more inferential questions in teacher's editions and classroom discussions has taught students that questions can, and often do, have more than one right answer, but older tests do not allow for alternative interpretations of the text;

- Newer curricular materials are giving increased attention to the teaching of purposeful, strategic reading behaviors, but older tests seldom include questions that evaluate students' awareness or flexible use of reading strategies;

- Successful reading demands the simultaneous use of many abilities that combine to help the reader actively construct meaning from text, but older measures of reading ability often fragment the reading process into a number of subskills that are tested in isolation (Wixson & Peters, 1987).

Clearly, teachers who are asked to teach students how to become active, strategic readers *and* how to do well on older standardized reading measures are confronted with the question "How do I implement a dual curriculum?"

Through their assessment development efforts, Pearson and Valencia (1987) began to address the mismatch between assessment and instruction. They introduced a new, large-scale reading assessment that is more closely aligned with newer curricular trends in reading instruction. The tests include questions that activate and measure students' entering knowledge, make use of "real" texts, focus comprehension questions on important elements of the text, include questions that assess students' sensitivity to and knowledge of reading strategies, and include affective measures related to reading.

Research Questions and Hypotheses

The purpose of this study is to evaluate the validity of the grade 3 reading portion of the Illinois Goal Assessment Program (IGAP), which was developed by Pearson and Valencia. The study investigates the IGAP's relationship to various other group and individual measures of reading comprehension via factor analyses and looks at the instrument's sensitivity to differing instructional practices via regression analyses. Three different instructional sites that reflect three differing, currently held beliefs about reading instruction (based on classroom observations) provide the data for the analyses.

The questions investigated are as follow:

Question 1

What is the factor structure of the four components of the 1987 IGAP grade-3 reading assessment? Do other formal measures of reading ability load on the same factors? (A complete list of measures to be evaluated is presented in the methods section of this report.)

Hypothesis. An unpublished study by Krug (1987) showed that the four components of the IGAP reading assessment load on four separate factors: construction of meaning, centrality reading strategies, problem-solving reading strategies, and literacy habits and attitudes. In the present study, it is hypothesized the IGAP constructing-meaning and topic-familiarity items will load on a factor with items from other comprehension measures, a factor that will include items from the individual reading inventories, group comprehension measures, and prior knowledge tests. The measures of reading rate and accuracy will either load alone or will be contrasted with reading-comprehension measures. The reading strategies section of the IGAP will load alone, or perhaps with items from the Circus--Think it Through (Educational Testing Service, 1976b) and the Open Court Error Detection Test (Open Court, 1987). The Literacy Survey will load on a separate factor, as well.

Question 2

Is the 1987 version of the grade-3 IGAP sensitive to instructional differences across three instructional sites? To wit, are there sound, research-based recommendations that can be given to schools about how they can adjust instruction to students' abilities to read and perform on the IGAP?

Hypothesis. After accounting for home influence, kindergarten ability, and the influence of grade-1 and grade-2 teachers, there will be differences in IGAP scores that will be explained by grade-3 differences in average time spent in reading and decoding, mean number of types of questions asked, and instructional versus non-instructional time. Increases in interactions that focus students' attention on the construction of meaning from text are likely to be associated with increases in student performance. Interactions which focus students' time on decoding tasks, to the exclusion of the process of constructing meaning, are likely to have a negative relationship with the dependent measure, the IGAP reading test.

Conclusion

This study has practical and theoretical significance. Tests are important to the extent that people have faith in the test results and allow their judgments and behavior to be affected by those results. It is therefore essential that researchers and test developers do all that they can to ensure that the test they have developed is valid for the purposes for which it is intended to be used. It is also important to determine that the instrument is sensitive to the behaviors and abilities it purports to measure. And it is important to know if the test is sensitive to modifications in instruction recommended by current research and instructional materials.

It is incumbent upon test developers to show that new measures meet the desired, conventional standards with respect to test reliability and test validity. If the measures do not, they will not be viewed as being competitive, from a measurement standpoint, with older instruments, and the mismatch between what takes place in reading instruction and how reading is tested will continue.

REVIEW OF THE LITERATURE

The present study relies on two areas of the literature: cognitive science, and reading research that supports the "new" definition of reading, and current traditions of validity research.

Research that Supports the "New" Definition of Reading

These newer assessments are reflective of changes in comprehension instruction suggested by four lines of applied research: methods of teaching students to activate and use relevant topic familiarity to aid comprehension, methods for teaching students to recognize and use text structure to facilitate comprehension, methods that teach students to master and use flexibly appropriate reading strategies to aid comprehension, and methods to improve teachers' selection and use of comprehension questions. All four lines grew, not surprisingly, from the research on schemas and metacognition. A discussion of the impact of these lines of research is presented below.

Prior Knowledge Research

Schemas are abstract knowledge structures which, if correctly activated, can aid the recipient of information in processing, prioritizing, proofing, and later recalling or using that information (Anderson, 1977; Mandler & Johnson, 1977; Rumelhart, 1975). Work that applied schema-theory to memory and reading recall illustrated the impact the activation of different schemas can have on the recall of reading passages (Anderson & Pichert, 1978; Bransford & Johnson, 1972; Pichert & Anderson, 1977). This work contributed to the development of the model of the "dynamic" reader, one who interacted with the text

to produce a recall, or model, which included components of the text as well as components of the reader's own knowledge of the topic. The more readers knew about a topic, the easier it was for them to construct more complete models.

Instructional research that followed this initial work looked at ways of teaching children to use their knowledge to aid them in recalling stories and to answer questions. Children's knowledge about a topic is a powerful predictor of their performance and understanding (Lipson, 1982; Pearson, Hansen, & Gordon, 1979; Stein & Glenn, 1979). Knowledge of topic was shown to influence young readers' abilities to make inferences and reason beyond the explicitly stated material in the text (Collins, Brown, & Larkin, 1980; Fielding, Anderson, & Pearson 1988; Pearson et al., 1979; and Raphael & Pearson, 1985).

Text Structure Research

A second area addressed by new testing efforts is fidelity to text structure. Texts have internal, organizational structures by which they are readily identifiable. Two of the more common structures, or genres, are narration and exposition. Studies by Bartlett (1978), Armbruster (1979), and Fitzgerald and Spiegel (1983) point to the following: not all students, even beyond the beginning grades, have a knowledge of text structure. Students can be taught to identify and use story or passage structures to facilitate their comprehension, and students can be taught to transfer and use this new knowledge beyond the text used in the lesson (Bowman, 1981; Spiegel & Fitzgerald, 1986).

Reading Strategies Research

A third area of reading that was not addressed by traditional forms of reading assessment is the area of metacognition. This omission is readily understandable if one recalls that until fairly recently, many people saw reading comprehension as the accurate reproduction of what the author had put down on the page. Until recently, reading was taught and measured as though it were a predominantly unidirectional process (Brown, Bransford, Ferrara, & Campione, 1983; Collins et al., 1980).

Baker and Brown (1984) described the characteristics of the "expert reader" and identified six behaviors that good readers use to help themselves read and remember text:

1. Understanding the purpose for reading as well as the demands of the task,
2. Recognizing the important information in the text,
3. Focusing attention on important rather than unimportant information,
4. Monitoring reading to make sure that comprehension is occurring,
5. Questioning one's understanding to make sure that the purpose for reading is being fulfilled,
6. Employing fix-up behaviors when comprehension has failed.

Baker (1979), Collins et al. (1980), and Brown, Armbruster, and Baker (1985) provided empirical evidence that expert readers employ a variety of strategies. Additionally, individuals differ in their selections of strategies. Further refinements of our knowledge of "expert" readers have shown that flexible use of strategies, as well as the confidence one has in one's selection of appropriate strategies, is a further hallmark of an expert (Seigler, cited in Snow, 1989). Research involving young readers has

shown that they differ markedly from this model of experts (Brown, Campione, & Day, 1981; Phillips, 1989).

There are developmental differences in how purposefully children monitor their understanding and how well they act on the information they do extract from this monitoring process (Flavell, Speer, Green, & August, 1981; Markman, 1977; Paris & Lindaur, 1976). Brown and Palincsar (1982) and Brown, Palincsar, and Armbruster (1984) demonstrated that children's ability to summarize and discriminate important from unimportant information can be improved by instruction. In a series of studies of "reciprocal teaching," Palincsar and Brown showed that children, particularly poor readers, can be taught to use strategies exhibited by experts: predicting, questioning, summarizing, and clarifying. These strategies are taught and practiced in the context of reading lessons. The process was first used with low-achieving seventh graders and has since been shown to be effective with children as young as 6 and 7 years of age (Baker & Brown, 1984). This work also refuted earlier studies such as those by Gall et al. (1975) and Winne (1979) which suggested that focusing students' attention on higher-order thinking skills such as inferencing, summarizing, and predicting would cause drops in literal comprehension and produce difficulties in learning. (See also Wixson, 1983.)

Research on Teachers' Selection and Use of Questions

Since Durkin's 1978 and 1979 studies of what teachers do in their classrooms under the heading of "reading instruction," much has changed in the way teachers ask questions and the informal assessment information they expect to derive from children's responses to their questions. The way we view the relationship between the questions teachers ask and the answers they receive has changed, as well (Pearson & Johnson, 1978). In their 1978 book on the teaching of reading comprehension, Pearson and Johnson presented a new way of looking at teachers' questions, a way which takes into account the role of the reader's background knowledge and experience.

Pearson and Johnson's taxonomy did much to bring the manner in which we design and analyze questions in line with the emerging dynamic definition of the reading process. Their categories--explicit, implicit, and scriptally implicit, along with their explanation that the Question Answer Relationship, not the isolated question stem is what must be analyzed--changed the way teachers and researchers, alike, looked at comprehension instruction. Studies which followed the development of the taxonomy have supported the instructional relevance and validity of its categories (Pearson, Hansen, & Gordon, 1979; Raphael, 1984; Raphael & Wonnacott, 1985; Thompson, Gipe, & Pitts, 1985; Wixson, 1981).

Additional work in the area of reading comprehension instruction has elaborated on comprehension-fostering questioning techniques which are not part of the taxonomies discussed thus far. Many of these questioning techniques are found in currently used basals and in newer tests of reading comprehension, as well. The category of textually implicit questions has been broken down into categories that differentiate between inferences which are made between one or two connected sentences (locally implicit) and inferences which occur across larger blocks of text (distant) (Hare & Pulliam, 1980). An additional question type, which is not directly related to topic knowledge or metacognitive awareness alone and is not part of a current taxonomy, is summarizing. The instructional validity, or "teachability," has been documented by Brown and colleagues in two studies which suggest that better readers produce better summaries, students can be taught to summarize, and instruction in summarizing facilitates comprehension (Brown et al., 1981; Brown et al., 1984).

A final area of research related to comprehension instruction, which has affected instruction and newer assessments, is question selection, or focus. The level of importance of teacher questions has been studied systematically by classroom observers. Allington (1984) and Durkin (1978-1979, 1984) agreed that teachers ask many questions of relatively little importance to the story line of the text. Beck, McKeown, McCaslin, and Burke (1979); Beck, Omanson, and McKeown (1982); and Fielding (1988)

reported that questions which follow the story line, or story map, produced better understanding of content than did questions which focused students' attention on irrelevant details. Taken together, these works suggest that the focus of the question is as important as the question type.

In summary, applied research related to the areas of schema theory and metacognition has helped to shape a more dynamic model of reading, one in which the reader interacts purposefully with a text to construct meaning. This model has led to research-based changes in the way we teach reading, the materials we use to teach reading and, more slowly, the tests we use to evaluate the efforts of our teachers and the progress of their students.

Issues of Test Validity

How do we assess the validity of a test? And what does it mean to say that a test has "validity?" Like the field of reading comprehension assessment, the area of test validity has recently gone through a redefinition of sorts. Until very recently, researchers in testing pointed to four general aspects of validity: concurrent, content, construct, and predictive (Angoff, 1988). The concept of validity, although always assumed to be incumbent upon the ethical researcher, was not clarified as a psychometric issue until relatively late in educational research history, the mid 1940s. At that time, validity was defined as "a correlation of scores on a test with some other objective measure of that which the test is used to measure" (Bingham, cited in Angoff, 1988, p. 214). The chief application of validity research during the late '40s and '50s was for purposes of prediction: the use of a test to screen for later performance such as college or the military service. Hence the term "predictive validity."

A second type of criterion-related validity, concurrent validity, was not recognized in its own right by the American Psychological Association until 1954. Concurrent validity evaluates the performance of one measure based on its correlation with another, already accepted measure of the same trait or construct. Concurrent validity was necessary to prove that a new test or shorter version of an existing test or battery was sufficiently correlated with older, accepted measures to warrant claims that the new version was testing the same trait or construct. Anastasi (1982) distinguishes between predictive and concurrent validity based on the different objectives for testing. "Concurrent validation is relevant to tests employed for diagnosis of existing status, rather than prediction for future outcomes" (p. 137 - 138).

