2004

An investigation of the validity of the Iowa Early Learning Inventory

Sigurgrímur Skúlason
University of Iowa

AN INVESTIGATION OF THE VALIDITY OF THE IOWA EARLY LEARNING INVENTORY

by

Sigurgrímur Skúlason

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Education
(Educational Measurement and
Statistics) in the Graduate
College of The University
of Iowa

May 2004

Thesis Supervisors: Professor Audrey L. Qualls
Associate Professor Timothy N. Ansley

ABSTRACT

The objective of the present study was to investigate the validity of a new teacher rating inventory, the Iowa Early Learning Inventory (IELI). It is intended to assess the young students' cognitive behaviors, behaviors related to or supportive of the learning process. It is intended to be useful for early identification of students who are likely to encounter learning difficulties later in their academic careers. The intended users are kindergarten and 1st grade teachers.

The construct theory of the IELI was empirically derived using qualitative analysis of behavior statements obtained from relevant sources. Six areas of cognitive behaviors emerged from the qualitative analysis. The IELI was constructed to measure these areas using representative behaviors for item content.

All six areas of cognitive behaviors had medium or strong correlations with each other and all had correlations of medium strength with ITBS achievement measures. Investigation of the relationship of individual cognitive behavior areas with selected measures of achievement supported that the IELI captures variance related to other measures of related constructs.

Confirmatory factor analyses of the structure of the IELI investigated five models consisting of the six areas. None of the models was evaluated as adequately fitting the data. The most promising model consisted of six correlated areas or factors, each defined by the items intended to measure the respective cognitive area.

Investigation of sex DIF identified a small number of potential problems in three of the six areas, but investigation of ethnic DIF was inconclusive but called for further investigation.

Overall, the results indicated that the IELI provides useful and valid information about the cognitive behaviors of early learners.

Abstract Approved: ________________________________
Thesis Supervisor

____
Title and Department

____
Date

AN INVESTIGATION OF THE VALIDITY OF THE IOWA EARLY LEARNING INVENTORY

by

Sigurgrímur Skúlason

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Education
(Educational Measurement and
Statistics) in the Graduate
College of The University
of Iowa

May 2004

Thesis Supervisors: Professor Audrey L. Qualls
Associate Professor Timothy N. Ansley

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Sigurgrímur Skúlason

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Education (Educational Measurement and Statistics) at the May 2004 graduation.

Thesis Committee: ______________________________
Timothy N. Ansley, Thesis Supervisor

____
Sheila Barron

____
Stephen Dunbar

____
Hiram D. Hoover

____
Eugene Madison

To Audrey L. Qualls

ABSTRACT

The objective of the present study was to investigate the validity of a new teacher rating inventory, the Iowa Early Learning Inventory (IELI). It is intended to assess the young students' cognitive behaviors, behaviors related to or supportive of the learning process. It is intended to be useful for early identification of students who are likely to encounter learning difficulties later in their academic careers. The intended users are kindergarten and 1st grade teachers.

The construct theory of the IELI was empirically derived using qualitative analysis of behavior statements obtained from relevant sources. Six areas of cognitive behaviors emerged from the qualitative analysis. The IELI was constructed to measure these areas using representative behaviors for item content.

All six areas of cognitive behaviors had medium or strong correlations with each other and all had correlations of medium strength with ITBS achievement measures. Investigation of the relationship of individual cognitive behavior areas with selected measures of achievement supported that the IELI captures variance related to other measures of related constructs.

Confirmatory factor analyses of the structure of the IELI investigated five models consisting of the six areas. None of the models was evaluated as adequately fitting the data. The most promising model consisted of six correlated areas or factors, each defined by the items intended to measure the respective cognitive area.

Investigation of sex DIF identified a small number of potential problems in three of the six areas, but investigation of ethnic DIF was inconclusive but called for further investigation.

Overall, the results indicated that the IELI provides useful and valid information about the cognitive behaviors of early learners.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Information about student achievement is important in planning instruction, evaluating students' learning, communicating students' growth to parents, and making decisions about interventions for students (Millman & Greene, 1989; Salvia & Ysseldyke, 1998). Such information can be obtained by many informal and formal methods. The informal methods of gathering information are primarily observations of everyday student behavior or classroom performance, incidental examination of their class work, and discussions, interviews or oral questioning of students. Often these methods involve little or no attempt to control what information is gathered (Anastasi & Urbana, 1997; Gronlund, 1993; Salvia & Ysseldyke, 1998).

The information teachers obtain by observation and interaction with their students is the most readily accessible of all information about students. Regretfully, the information that teachers observe and remember during the regular classroom activities may not be systematically collected. The information teachers assimilate about students in everyday interaction can be influenced by assumptions or prejudice, psychological processes, selective attention, existing opinions and prior knowledge. The teachers may note different things about different students. What they note may be evaluated in light of prior knowledge, and what they miss may be as relevant as what they note (Gleitman, 1992; Thorndike, Cunningham, Thorndike, & Hagen, 1991). The extent to which psychological processes and cognitive structures of this kind affect teachers' observations and other informal assessments constitutes the most serious threat to their validity (Anastasi & Urbana, 1997; Thorndike, et al., 1991).

Formal measurement methods control the process of information collection by presenting the students with specific questions or tasks that they respond to (Anastasi & Urbana, 1997). Numerous formal methods have been designed to collect more reliable and valid information about relevant student achievement than the informal methods yield. The most common are teacher made paper and pencil exams, commercial paper and pencil exams available as part of textbooks or instructional material, and planned grading of assigned projects and homework. These types of measurement tools are the most commonly used to gather information about students' cognitive performance (Stiggins, Conklin & Bridgeford, 1989). In addition, standardized tests are often administered at various grade levels as external information sources to supplement the information obtained in the classroom (Salvia & Ysseldyke, 1998; Stiggins & Bridgeford, 1985).

It is most important for teachers to obtain information about their students' present and potential academic achievement in the various curriculum domains that the classroom instruction covers. Academic achievement has been defined as the skills and knowledge that a student acquires during a course of study (Anastasi & Urbana, 1997; Messick, 1984; Snow & Lohman, 1984). The focus on knowledge or skills acquired as a result of learning or training in a given curriculum area distinguishes achievement from the related concept of aptitude. Aptitude, or ability, is defined in terms of individual differences on psychological characteristics such as intelligence or cognitive processing (Snow, 1980; Snow & Lohman, 1984). Aptitudes are prerequisite or readiness skills for acquiring new skills or knowledge. Achievement represents the skills or knowledge that the learner accomplishes mastery of as part of specific educational experience (Anastasi & Urbana, 1997; Snow, 1980; Snow & Lohman, 1984, 1989). While these concepts can be distinguished at the theoretical level, measures of achievement and aptitude differ primarily in how the construct underlying the measure is defined. The measurement tools themselves may differ only to a limited extent (Anastasi & Urbana, 1997).

What the teachers observe and learn about their students through informal methods is likely to cover aspects of both students' aptitude and achievement. It will capture various aspects of students' academic strengths and weaknesses in specific curricular areas, as well as an evaluation of their overall intellectual abilities. This type of information is also captured by formal measurement tools. But teachers are also likely to observe specific behaviors that enable the student to become a learner and characterize the learning activity of academically successful students. There are important cognitive behaviors that students must acquire at an early age to become successful learners. Cognitive behaviors are skills, work habits and other behaviors that students acquire that support further learning and make learning efforts more efficient. As learned skills and tools to acquire other new skills and knowledge, cognitive behaviors share the defining properties of both aptitude and achievement. Such cognitive behaviors as persisting on task, working independently, listening attentively, organizing, and being actively involved in classroom activities are emphasized in many curriculum guides as desirable and essential characteristics of the young learner (California Department of Education, 1988; Kentucky Department of Education, 1993a, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988). The assessment of learners' cognitive behaviors and work related skills is also important as an integrated part of the assessment of the young learners' academic achievement (Bergkamp & Rosegrant, 1992; Resnick & Resnick, 1992).

Differences in how well children acquire or master cognitive behaviors as they begin school may greatly affect the students' early academic success. Those who lack these behaviors are likely to experience academic difficulties early on, while the academic success of students that master those cognitive behaviors is likely to continue beyond the elementary school years. Numerous studies support the role of specific cognitive behaviors during the learning process. Furthermore, the lack of such behaviors has also been related to academic failure (Badian, 1982, 1988; Cartledge & Milburn, 1978; Cobb, 1972; Loyd & Loper, 1986; McKinney, Mason, Perkerson, & Clifford, 1975; Perkins, 1965; Soli & Devine, 1976). Information regarding the extent to which kindergarten and first grade students have or have not acquired those behaviors is therefore important for teachers. The earlier these deficiencies in behaviors are identified, the more likely it is that they can be successfully taught and reinforced (Cartledge & Milburn, 1978; Satz & Fletcher, 1988; White, 1986).

It is difficult to obtain relevant information about students' cognitive behavior with direct measurement tools. Direct measurement tools are designed to yield trustworthy information about students based on their completed academic work, such as responses to questions or completed projects. They are well suited to assess knowledge or skills that are on either end of the aptitude achievement continuum: either intellectual aptitudes or academic achievement in content matter areas (Anastasi & Urbana, 1997; Thorndike et al., 1992). However, they are not well suited to assess the cognitive behaviors and skills that students use during academic work. Teachers obtain information regarding this important aspect of their students' cognitive behaviors every day using informal methods in the classroom, but the subjectivity involved in the informal methods severely limits their validity and reliability. This limits the usefulness of a valuable source of information to reach sound instructional decisions (Anastasi & Urbana, 1997).

The validity of information obtained by teachers through observation and interaction with their students can potentially be enhanced using formal measurement tools to collect and summarize this knowledge. Such tools may increase the quality of the information from teacher-student interactions, resulting in better instructional decisions. Rating scales and inventories are structured measurement tools that could prove to be appropriate mechanisms for extracting information about students' academically relevant behaviors from the teachers; information that the teachers regularly gather and currently underutilize (Gronlund, 1993; Thorndike, et al., 1992). These types of measurement tools can provide a degree of standardization, avoid the selectivity of recall by systematically prompting the user to recall what he or she knows about the relevant dimensions of the students' behavior and skills, and minimize subjectivity by focusing on observable and concrete aspects of student behavior (Anastasi & Urbana, 1997; Gronlund, 1993).

Rating scales and inventories have been successfully designed to measure psychological attributes such as students' emotional development or behavior problems, but are less commonly used to assess cognitive behavior (Hoge, 1983; Nitko, 1996; Salvia & Ysseldyke, 1998). The primary justification for their use is that they offer an efficient way of collecting data from individuals who already have relevant information regarding the person, object or situation. They require relatively limited teacher time, and they do not demand student time (Anastasi & Urbana, 1997; Thorndike, et al., 1991). In addition, they can be used to gather information that would otherwise be difficult to gather or even inaccessible with direct student measures (Anastasi & Urbana, 1997; Gronlund, 1993). Cognitive behavior is one area where rating scales and inventories can potentially prove to be valuable assessment tools.

Cognitive or learning related behaviors are behaviors assumed to be supportive of successful learning or successful completion of the tasks engaged in for learning purposes in school. What behaviors would be most conducive for systematic assessment in order to yield valued information that could be used to maximize the students learning opportunities? Is there a set of cognitive behaviors indicative of successful achievement in school? What behaviors enhance learning; are essential elements in the learning process; or make learning activities more efficient? Are there behaviors that would yield useful information for instructional purposes? No explicit definition of cognitive behavior could be located in the literature, but any study of the relation between specific behaviors, or classes of behavior and achievement, operates on the basis of assumptions similar to the one above. It is therefore important to consider what cognitive behaviors are, and why certain behaviors are indicative of educational or academic achievement. For reasons related to the validity of measurements, behaviors used in a measurement tool as indicators of academic or educational achievement should be related to the theoretical conceptions of that construct. It is important to consider both whether the relationships of certain behaviors with academic achievement are theoretically sound, and also to identify empirical evidence supportive of those relationships (Messick, 1989). Current theories of learning and cognition could provide insight regarding which cognitive behaviors should be related to educational achievement. The relevant theories of learning commonly model learning in units of unobservable behaviors, and they may only inform us indirectly about observable cognitive behaviors (Eggen & Kauchak, 1999; Gleitman, 1992; Snow & Lohman, 1984). This renders the empirical support for the assumed relation between cognitive behavior and academic achievement all the more important.

The three current dominant theories of student learning and achievement are derived from behavioral psychology, cognitive psychology, and situated cognition (Greeno, Collins & Resnick, 1996). The first, based on behavioral psychology, views knowledge as a body of independent facts and skills connected into an organized system. Learning is both the acquisition of new facts or skills and new connections formed within the system. According to this view, learning occurs as knowledge and skills are associated with consequences of behaviors or objects and events in the environment. Behavioral theories of learning emphasize small or simple connections as the units of learning and assume that complex knowledge and patterns of behavior are built up from simpler skills and knowledge. Behavioral learning theory also incorporates observation and modeling as important aspects of learning in social contexts (Gleitman, 1992; Greeno, et al., 1996). According to this theory certain classroom behaviors would reasonably be expected to relate to learning. These would include paying attention to instruction and demonstrations, practice of recently learned material, and using current knowledge and skills as part of the process of completing assigned projects or tasks. Evidence of the possession of fundamental linguistic and quantitative concepts should also be regarded as cognitive behavior (Ferster, Culbertson, & Boren, 1975).

Cognitive theories of learning emphasize the role of cognitive structures in the storage and processing of information. Knowledge is commonly represented as consisting of declarative and procedural knowledge. The former consists of networks of facts and connections among them, the latter, of sequences of steps and operations required to solve a task. Learning is a goal directed activity according to the cognitive framework. Cognitive structures are modified, refined or reorganized as richer information is stored in memory; or as procedural knowledge structures are built from declarative knowledge; or as existing procedural structures are adapted to unfamiliar problems (Anderson, 1983, 1990; Eggen & Kauchak, 1999; Gleitman, 1992). Knowledge and learning are modeled as invisible mental events, but they are frequently accompanied by certain observable classroom behaviors that indicate learning. Examples of such observable cognitive behaviors are completing assigned tasks and projects, being involved in learning related activities, practicing acquired skills, or using solution strategies to solve new tasks or problems. Answering questions, asking clarifying questions, and discussing one's knowledge are also important behavioral indicators of communication and language skills (Anderson, 1990; Eggen & Kauchak, 1999).

The more recent and less well established theory, situated cognition, views aspects of the learner, other individuals, and the environment or learning situation as relevant parts of the learning process. Knowledge is viewed as the individual's ability to function in a situation. Learning represents the increased adaptation of learners to an environment or their increased effectiveness within an environment (Eggen & Kauchak, 1999; Greeno, et al., 1996). This view emphasizes skills that enable a learner to adapt and become proficient in new and unfamiliar situations, and the active involvement of the learner in the practices that are common to the situation. Examples of observable cognitive behaviors consistent with situated cognition theory are effective work habits and specific skills that allow a person to adjust to and master a new situation, and the establishment of situational specific skills. General knowledge or skills that are helpful in many unfamiliar or new situations would also be considered cognitive behaviors. Language, communication skills and quantitative concepts are such important general behaviors. Many of the cognitive behaviors that the behavioral and cognitive theories of learning identify are also consistent with the situated cognition view (Greeno, et al., 1996).

Cognitive behaviors can be better understood if one is willing to accept the assumption that they are behaviors involved in learning and supporting the learning process. For the purpose of the present study, the interest is in observable behaviors that are prominent and commonly used by children when they are engaged in learning activities in the classroom. Given the present focus, cognitive behavior can be defined as observable behaviors that enhance learning (i.e., planning one's work, communication and/or listening skills); behaviors that are part of a learning activity (i.e., attending to instruction or staying on task); and behaviors that make learning activities more efficient (i.e., using appropriate study skills). Furthermore, observable behaviors that indicate to the teacher that learning has occurred, such as completing assigned tasks or turning in a project, would also be considered cognitive behaviors. Finally, behaviors that indicate cumulated prior learning can also be viewed as cognitive behaviors. Examples of these behaviors indicating earlier learning knowledge of things or concepts prominent in the child's culture, such as linguistic fluency, knowledge of print, and numerical concepts.