Curricular and content validity issues were introduced and debated in the '40s and '50s as they related to achievement tests and proficiency tests. It was argued by Rulon and others that curricular tests needed no validity evaluations because, if they were exhaustive, then they fell into a class of "obviously valid." Rulon's claims were challenged in the late 1950s and eventually disproven by Loevinger and others who showed the difficulty associated with adequately representing and sampling the domains, objectives, and skills associated with any given subject matter. Not only would the number of items necessary to adequately test some domains be unwieldy, but since a bank of "all possible items" for a domain is rarely, if ever, written, a reliable sample of manageable size could not be obtained either (Angoff, 1988).

The concept of "face validity"--does the test look like what it purports to measure--was introduced in the late '40s, but was not pursued seriously because of its ambiguity. It is, however, still regarded as an important issue related to test use (Anastasi, 1982, p. 136). While face validity may not be proven empirically, it is necessary in order to guarantee acceptance by those who purchase and use the test. If a reading test does not look like a reading test, teachers will be less likely to use it. And, if its use is mandated, teachers are likely to place little faith in the results.

In 1954, the APA, AERA, NCME standards, for the first time, listed a fourth type of validity: construct validity. This new concept of validity changed the way we think of validity because it suggested an evaluation process, or cycle. In this process, the items that make up the test are shaped by the theory

or *construct* being measured, and the results from the test (not just answers, but all relevant data which emerge from other forms of validity checks) influence or shape the theory as well as future iterations of the test. Construct validation then involves a process and does not result in a single number or "answer" about a test. The underlying theory shapes the first selection of items; the information gained from testing modifies the theory, which in turn modifies future item selection, and so on. This work, introduced by Cronbach and Meehl in middle 1950s, is similar to Gulliksen's earlier description of intrinsic validity (Angoff, 1988).

Growing out of this new work in construct validity, Campbell and Fiske introduced an empirical test of construct validity known as discriminant validity. It subjected the test in question to a series of comparisons with two other groups of measures. One group of measures, though different in format and content, measures the same construct as the test being evaluated. The other group of measures, though similar in measurement method, differ with respect to the construct being tested. The test in question should correlate more strongly with measures of the same construct, regardless of format, and less strongly with tests measuring another construct. Furthermore, the sets of correlations among tests within a construct should be higher than the correlations of tests of like format, but different construct (Angoff, 1988). In his 1975 article, "The standard problem: Meaning and values in measurement and evaluation," Messick acknowledged that the work of Campbell and Fiske was a strong test of the theory of construct validity because it considers a plausible rival hypothesis for the relationships among test scores. Once that rival hypothesis can be discounted, the notion of construct validity gains empirical support. He went on to suggest that tests of construct validity should not be limited to correlational analysis, but should include experimental studies, evidence about interpretations and content coverage and relevance. Over time, what emerged from this work was a shift in the concept of the test validation process. Messick (1975, 1980) and others concluded that construct validity involved the integration of all other forms of validity into a common framework for evaluating theoretically relevant relationships.

Messick has gone on to challenge the idea of just exactly what is to be validated: the test, the situation, the match between user and test, the interpretation of test results, the use of those results, and/or the extent to which the use of test results is fair and ethical (Messick, 1980, 1988a). This broadening of the concept of validity was echoed by Cronbach in his 1988 keynote address to attendees at the ETS conference entitled "Test Validity for the 1990s and Beyond" when he challenged: "Validation speaks to a diverse and potentially critical audience; therefore, the argument (of validity) must link concepts, evidence, social and personal consequences and values" (p. 16). In his 1988 article on future issues in validity, Messick explains that, under this new "unified" view of validity described by Cronbach, all efforts toward the validation of an instrument must have as their foundation, construct validation, for there can really be no validity at all without it. However, because of the power of tests, and the potential influence they can have on peoples lives, construct validity alone does not answer the question "How valid is this test?" The social and political influences and implications must come into play as well. Test validation should be an overall evaluation of the "adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1988b, p. 42). The 1985 *Standards for Educational and Psychological Testing* (American Psychological Association, 1985) also reflects this emerging definition of validity.

It is in this newer context of validity that the current study is undertaken. The current study examines the IGAP Reading Test as it relates to other measures of a similar nature (reading comprehension tests) and of a somewhat different nature (tests of letter, sound, and word recognition). It goes on to look at the relevance of the IGAP Reading Test as it relates to classroom practices and the relationship of subjects' test performance to their classroom experiences. The present study is limited in scope and will not directly evaluate the impact of social and political forces on the Illinois Goal Assessment Reading test or, vice versa, the social and political impact of the Illinois Goal Assessment Reading Test.

METHOD

Data for the present study were collected as part of a much larger longitudinal study of how children develop the ability to comprehend written text and how they acquire science concepts. Data were collected from four sources: (a) throughout the school year during kindergarten and Grade 1, and Grade 2, 9 (6 during grade 3) full-day classroom observations were conducted by trained observers; (b) student work samples were collected during each of the observations; (c) beginning and end-of-year individual and group tests were administered; and (d) parent questionnaires were collected annually. The multi-site, longitudinal nature of the study and the use of multiple measures of reading ability offered rich and unique opportunities for the evaluation and validation of individual reading assessment instruments.

Subjects

Two cohorts of students from three school districts participated in the study. Only the data from students in cohort 1 are used in the present study because students in cohort 2 did not take the IGAP reading assessment in the spring of their grade-3 year. Students in cohort 1 entered kindergarten in the fall of 1983. The three school communities differ with respect to geographic location, SES, ethnic makeup, and professional diversity of parents. Schools were selected because of their differing educational approaches, philosophies, and curricular materials. The following descriptions of the three districts were taken from a Center for the Study of Reading Technical Report written by the study's principal investigators, Meyer, Wardrop, and Hastings (1989).

> District A is located in a fairly self-contained small town in central Illinois. The school participating in the study has a fairly homogeneous student body. The students entered kindergarten with mixed abilities. There are approximately 80 students per cohort divided into four classes for kindergarten and first grade and three classes for second and third grade. This district is known to have high student performance in reading, particularly in the very early grades, and average student performance in science. The district's educational philosophy includes whole-class instruction in all subjects beginning with an academic focus in kindergarten. Kindergarten classes are characterized by a high percentage of time spent in teacher-directed reading instruction. Teachers in district A almost never group children for instruction. This decision is due, in part, to the district policy which states that all students at a grade level will cover the same content each year. Regular classroom teachers in this district maintain primary responsibility for the children in their classes who have special needs. Some of these teachers often gear many of their instructional interactions and feedback to their lower performing students. Kindergarten teachers used the *Alpha K Time* program for beginning reading instruction (Reiss & Friedman, 1976). They introduced Houghton Mifflin's *Getting Ready to Read* at mid-year (Durr & Hillerich, 1981). In addition, they provided reading instruction for some children in three pre-primers. Reading instruction for all children in first, second and third grades continued in the Houghton Mifflin series (Durr et al., 1983). Some teachers also provided numerous opportunities for children to read supplementary books.
>
> District B is in a small town in Illinois. It is located just a few miles from a larger town to which many of its citizens commute to work. The school participating in this study has about 150 children in Cohort 1. This district has a tradition of average student achievement in reading in the lower grades and higher achievement in reading in the middle grades. Science achievement is characterized as better than average in the lower grades. Teachers in district B begin grouping for reading instruction in kindergarten, and this process continues through the elementary grades. Classes are

divided into as many as five or six groups for reading instruction in first, second, and third grades. Thus, students in these classes average considerable time each day doing independent seat work. Special needs teachers begin to play important roles with lower-performing children in second grade in this district. The Harcourt Brace Jovanovich reading program is used by all teachers in grades kindergarten through third in District B (Early, Cooper, & Santeusanio, 1983). Typically, teachers organize a variety of learning centers set up at tables. They introduce their classes to these activity centers on Monday mornings. Students complete work from these centers in the mornings while other students meet with the teacher for reading groups. In kindergarten, teachers meet with an average of two groups per day for teacher-directed instruction. In first, second and third grades, on average, teachers meet with all of their groups 4 out of 5 days a week. . . . The year that cohort 1 entered kindergarten was the first year that all teachers, district-wide, were directed to use the Harcourt Brace Jovanovich materials with every child in kindergarten. While teachers had some latitude about when and how they used the materials, all teachers were expected to provide similar instruction from the curriculum.

District C is located in a suburb of a major city in Illinois. The school participating in this study has many characteristics usually associated with an urban school. It has a very heterogeneous student population. Children are of mixed socioeconomic and ethnic backgrounds. About 40% of the children in the school are Black, 20% are Hispanic, and the remaining 40% are White. There are approximately 85 students per cohort. Students are in three "classes" which are grouped into self-contained kindergarten classes and combination first/second grade homerooms. Children are then regrouped to one of six teachers on a "team" who then typically has at least three groups for reading instruction. Special needs teachers play an important role in district C. Bilingual instruction begins in kindergarten for all children whose parents choose it. These children are pulled out of their regular classrooms for a half hour each morning and afternoon. Beginning in first grade, other teachers work with "low stanine" children in pull-out groups to provide additional reading instruction using the *Distar Reading Program* (Engelmann & Hanner, 1972). No basal reading materials were used in any of the three kindergarten classrooms. In fact, even after carefully tracking reading vocabulary introduced by the three kindergarten teachers, no common reading vocabulary words were found. Regular classroom reading instruction at the first, second, and third grade levels was from the Ginn reading series (Clymer et al., 1976).

Materials and Data Collection Method

The materials selected for inclusion in the analyses for the present study are taken from three of the four longitudinal study data sources: the individual and group evaluation tools used to measure grade-3 reading performance and an entering kindergarten ability measure, the parent questionnaires administered at the end of kindergarten, and the grade-3 classroom observational data. A summary of the evaluation measures included in these analyses is presented here.

As part of the data collection process, at the beginning and end of each school year, all students in the longitudinal study participated in a battery of individual and group tests. As was stated earlier, the number and range of group and individual tests administered to the students is large in comparison to most studies and offers a unique opportunity for evaluation of those instruments.

WRAT (Wide Range Achievement Test--Revised Level 1). The reading subtest of the WRAT was administered at the beginning of grade 3. The letter recognition (if needed) and word identification

sections of the WRAT were individually administered to all students. Students read aloud to the examiner a list of 75 increasingly difficult words. The examiner terminates the testing session as soon as the child has made 12 consecutive errors. Students' correct responses are summed to produce a raw score (Jastak & Wilkinson, 1984).

Woodcock Reading Mastery Test--Passage Comprehension, Form A. This test was administered at the beginning of grade 3. The Reading Comprehension portion of this test, a cloze test of comprehension, was administered individually to each child. This 85-item cloze reading measure is made up of items of increasing difficulty. Each item is presented on a card. The students are directed to read the items to themselves and provide an appropriate word or phrase to fill in the blank. A total raw score is computed by summing the number of correct responses (Woodcock, 1973).

IRAS (Interactive Reading Assessment System). The IRAS was administered at the beginning of grade 3. This individualized reading inventory was administered to each child. It is made up of a series of increasingly difficult word lists followed by seven reading passages and comprehension questions. Performance on the word lists determines the student's point of entry on the passage comprehension measures which follow. The examiner times the child's rate of reading for each passage, records oral reading errors and then asks the child the accompanying comprehension questions. The termination point is based on the number of oral reading errors and the number of comprehension questions which the student answers incorrectly (Calfee & Calfee, 1982). Four scores are obtained from the IRAS: average passage reading rate, average number of errors, average number of self-corrected errors, and total number of comprehension questions answered correctly.

CIRCUS Listening, Level D, Form X. This test was administered at the beginning of third grade. This is a group-administered, 40-item measure of listening comprehension. Children listen to a story read aloud by the examiner. Periodically the examiner stops reading and asks a question. The children are directed to mark the appropriate picture. A raw score is calculated by summing the number of correct responses (Educational Testing Service, 1976a).

CIRCUS Think It Through, Level D. This test was administered at the beginning of grade 3. This is a group-administered, 33-item measure of problem-solving skills, pattern recognition, seriation, and maze completion. It is similar in many ways to a group-administered I.Q. test for young children. Children listen to directions that pertain to an item or series of items; they mark their responses in their booklets. A raw score is calculated by summing the number of correct responses (Educational Testing Service, 1976b).

Diagnostic Science Reading Passages, Form B (DSRP). This test was administered individually to third graders in May. This individual reading inventory is similar to the IRAS, discussed above, except that there are no word lists for the student to read, and both of the passages are about some aspect of science. The two passages used in this test are much longer and of greater difficulty than the most difficult passages on the IRAS. The examiner times the child's rate of reading for each passage, records oral reading errors and self-corrections, and then asks the child the accompanying comprehension questions. Four scores are obtained for this test: number of oral reading errors across the two passages, number of self-corrections made while reading the two passages, rate of reading the two passages, and number of comprehension questions answered correctly (Meyer, Hastings, & Linn, 1987a).

Open Court Error Detection. This test was administered to all third graders in May. This is a group-administered measure of students' sensitivity to inconsistencies in text. Students read a series of passages and underline the part of the text which is inconsistent with the rest of the information presented in the passage. Texts become increasingly lengthy and complex. The inconsistencies become less "glaring" and, hence, more difficult to identify as the child progresses through the series. A raw score is calculated by summing the number of inconsistencies the child identified (Open Court, 1987).

STEP Science Test, Level E. This test was administered to all third grade students in May. This is a group-administered measure of students' ability to recognize and apply science vocabulary and concepts. This is a 40-item, multiple-choice test. The students read each item to themselves and then mark their response in their test booklets. A total score is calculated by summing the number of correct responses (Educational Testing Service, 1979).

NSF Third-Grade Science Test, Living Things. This test was administered to all third-grade students in May. The test was developed and piloted by the principal investigators of the longitudinal study in 1985. It was revised in 1987 to include more difficult items. It is a group-administered, 35-item test of students' knowledge of living things. The items and answer choices on this test are read aloud to students. Ten of the items are traditional multiple-choice items. The other 25 items are blocks of 3 to 12 questions nested within a stem. Answer choices are of three kinds for these items: yes/no, yes/maybe/no, and all/some/none. Students record their answers in their test booklets. A total score is calculated by summing the correct responses (Meyer, Hastings, & Linn, 1987b).

Reading Portion of the Illinois Goal Assessment Program. This test was administered to all grade-3 students in May. This is a group-administered measure of students' ability to use their own knowledge in concert with the information presented in a full-length text to construct meaning from the text. Each student reads one narrative and one expository passage. Each passage is accompanied by 45 items.

Twenty of the items are prior-knowledge questions that precede the passage. Five of the prior-knowledge items are single answer, multiple-choice vocabulary questions about words that are related to the concepts the students will encounter in the passage. The other 15 prior-knowledge questions are prediction questions. The prediction questions are preceded by a short summary of the story the students will later read. This summary is followed by 15 ideas. The students must decide how likely they think it is that a given idea will appear in the upcoming story or article. Their answer choices are yes, maybe, and no.

Following the prior knowledge sections, the students read a full-length passage that closely approximates the length and complexity of passages found in grade-3 materials. Fifteen constructing-meaning or comprehension questions follow the passage. A story map of each passage is used to guide item writing and selection. Thus, all items focus on information that is central to the understanding of the selection. Items are of seven general types: explicit, implicit (within and across paragraphs), application/transfer, characterization, author's craft, and vocabulary. Each of these multiple-choice items has five answer choices. Each item may have one, two, or three correct answers. The scripted directions encourage students to look back at the passage to help them answer the items. Ten reading-strategies questions follow the constructing-meaning section.

The reading-strategies section is comprised of two subsets of five items each. In each subset, students are presented with a scenario that describes a problem they might have encountered while reading the passage they have just finished. Each scenario is followed by five possible solutions. The students must indicate how helpful they think each solution would be in solving the problem presented in the scenario. Their answer choices are: will help a lot, will help quite a bit, will help a little bit, won't help at all.

The test is followed by a survey of students' in-school and out-of-school literacy habits and attitudes. The prior knowledge, vocabulary, reading strategies and literacy survey items were read aloud to the students by the examiner. Students read the passages and constructing-meaning items to themselves (Pearson & Valencia, 1987). Items on this test are not scored in the traditional, dichotomous, right/wrong tradition. Prior-knowledge and reading-strategies items are scored using a discrepancy scoring model. Students are awarded two points for matching the key, one point for the closest answer to the keyed response, and no points for a choice which is more than one "step" from the keyed response.

The scoring of the constructing-meaning section is somewhat more complicated. For each item, 12 points are divided equally among all right answers and 12 negative points are divided equally among the wrong answers. For example, if an item has two correct answers, each correct response carries a value of +6. Each incorrect response carries a value of -4. Students are given the assigned point value for each distractor they select. Using the example above, if a student selected a single correct response and two of the incorrect responses, then their score would be calculated as (6 + (-4) + (-4)) = -2. To convert these item scores to a positive scale with a range of 0 to 1, a linear transformation is performed. These values are then summed across items to produce a total score with a range of 0 to 15.

The literacy survey measures the frequency of students' participation in literacy-related activities. Answer choices for the items are "never," "some of the time," "a lot of the time." Students receive two points for activities which they report participating in "a lot of the time," one point for activities they report participating in "some of the time" and zero points for activities which they report participating in "never." A total literacy score is calculated by summing students' response values.

Nine scores are derived from this test: a topic familiarity raw score for passage one and passage two, a constructing-meaning raw score for each passage, reading-strategies raw scores for problem solving and for centrality for each passage, and a literacy-survey response score.

Kindergarten Measures

In addition to the grade-3 measures, an entering-ability measure was used in the regression modeling. This measure, Ability-zero, is a scale value that was calculated based on confirmatory factor analyses of all entering kindergarten measures (Meyer, Wardrop, & Hastings, 1990a). Factor analyses of the kindergarten data indicated that three measures, all administered in the fall of kindergarten, loaded on a single factor. These measures are Wide Range Achievement Test - Reading Subtest--Level 1 (Jastak & Wilkinson, 1984), Circus Listening Test--Level A (Educational Testing Service, 1976a), and the Analogies Subtest of the Language and Problem Solving Battery (Mason & Meyer, 1983). Because all three scores did not have equal factor loadings, a factor scale was created by standardizing each raw score and multiplying each by its factor score coefficient to arrive at a weighted value for each score (Kim & Mueller 1978; Mulaik, 1972). These weights are, respectively, .71, .67 and .74. These three new values were summed to arrive at a scale value, Ability-zero, for each student.

Wide Range Achievement Test--Reading Subtest--Level 1 (WRAT). This test was described previously in this section, as it was also administered as part of the grade-3 measures.

Circus Listening Test--Form X--Level A. This test was administered in the fall of kindergarten. This is a group-administered, 24-item measure of listening comprehension. Children listen to a story read aloud by the examiner. Periodically, the examiner stops reading and asks a question. The children are directed to mark the picture that answers the question. A raw score is calculated by summing the number of correct responses (Educational Testing Service, 1976a).

Analogies Subtest of the Language and Problem Solving Battery (Mason & Meyer, 1983). This test was administered in the fall of kindergarten. This is an individually administered 10-item subtest in which students must supply the missing word to complete the analogy. A total score is computed by summing the number correct.

Parent questionnaires. Parent questionnaires were given to each child at the end of each year of the study. The questionnaire is comprised of questions about general home and family characteristics and literacy behaviors. The questionnaires were designed to obtain an index of opportunities and support for informal learning and literacy activities. As part of the larger, longitudinal study, based on theoretical "relatedness" and item correlations, items were grouped into the following scales: the *parent*

reads to child scale, which is composed of 7 items; the *child participates in reading*, which is made up of 6 items; the *parental resources scale*, which is made up of 12 items; the *parental support of literacy and schoolwork*, which comprises 4 items; and the *parental instruction scale*, which also contains 4 items.

Classroom Observation and Coding Methods

To obtain observational data, each classroom was observed six times during the grade-3 school year. Observation dates were distributed evenly throughout the school year. Each observation was a full school day in length. A trained observer sat in the classroom and made a written record of everything that took place in the classroom. Each observer recorded the beginning and ending time for each lesson; the types of lessons; lengths of non-instructional time, for example, recess, transition or clean-up; and the percent of students on task during sustained seat work, as well as the number of praises and corrections made to individuals or groups of students. In addition to recording this overall description of how the day was broken up and how "on task" students were, observers kept a written record of all teacher-initiated interactions. Observers categorized and recorded the number and type of teacher-initiated interactions or questions, the identity of the student or students to whom each interaction was directed, and the nature of the teacher's feedback to the student or students. Using these data, it is possible to determine the number or proportion of questions of a given type a teacher directed to a given child or group of children in a single day, or throughout the school year. It is possible to look at differences in numbers or types of questions asked from classroom to classroom and from school to school. It is also possible to look at the amount of time, in minutes, spent in different types of instruction. The coding scheme for reading-related activities reflects the levels of questions identified in Pearson and Johnson's 1978 taxonomy as well as schema theory and story grammar research. Question categories include, but are not limited to, background knowledge, implicit plot, explicit plot, implicit and explicit characterization, word meaning, decoding, sentence production, sequencing, and recognition of text structure and study skills. To be sensitive to instructional changes that take place over time, the coding system was augmented when necessary.

Method of Data Analysis

For ease of reference, the research questions presented earlier are repeated here:

1. What is the factor structure of the 1987 grade-3 IGAP, and do other formal measures of reading ability load on the same factor or factors?

2. Is the 1987 version of the grade-3 IGAP sensitive to instructional differences across three instructional sites, to wit, are there sound, research-based recommendations that we can give to schools about how they can adjust instruction to increase performance in students' abilities to read and perform on the IGAP?

To answer question 1, a single, exploratory factor analysis was performed across all three sites. This analysis examined the underlying pattern structure across the ten evaluation measures outlined in the materials section. The correlation matrix used in the analysis contained correlations among the 22 scores associated with the measures described in the materials section. They are: the total WRAT score (Jastak & Wilkinson, 1984), the total Woodcock score (Woodcock, 1973), the total number of comprehension questions correct on the IRAS (Calfee & Calfee, 1982), the average reading rate per passage for the IRAS, the average number of oral reading errors per passage for the IRAS, the average number of self-corrected oral reading errors for the IRAS, the total correct for CIRCUS Listening Test (Educational Testing Service, 1976a), the total correct for CIRCUS Think-It-Through (Educational Testing Service, 1976b), the total number of comprehension questions correct on the Diagnostic Science Reading Passages (Meyer et al., 1987a), the average reading rate per passage for the Diagnostic Science Reading Passages, the average number of oral reading errors per passage for the Diagnostic Science

Reading Passages, the average number of self-corrected oral reading errors for the Diagnostic Science Reading Passages, the total number correct for Open Court Error Detection (Open Court, 1987), the total number correct for STEP Science Test (Educational Testing Service, 1979), the total number correct for NSF Third Grade Science Test (Meyer et al., 1987b), the prior knowledge score for passage one of the IGAP reading test (Pearson & Valencia, 1987), the prior knowledge score for passage two of the IGAP reading test, the constructing meaning score for passage one of the IGAP reading test, the constructing meaning score for passage two of the IGAP reading test, the reading strategies centrality score for the IGAP reading test, the reading strategies problem-solving score for the IGAP reading test, and the literacy survey score from the IGAP.

Following the extraction of the initial factors, Varimax and Oblimin rotations were evaluated for theoretical congruence and interpretability. The model was reduced and rerun by specifying the number of factors. The rotated factor structure which was most interpretable is the focus of the results and discussion sections. While this study has been described as an investigation of the relationship between various reading measures, there are two science tests on the list of measures to be included in the factor analysis. The decision to include these two tests was based on the well-documented role of background knowledge in reading comprehension (Lipson, 1982; Pearson et al., 1979; Stein & Glenn, 1979) and the contents of the IGAP reading test. The second passage on the IGAP reading test is an expository passage entitled "How Plants Help People." This allowed for the comparison of text-specific prior knowledge measures on the IGAP to other more general domain-specific content-area measures.

The second research question was examined using multiple regression techniques. A separate regression equation was developed for each of the three sites. This reduced the existence of "noise," or error variance, in the models. The three regression equations were developed using hierarchical and blocked multiple regression techniques. Within each of the blocks, stepwise regression techniques were used.

Question 2 was designed to investigate to what extent differences in third-grade teachers' instructional practices within each of three schools influenced student performance on the IGAP reading test. To answer this question as cleanly as possible, entering school ability, some index of parents' contributions, and the effects of grade-1 and grade-2 teachers were accounted for before attempting to look at the contribution made by grade-3 teachers. Criterion scaling was used to account for the effects of first- and second-grade teachers on students' performance on the dependent measure. The scale was constructed using students' IGAP scores. Two scales were developed, one for grade 1 and one for grade 2. Thus, there is a single scale value associated with each grade-1 and grade-2 teacher in the study. This method accounted for all possible combinations of teachers, within sites, far more conveniently than any other form of coding would have. The initial full regression model is presented below in Figure 1.

[Insert Figure 1 about here.]

This full model was evaluated for each site. To control, as far as possible, the spurious effects attributable to small sample sizes, the final block of variables was trimmed, and the full model rerun. Initially, the final block in the regression analysis contained approximately 52 variables to be evaluated for entry using the stepwise procedures. This number (52) varied slightly from site to site because some instructional variables did not occur within a given site. In order to reduce this number, following the first regression run, all of those variables from the final block which had F values less than one following step four and following the final step of the regression were deleted. Variables which were not significant from the first four steps were also deleted. The trimmed model was then rerun to produce the site-specific reduced model for each site. Because there were still too many variables in the District C model, where the number of subjects is smallest, the trimming procedure was repeated prior to producing the final reduced model.