Acquisition and development of cognitive behaviors are important for children's early experience in school. They are believed to be important because they allow the child to become a learner (Bergkamp & Rosegrant, 1992; National Association for the Education of Young Children, 1988; Nebraska Department of

Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988; Resnick & Resnick, 1992). How well children have acquired basic cognitive behaviors as they progress through the beginning years of school may greatly affect their academic success. Those who lack these behaviors are more at risk to experience limited academic success, but it might be possible to help a child acquire these cognitive behaviors if these deficiencies were known. A compelling rationale for early identification of cognitive behavioral deficiencies is that behavior patterns of younger children are more susceptible to change than those of older children. Therefore it may be that teaching cognitive behavior to children who lack these skills in the early grades could ensure that these students are more likely to become academically successful (Mercer, Algozzine, & Trifileti, 1988; Satz & Fletcher, 1988). Behavioral intervention has been more effective if introduced early in the students' schooling, while such efforts have proven less helpful as the students get older. If such difficulties are known already in kindergarten or first grade, they may be remedied relatively quickly and easily. They need not serve as impediments to learning (Salvia & Ysseldyke, 1998; Satz & Fletcher, 1988; White, 1986).

The purpose of the present study is to investigate the properties of a new teacher rating inventory, the *Iowa Early Learning Inventory* (IELI) (Hoover, Dunbar, Frisbie & Qualls, 2001), that has been developed to measure kindergarten and first grade students' cognitive behaviors. The IELI is intended to provide teachers with relevant information about the strengths and weaknesses of students' cognitive behaviors based on what the teachers observed and learned about their students. It is primarily intended to be useful for instructional planning and to be a source of reliable and valid information that can supplement what teachers already have access to through other formal and informal measurements tools. The IELI measures six areas of cognitive behaviors, each assessing a different aspect of students' cognitive behaviors and skills: general knowledge, oral communication, written language, math concepts, work habits, and attentive behaviors. These areas of cognitive behavior were identified based on analysis of the content of curriculum guides and empirical research regarding the relation between classroom behavior and achievement. The cognitive behaviors chosen for each area were selected from behaviors identified in that content analysis. If the assumption of the relation between achievement and certain cognitive behaviors is valid, an inventory such as the IELI could be used to collect such information regarding these behaviors in a systematic and reliable manner.

Of primary interest in this study are the validity and reliability of the IELI: its internal structure, its relation to a standardized achievement test, the absence of bias from its scores, and the reliability of the ratings. Do the behaviors assessing each of the six different aspects of cognitive behaviors have stronger relations among themselves than with the behaviors measuring other aspects of cognitive behavior? Do the scores reflect the theorized structure of cognitive behaviors, and are they correlated with scores on a standardized achievement instrument? How do the cognitive behaviors measured for students belonging to different demographic subgroups compare? The purpose of the present study is to provide answers to these questions.

CHAPTER II
REVIEW OF THE LITERATURE

There are a number of distinct literature areas that must be consulted in this review. It is necessary to consider the views of educators and the educational community related to cognitive behaviors, as well as the empirical literature on the relation between specific behaviors and achievement. In addition, the measurement issues relevant to rating scales and inventories will be reviewed. Finally, several specific teacher rating scales designed to assess achievement related concepts will be reviewed.

Knowledge and opinions regarding which cognitive behaviors are important for young learners' early academic success differ among educators and educational researchers. Several important sources in the identification of these opinions can be found in educational documents such as: curriculum guides, findings from observational studies in classroom situations, teacher ratings of student behaviors, and empirical studies investigating early learning skills. The knowledge gathered from these various sources has been used by many developmental researchers to develop assessment tools that represent a collection of behaviors important for early academic success. The views and results of these researchers must also be consulted. An important step towards the development of a comprehensive understanding of cognitive behaviors would begin with a review of these varied efforts. Such a review was carried out on a sample of documents representing all three of the above identified areas of beliefs and opinions, the educational community, the research literature and those who have previously attempted to construct rating instruments assessing cognitive behaviors.

Areas of cognitive behaviors emphasized in curriculum guides

The curriculum and classroom activities presented by individual teachers are greatly influenced by collections of decisions and instructional planning reflected in curriculum guides. Curriculum guides are written by instructional and curriculum matter experts as well as experienced teachers for the purpose of supporting good teaching practices in schools, establishing promising new practice, and providing coherent instructional frameworks for schools and teachers. They are intended to aid schools in planning and designing their own local curricula. They represent what is known to be good instructional and classroom practice and are influenced by the experience of teachers in the classroom (Biehler & Snowman, 1993; Nebraska Department of Education & Iowa Department of Education, 1993). Multiple aspects of instruction and classroom activities are discussed in curriculum guides. They commonly spell out certain content that teachers are expected to present to students; sometimes instructional methods are mentioned; and the guides often speak about valued student characteristics, attitudes or behaviors (Biehler & Snowman, 1993). A valuable discussion for the present purpose involves expectations regarding students' learning skills and achievement that are stated in terms of the learners' behaviors.

A sample of 12 curriculum guides from various state departments of education were obtained and examined to identify cognitive behaviors relevant for the young learner (Alabama State Department of Education, 1992; California Department of Education, 1988; Indiana Department of Education, 1987, 1992; Kentucky Department of Education, 1993a, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988, 1990; Tennessee Department of Education, 1992; Wisconsin Department of Public Instruction, 1986, 1991). Several classes of behaviors were systematically mentioned in the curriculum guides as important for the young learner. These areas of behavior can be categorized as oral communication, written language, general or personal knowledge, quantitative or math concepts, work habits, attentive behaviors, social behaviors, affective behaviors, and appropriate conduct. Of these areas, the first six were judged to be cognitive behaviors relevant for primary grade student learning and academic achievement. The latter three, although important aspects of student personal growth, were not evaluated to be cognitive behaviors, or goals easily amenable to objective measurement.

Behaviors categorized as oral communication consisted of speaking and communication skills, and expression of ideas. Among the mechanisms of communications were behaviors related to pronunciation, feedback provided throughout a conversation, use of complete sentences, sufficient vocabulary, and appropriate use of nonverbal cues such as tone of voice, maintain eye contact, and facial expressions (Kentucky Department of Education, 1993b; Oregon State Department of Education, 1986, 1990; Indiana Department of Education, 1992; Tennessee Department of Education, 1992; Wisconsin Department of Public Instruction, 1991). A second major communication theme was related to the message or topic conveyed, which consisted of expression of ideas, structure of the communication and the selection of topic. Behaviors exemplifying structure and appropriate topic are: provide ideas relevant to the topic, ask questions, recognize main idea and relevant details, understand and use chronological or logical order, retell a story, and describe a picture (Oregon State Department of Education, 1986, 1988; Indiana Department of Education, 1992; Tennessee Department of Education, 1992; Wisconsin Department of Public Instruction, 1991). This theme also included expression of ideas: share ideas, experiences and opinions, and the use of emotions and feelings as a topic of conversation. Many of the communicative behaviors were described both as expressive and receptive. They were also believed to be relevant for both classroom discussion and activities, and as part of conversation in general (Kentucky Department of

Education, 1993a,1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988, 1990; Wisconsin Department of Public Instruction, 1991). From these guides one could not find clear boundaries between oral communication and social behaviors. The two concepts must be arbitrarily dichotomized depending on the purpose of each investigation. These areas of behaviors merge into one another, as many behaviors are both part of appropriate communication and general social awareness.

A second area identified in the curriculum guides involved behaviors related to print. These can be described in terms of increasing mastery of the symbolism used to represent language in written form. Awareness of letters and words was the first level: recognize certain strings of symbols as print, know that print carries a message, scribble resembling letters, and interest in books and other print related material (Oregon State Department of Education, 1988, 1990; Indiana Department of Education, 1992; Nebraska Department of Education & Iowa Department of Education, 1993; Tennessee Department of Education, 1992). A second level focused on the recognition of individual letters and words: recognize one's own name, know phoneme-letter correspondence, recognize individual words, write individual letters or simple words (Oregon State Department of Education, 1986, 1988; Indiana Department of Education, 1992; Nebraska Department of Education & Iowa Department of Education, 1993; Tennessee Department of Education, 1992). Other more advanced levels involved writing of many words, mastering conventional spelling, reading strategies, writing meaningful sentences and writing short stories or letters. Fluent reading and writing represent still higher levels of print related behaviors. The curriculum guides emphasized the continuity of pre-literacy behaviors to fluent reading and writing. As a result, the distinction between communicative and written behaviors blurs.

Behaviors related to quantitative concepts constituted a separate area. One aspect of this area was behavior related to numbers: count, numeral literacy, recognize mathematical concepts such as more and less, and use ideas related to magnitudes. Other measurement and classification concepts are important behaviors in this area: know relative value of coins, recognize measures of different aspects of objects, recognize basic shapes, copy basic shapes, and classify objects according to a rule (Oregon State Department of Education, 1986, 1988; Indiana Department of Education, 1992; Nebraska Department of Education & Iowa Department of Education, 1993; Tennessee Department of Education, 1992).

An area of behavior that can be described as general knowledge consisted of personal behavior and behavior relevant in a variety of situations. Among the personal behaviors were knowing one's own first and last names, address and birthday. Other personal behaviors were related to independence: taking off and putting on one's coat or jacket and shoes. The general behaviors included recognition of the relative size of common objects, name colors, follow directions that use left or right, know units of time such as days, weeks and months (California Department of Education, 1988; Indiana Department of Education, 1987, 1992; Oregon State Department of Education, 1986, 1988; Tennessee Department of Education, 1992; Wisconsin Department of Public Instruction, 1991). This area can be described as consisting of behaviors relevant for the child in many different situations.

Work habits were a set of behaviors ranging from task monitoring strategies, to setting personal learning goals. Important elements or themes in work habits were planning and monitoring behaviors during work: staying on task, working on assignments despite difficulties, selecting effective strategies, identifying personal strengths, and setting realistic learning goals. The curriculum guides stressed the young learners' independent involvement in classroom activities. This was indicated by behaviors such as: completing assignments by themselves, requesting assistance when needed, asking clarifying questions, applying learning strategies depending on situation, turning assignments in on time, and following a study plan (Indiana Department of Education, 1987, 1992; Kentucky Department of Education, 1993a, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988, 1990; Tennessee Department of Education, 1992; Wisconsin Department of Public Instruction, 1991). Some of these may be unrealistic expectations for the primary student, but the curriculum guides did place an emphasis on supporting the young learner in attaining disciplined work habits.

The final area of cognitive behavior was attentive behaviors. Relatively few attentive behaviors were described in the guides. They can be described as listening behaviors and task orientation behaviors. Listening behaviors were: listening to teachers' instructions, stories and directions, and to other students talking. Task orientation behaviors were: staying on task, completing short assignments, and resisting distractions. It should be noted that there was very little distinction for this area and work habits (Indiana Department of Education, 1992; Kentucky Department of Education, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988,1990; Tennessee Department of Education, 1992; Wisconsin Department of Public Instruction, 1991).

In addition to the cognitive behavior areas, three other areas of behavior were identified: social behaviors, appropriate conduct behaviors and affective behaviors. Social behavior was the most prominent of these areas in the curriculum guides. The most prominent themes were relations with peers, communicative mechanisms, relations with adults and authorities, ability to work in groups, and attitudes such as respect for others and their ideas. Relations with peers in the classroom appeared to be the most heavily emphasized. Examples of such behavior were: taking turns, showing respect for others, initiating contact, and solving conflict (California Department of Education, 1988; Kentucky Department of Education, 1993a, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1988, 1990; Wisconsin Department of Public Instruction, 1986, 1990). Behaviors classified as appropriate conduct in the classroom were also frequently mentioned: conform to rules and authorities, be responsible for one's actions, and interact with peers without conflict (Indiana Department of Education, 1992; Kentucky Department of Education, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1988; Wisconsin Department of Public Instruction, 1986, 1990). Affective behaviors included behaviors such as independence and emotional stability, behaviors indicating anxiety, and adjusting to new situations (Kentucky Department of Education, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1988).

This review has shown several areas of classroom behavior that are either typical of young learners or expected of them. It is clear that not all these behaviors are integral parts of the learning process, and that some important behaviors, such as those in the affective dimension, fall outside the present focus on cognitive behaviors.

Empirical evidence of the relation of cognitive behavior with academic achievement

The studies reporting empirical data on the relationships between specific classroom behavior and academic achievement will be reviewed next. They were organized into three groups based on the methods used to collect data: observations of classroom behaviors, teacher ratings of student behaviors, and studies using direct measures of specific behaviors or skills. Studies using different data, such as observations, teacher ratings and direct measures, may be informative regarding the relationships of different areas of cognitive behavior with achievement. They may also have different strengths and weaknesses. It should be noted that a considerable number of studies sampling exceptional or clinically diagnosed students were excluded.

Observations of classroom behavior and academic achievement

Most of the studies that relate observation of classroom behavior with academic achievement were conducted between 1965 and 1980. A theoretical focus common to many of these studies was to identify distinct behaviors characterizing low achieving students. These behaviors were subsequently used to identify students at risk of developing learning difficulties or disabilities.

Behavior observations can assess a variety of characteristics or dimensions of behavior: frequency, duration, or intensity (Gresham, 1982; Thorndike, et al., 1991). Observational procedures known as time sampling were used in most of the studies reviewed. This is a frequency of behaviors approach; behaviors are counted during samples of short time intervals over an extended time. Commonly, each child is observed for several five minute intervals on each of a given number of days. The observer notes the child's behavior every 5 to 10 seconds, and checks the appropriate behavior category on the scoring sheet (Cone, 1982; Foster & Cone, 1980). Observations would be repeated for four or five days, in the same week, for each child in most of the studies reviewed (Cobb, 1972; Derevensky, Hart, & Farell, 1983; Lahaderne, 1968; McKinney, et al., 1975). A few studies relied upon observations over longer periods of time (Bryan, 1974; Perkins, 1965), or repeated short observation periods during the fall and spring of the same school year (McKinney, et al., 1975).

Several studies have investigated the relation between general behavioral categories and achievement. The observer notes either the attention and inattention or focus-on-task and off-task type of behavior categories (Bryan, 1974;

Bryan & Wheeler, 1972; Derevensky, et al., 1983; Lahaderne, 1968). The term molecular has also been used to describe these general behaviors (Hoge, 1985).

An important early study of attentive behaviors was Lahaderne's (1968) observations of 125 6th grade children. Attentive behaviors were defined as working on assigned activities, writing or reading the assigned subject matter; and inattentive behaviors as work on non-assigned tasks or subject matters, play with tools and materials, and fooling around. A separate category was used when the observer was uncertain or could not determine the focus of attention of the student. The frequency of attentive and inattentive behaviors correlated significantly with standardized tests scores in reading, arithmetic and language. The correlations between behaviors and the achievement measures were reported separately for boys and girls, and all were significant. They ranged respectively from $r = 0.37$ to $r = 0.53$ for attentive behaviors, and $r = -0.38$ to $r = -0.53$ for inattentive behaviors, with medians of the correlations, $r = 0.47$ and $r = -0.455$ respectively. The behavior achievement correlations remained significant after controlling for students' intelligence scores. Lahaderne & Jackson (1970) replicated this study as part of a later study and obtained correlations of similar magnitude.

Samules and Turnure (1974; Turnure & Samules, 1972) investigated the relationship between behaviors and achievement for a primary grade sample, observing 88 students in four 1st grade classrooms. They used an observational system similar to that used by Lahaderne (1968). Attention was defined as direction of gaze or visual orientation of the student. Their examples of attentive or task relevant behaviors were: eyes oriented to text, teacher or chalkboard, working on assigned tasks, and following the teacher's directions. Inattentive behaviors were: failing to attend to instruction or follow directions, playing, and working on non-assigned tasks. Attentive behaviors correlated significantly with a word recognition measure, $r = 0.44$. The importance of this study was the use of a 1st grade sample to examine the relationship between attentive behaviors and achievement. The findings indicate that meaningful relationships between attentive behaviors and achievement exist in the primary grades.

Gambrell, Wilson, and Gnatt (1981) were interested in the relationship between reading behaviors and achievement. They observed task-related behaviors among 35 good readers and 35 poor readers during reading instruction in 17 4th grade classrooms. Good and poor readers were identified based on both standardized test scores and teacher nominations. The authors attempted to control for the effects of general aptitude by excluding students who obtained a high or low score on a test of general intelligence. On-task behaviors were reading behaviors, while the off-task behaviors were all other behaviors, including work on other subject matters. Gambrell, et al. (1981) found a significantly greater percent of on-task behaviors among the good readers, and a significantly greater percentage of off-task behaviors for the poor reader group. An interesting aspect of this study was the focus on classroom behavior for a particular subject matter. If the students were working on another subject matter than the one assigned, those behaviors were classified as off-task behaviors. These findings indicate that working on each subject matter as assigned relates to higher achievement in the primary grades.