While there were a little over 300 students in the study, not all of them had been part of the study since kindergarten. Since it would have been impossible to partial out any sort of entering ability or first or second grade teacher effects from students who entered the study from other schools after fall of the kindergarten year, only those subjects who had been enrolled in their respective schools since kindergarten were included in this second analysis. While this decision did impose the most rigorous form of post hoc statistical control possible on such an investigation, it also reduced the number of subjects substantially. There were 52 complete cases available for analysis from District A, 83 from District B and 42 from District C.

Taken together, the two analyses described above begin to empirically evaluate the validity of the Illinois Goal Assessment Program reading assessment, based on its relationship to other measures of reading ability and its "fit" to three different instructional sites.

RESULTS AND DISCUSSION

Analysis 1--Factor Analysis

Descriptive Results

Before presenting the results of the factor analysis, the characteristics of the data will be discussed. The data set presents an opportunity to investigate a uniquely rich mix of variables. It includes expository comprehension measures that are matched topically to science content-area tests in addition to a particularly wide range of individual and group comprehension measures. The descriptive statistics for the 16 test scores included in the factor analysis are presented in Table 1. Table 1 presents the means, standard deviations, observed minimum and maximum and the possible minimum and maximum for each test for the entire population across the three sites. These values are not reported for reading rates, self-corrections, or reading errors.

[Insert Table 1 about here.]

Examination of the means and standard deviations of the 16 tests suggests that 10 of the measures are somewhat negatively skewed. The WRAT, the IRAS Comprehension, the Circus Listening Test, the Circus Think-It-Through, the Open Court Error Detection, the Step Science Test, the two IGAP Comprehension measures, and the two IGAP prior knowledge measures all have negatively skewed distributions. There are two explanations for these distribution characteristics--scoring methods and test difficulty. The IGAP comprehension and topic familiarity distributions are, to some extent, a function of the scoring method used. The multiple-answer comprehension measures have a chance level of 50% because, similar to a true/false test, students must decide to mark or not mark each of five distractors linked to each stem. The topic familiarity subtests have a slightly lower chance rate of approximately 33%. Answer choices for this section are yes, maybe and no, and students have a 66% chance of getting at least one out of a possible two points for each question. In the case where the keyed response is Maybe (approximately one third of the items), students are guaranteed at least one point. The other six negatively skewed distributions are, more directly, a function of test difficulty.

Of the remaining six tests, five are fairly normally distributed. The sixth, the Science Passages (DSRP) is somewhat positively skewed. While none of these distributions appears to be so skewed as to make its use in the analyses inadvisable, they create limitations which should be kept in mind in interpreting the data. The reduced variability brought about by the 11 skewed distributions is likely to suppress test reliabilities, as well as correlations among variables and correlations of variables with factors.

Reliabilities of the tests included in the factor analysis are presented in Table 2. These values are not reported for reading rates, self-corrections, or reading errors. Reliabilities were calculated using

Cronbach's Alpha. Of the 16 measures, 9 have high reliabilities of .75 or more. The remaining 7 measures reported in Table 2 have reliabilities less than .70. Of the 7 tests, four were identified earlier as having skewed distributions and reduced variability. The moderate reliabilities of the WRAT, the Circus Listening Test, and the two IGAP prior-knowledge measures are, no doubt, partially a function of their reduced variability. The reliability of the NSF Living Things test, .609, is not a function of range restriction. It is most likely a result of subject fatigue. The test includes 144 items nested within 35-item stems. A review of the other tests in this study reveals that this test is two to four times longer than many other group-administered measures.

[Insert Table 2 about here.]

The reliability of the two IGAP reading strategies measures is very low: .214 for Centrality, .215 for Problem Solving. These findings are not the result of restricted variability. The descriptive statistics in Table 1 suggest each is fairly normally distributed, with a slight suppression of high scores, but no apparent floor effect. There may have been a slight fatigue effect. (Each half of the test, narrative and expository, takes approximately 1 hour to administer, and strategy items are last.) However, fatigue has not been found to be a problem in any of the larger, statewide pilot efforts. Nor should the reading level of these items have presented a problem, as these items were read aloud to students. The most plausible explanation is the low number of items, ten centrality and ten problem-solving, and the fact that these items do not appear to work well with young students. The format of these items is novel, as are the response formats. Recall that this is the section of the IGAP in which students are presented with a scenario which includes a problem they might have encountered while reading the passage they just completed. The scenario is followed by five alternative solutions to the problem. The students must evaluate each solution and indicate how helpful it would be in solving the problem presented in the scenario. This explanation is supported by additional pilot and exploratory studies conducted by Krug (1987, 1989). In statewide pilot test analyses the reliability of the third-grade reading strategies item sets was lower than it was at the other three grade levels, but slightly higher than in the present study. The reliability of the reading strategies items on the statewide pilot tests, across six test forms at grade 3, ranged from .25 to .53. The median reliability at grade 3 was .42.

Factor Analysis Results

In the following section, the communalities of variables included in the factor analysis and the factor pattern matrix will be presented and discussed.

To analyze the underlying relationships among the 22 variables presented in the preceding sections, an exploratory factor analysis was conducted using SPSS/PC+, the Statistical Package for the Social Sciences (Norusis/SPSS, Inc. 1988). Exploratory factor analyses procedures were chosen over confirmatory procedures because there have been insufficiently conclusive results of factor analysis of reading measures to warrant a specific hypothesis of the latent structure of these particular measures. Initially, principal axis factoring extracted four factors; all had eigenvalues greater than 1.00. The Varimax rotation of these four initial factors produced a rotated factor matrix which was made up of a factor comprised of comprehension measures and prior-knowledge measures, a second factor that contrasted fluency and some of the comprehension measures, and the third and fourth factors, which were uninterpretable.

Based on these initial results, the factor analysis was rerun. This time, the number of factors to be extracted was specified as two. Factor loadings below .300 were not reported. Again, Varimax and Oblimin rotations were requested. Using these new specifications, both methods of rotation converged in fewer than 10 iterations. The Oblimin solution of this second analysis proved to be the more interpretable and the more theoretically congruent of the two. The communalities for the variables in the analysis are presented in Table 3.

Table 3 presents the proportion of variance in each variable that is explained by the two-factor model. Since the communality of a variable cannot exceed its reliability, three of the variables, the IGAP centrality and problem-solving reading strategies measures and the IGAP literacy survey, with communalities of .0250, .0195, and .0847 respectively, share very little with the rest of the variables in the analysis. The low reliabilities of the reading strategies measures make it difficult to interpret the meaning of these very low communalities.

[Insert Table 3 about here.]

The communality of the literacy survey presents a different picture. The reliability of the literacy survey is .812. The communality of the literacy survey, however, is .0847. This suggests that the literacy survey is measuring some latent aspect associated with reading which may not be being captured by the two factors identified in the factor analysis.

The factor pattern matrix produced by the Oblimin rotation is presented in Table 4. The first factor, which has a eigenvalue of 8.852 is best described as a comprehension factor which is defined most strongly by expository comprehension measures. Narrative comprehension measures all load on this first factor as well; however, the factor loadings of the narrative measures tend to be slightly less than those of the expository measures. Two narrative measures appear to have abnormally high loadings on factor 1, the Circus Listening and the IGAP narrative prior knowledge. Their strong association with factor 1 is due, in large part, to the fact that they differ from the rest of the narrative measures being analyzed because they are not read by the students, but by the examiner. Hence, any variance in silent reading rate and accuracy that might have affected students' performance in a manner that would have led the scores to vary in a manner more similar to the other narrative score distributions is not present. Because of this, these two narrative measures do not load on factors one and two, as do the rest of the narrative measures, but on factor one, alone.

[Insert Table 4 about here.]

The second factor contrasts fluency with word reading and narrative comprehension measures. This factor has been reflected for ease of discussion and interpretation. All six of the oral reading measures of rate and accuracy have large negative loadings. The other five measures which load on factor two are the WRAT, a measure of students' ability to read a list of increasingly difficult words, and four measures that rely largely on narrative comprehension: the Woodcock, the IRAS comprehension measure, the Open Court error detection, and the IGAP narrative comprehension measure. Each of these measures have positive loadings on factor 2. This result is easily interpretable. The slower students read and the more errors they make, the less successful they are in reading and comprehending. The factor loading for the WRAT, .62966, is the largest.

This finding, as well as the overall structure of the factor pattern matrix, may be explained by the LaBerge and Samuels human information processing model of reading, which was first reported in 1974, and is presented and discussed in its revised form by Samuels and Kamil in the 1984 Handbook of Reading Research. LaBerge and Samuels' model is based upon the idea that the brain is a limited cognitive capacity processor that, when confronted by competing demands, has to alternate between the two tasks to complete them both. Through a series of experiments, LaBerge and Samuels showed that the amount of available attention we "have" to bring to a task is finite. How we allocate that attention during the reading process is a function of our ability as readers. Two of the more basic elements of the reading process are letter and word recognition, or decoding. If these tasks are difficult and require a great deal of attention, the reader has less attention, or cognitive capacity, available to devote to more complex skills of sentence and passage comprehension which involve skills such as inferencing, evaluating, and generalizing--all necessary skills for successful comprehension. As decoding becomes more and more automatic, readers possess the necessary "available attention" for comprehension activities.

Two other sources of variance in text characteristics that tap the available attention of the reader are topic and genre. A topic which is difficult or unfamiliar to readers can place extraordinary demands on readers' attention, as they struggle to make sense of the information they read. Even though the vocabulary may be relatively simple, the topic and the concepts may require a great deal of attention. Likewise, an unknown genre will also draw an abnormal amount of readers' available attention. In the present study, four sets of contrasting characteristics among the measures included for analysis make those measures different from one another. LaBerge and Samuels' 1974 work suggests that these four contrasting characteristics compete for reader's limited cognitive resources: (a) reading words in isolation as opposed to reading connected text to construct meaning; (b) reading and comprehending narrative text as opposed to reading and comprehending expository text; (c) reading text, either orally or silently, as opposed to having the text and or questions read aloud by the test administrator; and (d) reading about topics with which students have a high level of familiarity as opposed to reading about topics with which students have less familiarity.

These four characteristics aid in the theoretical interpretation of the factor pattern matrix presented in Table 4. Factor 2 will be discussed first, because its interpretation has implications for the interpretation of factor 1.

As was stated earlier, factor 2 contrasts fluency with largely narrative comprehension measures. The strongest negative loading on this factor is associated with the WRAT, a measure of isolated word reading skill. Isolated word reading relates most strongly to measures of rate and accuracy because it is not subject to vitiating influences of efforts to comprehend, activate prior knowledge, or recognize and use a given genre structure.

As to the explanation for why narrative, but not expository, measures load on factor 2, the narrative comprehension measures are strongly related to the fluency measures because the words and concepts in the passages are relatively easy, hence rate and accuracy do not begin to break down. Interestingly, the two longest and most complex texts, the IGAP narrative measure and the Open Court error detection measure have the weakest factor loadings. The strong negative relationship presented in factor 2 suggests that students' fluency contributed fairly directly to their comprehension of narrative text.

The second explanation for the structure of factor 2 is one which includes the effects of topic and genre. Students' performance on the narrative measures in this analysis related fairly strongly to measures of rate and accuracy because there did not tend to be a random breakdown in rate or accuracy brought about by large, non-systematic fluctuations from student to student in topic knowledge or knowledge of genre.

Factor 1 is made up of comprehension measures, prior knowledge measures, and the Circus Think-It-Through test. The structure of factor 1 suggests that, across all 16 measures included in this analysis, there is a single, underlying comprehension trait. Consistent with the research on schema theory conducted in the 1970s, the structure of factor 1 suggests that a reader's comprehension performance is made up, partially, of prior knowledge. The measures which have the strongest factor loadings are the IGAP expository comprehension measure and the STEP science test, a measure of science prior knowledge and, indirectly, since students read the items to themselves, reading ability. The remaining expository measures have factor loadings which are slightly stronger than the loadings of the narrative measures.

There are two possible explanations for the lower factor loadings of the narrative comprehension measures (lower, when compared to the loadings of the expository measures). The first explanation is related to measurement characteristics; the second, to the cognitive capacity theory of LaBerge and Samuels (1974). Four of the five distributions of the narrative measures are negatively skewed, resulting in reduced variability. Reduced variability of a measure limits the size of correlations involving the measure and limits the maximum reliability of a measure. These in turn limit the communality and

factor loading of a variable. Therefore, the maximum possible factor loadings of the narrative measures are more restricted than are most of the expository measures. Second, in their 1974 model of the reading process, LaBerge and Samuels suggest that the level of prior knowledge required and the knowledge of genre are elements which detract from readers' ability to comprehend, thus changing the task somewhat. The consistently higher loadings for the expository measures on factor 1, taken together with their failure to load on factor 2, suggest that there is a subtle, yet consistent, difference between the expository and the narrative measures, although they still comprise a single trait.

Summary of the Factor Analysis Results

The factor structure in the present analysis offers evidence to suggest that the IGAP reading assessment does have construct validity. While it shares an underlying comprehension construct with the other measures in the analysis, it also includes unique elements which reflect the new dynamic definition of reading not reflected in more traditional measures. While the IGAP reading measure differs from more traditional measures in format, passage length, and types of questions, the underlying trait measured by the IGAP comprehension measure is shared by more traditional measures of reading comprehension. The IGAP is unique in that it comprises a topic familiarity measure, a component not featured by other reading tests in the present analysis, but clearly part of the larger underlying trait of comprehension. The construct validity of the IGAP is further bolstered by the fact that the factor loading is in congruence with the research-based theory of available attention described in LaBerge and Samuels' 1974 work. The IGAP is the only test analyzed which measures literacy habits, a component unique in the overall factor structure.

Analysis 2--Regression Analysis

For ease of discussion in the following sections, District A, the small homogeneous community with a philosophy of whole-group instruction will be referred to as "Univille." District B, the medium-sized community with a fairly traditional instructional philosophy including self-contained classrooms with three to seven ability-clustered reading groups will be referred to as "Middleburg." District C, the large urban district with the most diverse student population, will be referred to as "Divercity."

Descriptive Results

Before discussing the results of the regression analyscs, descriptive information regarding the observational ariables will be presented. Table 5 presents the descriptive statistics for the instructional variables included in the final, reduced regression models for each site.

[Insert Table 5 about here.]

Because the present study looks at within-site differences, as opposed to between, this table does *not* present the descriptive statistics for all of the approximately fifty instructional variables which were evaluated for inclusion in each of the three regression equations. Each of these eleven variables was significant in explaining within-site differences in performance in one or more of the District models. The first two values, Index of Parental Resources and Index of Parental Support for Schoolwork, are the mean, schoolwide values for the two parent survey scales. Parents from Divercity report having the highest average level of resources. The parent population in Divercity is also the most varied population in terms of resources. Parents in Univille report that they help their children with reading or schoolwork slightly more often than do the parents in Middleburg and Divercity. The lower standard deviation associated with the responses of Univille suggest that, as a whole, their performance tends to be more similar, schoolwide than are the performances reported by parents in Middleburg and Divercity.

The remaining nine variables presented in Table 5 are the classroom observational variables. The means reported in Table 5 represent the average number of interactions (of the type specified in the column headed "Variable") a child in that district received during the third-grade school year. The means reported here are averages *across* all of the classrooms within each district. The data values which were averaged to arrive at these values are the mean number of interactions a child receives across the six observational rounds. This number includes interactions directed specifically to the child, interactions directed to the child's entire reading group, and interactions directed to the entire class during whole-group instruction. For example, the mean of 1.59 background knowledge interactions for Univille may be interpreted to mean that, on average, a child in a third-grade classroom in Univille received 1.59 background knowledge questions per round. What is of particular interest in this table is the within-site variance. That is, how different are the teachers in each of the districts? It is the within-site variance, as well as the intercorrelation of these variables, that will be essential in the explanation of differences in student performance within each of the three sites on the Total IGAP reading comprehension score, not simply the differences in means across sites. There is additional evidence from work conducted by Meyer, Wardrop, and Hastings (1990b) that these teacher differences are systematic and that these models may be generalized. Based on the analyses by Meyer et al. (1990b), these teachers' types and frequencies of instructional interactions tend to be very stable from one year to the next.

Despite Univille's philosophy of gradewide uniform instruction (whole-class instruction and a mandate for identical material use and coverage across classrooms), Table 5 indicates that, while they appear to vary their behaviors slightly less than teachers in Middleburg and Divercity, teachers' behaviors in Univille vary substantially. Teachers in Middleburg seem to be the most varied in their classroom teaching behaviors. Just how these differences in teacher behavior affect students' performances will be discussed in the following section.

Multiple Regression Analysis Procedures

In order to analyze the instructional variables which affected students' performances on the IGAP reading comprehension test, three hierarchical, blocked multiple regressions, one for each district, were completed using SPSS-PC+, the Statistical Package for the Social Sciences (Norusis/SPSS, Inc. 1988). The initial full regression model was presented earlier in the methods section in Figure 1.

The full model was evaluated for each site. Plots of residuals were examined to determine whether or not the assumption of uncorrelated errors had been violated in any of the three models. All of the residuals appeared to be randomly distributed. Therefore, it cannot be assumed that the assumption of uncorrelated errors has been violated in any of the three models. In order to control, as far as possible, the spurious effects attributable to small sample size, the final block of variables was trimmed, and the full model rerun. Initially, the final block in the regression analysis contained approximately 52 variables to be evaluated for entry using the stepwise procedures. In order to reduce this number, following the first regression run, all of those variables from the final block which had F values less than 1 following step four and following the final step of the regression were deleted. Variables from the first four steps which were not significant were, of course, also deleted. The trimmed model was rerun to produce the site-specific reduced models for Univille and Middleburg. In the case of Divercity, where the number of subjects is small, the trimming procedure described for block 5 was repeated prior to producing the final reduced model. The intent of this second trimming was to reduce the likelihood of spurious results due to a small sample size and too many independent measures.

Results: District A--Univille

The summary table for the Univille final regression model is presented in Table 6. The model explains 58% of the variance in the IGAP scores of students in Univille. The model has 50 degrees of freedom

at step 1. In step 1, 15% of the variance in the dependent measure is explained by the students' entering kindergarten abilities.

[Insert Table 6 about here.]

Interestingly, none of the parent survey variables entered initially in step 2 remained in the equation. Re-examination of Table 5 indicates that the parents' responses from Univille have less variability than do the responses of parents in the other two districts. This suggests that parents tend to do the same sorts of things for their children, regardless of students' abilities or families' socioeconomic status (as indicated by the parental resource scale).

In the first run of the regression analysis, the criterion variables were entered to explain the contribution of first- and second-grade teachers. Neither of these measures was significant in explaining variance in the dependent measure. The final regression model seems to reflect the philosophy of the district. Students were not grouped in or across classrooms by ability, and teachers across classrooms presented the same material to students. Therefore, there is no systematic teacher variability for first- and second-grade teachers associated with IGAP performance.

The remaining four variables in the final regression model are instructional variables. All instructional variables were entered in one step and allowed to compete for entry. The first variable to enter, which explains an additional 26% of the variance in the dependent measure, is mean phrase or sentence production. "Mean phrase or sentence production" indicates the average number of phrase or sentence production questions a given student in a teacher's classroom received across the six observational rounds. A phrase or sentence production interaction is a language development interaction in which the teacher asks the student to make up and verbalize a sentence which meets the teacher's specification. Mean phrase or sentence production has a negative beta weight of -.5084. This suggests a negative relationship between the frequency of these interactions and comprehension, as measured by the IGAP. These figures are consistent with Brophy and Good's (1985) data; they reported that students who are poorer readers tend to spend more time in decoding word list reading activities and less time in activities that allowed them to practice and develop comprehension abilities.

The next variable which entered the equation is "mean background knowledge questions." This variable indicates the average number of questions that tap students' background, or prior knowledge, that a given student in a teacher's classroom received across the six observational rounds. Mean background knowledge interactions explain an additional eight percent of the variance in the model. This measure has a beta weight of .3076, indicating that it is associated with increased performance on the IGAP reading measure. Data from the present study are limited. They do not indicate whether teachers' instructional content and questions taught students how and when to use their prior knowledge or simply surveyed students' prior knowledge. They do, however, indicate that students who were asked more frequently about their background knowledge tended to comprehend more of what they read, as measured by the IGAP reading comprehension test.

The next variable that entered the equation is "mean time spent decoding without written text." This measure indicates the average amount of time a given student spent in decoding instruction which was not followed by a text-reading activity across the six observational rounds. This variable has a beta weight of -.3077. Again, these data support the findings of Brophy and Good (1985), which indicated that students who spend increased amounts of time in decoding activities tend to be poorer comprehenders. This relationship is borne out again in the final step of the regression.

The final variable in the equation is "mean letter sound interactions." This measure indicates the average number of questions about letter sounds a given student in a teacher's classroom received across the six observational rounds. "Mean letter sound interactions" has a positive beta weight of .4238. While this result appears anomalous, it can readily be explained by examining the correlation between the

variable which entered the equation just before "mean letter sound interactions," "mean time decoding, not followed by written text." The correlation between these two variables is .782. When two variables are highly correlated, the first variable to enter the equation may act as a suppressor to the second variable, causing it to have a beta weight sign which is the opposite of what would be expected (Pedhauzer, 1982). Thus, as would be expected from the teaching effectiveness research of Brophy and Good (1985), increases in letter sound interactions are associated with decreases in performance on the dependent measure.

In all, just over 58% of the variance in IGAP reading comprehension scores in Univille is explained by this model. The IGAP reading comprehension test is sensitive to differences in teachers' behaviors in this district. Despite Univille's philosophy of uniform, whole-group instruction across classrooms, there are variations in teacher behavior that affect performance on the IGAP reading comprehension measure. The influence of teachers' behaviors in this model is congruent with reading research. Increases in the frequency of background knowledge questions are associated with increases in comprehension, as measured by the IGAP comprehension test. The following discussions will show that these relationships between the IGAP and instructional practices are similar to those found in the models for Middleburg and Divercity.

Results: District B--Middleburg

The summary table for the Middleburg final regression model is presented in Table 7. Middleburg is a small commuter community located in central Illinois. It is larger than Univille and students come from slightly more diverse backgrounds. Children are grouped for reading instruction beginning in kindergarten. The regression model explains 58% of the variance in the IGAP scores. The regression model has 82 degrees of freedom at step 1. In step 1, 27% of the variance in the dependent measure is explained by the students' entering kindergarten abilities.

In step 2, parental support of schoolwork and literacy explains an additional 4.5% of the variance. This variable has a negative beta weight of -.2147. There are three possible explanations for this relationship: one, parents of children who are poorer readers spend more time working with children in an effort to help them improve; two, children who are poorer readers ask for more parental support than do their more able peers; or three, parents of poorly performing students tend to enhance themselves when given the opportunity to "self-report."

In step 3, the influence of the second-grade teacher is significant in explaining an additional 3.4% of variance in the model. The criterion scale values for the second grade teachers have a beta weight of .1870, which suggests that being in a second-grade classroom with a higher criterion scale value is associated with an increased score on the dependent measure. This finding suggests that some degree of ability grouping might be taking place.

The remaining four variables in the final regression model are instructional variables. The first variable to enter, which explains an additional 9.6% of the variance in the dependent measure, is "mean time spent decoding, not followed by written text." This measure indicates the average amount of time a given student spent in decoding instruction which was not followed by a text reading activity. This variable has a beta weight of -.3263. This finding is similar to the relationship between increased decoding activities and comprehension, which was discussed earlier in the description of the Univille regression model.

[Insert Table 7 about here.]

The fifth variable to enter the equation and explain an additional five percent of the variance is "mean request for explanation feedback." This measure indicates the average number of times, across the six

observational rounds, a teacher asked a student to explain an answer the student had given. This variable has a negative beta weight of -.2314, and it is negatively correlated with ability as measured by the IGAP (-.375) and the Ability-zero measure (-.163). This suggests that teachers tend to have to ask poorer readers to explain their answers more frequently than they do better readers. This relationship should be investigated more thoroughly.

The sixth variable which entered the equation and explains an additional 5.5% of the variance is "mean word comprehension interactions." This measure indicates the average number of times, across the six observational rounds, a teacher asks a student what a word means, either in isolation or in context. This variable has a beta weight of .2373. This indicates that students who were asked more word comprehension questions tended to be better comprehenders. This finding is consistent with research which underscores the importance of building an understanding of reading vocabulary as part of establishing prior knowledge during the prereading portion of a lesson (Anderson, Hiebert, Scott, & Wilkinson, 1985).

The seventh and final variable which entered the equation is "mean phrase or sentence production." It explains an additional 2.7% of the variance in the dependent measure. "Mean phrase or sentence production" indicates the average number of phrase or sentence production questions a student in a teacher's classroom received across the six observational rounds. As was stated earlier in the description of the Univille model, a phrase or sentence production interaction is a language development interaction. It is not an interaction which requires students to construct meaning or make inferences about text. Mean phrase or sentence production has a negative beta weight of -.1904. The alignment of these findings with the 1985 research of Brophy and Good was described in the preceding section on Univille and will not be repeated here.