Derevensky, Hart, & Farrell (1983) investigated behaviors differentially characterizing high and low-achieving students. Their sample consisted of the 15 highest and 15 lowest scoring students on a standardized achievement test in each of the 1st through 6th grades in one school. They used a seven category observation system of task related behaviors: on-task writing: apply pencil in task related to lesson; on-task oral: ask question, respond to question, and discuss lesson with teacher or peers; on-task covert: listen, read; off-task: daydream, socialize, misbehave. Significant differences between high and low achieving primary grade students were found for four behavior categories: off-task behaviors as well as on-task writing, on-task oral, and covert on-task behavior. Based on the comparison of high- and low-achieving students, these authors concluded that the higher-achieving students spent a "greater proportion of their time academically engaged in activities such as listening, asking questions, and seat-work, thought to be conducive to the acquisition of knowledge in the classroom setting" (Derevensky, et al., 1983, p. 333).

Bryan & Wheeler (1972) investigated different behaviors of students with learning difficulties and regular students in several subject matter classes. They observed ten students identified by the five classroom teachers as learning disabled and ten students identified as average by the same teachers. The behaviors observed were task-oriented behaviors, non-task-oriented, interactions and waiting. Examples of task-oriented behaviors were reading and participating in classroom activities; and non-task-oriented behaviors were staring out a window and moving in and out of their seats (Bryan, 1974; Bryan & Wheeler, 1972). Bryan & Wheeler (1972) reported learning disabled students as having a significantly lower percentage of task oriented behaviors (68.0%) than the typical students (87.7%), and a significantly greater percentage of non-task behaviors (33.0%) than the typical students (10.9%). Bryan (1974) replicated this result with a smaller sample of ten boys, but each student was observed more frequently over an extended period of five months. This study also found off-task behaviors more frequent for learning disabled students than typical students (70% versus 57%), and their off-task behaviors more frequent than the typical students (34% versus 26%). In addition, significant results, with the same direction of relationship, were found for observations during four specific subjects matters: arithmetic, language, arts and music; and for attending to teacher instructions (Bryan, 1974; Bryan & Wheeler, 1972).

The studies reviewed so far have supported the relationship between attentive and task-related behavior and achievement, as well as the relationship of deficiencies in cognitive behavior with achievement. High frequency of on-task behaviors related with high achievement while frequent off-task behaviors related with low achievement (Bryan, 1974; Bryan & Wheeler, 1972; Derevensky, et al., 1983; Lahaderne, 1968; Lahaderne & Jackson, 1970; Samuels & Turnure, 1974).

Studies using observation systems that code specific behaviors independently will be reviewed next. In an early study attempting to identify student behaviors that relate to achievement, Perkins (1965) observed the behaviors of 36 low achieving and 36 high achieving students in 14 classrooms. The observational system consisted of behaviors theorized to characterize high and low achieving students based on behavioral learning theory. Students appeared to have been observed over a five month period, with a total of about 70 minutes of observations for each student. Classes of behaviors theorized to characterize low achieving students were: work on other academic tasks: work on any other school activity than those assigned; non-academic tasks: non-academic work, prepare for assignments, frequently clean desk; and a class of behaviors labeled withdrawal: not involved in a classroom situation, out of contact with people or ideas, detached, and daydream. Perkins (1965) combined all behaviors theoretically related to low achievement, and found a significantly greater percentage of such behaviors among the low achievers, 21.9%, than the high achievers, 17.7%. In addition, significant differences were found to the effect that low achievers discuss academic matters less with peers, and are less involved in classroom activities. Interestingly, neither a hypothesis nor a significance test was reported regarding the high achievement behaviors. But the differences between achievement groups for the combined classes of behaviors hypothesized to characterize high achievers were almost as great as the differences on the low achieving behaviors, and greater than differences reported to be significant between gender groups.

In an important study, Cobb (1972) investigated whether relations previously obtained with broad behavior categories held up when more specific behaviors were observed. Observations of 103 4th grade students were conducted during arithmetic classes in two schools. A 14 category behavioral observational system was used. Behavior categories relevant for the present study were: attention: attend to instruction, write or work on an assignment, listen to other students recite; compliance: open book or turn in paper at teacher's request; talk-to-peer: talk about academic matters with a peer; self-stimulation: attend to non-academic tasks, play with pencil or erasers, tamper with clothes, fidget, scratch; out-of-chair: walk around room, wait for assistance; volunteer: raise hand in class when teacher asks questions, indicate intent to make academic contribution; and finally look around: stare out a window, stare blankly into space, or look around the classroom. These seven classes of behavior had significant correlations with reading or math scores on a standardized achievement test in one or both schools. The absolute value of the significant correlations ranged from $r = 0.30$ to $r = 0.48$. Positive relations were found for attention, compliance, and talk-to-peer; negative relations for self-stimulation, out-of-chair behaviors, looking around, and non-attentive behaviors. One class of behavior, volunteering, had negative correlations with achievement in one school but positive in the other (Cobb, 1972). Two important issues regarding these results should be noted. The first issue is the instability of the relationships, only attention and out-of-chair behaviors had significant relations with both achievement measures in both schools. The second issue regards the generalizability of cognitive behavior across situations. Cobb's (1972) observations of behaviors were only conducted in arithmetic classes, but they also predicted reading achievement in both schools ($R^2 = 0.56$ and $R^2 = 0.41$, respectively), indicating that these behaviors were relevant for both subject matters.

Soli & Devine (1976) investigated whether the same or different behaviors characterize high and low achieving students for a sample of 312 3rd and 4th grade students in 14 classrooms during arithmetic and verbal skills instructions. They obtained 125 observations per student using Cobb's (1972) observational system described above. Two additional behavior categories used need to be described: initiating contact with the teacher: requesting academic assistance from the teacher; and playing: play with another student when not appropriate (Cobb, 1972; Soli & Devine, 1976). Soli and Devine (1976) reported the correlations of the observed behaviors with standardized reading test scores for the total sample and samples of high and low achieving students. They also reported a multiple regression of behavior observations on achievement scores for the same sample. The behavior categories that correlated with achievement scores for the total sample were talk-to-peer (r = 0.23), attending (r = 0.17), playing (r = -0.18), self-stimulation (r = -0.14), looking around (r = -0.14), and not attending (r = -0.21). Splitting their total sample at the mean of the composite achievement score to obtain separate high and low achievement groups, Soli & Devine (1976) found that for the high achievement (n=165) group only talk-to-peer correlated with reading (r = 0.32). Significant correlations were found for six classes of behaviors in the low achieving group (n=147): complying (r = 0.18), attending (r = 0.24), playing (r = -0.24), self-stimulation (r = -0.19), looking around (r = -0.18), and not attending (r = -0.19). A multiple regression analysis with achievement scores as the dependent variable yielded a squared multiple correlation of R^2 = 0.45 for the total sample. Behavior categories that entered the regression equation were: talk-to-peer, attending, self-stimulation, and initiating communication with teacher. For the high achieving group talk-to-peer, initiating to teacher, and out-of-chair behavior entered a similar regression equation (R^2 = 0.44). Behaviors entering the regression equation for the low achieving group were: playing, not attending, self-stimulation and complying (R^2 = 0.44). Furthermore, when the separate regression equations obtained for the high and low achieving groups were cross-validated on the other achievement sample, they yielded near zero and non-significant multiple correlations (R^2 = -0.06 and 0.04). These results are interesting for the clear indication that behavior characteristics of academically successful students are absent in lower achieving students, supporting the notion that lower achieving students have cognitive behavior deficiencies.

McKinney et al. (1975) attempted to identify behaviors that characterize students having learning difficulties. They observed 90 2nd graders in two schools. A multicategory observational scheme was used, and separate observations were conducted during both the fall and spring semesters. The behavior categories that were found to relate to achievement were as follows: 1) attentive behaviors: attend to instruction, respond to teachers questions, listen to a story, raise hand to answer questions; 2) constructive play: draw, puzzle; 3) self directed activity: work independently, read, write, complete assignments, work on assigned seat work; 4) task oriented peer relation: converse about a lesson, discuss an assignment, and work with peer on assignment; 5) distractibility: look out a window, look up when another child moves, stare into space; 6) passive responding: nonattentive in group activity, wait for directions or assistance; 7) dependability: wait for teachers' attention, ask another child for information or assistance, request teacher to check one's work; and finally 8) aggression: refuse to carry out instruction. The first four classes appear to have had positive relations with achievement, but the latter four had a negative relationship. McKinney, et al. (1975) reported separate regressions for the fall and spring observations with the standardized achievement test scores as the dependent variable. Distractibility, passive responding, dependability, and constructive play entered the regression equation for fall observations with fall achievement used as the dependent variable, $R^2 = 0.40$. When fall observations were used to predict spring achievement, the same four behaviors were significant predictors ($R^2 = 0.36$). Using spring achievement as the dependent variable, four different behaviors from spring observations were significant predictors: self directed activity, aggression, attending, and task orientation ($R^2 = 0.23$). McKinney (1989) summarized several studies using the same observation system with samples of learning disabled or special education students and normal controls. The higher achieving students were found to be more attentive, task oriented and more independent than lower achieving students.

In summary, these classroom observational studies support a relation between cognitive behavior and academic achievement. Specific cognitive behaviors such as work habits, attentive behaviors, and oral communication have at least a tentative relation with achievement. Similar conclusions have been drawn in earlier evaluations of classroom observation studies. Hoge & Luce (1979) concluded that there was "at best, a moderate relation of classroom behaviors and academic achievement" (p. 493). McKinney (1989), on the other hand, interpreted the evidence as indicating that students exhibiting certain behaviors "were more likely to succeed" (p. 142) academically than students who were deficient in those behaviors.

Teacher ratings of behavior and academic achievement

The second source of empirically based information regarding student cognitive behavior is teacher ratings of their students. Teacher ratings or judgments are commonly used for research purposes, either as informal reports or through the use of structured rating instruments (Hoge, 1983). Of primary relevance for the present study is research relating teacher ratings to the academic achievement of students in the early grades. Teacher ratings of students can provide a fuller picture of cognitive behaviors that students actually need and use. An important factor that contributes to the richness of this source of information is that teacher ratings reflect students’ behavior over longer time periods (Gresham, 1982). Teacher ratings also reflect information gained from discussions with students and evaluations of their work. First, several studies using informal reports or non-standardized rating tools will be reviewed, followed by reviews of studies using structured rating tools.

An early study using, among others, teacher ratings of kindergarten classroom behavior with academic achievement in 5th grade was reported by Meyers, Attwell & Orpet (1968). This study utilized a collection of existing measurements, among them kindergarten teachers' global ratings of students' attentive behaviors. The correlations of the ratings of attentive behaviors with 5th grade standardized achievement test scores ranged from $r = 0.26$ to $r = 0.43$, with a median correlation of $r = 0.40$. The ratings of attention had a stronger correlation with achievement than both readiness tests and ratings of other student characteristics.

Several studies investigating the potential use of teachers' global ratings to provide additional information to standardized tests were identified. Bolig and Fletcher (1973) investigated teacher ratings of 625 kindergarten students as predictors of 1st grade reading achievement. Their rating procedure required several global ratings of student ability or cognitive behaviors: verbal concepts, visual perceptions, listening, alphabet, numbers and copying. They reported a significant correlation between the teacher ratings and standardized achievement test total scores, $R^2 = 0.61$. For this same purpose, Sharpley and Edgar (1986) obtained teacher ratings of students' present levels of achievement in reading vocabulary, reading comprehension, and math; achievement in the same content areas were assessed by a standardized test. Results were reported separately for 120 boys and 110 girls, with the correlations of ratings with achievement scores in the same subject matter ranging from $r = 0.38$ to $r = 0.56$ (median of six correlations, $r = 0.445$), and $r = 0.35$ to $r = 0.54$ with achievement in different subject matter (median (of 12), $r = 0.465$). Quay and Steele (1998) obtained global ratings of five developmental and behavioral areas: academic, communication, physical, self-help, and social. They reported significant correlations of kindergarten ratings with 2nd grade standardized tests in reading ($r = 0.61$) and math ($r = 0.52$) for a sample of 192 students. Furthermore, they factor analyzed the ratings with scores from a standardized readiness test, measuring the same five dimensions. The factor analysis failed to support that ratings and direct measures of the same dimensions were closely related. All direct measures and all ratings loaded on separate factors. These results indicate that teacher global ratings provide global information, but fail to provide information regarding specific subject matter.

Feshbach, Adelman and Fuller (1974) attempted to identify specific behaviors of kindergarten students who were later identified as having difficulty in reading in the primary grades. Teachers in 10 schools rated 888 kindergartners on the *Student Rating Scale* (SRS) (Adelman & Feshbach, 1971; Feshbach, Adelman & Fuller, 1974). Standardized reading test scores were available for these students. The SRS is an instrument intended to assess skills and behaviors required for successful early reading, and also behaviors interfering with reading achievement (Feshbach, Adelman & Fuller, 1974). The SRS was constructed to have five factors: classroom management: follow simple directions, work independently, listen attentively, maintain seatwork, answer questions about a story, and tell stories from pictures; verbal: speak in complete sentences, speak clearly, have adequate vocabulary, and express ideas, thoughts and feelings; letter/sound discrimination; and recall of recent information; and finally perceptual-motor skills exemplified by pencil use (Adelman & Feshbach, 1971; Feshbach, Adelman & Fuller, 1974). All five factors of the SRS correlated significantly with the reading scores, with correlations ranging from $r = 0.41$ to $r = 0.55$.

Glazzard (1980) compared the use of a reading readiness test and kindergarten teacher ratings as predictors of later achievement. The original kindergarten sample consisted of 107 students experiencing reading difficulties. Sixty two students remained in the four schools in 4^{th} grade. A modified *Teacher Estimate of Kindergarten Readiness Skills* (TEKRS), originally developed by Kirk (Glazzard, 1980), was used to obtain ratings. TEKRS requires global ratings on six dimensions, four of which assess cognitive behaviors. Several examples of cognitive behaviors, skills or abilities are provided to define the dimensions: verbal comprehension; listen to stories, understand directions, and vocabulary (comprehension); verbal expression: expressive vocabulary, use of sentences, expression of ideas and needs; numbers and space relations: count objects, use quantitative concepts such as more and less, puzzle; reasoning: use planning in problems solving, use logical order in communication, identify similarities and differences in common objects, and categorize objects (Glazzard, 1980; Selders, 1981). A standardized test was used to measure reading achievement. All four ratings based on cognitive behaviors entered regression equations with 1st and 2nd grade reading comprehension and vocabulary scores as dependent variables. In addition, verbal comprehension and numbers and space relations entered the regression equations with 3rd and 4th grade reading achievement as the dependent variable. The squared multiple correlations ranged from $R^2 = 0.73$ and $R^2 = 0.84$, with the median (of eight) $R^2 = 0.765$. Tollefson, Rodriques and Glazzard (1985) obtained similar results with TEKRS kindergarten ratings on a sample of 273 regular students. They found significant squared multiple correlations between 2nd grade TEKRIS ratings and standardized achievement test reading scores that ranged from $R^2 = 0.46$ to $R^2 = 0.57$ (median $R^2 = 0.53$).

Hightower et al. (1986) investigated the potential use of teacher ratings to identify young learners with school adjustment problems. A rating tool constructed for this purpose, *Teacher-Child Rating Scale* (T-CRS) assesses six factors, two of which appear to assess lack of attentiveness and work habits: learning and task orientation. Both learning and task-orientation correlated significantly with a standardized achievement test reading and math scores. Correlations of learning with achievement ranged from $r = -0.42$ to $r = -0.55$, and $r = 0.27$ to $r = 0.54$ for task-orientation. Correlations of the T-CRS with two other teacher rating tools, *Classroom Adjustment Rating Scale* and *Health Resources Inventory*, were also reported as evidence of the validity of T-CRS as a measure of school adjustment problems.

Alexander, Entwisle and Cauber (1993) investigated the long term relation between kindergarten behavior and primary grade achievement. They constructed a teacher rating tool from teacher descriptions of student behavior. They found that teacher ratings clustered in three areas, interest-participation represented work habits and interest in school; cooperation-compliance were social and affective behaviors; and attention-restlessness were attentive and affective behaviors. They found significant correlations between interest-participation and attention-restlessness and both standardized achievement test scores and teacher grades for 1^{st}, 2^{nd} and 3^{rd} grade. They did not report correlations for individual variables.

The studies reviewed above support that a relationship exists among certain cognitive behaviors and achievement measures. All the studies suffer from a lack of information regarding the validity of the measures of cognitive behaviors. Those instruments were as a rule constructed by the author for the purpose of each individual study, and limited, if any information regarding their validity is available. This renders them methodologically weaker than studies using structured rating tools to measure cognitive behavior.