In all, 58% of the variance in IGAP comprehension scores in Middleburg is explained by this model. The IGAP reading comprehension test is sensitive to differences in teachers' behaviors in this district. There are variations in teacher behavior that affect performance on the IGAP. What's more, as was the case with the Univille model, the influences of differences in teachers' behaviors are all in the direction that would be expected, based on recent research in reading instruction. Recall, however, that it was noted that questions remain regarding the relationship between teachers' requests for clarifications and the dependent measure. This relationship requires further investigation. The relationships between the IGAP and instructional practices reported thus far for Univille and Middleburg are similar to those found in the model for Divercity.

Results: District C--Divercity

The summary table for the Divercity final regression model is presented in Table 8. It is important to keep in mind that the sample size is somewhat small. There are only 40 degrees of freedom at step 1 of the regression.

Divercity is a fairly urban suburb located in northeast Illinois. It is larger than Univille and Middleburg and is far more heterogeneous. Compared to students in districts A and B, Divercity children come from slightly more diverse backgrounds, socioeconomically and ethnically.

[Insert Table 8 about here.]

The regression model explains 79% of the variance in the IGAP scores. In step 1, 42% of the variance in the dependent measure is explained by the students' entering kindergarten abilities. This is substantially more variance than the Ability-zero measure explained in the previous two models. However, students began school in Divercity with a larger range of entering abilities and a much larger standard deviation in Ability-zero scores, so this result is not altogether surprising.

In step 2, the parental resource index explains an additional 10.8% of the variance. This variable has a beta weight of .3806. This suggests that increased parental resources are associated with increased scores on the IGAP. A logical question at this point is, "Why only Divercity? Why is socioeconomics not significant in the other two districts?" There is more variance in the socioeconomic status of parents in Divercity. The variance in socioeconomic status in Divercity is more than twice as great as the variance in socioeconomic status of parents in Univille and Middleburg. A 1984 review chapter by Wigfield and Asher on the social and motivational influences on reading presents several studies, all of which conclude that socioeconomic level of the family has a strong influence on students' success in school.

The remaining four variables in the final regression model are instructional variables. All instructional variables were entered in one step and allowed to compete for entry. The third variable to enter the equation, which explains an additional 10.7% of the variance in the dependent measure, is "mean plot, text-implicit interactions." "Mean plot, text-implicit interactions" indicates the average number of implicit comprehension questions about the plot of a story a student received across the six observational rounds. This is an interaction which requires students to construct meaning and make inferences about text. As the beta weight of .3972 indicates, this is an activity which is associated with increased performance on the dependent measure. This finding is consistent with research which underscores the impact that good comprehension questions can have on the development of successful comprehenders (Beck et al., 1982).

The fourth variable which entered the equation and explains an additional 7.1% of the variance is "mean word comprehension interactions." This measure indicates the average number of times, across the six observational rounds, a teacher asks a student what a word means, either in isolation or in context. Mean word comprehension has a beta weight of .3044. This relationship indicates that students who were asked more word comprehension questions tended to be better comprehenders. The effect of this variable is also consistent with the effect observed for the same variable in the regression model for Middleburg. This finding is consistent with research which underscores the importance of building an understanding of reading vocabulary as part of establishing prior knowledge during the prereading portion of a lesson (Anderson et al., 1985).

The fifth variable to enter, which explains 4.8% of the variance in the dependent measure, is mean setting, text-explicit interactions. "Mean setting, text-explicit interactions" indicates the average number of explicit comprehension questions about the setting of a story a student received across the six observational rounds. This is an interaction which requires students to recall information explicitly stated in the text. Based on the 1982 work of Beck et al., it is reasonable to expect this variable to have a positive relationship with the dependent measure. However, the beta weight -.2444 seems to suggest the contrary; this activity appears to be negatively related to performance on the dependent measure. The true relationship may be being suppressed. This suppressing effect is due to the correlation between "mean word comprehension interactions" entered in the previous step and "mean setting, text-explicit (.403). Thus, as previous research suggested, increases in the number of comprehension questions are related to increased performance.

Finally, the last variable to enter the equation and explain 3.6% of the variance is "mean word production interactions." "Mean word production interactions" indicates the average number of word production questions a student in a classroom received across the six observational rounds. A word production interaction is very similar to the phrase or sentence production interactions described earlier. It is a language development interaction in which the teacher asks the student to generate a word which meets the teacher's specification. It is an interaction that can take place in the absence of text. Mean word production has a negative beta weight of -.1986. As could be expected by the similarity of the two measures, this variable performs very much like "mean phrase or sentence production" did in the Univille and Middleburg models. Once more, the results are consistent with Brophy and Good (1985), who reported that students who are poorer readers tend to spend more time in decoding word list

reading activities and less time in activities which allowed them to practice and develop comprehension abilities.

In all, 79% of the variance in IGAP reading comprehension scores in Divercity is explained by this model. The seemingly large difference between the amount of variance explained in the models for Univille and Middleburg, 58%, and Divercity, 79%, may be due, in part, to the low number of subjects in the Divercity model. However, it appears to be primarily due to the large amount of variance explained by the Ability-zero measure in Divercity, 42%, as compared to 15% and 27% in Univille and Middleburg. As was the case for Univille and Middleburg, the IGAP reading comprehension test is sensitive to differences in teachers' behaviors in Divercity. There are variations in teacher behavior that affect performance on the IGAP reading comprehension measure. Once again, as was the case with the Univille and Middleburg models, the influences of differences in teachers' behaviors are all in the direction that would be expected, based on recent research in reading instruction. Students who are involved in increased numbers of text and word comprehension measures tended to perform better on the IGAP.

Overall, it seems reasonable to conclude that the IGAP comprehension measure is sensitive to some of the instructional behaviors that research has shown to contribute to increased comprehension. However, as will be discussed in the next section, it is wholly inappropriate to conclude that instruction in decoding and word recognition should not take place.

DISCUSSION

Overview and Summary of Analyses

The primary focus of the present study was the evaluation of the validity of a new reading assessment instrument. Two methods were used to evaluate the validity of the reading portion of the Illinois Goal Assessment Program (IGAP). First, factor analysis was used to determine the factor structure of the IGAP scores; what elements the IGAP shared with nine other, more traditional measures; and what elements, if any, were unique to the IGAP. Second, multiple regression analyses were used to find out if the IGAP was sensitive to instructional differences within three instructional settings with differing educational philosophies, geographic locations, and socioeconomic complexions. Following a brief summary of the results of the analyses, the contributions of each of these analyses and the implications for future research are presented.

The results of the factor analysis attest to the construct validity of the IGAP reading assessment. The factor analysis indicates that while some portions of the IGAP are similar to other measures, other portions are unique to the present factor structure. While the IGAP reading measure does differ from more traditional measures in format, passage length, and types of questions, the underlying trait measured by the IGAP comprehension measure is the same as that measured by more traditional measures of reading comprehension. The IGAP is unique, in that it measures an element of the comprehension factor not measured by other reading tests in the present analysis: prior knowledge. In the present analysis, domain-specific, content-area tests proxied for measures of topic familiarity for the other expository comprehension tests. There were no comparable topic familiarity measures for the traditional narrative measures. It is important to keep in mind that prior knowledge measures loaded on the comprehension factor. They did not make up a separate factor. Thus, the results of the present analysis suggest that prior knowledge is an element of comprehension. The IGAP is the only test analyzed which has a literacy habits and attitudes survey, a component which is unique in the overall factor structure. Possibly due to low reliabilities, the IGAP reading strategies centrality and problem-solving measures failed to load on a factor in the present study. It is difficult to draw conclusions from the present analysis regarding these reading strategies items. With the exception of

the failure of the reading strategies measures to load on a factor, the results reported here are in agreement of the multi-form, cross-grade factor analysis of the IGAP reported by Krug in 1987.

Like the other group-administered measures in the analysis, the IGAP has no direct measure of fluency. Only the two individual reading inventories, the IRAS and the Science passages, measure fluency directly. The IGAP narrative comprehension and other narrative measures analyzed here do share a strong negative relationship with the direct measures of fluency contained in the IRAS and the Science passages measures.

In analysis 2, there are striking consistencies in the effect of instructional variables across sites. In all cases, instructional interactions which focus students on decoding and language production, as opposed to comprehension, are associated with lower performance on the dependent measure. Those variables include mean phrase or sentence production, mean time spent decoding not followed by written text, mean letter/sound interactions, and mean word production interactions. Also in this category is the feedback interaction, "asks for explanation or clarification of a response." The latter is not typically considered an instructional variable, but it is clear that students who must be asked more frequently to explain or clarify answers they have given, perhaps because the answer was incorrect or did not make sense, are likely to be poorer performers.

There are also consistencies across the three models with respect to teacher behaviors which are associated with increased performance on the dependent measure. Instructional variables which require students to make sense of text or to construct meaning from text are consistently associated with increased performance on the dependent measure. These variables include mean background knowledge interactions; mean word comprehension interactions; mean plot, text-implicit interaction; and mean setting, text-explicit interactions. Overall, it seems reasonable to conclude that the IGAP comprehension measure is sensitive to some of the instructional behaviors which research has shown to contribute to increased comprehension.

Contributions to the Field of Reading

Factor Analysis

The implications which arise from the factor analysis relate most directly to the ongoing validation of the test itself and to the construction of future reading assessment measures. First, while the IGAP comprehension, or constructing meaning section, differs in appearance from more traditional measures of reading comprehension, it measures the same underlying trait. Should this finding be interpreted to mean that it does not matter which measure is used? Not if one keeps the issue of validity in mind when responding to the question. One of the issues of validity is congruence with instructional materials and methods. The IGAP constructing meaning section must measure, in large part, the same trait that is measured by older, more traditional measures; however in doing so, it makes use of texts and questions which are more similar to the instructional materials used by today's teachers. Additionally, the multiple regression analyses support the validity of the IGAP approach to comprehension assessment as it relates to teachers' classroom behaviors.

Second, the IGAP is unique in its inclusion of prior-knowledge measures. This cannot be assumed to be simply a function of test construction. If only the IGAP prior-knowledge measures loaded on the comprehension factor, this might be the case. However, recall that the science tests, which were proxy measures of prior knowledge (because they were not part of a "reading test," but measured topics covered by other reading test texts) loaded on the comprehension factor. The reliabilities of the two prior-knowledge sections of the IGAP are marginal at the individual student level and should be addressed in future research.

Third, it is difficult, if not impossible, to arrive at large-scale fluency measures. While the factor analysis clearly indicates that fluency is an important element in reading, none of the group-administered measures include measures of fluency. Neither are timed group tests adequate substitutes for fluency measures because they force readers to change their reading behaviors to meet artificial time constraints. However, the results of the present study suggest that measures of narrative comprehension do at least provide a more direct indication of fluency than do expository measures. Recall that there was a tendency for the narrative measures which loaded on the fluency factor to be slightly negatively skewed. Thus, it is somewhat unclear as to whether the comprehension loadings on factor two reflect a genre effect or a "level of difficulty" effect. This relationship requires further investigation which will be elaborated on in the section on further research.

Fourth, the IGAP is unique in its inclusion of literacy survey measures. The analysis indicates that although the literacy survey is very reliable, it does not load with the other measures. The accompanying low communality for literacy habits and attitudes suggests that a separate factor is possibly being measured by the IGAP. Given the role of interest and attitude in learning, test constructors and teachers alike should include measures of affect in their evaluation and diagnosis of students' strengths and weaknesses. Further, from a teacher's perspective, information about students' habits, interests and attitudes can be a useful aid in selecting maximally effective instructional materials.

Fifth, the factor analysis of the IGAP conducted by Krug (1987) across 21 forms at four grade levels indicated that the reading strategies measures loaded on separate factors, one for centrality and one for problem solving. Krug's factor structure was more compelling at grades six, eight, and ten than it was at third grade. At third grade, the two different types of strategy item sets did not separate as cleanly onto separate factors. Krug's 1987 results were not replicated in the present analysis. There are numerous possible explanations for this anomaly. The lack of reliability of the measure at third grade is clearly one of them, as is the unfamiliarity of the item format. A possible interpretation (requiring further investigation) is that Krug's 1987 findings, when viewed in the context of instructional research on metacognition, indirectly support the idea that reading strategies can be taught and measured and that they do contribute to increases in comprehension. Given this, and the increased attention given to strategy instruction by basal series authors, a valid test of reading in the 1990s should include measures of students' awareness of and facility with various reading strategies. The present study indicates that the most reliable way to accomplish this is an issue for future research.

Multiple Regression Analyses

The second set of analyses was designed to investigate the instructional sensitivity of the IGAP reading comprehension measure. The IGAP was designed to do what older, more traditional measures of reading no longer did: reflect current reading instruction. To accomplish this, the authors incorporated many of the elements that research indicates are central to good reading instruction: the activation of prior knowledge; the use of ecologically valid texts with recognizable structures; questions which focus on information that is central to the purpose and content of the text; questions that require the reader to make inferences, restructure information, and reason beyond the text; and questions that require readers to give evidence of their knowledge and control of reading strategies. In many ways, the test models the way good readers *learn* from text. However, to be a truly valid measure of reading, a test must do more than simply look like a good test. It should be sensitive to instruction. The resulting scores should indicate to users whether or not teachers are teaching students to work actively with their own knowledge and the information on the page to construct meaning. Thus, teachers and students should be able to improve performance by learning and practicing those things that research tells us good readers do when they read.