The most prominent concepts measured by the existing teacher rating tools are psychological in nature, such as learning difficulties, conduct, social and emotional problems (Hoge, 1983, Spivack and Swift, 1973). Few rating instruments are primarily devoted to educational purposes, such as the assessment of cognitive behaviors. The majority have a broad psychological focus, with some of them assessing a few constructs related to the cognitive behaviors. In their early review of teacher rating instruments, Spivack and Swift (1973) noted that both the measured concepts and the criterion variables used in validity studies indicated that "measurement of overt behavior in the classroom has largely been motivated by mental health interest" (p. 84). They added that educational or instructional purposes had been served only to a limited extent. Similar conclusions can be inferred after reading more recent reviews (Bracken, Keith & Walker, 1998; Hoge, 1983; Naglieri & Flannagan, 1993). Reviewing the entries for the instruments listed under 'Behavior Assessment' in the Classified Subject Index of the two most recent editions of the *Mental Measurement Yearbook* (Conoley & Impara, 1995; Impara & Plake, 1998) does not indicate that this situation has changed. Studies using the few existing teacher rating tools assessing constructs related to cognitive behavior and using academic achievement measures as dependent variable will now be reviewed.

Swift and Spivack (1969) investigated the relationship between teacher ratings of classroom behavior and academic achievement of 106 5th grade students, using the *Devereux Elementary School Behavior Rating Scal*e (DESBRS) to assess classroom behavior. Five of the DESBRS dimensions were judged to assess cognitive behavior. External-reliance assesses the student's reliance on structure provided by the classroom situation; comprehension assesses the student's comprehension of daily instruction and directions; inattentive-withdrawn assesses deficient attention; irrelevant responses and creative-initiative, respectively, assess the relevance of the student's contributions to classroom discussions and activities in general (Spivack & Swift, 1967). Three achievement groups were identified based on two achievement measures, standardized test scores and school grades: overachievers were identified as students with low test scores and high school grades; underachievers had high test scores and low school grades, and consistent achievers scored similarly on both measures. The profiles of the mean DESBRS factor scores for the three achievement groups were compared, and differences were found in the behavior dimensions. Compared with consistent and overachievers, the underachievers were less attentive, lacked comprehension of day to day instruction, lacked independence, and rarely made relevant contributions to discussion or activities (Swift & Spivack, 1969). Several methodological weaknesses of this study were noted. There were questionable classifications of the students into achievement groups and potential subjectivity in the analysis of the profiles.

Wallbrown, Wallbrown, & Engin (1977) investigated the relationship between teacher ratings on the DESBRS obtained in kindergarten and standardized achievement test scores at the end of first grade for 299 students in 15 classes. They used a multiple regression analysis to identify the best behavioral predictors for the four reading and math measures. The only DESBRS dimension that consistently entered the regression equations was comprehension. Other dimensions that entered the regression equations of individual achievement measures were irrelevant-responses and two DESBRS dimensions assessing affective behaviors. In addition, two of the DESBRS single item measures reflecting deficient work habits were among significant predictors of three of the achievement measures. This study was methodologically stronger than the Swift and Spivack (1969) study discussed above. But the intercorrelations of the DESBRS dimensions were substantial, and only one of several highly correlated dimensions entered the regression equations while the relation between the other variables and achievement was obscured (Reynolds & Bernstein, 1982; Schaefer, Baker, & Zawel, 1975).

Schaefer (1975) investigated the relation between teacher ratings of student behaviors and academic achievement. Teacher ratings of the six scales the *Child Behavior Inventory* (CBI) for 72 primary grade students were related to their 1st and 3rd grade academic achievement assessed by a standardized achievement test. The CBI has been commonly used for research purposes, even though information regarding its validity is very limited. Two of the scales appear to assess constructs related to cognitive behaviors: task-orientation and distractibility. The correlation of task-orientation with the 1st grade test composite was $r = 0.56$ and with 3rd grade composite, $r = 0.64$; corresponding correlations for distractibility were $r = -0.50$ and $r = -0.56$, respectively. Kohn and Roshman (1974) investigated the relation between several ability, affective and behavioral measures and later achievement. They obtained, among other measures, CBI ratings for 209 kindergarten students. The correlation between the CBI Task-Orientation scale and three 2nd grade achievement test scores ranged from *r* = -0.33 to *r* = -0.44. The negative correlations are explained by the scoring procedure; the scoring of task orientation was reversed to render it consistent with other teacher rating tools used. After intelligence test scores and experimental cognitive tasks had been controlled for, its partial correlation with reading was *r* = -0.22. McKinney (1989) reported that several studies using the CBI along with classroom observations found that the cognitive behavior of regular and learning disabled students differed, and described the regular students as more attentive and task oriented.

Lindholm, Touliatos, and Rich (1976) investigated the relationship between teacher rating of classroom behavior and achievement. The behavior ratings for 425 3rd grade students were obtained using the *Behavior Problem Checklist* (BPC) and the achievement measure was a standardized achievement test. Canonical correlations indicated that the BPC inadequacy-immaturity scale was the only one related with the achievement measures. This scale consists primarily of behaviors that are related to attention.

Two studies were identified that used the *Pupil Rating Scale* (PRS) to investigate behavior-achievement relationships. In a study investigating the properties of this instrument, Bryan & McGrady (1972) identified 42 students with low PRS Auditory comprehension scores from 183 students referred for special evaluation as potentially having learning difficulties. The reading achievement of these 42 students was compared with the achievement of 42 students randomly selected from the remaining referred students. The low auditory comprehension students scored significantly lower on both standardized reading comprehension and word knowledge tests. In an investigation of early literacy skills, Collingen (1979) obtained teacher ratings of the PRS Auditory comprehension and spoken language scales for a 1st grade sample. Their correlations with 2nd grade standardized achievement test reading and math scores ranged from *r* = 0.53 to *r* = 0.72.

Three relevant studies using Conners *Teacher Rating Scale* (TRS) were identified (Camp and Zimet, 1974; Day and Peters, 1989; Lam and Beale,1991). TRS is a 28 item rating tool assessing three constructs: inattention, hyperactivity, and conduct problems (Goyette, Conners, Ulrich, 1978). Camp and Zimet (1974) obtained TRS ratings as well as several achievement test scores for 47 1st grade students. The inattentive scale of the TRS correlated significantly with four achievement measures, $r = -0.59$ to $r = -0.76$. Day and Peters (1989) investigated cognitive and behavioral differences of 22 underachieving and 10 normally achieving 3rd and 4th grade students. Underachievers were identified as students with IQ in the normal range, but more than a year behind students of the same age in language and math. The underachievers were rated as having significantly higher inattentive-passive TRS scores (14.2) than the normally achieving students (5.5). The groups did not differ significantly on intelligence test scores. In this study, cognitive behaviors were not directly related with achievement, but differences in important cognitive behaviors were found between groups that differed in achievement measures but not on ability measures. Lam and Beale (1991) used the TRS in a study of attention and achievement with 190 seven to ten year old Australian students. They found the inattentive scale to correlate with comprehension and vocabulary on a standardized achievement test, $r = 0.27$ and $r = 0.42$. These three studies support the relation between teacher ratings of attentiveness and academic achievement.

The studies using both informal and structured teacher ratings that were discussed in this section supported the relationship of oral communication, work habits and attentive behaviors with achievement. In addition, quantitative and written language behaviors were related with achievement in a few studies.

Direct measures of cognitive behavior and academic achievement

The third type of empirical methods providing information regarding the relation between cognitive behavior and academic achievement involves studies using direct measures of student behaviors or skills. The selected studies investigated the relation between measures of children's cognitive behavior or skills at the time they enter school and their academic achievement after one or several years.

Bond and Dykstra (1997/1967) summarized an extensive research program examining instruction and assessment of reading in the primary grades. They concluded that familiarity with print and auditory and visual discrimination are the most important behaviors or skills relevant for successful reading achievement in the primary grades. The most important kindergarten behaviors were found to be knowledge of letter names. The most prominent reading behaviors at the end of 1^{st} grade were found to be recognition of words and associating meaning to these words. Several studies have found similar results regarding the relation between written language behaviors and achievement. Lowell (1971) studied 161 1^{st} grade students, and found that correlations between measures of letter knowledge and beginning and end of year reading achievement were $r = 0.65$ and $r = 0.63$, respectively. Correlations between other behaviors, such as discrimination of phonemes, letters and words, and achievement ranged from $r = 0.34$ to $r = 0.51$. Telgredy (1975) investigated the correlation between several reading readiness tests and 1^{st} grade reading achievement for a sample of 56 students. He found measures of knowledge of letters, copying figures, and knowledge of numbers to be the most consistent predictors of reading at the end of 1^{st} grade. Busch (1980) measured early reading behaviors of 1,052 students as they entered school and reading achievement at the end of 1st grade. He found naming of upper and lower case letters, and identifying beginning sounds of words to be the best predictors of end of 1st grade reading achievement.

Badian (1982, 1988) investigated the long term predictive value of several measures of cognitive behavior and skills taken at beginning of kindergarten. An original sample of 180 children was given the Holbrook Screening Battery before kindergarten, and reading achievement was assessed in 3rd and 8th grade using a standardized achievement test. The kindergarten measures included several tests assessing constructs related to cognitive behaviors: shapes: copying geometric forms; counting: rote counting and count ten objects; naming shapes; naming colors; naming letters; name writing; information: know address and birthday; and verbal expression; telling a story about a picture. Badian (1982) found significant correlations between several of the kindergarten measures and 1st, 2nd and 3rd grade reading scores: shapes, counting, naming shapes, naming colors, naming letters, name writing, information, verbal expression, and sentences. The 27 significant correlation coefficients ranged from $r = 0.31$ to $r = 0.61$ (median $r = 0.45$). In a follow-up study, Badian (1988) found that shapes, counting, naming colors, naming letters, information, and sentences entered a regression formula with the 8th grade reading score as the dependent variable, yielding a squared multiple correlation of 0.45.

Kelly and Pevelry (1992) also investigated the predictive validity of direct measures of kindergarten cognitive behavior and skills for later achievement. The *Kindergarten Screening Battery* was used to obtain 12 measures of kindergarten behaviors and skills. Among them were several behaviors and skills relevant for the present study: personal information: knowing birthday, address, days of week; naming letters; blending: letter-phoneme relationships; story retelling; telling story from picture; syntax; word copying; word matching, and copy geometric designs. Kelly and Pevelry (1992) used a total of seven 1st and 2nd grade reading and math scores as dependent variables. Each of the seven regression analyses yielded a significant squared multiple correlation. These ranged from $R^2 = 0.16$ to $R^2 = 0.51$ (median $R^2 = 0.36$). The most consistent results were that either naming letters, word copying, or word matching entered each of the seven regression equations; personal information entered six of the regression equations; and either story retelling or telling story from a picture entered five equations. In addition, copy geometric designs, blending, and syntax entered two or three equations each.

Jorm, Share MaClaren & Matthews (1986) investigated the relation between family background and early cognitive skills or behaviors and 2nd grade academic achievement. They measured, among other constructs, reading skills and cognitive behavior of 543 children as they entered kindergarten. Three years later teachers identified the children as normal readers, backwards readers or as reading retarded. Such kindergarten cognitive behaviors as naming letters, writing own name, recognizing own name, copying letters, naming colors, naming pictures, and memory for complex sentences were found to be less frequent among 2nd grade backwards readers than among the normal readers.

Butler, Marsh, Sheppard and Sheppard (1982, 1985) investigated the extent to which reading achievement could be predicted from students' early skills and behaviors, intelligence, and demographic information. Their study was conducted on an Australian sample of 320 students in 3rd grade and 286 in 6th grade. The Sheppard School Entry Screening Test (SSEST) was used to obtain kindergarten measures of cognitive behaviors. Three factors were identified among the SSEST subtests: figure drawing: copying geometric shapes, drawing geometric shapes from memory; language: distinguishing word pairs, repeating syllables, repeating a short paragraph, describing a picture, telling a story; perceptual-motor skills: hopping, throwing, catching, walking a beam. Significant correlations were found between SSEST language and composite reading achievement scores in 1st ($r = 0.38$), 2nd ($r = 0.44$), and 3rd grade ($r = 0.48$). The correlation of figure drawing with reading achievement was $r = 0.37$, $r = 0.40$ and $r = 0.45$, for the respective grades (Butler, et al., 1982). Different reading achievement tests adopted at the nationwide level in Australia were used as dependent measures for each grade. Following the same sample through 6th grade (Butler, et al., 1985) found a significant correlation between kindergarten SSEST language and 6th grade reading achievement, *r* = 0.48. Additional conclusions from path analyses indicated that the kindergarten measures primarily influenced the 1st grade achievement scores, but the 1st grade achievement scores had strong influence on the later achievement scores.

Daly, Wright, Kelly and Marhtens (1997) reported the findings from a study investigating the relation between early reading behaviors and later reading achievement. Among the skills they measured for the 30 1st grade students were reading of letters, words, and numbers, and counting. They found significant correlations between letter and word reading and two curriculum based reading achievement measures ranging from *r* = 0.69 to *r* = 0.73, and from *r* = 0.43 to *r* = 0.59 for two math achievement measures. Number reading and counting had significant correlations with the reading scores, *r* = 0.44 to *r* = 0.48, but not with the math scores. Grogan (1995) investigated the correlations between reading behaviors at age four and reading achievement at age seven for 51 English students. Skills that correlated with achievement were knowledge of letter-sound correspondence for upper case, *r* = 0.49, and lower case letters, *r* = 0.40, and copying own name, *r* = 0.32.

Studies relating direct measures of student cognitive behavior and skills provide additional support for the relation between cognitive behavior and academic achievement measures. Of particular importance is that the direct measures of skills similar to written language behaviors were repeatedly found to relate with academic achievement, and several studies provided support for the relation between general knowledge and achievement. Neither the classroom observations nor teacher ratings related these cognitive behavior areas with achievement. The studies reviewed also found additional support for the relation between skills and behaviors from the quantitative and oral communication cognitive behavior areas and academic achievement. These studies using direct measures represent a selection of studies assessing constructs related to cognitive behaviors. In a more general review of research on the prediction of academic achievement in the primary grades, Tramonta, Hooper, and Selzer (1988) also found kindergarten and preschool language related skills or behaviors to be consistently related with academic achievement.

Summary of research on classroom behavior and academic achievement

The review of empirical research regarding the relations of cognitive behavior with academic achievement was divided into three sections based on methodological considerations. Studies using classroom observations, teacher ratings, and direct measures were identified. Each of the empirical methods provided valuable information regarding some of the cognitive behavior areas, while remaining mostly silent on others. Empirical evidence of the relation between attentive behaviors and work habits and achievement was found in studies using both classroom observations and teacher ratings. Similarly, evidence of the relation between quantitative and oral communication behaviors and achievement was found in studies using both teacher ratings and direct measures, and the support for the relation between general knowledge and written language behaviors with achievement came from a number of studies using direct measures. Some findings fell outside this pattern, but these represent the most consistent findings.

Cognitive behavior and three theories of learning

A fundamental assumption underlying the IELI is that certain student behaviors relate to academic achievement, and that these behaviors enhance learning. Empirical research supporting that cognitive behaviors actually related to achievement has already been reviewed. Cognitive behaviors' emphasized in curriculum guides were also described. Current validity theory requires a theoretical rationale to support the relation between cognitive behavior and achievement. The next few paragraphs outline three major theories of learning, and attempt to place the relation between cognitive behaviors and achievement within the framework of each of them.

Behavioral learning theory was very influential in educational psychology and instructional practice during the 20th century. Its views of learning are still very influential and the fundamental learning mechanisms set out by this theory are commonly accepted, despite supported claims that behavioral learning theory presents a limited view of human learning (Greeno, et al. 1996). Behavioral learning theory defines learning as a change in the probability that a certain behavior will be repeated, and it relates this change in probability of behavior to external or environmental variables by several mechanisms of learning. One is classical conditioning. When an event occurs at the same or similar time as a stimulus eliciting a behavior, the organism associates the event and the behavior. At a later time the event may elicit the behavior (Gleitman, 1992). Another fundamental learning mechanism is operant conditioning; which occurs when the consequences of a behavior influence the probability that the behavior will be repeated. The frequency of a behavior increases when it is followed by desirable consequences, but its frequency decreases when it has undesirable consequences. A powerful variant of the operant mechanism occurs when fixed sequences of behaviors are offset by a stimulus, and an entire sequence of behaviors culminates in a reward. These sequences can be shaped into well practiced sets of behaviors as slight variations that are differentially reinforced (Ferster, et al., 1972; Gleitman, 1992). Still other mechanisms are learning from the example of others, or modeling observed behavior (Eggen & Kauchak, 1999; Gleitman, 1992).

How can the learning mechanisms outlined by behavioral learning theory provide information about cognitive behaviors? One of the fundamental principles of behavior theory is the role of practice and repeatedly trying existing behaviors in new and slightly different situations. Such behaviors are clearly seen in any classroom.

One of the criticisms of behavioral learning theory is its sole emphasis on external variables as motivating or driving forces behind learning. According to cognitive learning theory, learning takes place as students' cognitive or mental structures are altered, reorganized or new structures created. Models of how information is processed are central to the cognitive theories (Anderson, 1990; Gleitman, 1992). These models commonly consist of three main units and processes that link them. The main units are: sensory memory as a gateway, working memory as an active processor, and long term memory as a stable and permanent storage system. Attention and perception are processes that link the sensory memory with working memory. They identify limited parts of the sensory input as relevant. Cognitive processes retain, change or restructure information that is held in working memory. Encoding and rehearsal processes store information in long term memory, while retrieval processes can recall that information. In addition, metacognitive processes are believed to control or monitor these information processes (Anderson, 1990; Biehler & Snowman, 1993; Eggen & Kauchak, 1999; Gleitman, 1992). Long term memory is believed to consist of declarative knowledge, units of information and connections among them. Structured blocks of connected knowledge, declarative knowledge, can be retrieved into working memory to restructure or change information (Anderson, 1983, 1990).

How can the learning model presented by cognitive learning theory provide information about cognitive behaviors? The entire information processing model is spelled out in terms of invisible mental behaviors. But behaviors that help the learner to focus on relevant sensory input are cognitive behaviors; for example, attending to the teacher during lecture. Behaviors that aid rehearsal and encoding of newly learned information are cognitive behaviors;

for example practicing letter sound relationships, or writing letters. The metacognitive process provides theoretical rationale for considering work habits as cognitive behaviors. Asking relevant questions, classifying according to a rule, sustaining work on a difficult task, and resisting distractions should be regarded as cognitive behavior according to this view.

The information processing framework models memory, knowledge acquisition, problem solving and planning as rational processes occurring more or less in isolation. Sensory input activates the cognitive system, and there is a reaction after processing is completed. Situated cognition models the learner as actively seeking out information and creating her/his own knowledge through her/his actions and interactions with the environment (Greeno, et al. 1996).

One view within the situated cognition field emphasizes processes acting to resolve conflicts between the learners' presently accepted knowledge and new information or experiences as a central or driving force in learning. When new experiences are inconsistent with presently held knowledge, the learner actively seeks out more and more information to either confirm the previously held knowledge, the new information, or to construct a new knowledge that takes both into account (Biehler & Snowman, 1993; Eggen & Kauchak, 1999; Greeno, et al., 1996). Learning from others is central to an alternative view within the situated cognition framework. According to this view, language and other fundamental cognitive concepts are social constructions, and learning from others is the primary means of the young learner to acquire this knowledge (Biehler & Snowman, 1993; Eggen & Kauchak, 1999).

How does the situated cognition theory incorporate cognitive behaviors? Independence and disciplined work habits are essential tools that would characterize the active learner portrayed by some proponents of the situated cognition view. Efficient communication skills would be valuable according to others. Knowledge that is presently held and believed by the learner also plays an important role in learning according to the situated cognition view. Cognitive behaviors resting on such knowledge that are important for the primary grade student are language, knowing and using names of common colors, values of coins, recognizing the relative size of common objects, fundamental geometric shapes, and numbers and magnitude.

These three basic learning theories provide theoretical rationale, each within in its own framework, that certain behaviors are cognitive behaviors. Each provides rationale regarding some cognitive behaviors, while being silent about others. The present study is informed by all three theoretical frameworks. Its purpose is to investigate the value of a measurement tool based on cognitive behavior for practical use in the classroom, and the theoretical underpinnings provided in this section are purposefully laid out broadly.

Summary of empirical and theoretical views on cognitive behaviors

Six areas of cognitive behavior were identified in a sample of curriculum guides, which both reflect and influence the classroom activities that children experience in schools. Empirical literature representing diverse methodologies provided support for the relation between these areas of cognitive behavior and academic achievement. Classroom observations supported the relation between work habits and attentive behaviors and academic achievement. Teacher ratings supported the relation between oral communications and quantitative behaviors and achievement in addition to work habits and attentive behaviors. And studies using direct measures of students' skills and behaviors supported the relation between general knowledge, oral communication, written language, and quantitative behaviors and academic achievement. Finally, the assumption that certain cognitive behaviors enhance student learning was shown to be consistent with each of the three current frameworks within the psychology of learning. This is the empirical and theoretical basis on which the interpretation of the scores from teacher ratings of cognitive behaviors can be based.

Measurement issues relevant to rating scales and inventories

Validity issues

Validity, the appropriateness of inferences drawn from test scores, is the most important score property that must be considered in the use of educational measurement tools (Messick, 1989). Validity issues must be addressed beginning with the test construction process. Many of the important properties that support and justify the use of assessment results for a particular purpose must be built into the measurement tool. This overview of validity will address several main aspects of validity and relate them to issues that need to be considered during the construction of a teacher behavior inventory such as the Iowa Early Learning Inventory (IELI).

Validity represents the accuracy of inferences about test takers drawn from test scores (Messick, 1989). Validity is defined in the *Standards for Educational and Psychological Testing* as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association/ American Psychological Association/ National Council on Measurement in Education (AERA/APA/NCME), 1999, p. 9). The conceptualization of validity in the Standards is consistent with a trend in theoretical discussions regarding the focus on the theoretical construct presumed to capture the nature of the attribute being measured at the centre of the psychometric concept of validity (AERA/APA/NCME, 1999; Messick, 1989; Moss, 1992; Shepard, 1994). The fundamental idea of this rationale is that an interpretation of test scores involves inferences from the theoretical construct of the attribute measured to specific attributes or characteristics of the student. For such inferences to be valid, differences in the student attributes must be reflected in the test scores in a manner consistent with the underlying theoretical conceptualization of the attribute (Messick, 1989). According to this view, known as the unitary theory of validity, aspects of validity that were previously considered as distinguishable types of validity are subsumed under the central concept of construct validity (Crocker & Algina, 1986; Messick, 1989). The traditional distinguishable validity frameworks were content, criterion, and construct related validity. The traditional content validity referred to how adequately the items on a measurement tool represented the ability, knowledge, skill or behaviors that the test had been constructed to measure. Content validity was viewed as sampling from a domain of interest, where the domain was conceptualized either as instructional content, a domain of behaviors, or cognitive processes (Messick, 1989; Sireci, 1998). Criterion related validity reflected the empirical relation between the resulting scores and a real life independent measure of the same or related construct. The test score was seen as a replacement for a measure that was more difficult to obtain or as a useful prediction of a measure available only at a later time (Anastasi & Urbana, 1997; Crocker & Algina, 1986). The original concept of construct validity in the traditional three faceted view of validity introduced a theoretical construct underlying or explaining the measured attribute or characteristic. This concept of construct validity placed the test score in a network of theoretical constructs connected by logical relationships. Empirical data were then required to show the relations of the measurement within this network (Crocker & Algina, 1986; Messick, 1989). Increasingly, the theoretical construct has become central to all validity issues, and in the unitarian view of validity, it is the cornerstone on which interpretations of the resulting scores are built (Messick, 1989).

Validity is presently viewed as a judgment of how appropriate the inferences drawn from the test scores are (Messick, 1989). The unitarian view is well captured in a definition of validity given by Messick (1989) as "...an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on the test scores or other modes of assessment" (p. 13). In this view, the theoretical construct underlying the attribute or characteristic that is measured plays the central role in all arguments supporting the validity of tests, and the importance of any validity considerations hinges on the potential theoretical support for the chain of inferences that are involved in score interpretation. The unitarian view sees the measurement tool as one of many operationalizations of the theoretical construct. The resulting scores then borrow their meaning from this construct, and any interpretation of them builds on relationships with other constructs and measures derived from the construct theory. Validating the interpretation then becomes a task of supporting or refuting these hypotheses regarding the relation between the scores and other measures or constructs (Messick, 1989). Empirical evidence plays a major role in displaying whether the assumed relationships spelled out by the theoretical construct are systematically present or absent according to the theoretical expectations and rationale. Student variability in the theoretical construct must be reflected in the actual test scores if such inferences are to be valid (Anastasi & Urbana, 1997; Crocker & Algina, 1986; Messick, 1989).

The interpretations of test scores always involve inferences from the theoretical construct to the test takers' potential performance in a variety of situations different from the testing situation. The recognition of the omnipresence of the theoretical construct for all aspects of a measurement tool has been a driving force towards a conceptualization of the traditional components of validity as aspects of a unitary validity, often discussed as construct validity (Moss, 1992; Shepard, 1993). Such theoretical underpinnings are required to provide a rationale for both the relevance of content considerations as well as for the relevance of the criterion related evidence. The construct theory actually determines what counts as relevant content. It also provides rationale for expecting certain relationships with potential criteria to actually exist (Messick, 1989).

Content validity is subsumed under construct validity on the grounds that the construct theory determines what domain a particular measurement tool should represent, and thus what behaviors or items are acceptable as representations of the larger domain. A faithful content representation is an essential link to support the chain of inferences drawn from the resulting scores to the construct theory and then to other situations of interest. The primary role of content related evidence is the determination of the extent to which items in the test score accurately represent the skill or behavior that the intended interpretive inferences involve (Messick, 1989; Sireci, 1998).

Criterion related validity is also subsumed under construct validity. The theoretical construct defines what meaningful relations between the test scores and other constructs and measures should exist. This theoretical framework provides a rational basis to identify potential relationships of the test scores with other constructs and variables. The primary purpose of criterion related evidence is to establish or refute whether these theoretically meaningful relationships actually exist. The criterion variables illuminate the meaning of the test scores to the extent that theoretically meaningful relationships are supported. (Anastasi & Urbana, 1997; Messick, 1989).

A number of issues relevant for the development of the IELI become clear from the above discussion. The prominent theoretical concepts underlying the IELI are theories of learning and achievement. These theories must be considered to determine what cognitive behaviors are important, and to frame empirical investigations of teacher ratings on the inventory to later achievement. Such theories and their implicit values reflected in early grade curricula and instructional materials are also relevant, and provide an operationalization of the broad domain of cognitive behaviors that is relevant and can potentially be included on a rating inventory. It is also important to consider whether empirical evidence supports a relation between these behaviors and achievement.

A second issue of importance that emerges from the discussion above is the question of whether the structure of the IELI as an instrument consisting of six distinguishable areas of cognitive behavior is reflected in empirical data. Do students actually vary over these six areas? Can teachers reliably distinguish among these six areas in their ratings? This issue is at the heart of construct validity; how well does the measurement tool reflect the structure of the construct? But it is complicated by the fact that the construct theory and the measurement process are confounded in the scores. If students' cognitive behaviors can be subdivided into six independent areas, then the teacher ratings might reasonably be able to distinguish these behaviors. These issues can be addressed jointly with a factor analysis of a sample of IELI scores.

Another issue where empirical support is of primary importance is the relation between IELI scores and measures of related constructs, such as academic achievement. Of primary importance in this respect is their relation to the *Iowa Tests of Basic Skills* (Hoover, et. al. 1993). This is a complex issue where a multiplicity of relations is relevant to support inferences regarding students' academic standing from IELI scores. To begin with, it is of interest to assess their relation to direct measures of achievement obtained both concurrently and at a later time point. The relation between IELI scores in kindergarten or first grade and the probability that a student receives remedial education or other support may also be important. However, all those issues can not be addressed in the present study.

Collection of data by rating scales

Rating scales and inventories are measurement tools that allow a knowledgeable person to evaluate the quality, magnitude or frequency of some behavior or characteristic of another person or an object. These instruments consist of collections of descriptive statements and a response scale. The statements describe attributes, behaviors or symptoms that define the construct that is measured. Various dimensions or aspects of the construct can be measured by anchoring the scale points with numerical, behavioral or affective descriptors. The rater is required to indicate the congruence of each statement with her/his knowledge and perception of the ratee with the anchors defining the response scale (Gronlund, 1993; Saal, Downey, & Lahey, 1980). Rating scales and inventories differ from other formal measurement tools in important aspects regarding how they collect information. Classroom exams, class projects, and standardized tests of educational achievement are designed to directly measure student skills and knowledge in the relevant curriculum domains. Rating instruments are indirect measures in the sense that they rely upon teacher recollection or judgments of student behavior, and they target information that teachers obtain about their students during everyday work and interaction with students in the classroom. They may therefore provide valuable complementary information to that obtained from direct measures of student abilities (Salvia & Ysseldyke, 1998; Thorndike, et al, 1991).

Rating instruments are prone to a variety of methodological problems. They are subject to biases on behalf of the rater, ambiguity regarding the behavior or characteristic to be rated, raters' interpretation of the meaning of the scale points, and involvement of judgment or evaluation in the rating process (Anastasi & Urbana, 1997; Landy & Farr, 1980; Saal, et al., 1980; Thorndike, et al, 1991). The effects of each of these potential problems can be minimized if they are addressed during the construction of the measurement tool, by carefully selecting observable behaviors, clearly defining the response scales and carefully wording the instrument.

The subjectivity involved in the rating process can lead to a bias in several different ways. Certain biasing effects have been recognized and defined: halo effect, central tendency, and leniency or severity effects (Landy & Farr, 1980; Saal, et al., 1980; Thorndike et al, 1991). Explanations of these biases emphasize the raters' reliance on a general impression during the rating process, the raters' inability to discriminate between the behaviors or characteristics rated, or the raters' avoidance or inability to discriminate between ratees (Anastasi & Urbana, 1997; Saal et al., 1980). These biasing effects can best be challenged by minimizing the subjective element involved by using observable and concrete behavior statements, clear descriptions of rating points, and by providing raters with appropriate directions or training (Landy & Farr, 1980). Such precautions minimize the judgment and interpretation involved in the rating task.

Clear descriptions or definitions of the behaviors described by each statement are also essential to minimize ambiguity about the concept being measured. The statements describe a behavior that represents a class of behaviors which in turn defines the construct measured. Therefore, it is preferable to state relatively concrete, observable behaviors, in a manner that requires only limited judgment or inference regarding whether an observed behavior belongs to the defined class of behaviors (Anastasi & Urbana, 1997; Gronlund, 1993; Thorndike, et al, 1991).

Two aspects of the response scales need to be considered, the meaning and the number of the scale points. The scale points can be given a number of different meanings by anchoring them with descriptors defining the frequency of a behavior, quality of response, duration or magnitude of behavior, severity of a condition or attribute, or affective reaction to behavior. The anchors determine what aspects of a construct are measured. The quality of the ratings has been shown to increase as the meaning of scale points is made clearer. Clear definition of the anchors is essential to establish a clear meaning and a common frame of reference for all the raters (Anastasi & Urbana, 1997; Landy & Farr, 1980). The number of scale points must also be considered. The response scales used in a rating tool represent a continuous construct or attribute by a limited number of discrete points. Investigation on the psychometric properties of response scales having different numbers of points indicates that four to seven scale points suffice to measure a construct reliably for most situations (Anastasi & Urbana, 1997; Landy & Farr, 1980).

Teacher rating scales assessing achievement related student behavior

A limited number of teacher rating scales have been developed to assess academic achievement or related constructs. A few of these measurement tools have been developed to assess general achievement constructs, while others have focused on limited academic areas such as reading.

The *Classroom Behavior Rating Scale* (CBRS) is a teacher rating tool intended to measure classroom behaviors that relate to learning (Reynolds, 1978, 1979). The CBRS was developed to make the information teachers obtain as part of their daily interaction with students easily available to school psychologists. The CBRS was developed on the basis of teachers' descriptions of classroom behaviors. No theoretical rationale supporting the relevance of these behaviors is offered. It consists of 40 items describing classroom behaviors, and a six point response scale anchored with frequency descriptors. The validity data reported by Reynolds (1979) are limited. The relation between CBRS and achievement measures indicated that the correlations between the CBRS and standardized reading and math test scores ranged from $r = 0.65$ to $r = 0.87$ for a sample of 1st through 5th grade students. A factor analysis investigation showed that all items of the CBRS loaded on a single factor. The internal consistency reliability of the CBRS was reported as $r_{xx'} = 0.98$ (Reynolds, 1979). Additional normative data were not provided for this instrument.

The *Children's Behavior Rating Scale* (CBR) was developed by Neeper and Lahey (1984, 1986) to provide assessment of cognitive deficiency among school children in addition to the emotional and behavioral dimensions prominent in the teacher rating scales existing at the time of its development. The CBR consists of 102 items assessing seven dimensions or subscales. Two of the dimensions assess cognitive behaviors. These are labeled inattention-disorganization and linguistic-information processing deficits. The CBR consists of relatively concrete and observable cognitive behaviors and a five-point response scale. Neeper and Lahey (1985) provided factor analyses and mean differences between selected and unselected samples of students as validity evidence for the CBR scores. All the CBR items had high loadings on their respective factors in a factor analysis. Item loadings on other factors were not reported. In addition, Neeper and Lahey (1986) reported that a sample of students placed in a special education program were rated significantly more deficient than regular students on all six dimensions or scales. Test-retest reliability with two week intervals ranged from $r_{xx'} = 0.84$ to $r_{xx'} = 0.97$, with the linguistic-information processing deficits scale being the most stable. Given that correlations between the two cognitive behavior factors were $r = 0.85$, the claim of separate dimensions is questionable. Limited normative information was provided for this scale. The normative sample consisted of 586 five to fifteen year old students of mixed ethnic background from three southeastern states. Despite the age range being eleven years, the normative data were not broken down by student age.