The multiple regression analyses completed in the present study evaluated the sensitivity of the IGAP comprehension measure to teacher behaviors in three different sites. Each site had a different

educational philosophy, used different instructional materials and had a different ethnic and socioeconomic make-up. It is important to bear in mind that the present study is limited to third grade; it does not include students who are just beginning to read. In each of the three sites, regardless of instructional materials or educational philosophy with respect to instructional method or grouping, there are activities that third-grade teachers use in their classrooms that explain systematic variance in students' performances on the IGAP comprehension measures. Put simply, the teachers in this study did make a difference in student performance, as measured by the IGAP. What is equally important from a validity standpoint is which teacher behaviors are associated with increased IGAP scores and which are not.

The results here are consistent across the three models. Teacher interactions which focus on constructing meaning from text, for example, word comprehension and background knowledge interactions, are associated with higher IGAP performance. Teacher interactions which focus students' attention on sub-skill mastery, such as mean time spent decoding, not followed by written text and word, phrase, and sentence production tasks are associated with lower student performance. These findings have particularly important implications for educators and administrators. These implications relate closely to some important validity issues.

One of the key elements of validity is the match between what is taught and what a test measures. These findings suggest that the IGAP is a valid measure for reading programs that emphasize comprehension. It may not be a particularly valid evaluation measure for decoding emphasis programs. The more time teachers spend in activities which teach students to construct meaning from text, and the more they focus their efforts on developing strategic, active readers, the more their efforts will produce increased results as measured by the IGAP.

Teachers have a finite amount of time to devote to reading instruction. Teachers make choices daily in how they allocate their time. The choices they make can influence the validity of their students' test results. Some of those choices are driven by the curriculum they use; others reflect their own preferences within the curriculum. When teachers choose to spend time in decoding instruction, they make an implicit choice not to spend time in comprehension instruction. As the present study indicates, these choices influence results. If a teacher chooses to focus almost exclusively on decoding activities at the third-grade level, comprehension measures such as the IGAP are not likely to be sensitive to that teacher's instructional efforts. This point is particularly important with respect to low achievers, who tend to spend a great deal of their time in subskill instruction. Teachers and administrators need to be aware of these differences in order to guide their selection of evaluation instruments as well as their interpretation of test results.

A final word of caution regarding the interpretation of these results: The results from the present study should not be interpreted to mean that instruction in decoding is negatively related to learning to read. The present study was *not* designed as an instructional study. There is a vast body of research results that point to the important and positive role phonics can play in students' later efforts to construct meaning. That relationship was not evaluated in the present study. The research is, however, summarized in *Learning to Read: Thinking and Learning about Print--A Summary* (Stahl, Osborn, & Lehr, 1990). What the results of the present study *do* highlight is that phonics instruction is not a substitute for comprehension instruction, particularly by grade 3, when many, if not most, children have achieved a degree of automaticity in decoding. While fluency and decoding automaticity are skills which good readers possess, these skills alone do not produce strategic, fluent readers. The present study underscores the fact that subskills instruction alone is not enough. Subskill instruction tends to perpetuate students "roles" as good or poor readers. Quite often, poor readers require instruction in a particular area and thus, they miss out on comprehension instruction which they desperately need. Meanwhile their more capable peers, who do not require any subskill reinforcement, spend even more time in constructing meaning or comprehension instruction. This study suggests that in order to improve

comprehension as measured by the IGAP, students need to be given opportunities to practice the orchestration of those skills that are essential to the process of constructing meaning.

Conclusions

The results of the present study are encouraging. They suggest that, while the underlying construct of comprehension measured by the IGAP is somewhat similarly measured by other measures, the IGAP reading assessment measure is unique in several ways. The IGAP reading assessment measures comprehension, the same underlying trait measured by other more traditional measures of reading, but it also measures prior knowledge, a component of the comprehension trait not directly measured by other measures in the analysis. The IGAP reading assessment is the only measure in the analysis which measures the literacy habits and attitudes of students. The literacy survey represents an underlying trait, or factor, which is unique in the analysis to the IGAP. The IGAP reading assessment is the only measure in the analysis which attempts to measure students' awareness of reading strategies. Reading strategies may represent another underlying trait, or factor, which is unique to the IGAP. Further work is needed in this area, as results of the present study concerning reading strategies are inconclusive.

The present study indicates that the IGAP comprehension measure is sensitive to teachers' instructional behaviors. Increased numbers of comprehension interactions are associated with increased performance on the IGAP. These results are consistent across three instructional sites which have different instructional philosophies and use different instructional materials.

Taken together, these results provide increasing evidence of the validity of the IGAP reading assessment. The structure of the test, as identified in the factor analysis, is congruent with reading research, and as the second analysis indicates, the structure is also sensitive to good instructional practices. It is, however, important to keep in mind the new concept of validity described earlier in the literature review. Under this new "unified" view of validity described by Cronbach, all efforts toward the validation of an instrument must have as their foundation, construct validation, for there can really be no validity at all without it. However, because of the power of tests, and the potential influence they can have on people's lives, construct validity alone does not answer the question "How valid is this test?" The social and political influences and implications must come into play as well. Test validation should be an overall evaluation of the "adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1988b, p.42).

References

Allington, R. L. (1984). Content coverage and contextual reading in reading groups. *Journal of Reading Behavior, 16*, 20-31.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1982). *Psychological testing* (fifth edition). New York: Macmillan.

Anderson, R. C. (1977). The notion of schemata and the educational enterprise. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 415-431). Hillsdale, NJ: Erlbaum.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the commission on reading.* Washington, DC: The National Institute of Education.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learing and Verbal Behavior, 17*, 1-12.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Erlbaum.

Armbruster, B. A. (1979). *An investigation of the effectiveness of "mapping" text as a studying strategy for middle school students* (Doctoral dissertation, University of Illinois, 1979). *Dissertation Abstracts Internationl, 40*, 5369A.

Baker, L. (1979). *Comprehension monitoring: Identifying and coping with text confusions* (Tech. Rep. No. 145). Urbana: University of Illinois, Center for the Study of Reading.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson, M. Kamil, R. Barr, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 353-394). New York: Longman.

Bartlett, B. J. (1978). *Top-level structure as an organizational strategy for recall of classroom text* (Doctoral dissertation, Arizona State University, 1977). *Dissertation Abstracts Internationl, 41*, 5689A.

Beck, I. L., McKeown, M. G., McCaslin, E. S., & Burke, A. M. (1979). *Instructional dimensions that may affect reading comprehension: Examples from two commercial reading programs.* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Beck, I. L., Omanson, R. C., & McKeown, M. G. (1982). An instructional redesign of reading lessons: Effects on comprehension. *Reading Research Quarterly, 17*, 462-481.

Bloom, B. S. (1966). The role of the educational sciences in curriculum development. *International Journal of Educational Sciences, 1*, 5-16.

Bowman, M. A. (1981). *The effect of story structure questioning upon the comprehension and metacognitive awareness of sixth grade students*. Dissertation Abstracts International, 42, 626A. (University Microfilms No. 81-16, 456).

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding. Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Brophy, J. E., & Good, T. L. (1985). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328-375). New York: Macmillan.

Brown, A. L., Armbruster, B. B., & Baker, L. (1985). The role of metacognition in reading and studying. In J. Orasanu (Ed.), *Reading comprehension: From research to practice* (pp. 49-75). Hillsale, NJ: Erlbaum.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning remembering and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Handbook of child psychology* (Vol. 3, pp. 77-166). New York: Wiley.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher, 10*, 14-21.

Brown, A. L., & Palincsar, A. S. (1982). Inducing strategic learning from texts by means of informed self-control training. *Topics in Learning and Learning Disabilities, 2*, 1-17.

Brown, A. L., Palincsar, A. S., & Armbruster, B. B. (1984). Instructing comprehension-fostering activities in interactive learning situations. In H. Mandl, N. Stein, & T. Trabasso (Eds.), *Learning from texts* (pp. 255-286). Hillsdale, NJ: Erlbaum.

Calfee, R. C., & Calfee, K. H. (1982). *Interactive reading assessment system (IRAS)*. Austin, TX: Southwest Educational Development Laboratory.

Clymer, T., Wolfe, E. V., Blanton, W. E., Jacobsen, M. D., Johnson, K., Shuy, R. W., & Torrance, E. P. (1976). *Reading 720*. Lexington, MA: Ginn.

Collins, A., Brown, J. S., & Larkin, K. M. (1980). Inference in text understanding. In R. Spiro, B. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 385-407). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly, 7*, 628-678.

Durkin, D. (1978-1979). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly, 14*, 481-533, *15*, 624-249.

Durkin, D. (1984). Is there a match between what elementary teachers do and what basal reader manuals recommend? *Reading Teacher, 37*, 734-744.

Durr, W. K., & Hillerich, R. L. (1981). *Getting ready to read*. Boston, MA: Houghton Mifflin.

Durr, W. K., Pikulski, J. J., Bean, R. M., Cooper, J. D., Glaser, N. A., Greenlaw, M. J., & Schoephoerster, H. (1983). *Houghton Mifflin Reading K - 8.* Boston: Houghton Mifflin.

Early, M., Cooper, E. K., & Santeusanio, N. (1983). *Harcourt Brace Jovanovich Bookmark Reading Program, Eagle Edition.* New York: Harcourt Brace Jovanovich.

Educational Testing Service. (1976a). *Circus--Listen to the story--Levels A & D, Form X.* Menlo Park, CA: Addison-Wesley.

Educational Testing Service. (1976b). *Circus--Think it through.* Menlo Park, CA: Addison-Wesley.

Educational Testing Service. (1979). *Sequential test of educational progress (STEP).* Menlo Park, CA: Addison-Wesley.

Engelmann, S. E., & Hanner, S. (1972). *Distar Reading: Reading to learn.* Chicago: Science Research Associates.

Fielding, L. G. (1988). *The role of discussion questions in children's story comprehension* (Doctoral dissertation, University of Illinois, 1988). *Dissertation Abstracts International, 49,* 2599A.

Fielding, L. G., Anderson, R. C., & Pearson, P. D. (1988). *The role of discussion questions in children's story comprehension.* Urbana-Champaign: University of Illinois, Reading Research and Education Center.

Fitzgerald, J., & Spiegel, D. L. (1983). Enhancing children's reading comprehension through instruction in narrative structure. *Journal of Reading Behavior, 15,* 1-17.

Flavell, J. H., Speer, J. R., Green, F. L., & August, D. L. (1981). The development of comprehension monitoring and knowledge about communication. *Monographs of the Society for Research in Child Development, 46* (5, Serial No. 192).

Gall, M. D., Ward, B. A., Berliner, D. C., Cahen, L. S., Crown K. A., Elashoff, J. D., Stanton, G. C., & Winne, P. H. (1975). *The effects of teachers' use of questioning techniques on student achievement and attitude.* San Francisco: Far West Laboratory for Educational Research and Development.

Hare, V. C., & Pulliam C. P. (1980). Teacher questioning: A verification and extension. *Journal of Reading Behavior, 12,* 69-72.

Jastak, S., & Wilkinson, G. S. (1984). *Wide range achievement test.* Wilmington, DE: Jastak Associates.

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Beverly Hills, CA: SAGE.

Krug, S. E. (1987). *An analysis of the underlying factor structure of the IGAP reading assessment.* (Unpublished report to the Illinois State Board of Education). Champaign, IL: MetriTech.

Krug, S. E. (1989). *Guide to the 1988 Illinois state assessment.* Springfield: Illinois State Board of Education.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

Lipson, M. Y. (1982). Learning new information from text. The role of prior knowledge and reading ability. *Journal of Reading Behavior, 14*, 243-262.

Mandler, R. M., & Johnson, N. S. (1977). Remembrance of things parsed. Story structure and recall. *Cognitive Psychology, 9*, 111-151.

Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child Development, 48*, 986-992.

Mason, J., & Meyer, L. A. (1983). *The language and problem solving battery*. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Messick, S. E. (1975). The standard problem: Meaning and values in measurement. *American Psychologist, 30*, 955-966.