The *Study of Children's Learning Styles* (SCLS) instrument was developed by McDermont and Beitman (1984) to assess observable learning related behaviors, and was intended to be informative regarding remedial education and training of learning skills. The SCLS has three factors or subscales identified in a factor analysis: task aversion or lack of attentive behaviors, impulsive and inattentive behaviors, and the third is non-cooperative or overly independent behaviors. The SCLS consists of 16 items describing more or less concrete cognitive behaviors, such as "Accepts help when he cannot manage himself" or "Acts without taking time to look or think things out", but the authors refer to such behaviors as learning styles. The validity evidence reported by McDermont and Beitman (1984) consists of correlations of SCLS scores with six standardized achievement test scores, ranging from r = -0.33 to r = -0.57 for the total score (median r = -0.525), and r = -0.24 to r = -0.59 (median r = -0.52) for subscales. The internal consistency of the SCLS scales ranged from $r_{xx'}$ = 0.82 to $r_{xx'}$ = 0.93, and 2-month test-retest reliability ranged from $r_{xx'}$ = 0.80 to $r_{xx'}$ = 0.88. McDermont and Beitman (1984) relied upon a exploratory factor analytic study conducted as part of the construction of the SCLS as validity evidence. Normative information was not provided.

The *Academic Performance Rating Scale* (APRS) is a measurement tool that uses teachers' ratings to identify students with deficient academic skills (DuPaul, Rapport, & Perriello, 1991). Based on a factor analysis, three subscales were defined: academic success, academic productivity, and impulse control. The APRS consists of 19 items, requiring the teacher to rate the frequency of disciplined work habits and attentive behaviors on a five-point frequency-based response scale. The quality of students' work and academic success in several subject areas are also rated using qualitative anchors. The validity evidence reported by DuPaul, Rapport, and Perriello (1991) included a factor analysis showing three factors, and significant correlations with standardized achievement test scores ranging from r = 0.28 to r = 0.62. They also reported significant correlations of the APRS with classroom observations of student on-task behavior. Internal consistency reliabilities of the APRS total and subscale scores were reported to range from $r_{xx'}$ = 0.72 to $r_{xx'}$ = 0.95; 2-week test-retest reliabilities were $r_{xx'}$ = 0.88 to $r_{xx'}$ = 0.95. Normative information was not provided by DuPaul, et.al (1991).

The *Classroom Performance Profile* (CPP) is a teacher rating tool developed to assess the frequency of students' behaviors related with learning (Patterson, 1997). The CPP was constructed based on a collection of behaviors sampled from teacher report cards and from interviews with teachers. Its theoretical underpinnings are unclear. Five overlapping subscales are defined. An example of this overlap between subscales is the use of 13 common items in two of the identified subscales: creative, a 16 item scale, and ecological, a 14 item scale. The CPP consists of 29 items describing either specific classroom behavior or student ability and achievement, using a five-point frequency or norm referenced response scale. The validity related evidence reported by Patterson (1997) consists of significant correlations of CPP scores with standardized reading and writing test scores, $r = 0.40$ to $r = 0.62$. No analysis of the structure of the inventory was reported. Internal consistency coefficients range from ${}_{\alpha}r_{XX'} = 0.86$ to ${}_{\alpha}r_{XX'} = 0.97$, while one- and two-year test-retest reliability coefficients ranged from $r_{XX'} = 0.56$ to $r_{XX'} = 0.74$. Normative information was not provided for this instrument.

Two teacher rating measurement tools developed specifically to assess reading behaviors were identified. Kenny and Chekaluk (1993) developed a teacher rating scale intended to measure three dimensions relevant for early reading: academic performance, attention/behaviors, and cognitive/language ability. This rating tool consists of 15 items, and requires ratings of specific behaviors or abilities on a three-point response scale. Kenny and Chekaluk (1993) presented evidence to support its validity for research on reading, including significant correlations with 1st and 2nd grade standardized test scores, $r = 0.50$ and $r = 0.57$, and correlations with reading skills tests ranging from $r = 0.27$ to $r = 0.56$ in kindergarten through 2nd grade. Kenny and Chekaluk (1993) did not report any investigation regarding the internal structure of the questionnaire. Internal consistency reliability appears adequate, ${}_{\alpha}r_{XX'} = 0.92$. This is primarily a research instrument, and no normative information was given.

Flynn (Flynn & Rahbar, 1998a, 1998b) developed a teacher rating scale as part of a battery of measures intended for screening children for risk of reading failure. This instrument was intended to assess five domains identified as relevant according to theories of reading acquisitions: vocabulary, syntax, phonetic segmentation, alphabet and sounds. The teacher ratings reflect the same theoretical structure and consist of ten reading skills such as letter names, sight words, expressive vocabulary, and decoding. A ten point response scale with frequency based anchors at the endpoints is used. Investigations of the instrument's construct validity failed to support that the scores reflect the theoretical construct (Flynn & Rahbar, 1998a). The alphabet dimension was the only theorized dimension that emerged in a factor analysis where both teacher ratings and screening test subscales were analyzed. All other teacher ratings loaded on a separate factor that was evaluated as uninterpretable by the authors. The correlations of the TRS with a student percentile ranks in reading on a standardized achievement test were $r = 0.15$ for 1st and 3rd grade samples and $r = 0.14$ for a 2nd grade sample (Flynn, note 1).

Several instruments assess a few constructs related to cognitive behaviors. These instruments have a primary mental health focus and most commonly assess constructs related to attentive behaviors or deficient work habits.

The Devereux Elementary School Behavior Rating Scale (DESBRS) assesses children's classroom academic, conduct and emotional behaviors that interfere with their academic progress. It consists of 44 items that assess 11 concepts or dimensions that experienced teachers had judged to be relevant for students' academic achievement (Spivack & Swift, 1967). Five of the dimensions were judged to assess cognitive behaviors: external-reliance relates to work habits, comprehension and inattentive-withdrawn measures attentive behaviors, and irrelevant-responses and creative-initiative relate to oral communicative behaviors. Swift and Spivack (1968) reported correlations of the DESBRS with classroom grades that were supportive of the interpretation that the five dimensions assessing cognitive behaviors related to achievement. The range and medians of the significant correlations in 1st, 2nd, and 3rd grade with reading and math were: external reliance $r = -0.38$ to $r = -0.72$, median $r = -0.545$; inattentive-withdrawal $r = -0.32$ to $r = -0.74$, median $r = -0.505$; irrelevant responses $r = -0.20$ to $r = -0.5$, median $r = -0.39$; comprehension $r = 0.50$ to $r = 0.89$, median $r = 0.725$;

creative-initiative $r = 0.39$ to $r = 0.60$, median $r = 0.43$. The structure of the DESBRS as assessing 11 dimensions has not been supported empirically. Two factor analytical studies have shown the instrument to consist of three factors or dimensions (Reynolds & Bernstein, 1982; Schaefer, Baker, & Zawel, 1975). This teacher rating tool has been used in a number of empirical studies that have related teacher ratings of student behavior with academic achievement. These were previously reported in this chapter (Swift & Spivack, 1969; Wallbrown et al. 1977).

The *Devereux Elementary School Behavior Rating Scale - School Form* (D-SF) assesses children's and adolescents' emotional problems (Naglieri, Bardos, & LeBuffe, 1995). Although the D-SF is a revision of the *Devereux Elementary School Behavior Rating Scale* (Spivack & Swift, 1967), the instruments differ in their structure. The D-SF assesses four clinically defined dimensions: interpersonal problems, inappropriate behaviors/feeling, depression, and physical symptoms/fears. The interpretation of the D-SF scores has a clinical focus. It is intended for evaluation of conduct and emotional behavior disorders (Naglieri, Bardos, & LeBuffe, 1995), while the DESBRS scores were interpreted in terms of behaviors that interfered with the students' learning (Spivack, & Swift, 1967). Consistent with the intended clinical interpretations, the validity evidence provided with the D-SF consists of comparisons between clinically identified and comparison samples (Naglieri, Bardos, & LeBuffe, 1995). No relation with achievement measures was reported for the scores from this instrument. The most recent normative information existing for this tool was obtained by Naglieri Bardos, & LeBuffe (1995). Their normative information was based on 364 five to twelve year old children and 140 thirteen to eighteen year old children. The sample contained both regular and special education children. Normative information was provided in two age groups only.

The *Teacher's Rating Scale* (TRS) was originally designed as an instrument useful for medical research with school children. Its primary focus is on concepts related to attentive behaviors and conduct problems (Conners, 1969; Goyette, Conners, & Ulrich, 1978). The TRS consists of 37 items and measures three constructs: conduct problems, hyperactivity, and inattentive-passive. The inattentive-passive scale assesses attentive behaviors. Goyette, Conners, and Ulrich (1978) supported that the three dimensional factor structure was present in teacher ratings. Test-retest reliability ranged from $r_{xx'}$ = 0.70 to $r_{xx'}$ = 0.90 (interval not reported). Teacher – parent interrater reliabilities ranged r = 0.33 to r = 0.49. The authors did provide normative data. The normative sample consisted of 570 three to seventeen year old children from 277 families in a north-eastern state. Ethnic background was mixed. Norms were provided in three year age intervals. No correlations with achievement scores were reported for the TRS.

In addition to the teacher rating instruments already reviewed, two comprehensive behavior assessment systems that include a teacher rating component along with a parent and student self assessment are available. These instruments do assess a few cognitive behavior constructs.

The *Child Behavior Profile* (CBP) (Achenbach, 1978, 1992) measures a total of eight constructs using teacher and parent ratings and student self-report components. The instrument is intended for clinical assessment of conduct and affective disorders among four to 18 year old children and adolescents. Attentive behavior is the only scale of this complex instrument that assesses a construct related to cognitive behaviors. The behaviors rated are not specific to the classroom. Achenbach (1992) reported correlation between the CBP attention scale and the *Teacher's Rating Scale* inattention scale of r = 0.80. The test-retest reliability over 1 to 4 week intervals has been found to range from $r_{xx'}$ = 0.43 to $r_{xx'}$ = 0.99, while interrater reliabilities reported for pairs of teachers have ranged from r = -0.05 to r = 0.81 (Sattler, 2002). The items on the CBP are general and not specific to the classroom or learning related behaviors. This is consistent with the instrument's purpose of identifying and evaluating clinical problems.

Normative information based on an ethnically mixed sample of 1398 children aged four through eighteen exists (Sattler, 2002).

The *Behavior Assessment System for Children* (BASC) (Reynolds & Kamphaus, 1992) assesses children's cognitive, emotional and problem behaviors using separate rating formats for teachers, parents and student self assessment. Separate forms of the BASC exist for two to six year old children, for six to 11, and for 11 to 18 year old children. Its intended use is diagnosis and classification of children's emotional, behavior, and learning disorders. The BASC is based on a complex construct theory relating children's present behaviors with their emotions and self-perception, their environment, and with their developmental history (Reynolds & Kamphaus, 1992). The BASC teacher rating scale assesses 14 behavior areas. Two of these areas assess cognitive behaviors, attention and study skills, and a third, learning, requires teachers to rate a student's current achievement level. The construct related validity evidence for the relevant the BASC scales partially supports its intended use. Study skills were originally conceived of as loading on adaptive skills, while both attention and learning behaviors were believed to target a broad factor labeled externalizing problems. Confirmatory factor analysis indicated that all three areas loaded on externalizing problems, and the three were found to be highly correlated, $r = 0.82$, -0.73, and -0.83. In addition, Reynolds and Kamphaus (1992) reported significant correlations of the BASC cognitive behavior scales and several other teacher rating scales, *Teacher's Rating Scale*, the *Behavior Problem Checklist*. Relationships between this instrument and other measures of academic achievement were not reported. The internal consistency reliabilities of the BASC cognitive behavior scales ranged from $r_{XX'} = 0.83$ to $r_{XX'} = 0.89$; and two- to eight- week test-retest reliabilities were in the $r_{XX'} = 0.92$ to $r_{XX'} = 0.94$ range. Normative information based on a sample of 2068 children aged six through eighteen exists (Sattler, 2002).

Other comprehensive behavior rating systems do exist and assess primarily emotional or conduct related constructs (Salvia & Ysseldyke, 1998). The cognitive dimensions appear to be too limited to assess individual students' instructional needs. The instructional limitations of these other instruments may also be a weakness of several of the instruments discussed, like the CBP (Achenbach, 1992) or BASC (Reynolds & Kamphaus, 1992) that do have relatively strong evidence supporting the validity of their interpretations for clinical use.

Summary of the review of existing cognitive behavior teacher rating instruments

The existing teacher rating instruments assessing cognitive behaviors were found to lack the qualities required to justify their use for instructional purposes. Those intended for instructional purposes fail to assess the full spectrum of six cognitive behavior areas identified earlier. They also tend to have limited psychometric properties. The reliability is adequate for most instruments, but evidence of the validity of intended interpretations is limited. The psychometrically sound instruments were found to serve clinical purposes, and their assessment of cognitive behavior was therefore limited.

Concluding remarks

Cognitive behavior was defined at the outset as certain behaviors that enhance students' learning and academic achievement. Six areas of cognitive behavior were identified in an analysis of curriculum guides, measurement instruments and empirical research. Empirical support for the relation between each of the six areas and academic achievement was found. Empirical evidence was sought in classroom observations, teacher ratings, and direct measures of constructs related to the six areas of cognitive behavior. Furthermore, existing teacher rating tools were found to be of limited use for providing teachers with information regarding student cognitive behaviors. Each assesses only limited aspects of the six areas of cognitive behavior, and evidence of the validity of each was generally limited. Teacher rating instruments with acceptable psychometric properties were designed for clinical or mental health purposes, and appear to be poorly suited for instructional purposes.

CHAPTER III
METHODS

This chapter describes the research questions and the statistical analyses conducted. It also describes the data and the measurement instruments used. The samples used will be discussed first, followed by descriptions of the instruments used, and finally the hypotheses tested and their rationale will be described along with the statistical procedures used.

Data

Three samples of data were used for this study. The first of these consists of the Iowa Early Learning Inventory (IELI, Hoover, et. al. 2001) scores for a total of 1078 students obtained from a spring 1999 national tryout. From this total sample, which will be referred to as sample A, two separate subsamples were obtained to examine specific research questions. These subsamples will be described later as samples B and C. An overview of the samples is given in table 1, focusing on students with complete responses.

Sample A was obtained from 108 schools in 82 school districts. On average, 2.9 teachers rated 10.0 students in each school. A total of 310 teachers rated their students as part of the study, with each teacher rating on average 3.7 students. The modal number of students rated was 3, with 216 teachers rating 3 students. The instructions to the teachers laid out a sampling strategy to select students within the classroom and directions on scoring. Sampling within classrooms was obtained by providing each teacher with a random number. Teachers were instructed how to use that to select the first student and then to select every fifth student from the class list. According to the rating directions the teachers were to think about each student as they responded to the IELI and rate this student according to his or her behavior during the last week. Complete responses to all the IELI items were available for 942 of the 1078 students.

The sample consisted of 810 white students, 109 African American students, 83 Hispanic students, 31 Native American students, 18 Asian/Pacific, and 12 reporting mixed ethnic background. Also, the majority of the students, 1022, had English as their first language. Of the remaining 52 students 39 had Spanish as their first language. Finally, 121 students were classified according to special needs or conditions, 55 were classified with learning difficulties or needing special education, 28 were receiving speech therapy or speech aid, 11 had problematic motor skills, 10 had attention deficit or were hyperactive, 8 had hearing or visual difficulties, 5 were listed as gifted, and 4 had some medical condition.

The second sample, sample B, consisted of a sample of 141 kindergarten and 172 first grade students from Iowa for whom the primary battery Iowa Tests of Basic Skills (ITBS, Hoover, et. al, 1993) scores were available as well as the complete IELI data. This sample was obtained as part of a national item tryout. The sample was composed of primarily white students. Two of the kindergarten students were Hispanic, and among the first grade students there was one African-American and one Hispanic. The teachers of the students in this sample filled out the IELI during the spring of 1999. The ITBS primary battery Forms K and L were administered to the students as part of their district wide testing program, also in the spring of 1999. Level 5 was used for the kindergarten students, while Levels 6 or 7 were used for the 1st grade students.