Messick, S. E. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. E. (1988a). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. E. (1988b). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Meyer, L. A., Hastings, C. N., & Linn, R. L. (1987a). Diagnostic science reading passages (DSRP). Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Meyer, L. A., Hastings, C. N., & Linn, R. L. (1987b). *National Science Foundation Living Things Test*. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Meyer, L. A., Wardrop, J. L., & Hastings, C. N. (1989). *Interim report of trends from a longitudinal study of the development of reading comprehension ability*. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Meyer, L. A., Wardrop, J. L., & Hastings, C. N (1990a). *The development of reading ability in kindergarden*. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Meyer, L. A., Wardrop, J. L., & Hastings, C. N (1990b). *Trends from a longitudinal study of the development of science knowledge in kindergarten through second grade*. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Michigan Reading Association (1987). *Reading test blueprint* (2nd ed.). Lansing: Michigan Department of Education.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Norusis, M.J., /SPSS, Inc. (1988). *Stastical package for the social sciences / Personal computer plus* (SPSS/PC+), Version 3.0. Chicago: M.J. Norusis / SPSS, Inc.

Open Court (1987). *Open Court error detection test*. LaSalle, IL: Open Court.

Paris, S. G., & Lindaur, B. K. (1976). The role of inference in children's comprehension and memory for sentences. *Cognitive Psychology, 8*, 217-227.

Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Reading Behavior, 9*, 201-210.

Pearson, P. D., & Johnson, D. (1978). *Teaching reading comprehension*. New York: Holt, Rinehart, & Winston.

Pearson, P. D., & Valencia, S. W. (1987). *Illinois goal assessment program reading assessment.* Springfield: Illinois State Board of Education.

Pedhauzer, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: CBS College.

Phillips, L. M. (1989). *Developing and validating assessments of inference ability in reading comprehension* (Tech. Rep. No. 452). Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology, 69*, 309-315.

Raphael, T. E. (1984). Teaching learners about sources of information for answering comprehension questions. *Journal of Reading, 27*, 303-311.

Raphael, T., & Pearson, P. D. (1985). Increasing students' awareness of sources of information for answering questions. *American Educational Research Journal, 22*, 217-236.

Raphael, T., & Wonnacott, C. A. (1985). Heightening fourth-grade students' sensitivity to sources of information for answering comprehension questions. *Reading Research Journal, 20,* 282-296.

Readence, J. E., & Moore, D. W. (1983). *A meta-analysis of the effect of adjunct aids on learning from text*. Paper presented at the annual meeting of the American Educational Research Association.

Reiss, E. & Friedman, R. (1976). *The Alpha - K - Time Reading Program*. Plainview, NY: New Dimensions and Education.

Rumelhart, D. (1975). Notes on a schema for stories. In D. G. Bobrow, & A. M. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 211-236). New York: Academic Press.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18*, 8-14.

Spiegel, D. L., & Fitzgerald, J. (1986). Improving reading comprehension through instruction about story parts. *The Reading Teacher, 39*, 676-682.

Stahl, S., Osborn, J., & Lehr, F. (1990). *Beginning to read: Thinking and learning about print: A summary*. Champaign, IL: Center for the Study of Reading, University of Illinois.

Stein, N. L., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing: Vol. II. Advances in discourse processes* (pp. 53-120). Norwood, NJ: Ablex.

Thompson, B., Gipe, J. P., & Pitts, M. M. (1985). Validity of the Pearson-Johnson taxonomy of comprehension questions. *Reading Psychology, 6*, 43-49.

Valencia, S. W., & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher, 40*, 726-732.

Wigfield, A., & Asher, S. R. (1984). Social and motivational influences on reading. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 423-452). New York: Longman.

Winne, P. H. (1979). Experiments relating teacher's use of higher cognitive questions to students achievement. *Review of Educational Research, 49*, 13-50.

Wixson, K. K. (1981). The effects of postreading questions on children's comprehension and learning. In M. L. Kamil (Ed.), *30th Yearbook of the National Reading Conference*. Washington, DC: National Reading Conference.

Wixson, K. K. (1983). Postreading question-answer interactions and children's learning from text. *Journal of Educational Psychology, 30*, 413-423.

Wixson, K. K., & Peters, C. W. (1984). Reading redefined: A Michigan Reading Association Position Paper. *Michigan Reading Journal, 17*, 4-7.

Wixson, K. K., & Peters, C. W. (1987). Comprehension assessment: Implementing an interactive view of reading. *Educational Psychologist, 22*, 333-356.

Woodcock, R. W. (1973). *Woodcock reading mastery tests*. Circle Pines, MN: American Guidance Services.

Table 1

Descriptive Statistics of Test Scores Analyzed in the Factor Analysis

Test Name	*N*	*M*	*SD*	Observed Minimum	Observed Maximum	Possible Minimum	Possible Maximum
WRAT	305	62.23	8.22	32.00	87.00	0	100
WOODCOCK	305	38.99	11.41	1.00	73.00	0	85
IRAS							
Comprehension	304	57.44	14.21	2.00	77.00	0	80
CIRCUS Listening	311	32.61	4.03	15.00	40.00	0	40
CIRCUS Thinking	311	24.33	5.47	9.00	33.00	0	33
Science Passages	307	9.65	2.43	3.00	15.00	0	27
Open Court							
Error Detection	307	14.90	5.04	2.00	24.00	0	19
NSF Science	307	102.43	9.14	72.00	123.00	0	190
STEP Science	307	42.09	5.32	21.00	50.00	0	50
IGAP Comprehension Narrative	309	11.62	1.75	6.58	14.42	0	15
IGAP Comprehension Experience	309	11.40	1.71	6.71	14.58	0	15
IGAP Prior Knowledge Narrative	308	29.45	4.31	13.00	37.00	0	40
IGAP Prior Knowledge Experience	309	26.48	3.46	11.00	36.00	0	40
IGAP Centrality	309	9.78	2.14	3.00	16.00	0	20
IGAP Problem Solving	309	8.59	2.24	3.00	15.00	0	20
IGAP Survey	309	89.46	9.54	48.00	112.00	36	120

Note. Statistics are reported across the three sites for the entire population.

Table 2

Cronbach's Alpha Reliabilities for all Tests Included in Factor Analysis

Test Name	Reliability
WRAT	.622
Woodcock	.916
IRAS Comprehension	.803
Open Court Error Detection	.855
CIRCUS Listening Comprehension	.694
CIRCUS Think-It-Through	.843
Science Passages	.779
NSF Living Things	.609
STEP Science	.850
IGAP Prior Knowledge - Narrative	.618
IGAP Prior Knowledge - Expository	.461
IGAP Narrative Comprehension	.843
IGAP Expository Comprehension	.839
IGAP Reading Strategies - Centrality	.214
IGAP Reading Strategies - Problem Solving	.215
IGAP Literacy Survey	.812

Table 3

Communalities of all Variables Included in the Factor Analysis

Test Name	Communality
WRAT	.5623
Woodcock	.7684
IRAS Comprehension	.5022
IRAS Error Rate	.7625
IRAS Self Corrections	.3970
IRAS Reading Rate	.7907
Science Passages	.3960
Science Passage Error Rate	.5933
Science Passage Self Corrections	.2606
Science Passage Reading Rate	.7828
Open Court Error Detection	.5248
CIRCUS Listening Comprehension	.4591
CIRCUS Think-It-Through	.4589
NSF Living Things	.5057
STEP Science	.7225
IGAP Prior Knowledge - Narrative	.3840
IGAP Prior Knowledge - Expository	.3651
IGAP Narrative Comprehension	.6126
IGAP Expository Comprehension	.6214
IGAP Reading Strategies - Centrality	.0250
IGAP Reading Strategies - Problem Solving	.0195
IGAP Literacy Survey	.0847

Table 4

Factor Pattern Matrix and Factor Correlations for Third Grade Tests

Test Name	Factor 1	Factor 2
IRAS Error Rate		-.89921
IRAS Self-Corrections		-.67538
IRAS Reading Scale		-.84348
Science Passages Error Rate		-.75110
Science Passages Self Corrections		-.50943
Science Passages Reading Rate		-.85640
WRAT		.62966
Woodstock	.48875	.56799
IRAS Comprehension	.32733	.51829
Open Court Error Detection	.51283	.35506
IGAP Narrative Comprehension	.58769	.34237
CIRCUS Listening Comprehension	.69062	
CIRCUS Think-It-Through	.67700	
Science Passages	.59871	
NSF Living Things	.69863	
STEP Science	.70855	
IGAP Prior Knowledge - Narrative	.57538	
IGAP Prior Knowledge - Expository	.61043	
IGAP Expository Comprehension	.71808	
IGAP Reading Strategies - Centrality		
IGAP Reading Strategies - Problem Solving		
IGAP Literacy Survey		

Factor Correlation Matrix		
	Factor 1	Factor 2
Factor 1	1.00000	
Factor 2	-.37271	1.00000

Table 5

Descriptive Statistics for the Instructional Variables Which are Significant in any of the Three School Specific Regression Models

Variable C	District A *M*	District A *SD*	District B *M*	District B *SD*	District C *M*	District C *SD*
Parent Survey Scales						
Index of Parental Resources	108.75	21.11	114.04	23.28	* 119.75	35.3
Index of Parental Support of Schoolwork	1.62	.63	* 1.48	.74	.95	.83
Classroom Instruction Variables + +						
Mean Background Knowledge Interactions	* 1.59	.75	1.01	1.22	1.22	.93
Mean Asks Explanation Feedback	.09	.13	* .15	.31	.13	.18
Mean Letter Sound Interactions	* .34	1.01	1.56	2.01	.19	.32
Mean Plot, Text Implicit Interactions	.04	.12	.43	.43	* .39	.56
Mean Setting, Text Explicit Interactions	7.87	.26	1.99	1.99	* 10.62	1.54
Mean Phrase or Sentence Production Interactions	* .06	.30	.14	.14	.03	.11
Mean Time Decoding Without Text	* 1.59	2.70	3.68	3.68	.84	.79
Mean Word Comprehension Interactions	.63	.80	1.00	1.00	* 1.14	1.17
Mean Word Production Interactions	.35	.53	.09	.09	* .02	.07

Note. An * in the District Mean column for a given variable indicates that the variable was significant in that district's final regression model.
+ + Values reported for classroom instructional variables are means, across students and classrooms, of the individual student means calculated across the six observational rounds.

Table 6

Final Model of Regression Analysis for District A - Univille

Variable	R Square	R Square Change	Beta	*F* associated w/R square change	Significance of *F* for R sq. change
Degrees of Freedom at Step 1: 50					
Ability - zero	.1523	.1523	.3903	8.805	.005
Mean Phrase or Sentence Production	.4096	.2573	-.5084	20.922	.000
Mean Background Interactions	.4936	.0839	.3076	7.790	.008
Mean Time Spent Decoding Not Followed by Text	.5439	.0503	-.3077	5.075	.029
Mean Letter Sound Interactions	.5817	.0378	.4238	4.068	.050

Note. Dependent Measure: IGAP Comprehension, Total Score

Table 7

Final Model of Regression Analysis for District b - Middleburg

Variable	R Square	R Square Change	Beta	*F* associated w/R square change	Significance of *F* for R sq. change
Degrees of freedom at Step 1: 82					
Ability - zero	.2701	.2701	.5197	29.973	.000
Parental Support of Schoolwork	.3148	.0447	-.2147	5.225	.025
Second Grade Teacher Effect	.3492	.0343	.1870	4.166	.045
Mean Time Spent Decoding not Followed by Text	.4454	.0962	-.3263	13.535	.000
Mean Requests for Explanation Feedback	.4966	.0512	-.2314	7.827	.006
Mean Word Comprehension Interactions	.5518	.0552	.2373	9.364	.003
Mean Phrase or Sentence Production Interactions	.5790	.0272	-.1904	4.845	.031

Table 8

Final Model of Regression Analysis for District C - Divercity

Variable	R Square	R Square Change	Beta	*F* associated w/R square change	Significance of *F* for R sq. change
Degrees of Freedom at Step 1: 40					
Ability - zero	.4208	.4208	.6487	28.329	.000
Index of Parental Resources	.5287	.1080	.3806	8.707	.005
Mean Plot, Text Implicit Interactions	.6361	.1074	.3972	10.923	.002
Mean Word Comprehension Interactions	.7071	.0709	.3044	8.718	.006
Mean Setting, Text Explicit Interactions	.7756	.0485	-.2444	6.953	.012
Mean Word Production Interactions	.7922	.0365	-.1986	5.977	.020

Note. Dependent Measure: IGAP Comprehension Total Score.

Dependent Measure:	Individual's performance on the constructing meaning portion of Form 322 of the IGAP.
STEP 1:	Ability-zero, factor scale value made up of three weighted scores from fall, kindergarten testing.
BLOCK 2:	The five parent questionnaire scales that reflect parents reading to their child, child's participation in literacy activities, parental resources, and parental support of schoolwork.
STEP 3:	Criterion scale values for grade-1 teachers (based on IGAP reading test score).
STEP 4:	Criterion scale values for grade-2 teachers (based on IGAP reading test score).
BLOCK 5:	Classroom interaction variables, teacher feedback variables, and instructional and non-instructional time.

Figure 1. Initial Full Regression Model