The third sample, sample C, consisted of IELI scores for 629 students, 280 kindergarten and 349 1st grade students. Of the kindergarten students, 159 were girls and 121 boys; the 1st grade students were 170 girls and 179 boys. Data on ethnic background were available for 257 of the kindergarten students: 46 African-American, 33 Hispanic, and 178 white. The first grade sample consisted of 319 students for which data on ethnic on background were available: 44 African-American students, 35 Hispanic, and 240 white. The teachers of the students in this sample filled out the IELI during the spring of 1999.

Table 1. Overview of the three samples according to grade and ethnicity or sex.

Sample	Description	n	Grade	n	Sex / Ethnicity	n
A	Total sample:	942	K	421	F	233
	All students with				M	188
	complete records.		1st	521	F	266
					M	255
B	Iowa tryout[1]:	313	K	141	F	74
	Student with IELI				M	67
	and ITBS scores.		1st	172	F	96
					M	76
C	National tryout[1]:	629	K	280	F	159
	Students with complete IELI				M	121
	records and gender or		1st	349	F	170
	ethnic identity.				M	179
		576	K	257	AA	46
					H	33
					W	178
			1st	319	AA	44
					H	35
					W	240

Note: K = Kindergarten; 1st = First grade; F = female; M = male; AA = African American, H = Hispanic; W = White.

[1] Samples B and C are subsamples of sample A.

The current data were clustered in the sense that the original sampling unit was the school. The kindergarten and first grade teachers then rated a sample of students from their classroom. It was argued that despite the clustered nature of the sample, an analysis based on individual student data was appropriate. Because the primary purpose of the study was to investigate the properties of an individual students' test score, the individual level was the focus of interpretation.

Instruments

Two measurement instruments were used in this study, each assessing either students' academic skills or cognitive behaviors. The instrument that is of primary interest in this study is the *Iowa Early Learning Inventory*, a measurement tool assessing students' cognitive behaviors. The other is the *Iowa Tests of Basic Skills*, a test battery measuring student achievement. They will be described in detail in the next sections.

The Iowa Early Learning Inventory

The *Iowa Early Learning Inventory* (IELI) was constructed based on the assumption that certain behaviors of primary grade students are strongly related to their academic success. Such behaviors were labeled cognitive behaviors. For the present study cognitive behaviors were defined as behaviors enhancing and supportive of learning for kindergarten and 1st grade students. The IELI was designed as an instrument assessing those behaviors in primary grade students. It was designed as a rating tool to which teachers respond. The behaviors measured were required to be objective, reliably observable in classroom situations, and characteristic of primary grade students.

Identification of behavior statements

The construction phase of the IELI started with an investigation of cognitive behaviors represented in literature related to classroom work of primary grade students. Sample of documents from three different literature sources describing behaviors believed to be cognitive were obtained: a sample of curriculum guides published by state departments of education (Alabama State Department of Education, 1992; California Department of Education, 1988; Indiana Department of Education, 1987, 1992; Kentucky Department of Education, 1993a, 1993b; Nebraska Department of Education & Iowa Department of Education, 1993; Oregon State Department of Education, 1986, 1988, 1990; Wisconsin Department of Public Instruction, 1986, 1991), empirical research related to early classroom behavior and academic skills (for example Bryan & Wheeler, 1972; Cobb, 1972; Derevensky, et al., 1983; Fad & Ryser, 1993; Lahaderne, 1968; McKinney, et al., 1975; Perkins, 1965), and finally existing measurement tools intended to assess behaviors and skills relevant for children as they enter school (Achenbach, 1978, 1992; Chiu, 1997; Goyette, Conners, & Ulrich, 1978; Harrison, 1985; Kim, Anderson, & Bashaw, 1968; Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984a, 1984b, 1984c; Reynolds, Kamphaus, 1992; Ross, Lacey, & Parton, 1965; Seidman, Linney, Rappaport, Herzberger, Kramer, & Alden, 1979; Sparrow, Balla, & Cicchetti, 1985; Spivack, & Swift, 1967). The sample of documents from all three areas was analyzed and descriptions of cognitive behaviors believed by the authors to be relevant for the young learners' early academic development were documented. Descriptions of behaviors indicative of cognitive deficiencies in were also documented. Descriptions relying solely on traits or dispositions and representing attitudes were excluded. This process yielded a total of 427 statements describing cognitive behaviors. In addition, numerous other documents not referenced, but within the same three categories, were reviewed and either found to only contain behaviors already identified or not to provide additional behavioral descriptions.

Analysis of behavior statements and the construction process

The collection of behavioral descriptions was analyzed using qualitative methodology (Gene & Peshkin, 1992). The entire pool of behavior statements was read a number of times and all statements were classified based on the similarity of targeted behaviors into several categories. A coding system was developed as the analysis proceeded; new codes were defined when behaviors that stood aside from others were encountered. During periodic reviews, categories were occasionally combined or divided into a set of related categories. Occasionally a statement was judged to be more similar to statements that had been assigned to a different classification code. These statements were then reclassified and assigned the category code of the more similar statements to yield consistent categories.

When continuous review indicated that most of the behavioral descriptions additionally being encountered were similar or identical to those already documented, data collection was effectively ended. This indicated that the descriptors of cognitive behavior thus far collected gave a fair representation of the opinions and beliefs regarding cognitive behaviors found in the type of documents sampled. This is known as saturation in qualitative methodology and is taken as an indication that the process of data collection can be terminated (Gene & Peshkin, 1992; LeCompte & Goetz, 1982).

The areas of behaviors that emerged and the coding system that had been developed at the time of saturation is shown in table 2. The areas were reviewed with respect to which of them contained cognitive behaviors that could be expected to be prominent in classroom situations. These were found to be captured by the following six broad groupings: general knowledge, oral communication, written language, quantitative knowledge, work habits and attentive behaviors. The behavioral statements classified as social behaviors, affective and conduct problems, or motor behaviors were also reviewed and carefully considered in relation to cognitive development. Given the limited scope of the IELI, these behaviors were found to be outside the intended focus. They were excluded from the analysis at this point.

The classification of statements was reviewed after the study had been limited to the areas consisting of cognitive behaviors. The resulting refinements led to the final model of six areas of cognitive behavior from which the IELI was constructed. The relations of the IELI cognitive behavior areas to the earlier categories are shown in table 3. The number of behavioral statements that had been identified for each of the cognitive behavior areas ranged from 10 to 100.

A few behaviors were selected from each of the areas and items describing these behaviors were written in a way that was suitable for a behavior rating

Table 2. Classification scheme derived from qualitative analysis of the behavioral statements identified.

	Descriptive title and sub-classifications
1.	*Language related behaviors*
	a. Language use: vocabulary, syntax, pronunciation
	b. Written language
	c. Comprehension, story structure, content of communication
	d. Communicative; effective use of language
2.	*Quantitative behaviors*
	a. Quantitative ability
	b. General knowledge
3.	*School work and work habits*
	a. Application of writing tools
	b. School work; work habits
	c. Independence
4.	*Social behaviors*
	a. Social interaction
	b. Group behavior
	c. Rules and order
	d. Physical independence
5.	*Affective behaviors & conduct disorders*
	a. Stability; instability, self reliance; independence
	b. Disruptive behavior
	c. Temper; anger
	d. Anxiety
6.	*Attentive behaviors*
	a. On / off - task – behavior
	b. Distractibility

scale. This constituted the first version of the IELI. When only a limited numbers of behaviors statements belonged to a cognitive behavior area, several additional items were constructed incorporating behaviors judged to be similar to or characteristic of that area. The IELI underwent a number of revisions as behaviors were added or removed and the wording of the statements was revised.

Content theory of the IELI

The analysis of descriptions of behaviors related to learning and cognition that were identified in the qualitative analysis described above provided a theoretical framework or construct theory for the cognitive behavior areas

Table 3. The six areas of cognitive behaviors of the IELI.

Area	Descriptive title	Relation with 1st classification scheme
1	General knowledge	2.b
2	Written language	1.b
3	Oral communication	1.a, 1.c, 1.d
4	Math concepts	2.a
5	Work habits	3.b, 3.c
6	Attentive behaviors	6.a, 6.b

measured by the IELI. This construct theory can be described generally as consisting of six areas of cognitive behaviors. These areas measure behaviors that are integral parts of the learning process and observable in the classroom setting. The content of each of the six areas is described below.

The general knowledge area can be described as behaviors or skills that are useful in multiple situations. This area emerged as the more general aspect of the quantitative behaviors as seen in table 2, but during the refinement of categories they were found to be more similar to some general behaviors and thus the area evolved away from the math/quantitative notion. Examples are knowledge of personal information, recognize common opposites, know names of colors, and know colors of common objects. During the item construction phase many behaviors that were considered for potential inclusion in this area were rejected if teachers were unlikely to observe those behaviors in the classroom setting.

The oral communication behaviors ranged from the pronunciation and grammatical structure to the relevance of the communicated message and personal involvement. Behaviors characterizing this area were: grammar, sentence structure, word use, retelling a story, telling story from a picture, describing a thing or an event, identifying main characters and events of a story, recognizing main ideas and supportive details, using logical or chronological order, sharing ideas, choosing relevant topics, and discussing personal experiences and opinions. This area of cognitive behavior captured the mechanisms of speech and listening as well as the logical structure and appropriateness of the communicated message.

The written language behaviors are any print related behaviors and skills. These include recognizing letters, letter-sound relationships, and writing letters or words. It can be described as the manipulation of the letters and words as tools for symbolic representation of language. This area contained a number of behaviors that were judged to extend beyond the general limits of students in kindergarten and first grade.

Math concepts behaviors relate to the quantitative aspects of objects and ideas. They capture student's work with numbers of objects, quantitative relations among objects in space and time, and numerical values attributed to objects. Examples are recognize likeness and differences of shapes and objects, follow progressions from left-to-right or right-to-left, know values of coins, and state day of week when asked. This area captured both the physical manipulation of objects to the mental representation of their quantitative aspects.

The work habits area represents how students approach the task at hand, how they plan their work and carry it out. Examples are persisting on task, completing assigned work on time, asking clarifying questions, following teachers' requests, initiating and completing activities, working independently, and organizing work. Negative examples of work habits were frequent: start to work on task before understanding instructions, begin another task before finishing the current assignment, and rush through a task without attending to the correctness or quality of the work. These negative behaviors indicate deficient work habits. This area of behaviors represents the disciplined execution of skills as well as the carefulness and persistence in the process of working on or solving a task.

Attentive behaviors are the students' attendance to the instruction, learning tasks, or ongoing group work. Examples of these behaviors are attending to or working on the assigned task, following instructions consisting of a number of steps, and sustaining attention on task. Negatively worded descriptions, or deficient attention, were more frequent than attentive behaviors:

easily distracted, short attention span, stare into space, make remarks that are out of context, look out a window, and forget what was learned before. It can be described as general attentiveness and readiness to learn from instruction or the task at hand.

The behavior areas can be viewed as being either subject matter related cognitive behavior areas or general cognitive behavior areas depending on their direct relationship or similarity with a specific academic subject. The subject matter related areas consist of behaviors central to specific academic subject matters, such as the manipulations of symbols representing language in writing, using language to communicate or working with quantitative or mathematical understanding. These areas, oral communication, written language, and math concepts, correspond to the core academic subjects at the primary grades: language arts and math. The general cognitive behavior areas, on the other hand, measure behaviors indicative of student characteristics relevant for learning in any academic setting. These pertain to the ways a student approaches the academic tasks presented in school settings and her/his general learner characteristics or knowledge brought to these tasks.

A fundamental assumption implicit in the IELI construct theory is that student status on the cognitive behavior areas relates with academic achievement. The value of measuring the cognitive behaviors is that students' deficiencies in these behaviors will translate into difficulties in the academic subjects during the students' educational careers. This construct theory thus implies that a relation exists among the cognitive behavior areas and measures of achievement.

Based on the construct theory outlined above, predictions regarding theoretically relevant relationships of each of the IELI cognitive behavior areas can be derived. Each of the IELI areas should provide specific information regarding its theoretically relevant academic subjects. Oral communication has theoretically relevant relationships with measures of reading, listening and language usage. Written language relates strongly with measures of listening and reading. Math concepts has a theoretically relevant relationship with measures of math achievement. The three general cognitive behavior areas should relate to all academic subjects.

It is believed that the IELI measures behaviors that are relevant for successful progress during the early years of schooling and are an integral part of academic achievement. The information provided by the IELI is viewed as supplementary to what is provided by standardized achievement test batteries. The IELI 's intended use is to provide information for instructional planning, and for identification of student cognitive behavioral deficits in order to allow teachers to provide remedial instruction at an early stage. It is designed for use along with standardized test results.

Response scales for the IELI items

Each of the IELI items consists of a statement describing a cognitive behavior and a three or four point response scale; each statement asking whether the student is able to perform a certain behavior. Examples are, "identify and name common colors", "recall facts from a story read aloud", and "pay attention in class". For 34 of the items, a four-point response scale is used. The remaining items use a three-point response scale. The anchors accompanying the response scales indicate the frequency of occurrence of the behaviors, with the exact wording chosen to be semantically consistent with the behaviors rated. The

Table 4. Number of items and range of possible scores for the IELI cognitive behavior areas.

Scale	Number of items	Minimum score	Maximum score
General knowledge	5	5	17
Oral communication	8	8	32
Written language	7	7	26
Math concepts	10	10	35
Work habits	7	7	28
Attentive behaviors	7	7	28
IELI composite		44	166

anchors most commonly used with the four point scales are "never", "rarely", "occasionally", and "frequently". The anchors accompanying the three point response scale vary. One example is "none", "some" and "most". The teachers responding to the inventory indicate how frequently the student exhibits each of the behaviors by marking the appropriate rating circle. The IELI scores are obtained by adding the numerical values attached to each of the scale points, one through three or four. The score point representing the least magnitude of behavior is always scored as one. The IELI composite score is then obtained as the sum of the cognitive behavior area scores. The number of items and score range for each of the cognitive behavior areas and the composite are reported in table 4.

ITBS

The *Iowa Tests of Basic Skills* (ITBS) is a battery of achievement tests that assess the knowledge and skills required for successful academic achievement. The purposes of the ITBS primary level batteries are assessment of individual students' academic development as well as the assessment of specific knowledge and skills important for further successful academic achievement (Hoover, et al., 1993). It is intended, among other things, to be used to aid teachers in instructional planning, to identify students' relative strengths and weaknesses, and to provide information useful for reporting to parents (Hoover, et al., 1993). The psychometric properties of the ITBS have been evaluated and judged to be excellent (Brookheart, 1998; Cross, 1998; Lane, 1993). It has been judged to be among the best existing measures of achievement for elementary schools (Thorndike, et al., 1991).

Three levels (5, 6, and 7) of the ITBS Forms K and L were used in this study. Level 5 was administered to the kindergarten students, while levels 6 and 7 were administered to the 1st grade students. Level 5 consists of five tests: vocabulary, word analysis, listening, language, and math. It contains a total of

146 questions, and on average requires two hours to complete. Level 6 includes 207 questions in six tests: vocabulary, word analysis, reading, listening, language, and math. The students have about two hours and 45 minutes to complete these tests. All items are multiple choice items, and students mark the answers in the test booklet (Hoover, et al., 1993).

All of the content measured at level 5 of the ITBS is also measured at level 6. In addition, level 6 also measures some skills not measured at level 5. Similarly, all skills measured at level 6 are also measured at level 7 in addition to some skills not measured at level 6 (Hoover, et al., 1993). The vocabulary tests measure listening vocabulary; the word analysis tests measure early literacy skills like letter recognition and letter sound relationships, and level 6 also focuses on simple words. The listening tests measure skills used in oral comprehension of instruction and other material. The language tests assess the students' understanding of and skills in using language as a medium to express ideas, opinions and feelings. The math tests of both levels emphasize math concepts, and the level 6 test additionally assesses problem solving. The level 6 reading test measures early reading skills (Hoover, et al., 1993).

Hypotheses and statistical analysis

The purpose of the present study was to investigate the psychometric qualities of the IELI. This investigation will be fivefold, focusing on four aspects of validity as well as the internal consistency reliability of the IELI. The aspects of validity that were investigated relevant to the intended use of the IELI are the concurrent relations between the IELI cognitive behavior areas and theoretically relevant standardized achievement test scores. Additional relations of the IELI cognitive behavior areas with achievement test scores, the validity of the constructs underlying the IELI, and issues related to potential sex and racial/ethnic performance bias in IELI were also investigated. The reliability of the IELI scores was investigated. Each of these is considered in more detail in the next sections, along with considerations of the statistical analyses required.

Construct validity: IELI and related academic achievement measures

A recurrent theme in any discussion of validity is the role of a theory underlying the construct being measured. The relevance of any piece of empirical evidence regarding the validity of an interpretation of the scores depends on how the empirical data connect properties of the scores with aspects of the theory. Of primary importance is empirical evidence supporting or refuting the relation between the scores and measures that, by theory, capture related aspects of the same construct or measure related constructs. The empirical evidence relevant to test such hypothetical relationships typically involves correlation coefficients between two or more measures (Messick, 1989). The IELI is intended to assess cognitive behaviors; behaviors that supposedly are related to and enhance student academic achievement. The construct theory underlying the IELI dictates that certain relationships with a measure of academic achievement should emerge, and furthermore, patterns of relationships among the cognitive behavior areas should be apparent. Such evidence can be obtained by investigating the relation between IELI scores and students' academic achievement. The measure of academic achievement chosen for this study was the ITBS.

The correlation between the test scores and a theoretically related variable supports the validity of a given interpretation of the test scores (Messick, 1989). Theoretically important relationships among the IELI cognitive behavioral areas and ITBS scores were derived from the underlying construct theories for each of the IELI areas. These will be detailed next, along with the theoretical rationale supporting these hypothetical relationships.

Of primary relevance regarding the validity of the interpretation of the total IELI scores as measures of general or overall cognitive behaviors was the correlation between the IELI composite score and the ITBS composite score. The rationale for this relation is the role of each in providing useful general understandings of the students' academic achievement.

The construct theory argued for several important relationships of oral communication with the language related measures of achievement. The most direct correspondence is between oral communication and the language test, because both assess the role of language as a medium of communication. Oral communication should also correlate with the listening test as both assess aural reception of communicated messages. Finally, oral communication should also show a modest correlation with the 1st grade reading test.

The theoretically most important relation for the written language area was with the ITBS word analysis test. Analysis of the content of the written language area revealed that it primarily assesses manipulation of the symbolic representation code of written language - letters and words. The ITBS word analysis test measures skills related to letters and words (Hoover, et al., 1993). Superficially, the relation of IELI written language with the language test appeared to be of equal importance, but the ITBS language test assesses the use of language to express ideas (Hoover, et al., 1993). It measures a different construct and it was not expected that a strong relation with the IELI written language would be found. Furthermore, as the majority of students had mastered the behaviors related to letters and words in 1st grade, a ceiling effect on the written language area was expected as result. This was expected to suppress the relation of written language with ITBS scores in the 1st grade sample. The construct theory underlying the IELI also predicts a relation of written language with the 1st grade ITBS reading test scores.

The relevant relation for the IELI math concepts area was judged to be with the ITBS math score. IELI math concepts assess quantitative, numerical and measurement related constructs. The ITBS math subtest is described similarly (Hoover, et al., 1993).

The IELI general knowledge, work habits and attentive behaviors assess behaviors that are of importance for achievement in general. Therefore the ITBS composite score was the theoretically most relevant criterion variable for these three scores. In addition, all three general behavioral areas were expected to correlate with all of the ITBS scores discussed in relation with the content specific behavior areas. The rationale for this hypothesis is the same; that these cognitive behaviors are relevant for all aspects of academic achievement.

The relations of the IELI scores with a measure of academic achievement that were of primary relevance in supporting the validity of intended score interpretations have been outlined above. Statistical analysis relevant for this part of the investigation requires that the correlations between the IELI and the relevant ITBS criterion variables be obtained. Sample B, consisting of students with both IELI and ITBS scores, was used for these statistical analyses. The theoretically relevant correlations were expected to be higher than other correlations among the IELI areas and ITBS test scores.

Construct validity: Specific relations with academic achievement measures

A further issue related to the validity of the IELI scores was the extent to which each of the IELI cognitive behavior areas provided information regarding achievement measures independently of the other behavior areas. It is well known that measures of the various aspects of achievement tend to correlate to a certain extent with each other. Similar correlational patterns can be expected for the relationships of the cognitive behavior areas with the achievement measures. This raises the question of whether the subject matter specific behavior areas have specific relationships with the respective achievement measures. In this situation the question may be raised whether a correlation between two variables, which on theoretical grounds are expected to be related, represents anything other than such a general inter-correlation. The validity of the interpretations of the respective scores could be questioned in such situations on the grounds that the correlation does not represent the relation of interest. This is an example of what Messick (1989) discussed under the label of construct relevant variance. An investigation of these validity issues was conducted using multiple regression methodology. The dependent variables in the regression equations were the ITBS scores and two IELI cognitive behavior areas were independent variables. The cognitive behavior area that had a hypothetical relationship with the ITBS score was entered second, after the other behavior area, or construct irrelevant area, in each equation. Using this method the construct relevant variance can be separated from the general covariation among all the behavior areas and achievement measures. This was done by subtracting the variance explained by the construct irrelevant area from the variance explained by both areas.

As was implied in the discussion above, three of the IELI cognitive behavior areas have direct correspondence with specific ITBS content matter measures. These are oral communication, written language, and math concepts. The other three IELI areas assess more general cognitive behaviors that are expected to be important for all aspects of the primary grade curriculum. These are general knowledge, work habits, and attentive behaviors. The theoretically important correspondence of the more general IELI cognitive behavior areas identified above was with the ITBS composite score. An important question regarding the interpretation of the IELI scores is to what extent the general cognitive behavioral areas provide additional information regarding the ITBS content matter, over and above that provided by the content specific cognitive behavioral areas. Finally, some content specific cognitive behavior areas may provide additional information regarding the other achievement areas, for example oral communication may provide additional information regarding

ITBS math scores. These issues were addressed using multiple regression techniques as discussed above.

Construct validity: Confirmatory factor analysis of the IELI

A further purpose of this study was to assess the extent to which the internal structure of the IELI scores is congruent with the construct theory underlying the inventory. Evidence of this sort is a fundamental element in any argument supporting the construct validity of a measurement tool (Messick, 1989). The statistical methods of primary relevance to support or refute whether the construct underlying the IELI is represented in the teacher ratings is a factor analysis of the responses to each item on the IELI. Such statistical techniques are useful to investigate the relation among multiple variables (McDonald, 1985). They are frequently used to provide evidence regarding the validity of the constructs underlying measurement tools (Anastasi & Urbana, 1997; Crocker & Algina, 1986; Messick, 1989).

Currently two types of factor analytic methods are in common use, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Both can be regarded as special cases of structural equation modelling (SEM). SEM is a statistical technique helpful for the analysis of the relations among outcome variables (dependent variables), explanatory variables (independent variables), and unmeasured theoretical constructs (latent or underlying variables), relating each set of the variables to the other sets (Bollen, 1989; McDonald, 1985). The CFA variant of this class of techniques was used in the present study. The construct theory underlying the IELI discussed earlier in this chapter was used to identify several potential theoretical models investigated in this analysis. The analysis was conducted using sample A.

Criteria for evaluation of CFA results.

A persistent problem in CFA is the evaluation of the quality of the models under investigation, commonly referred to as model fit (Bollen, 1989). Quality, in this context, has to do with how well the theoretical model represented in the statistical model mirrors the real world phenomena it presumably explains as they are represented in the data at hand (Bentler & Bonett, 1980; Bollen, 1989), or, in statistical terms, "a plausible representation of the data" (Bentler & Bonett, 1980, p. 591). The most basic and commonly used method of evaluating model fit is a χ^2 test statistic, where a large χ^2 indicates great discrepancies between the model and the data, and therefore a model that fails to capture the relationship among variables (Bentler & Bonett, 1980; Bollen, 1989). The χ^2 test statistic has certain limitations. In addition to the model-data discrepancy, its value depends on the size of the sample, method of estimation, and violation of underlying assumptions (Bentler & Bonett, 1980; Hu & Bentler, 1998). Because the χ^2 test statistic has been seen as insufficient to guide model evaluation, several additional statistical fit indices have been developed. In essence, all these fit indices attempt to capture discrepancies between the empirical covariance matrix and a population covariance matrix that is implied by the model under investigation (Bollen, 1989; Hu & Bentler, 1998). Several such indices have been developed to aid the evaluation of the results of confirmatory factor analysis, but some of them have been shown to be unreliable in recent simulation studies (Hu & Bentler, 1998). In addition, many of these indices are closely related and yield almost identical information; investigations of their correlations suggest that with a careful selection from different 'families', a few indices are sufficient for the evaluation of structural models (Hu & Bentler, 1998, 1999). The current study reports three fit indices, in addition to the χ^2 test, selected based on the advice of Hu and Bentler (1998, 1999): the root mean square error of approximation (RMESA), the standardized root mean square residual (SRMR), and the comparative fit index (CFI).

Historically, model evaluation has been further complicated because the criteria for deciding when a model exhibited an acceptable fit had neither theoretical nor empirical basis. Traditionally only conventional 'rules of thumb' existed (Hu & Bentler, 1998; McDonald & Ho, 2002). Several recent simulation studies seem to have remedied this situation to an extent, and established empirically based cut-off criteria to guide model evaluation (Hu & Bentler, 1998, 1999; Marsh & Hau, 1996; Yu & Muthen, 2002). The current study uses evaluation criteria based on those simulation studies, using the following values as criteria for acceptable fit: SRMR ≤ .08; RMESA ≤ .06; CFI ≥ .95.

Description of the models estimated

The construct validity of the IELI was assessed using confirmatory factor analytical methods. The purpose of this part of the analysis was to evaluate and compare several statistical models that represent different conceptualizations of the construct of cognitive behavior as measured by the IELI. Five models of the underlying construct were constructed and evaluated. These models were derived from the IELI construct theory by considering various combinations of six specific factors corresponding to the six cognitive behavior areas outlined above and/or a general cognitive behavior factor. The five models are: a model where one general cognitive behavior accounts for the relation among all 44 items (the one factor model), a model where six independent factors account for the relations among items (the six independent factors model), a model where six related factors account for the relations among items (the six related factors model), a model where six independent factors and a second order factor are needed to account for the relations among items (the second order factor model), and finally a model where six independent factors and an independent general factor are needed to account for the relations among items (the general and specific model). The models considered represent five combinations of six specific factors and/or a general factor to account for the relations among the 44

IELI items. Four of the models contain specific factors, two of them in combination with general factors. Three models contain a general factor, of which two also involve specific factors. An overview of the models considered is presented in table 5.

The models considered in the present study were limited in several ways. Only models assuming a combination of one general factor and six specific factors were considered. Models assuming different number of factors that might be formed by combining some of the behavior areas were not considered. Correlated errors among pairs of items were not included in the models, an exception was made in the case of two item pairs, 17a and 17b and 20a and 20b. These item pairs explicitly aim at different developmental complexity of the same behavior. Related item content may in a number of instances justify the inclusion of correlated errors, but the investigation of their potential role was excluded from the original set of analyses. Only models assuming that each item relates to one and only one specific factor were evaluated. No item was allowed to relate to a second factor despite similarity with items belonging to another behavior area. A short description of the models investigated follows.

Table 5. Relations among specific and general factors in the five models under study.

	Specific factors	Relation among specific factors	General factor	Relation of general and specfic factors
One Factor model	-	-	1	-
Six Independent Factors model	6	Unrelated	-	-
Six Related Factors model	6	Related	-	-
Second Order Factor model	6	Unrelated	1	Related
General and Specific Factors model	6	Unrelated	1	Unrelated

The one factor model

The one factor model consists of one general factor only. This factor is assumed to account for any relationship among the IELI items. Statistically, the model consists of 44 loadings of individual items on the general factor plus the two correlated errors. According to this model, there exist no clusters of cognitive behaviors that have stronger relationships among themselves than with behaviors outside the cluster. The model is shown in figure 1.

Six independent factors model.

The second model, the six independent factors model, postulates six unrelated factors (where each item loads only on its respective behavior) as follows: items 1 through 5 loading on general knowledge, items 6 through 13 on oral communication, items 14 through 19 on written language, items 20a through 28 on math concepts, items 29 through 35 on work habits, and items 36 through 42 on attentive behaviors. Thus the model postulates six latent constructs that are assumed to be uncorrelated among themselves and two correlated error terms among the item pairs 17a and 17b and 20a and 20b. Statistically, the model consists of 44 loadings of individual items on one of the specific factors and the two correlated errors. Correlations among factors are set to zero. The six independent factors model postulates six isolated areas or factors of cognitive behaviors, isolated in the sense that the absence or predominance of one area in a child's behavior is unrelated to the presence or absence of the other areas. The model is shown in figure 2.

Six related factors model.

The third model, the six related factors model, consists of the 44 items relating to their respective behavior areas as discussed under the six independent factors model, but the behavior areas or latent constructs are now postulated to correlate among themselves, and the two correlated errors are assumed as before. This model thus assumes that the behaviors represented by each item relate directly only to their respective scales or latent constructs, but that the latent constructs are allowed to correlate. According to this model the IELI measures six related cognitive behavior areas, where the absence or presence of each is postulated to co-vary with the presence or absence of the others. Statistically, the six related factors model consists of 44 loadings of individual items on the specific factors, 15 correlations among the factors, and the two correlated errors. The model is shown in figure 3.

Second order factor model.

The fourth model, the second order factor model, postulates six latent factors representing the behavior areas, and a second order factor that is postulated to account for any covariation among the first order factors. The two correlated errors are included as in previous models. A major difference between the second order factor model and the six related factors model is that the second order factor model restricts to a greater extent the potential relationships among any set of three or more behavior areas than does the six related factors model. Apart from that, the general description of the six related factors model also holds for the second order factor model. The second order factor model postulates the same loading of each item on its respective first order factor as described under the six independent factors model, plus six loadings of each specific factor on the general or second order factor as well as the two correlated errors. The model is shown in figure 4.

General and specific factors model.

The final model considered, the general and specific factors model, postulates six unrelated specific factors plus a general factor that is also unrelated to the specific factors. The two correlated errors are assumed as in the previous models. Statistically this model consisted of the loadings of 38 items on the six specific factors and the loading of all 44 items on the general factor. One item

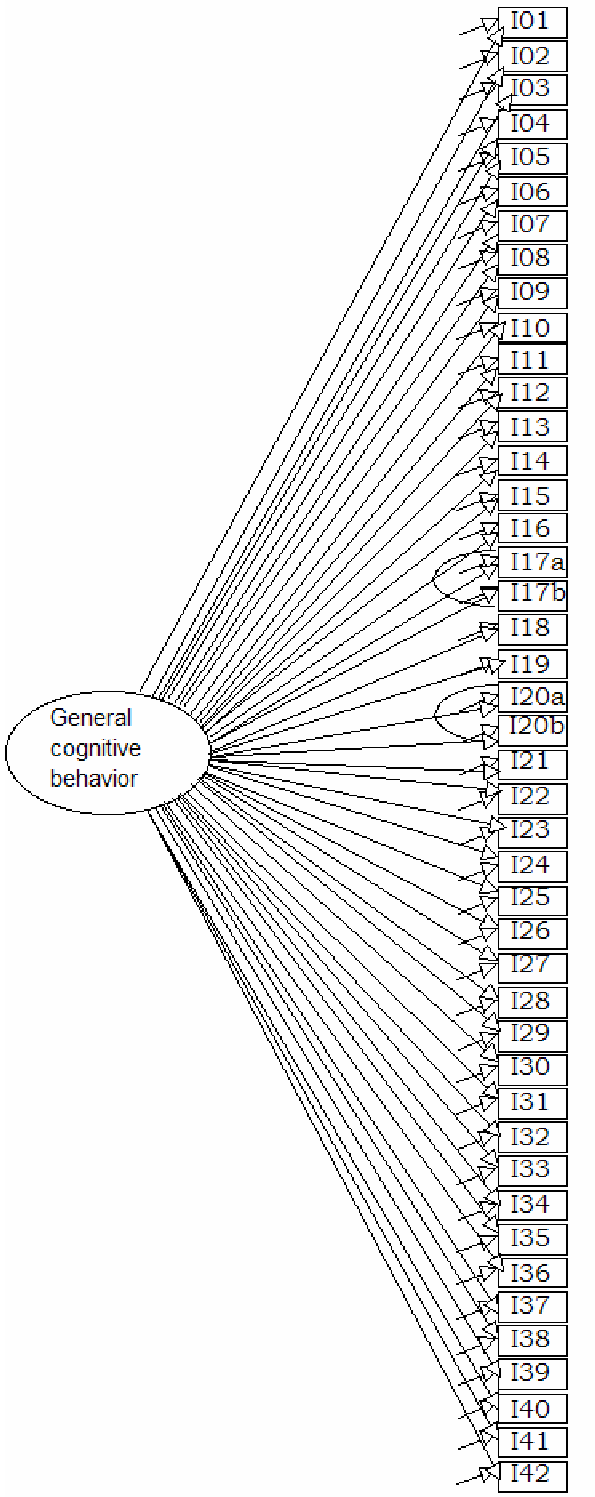

Figure 1. Model A, one general factor.

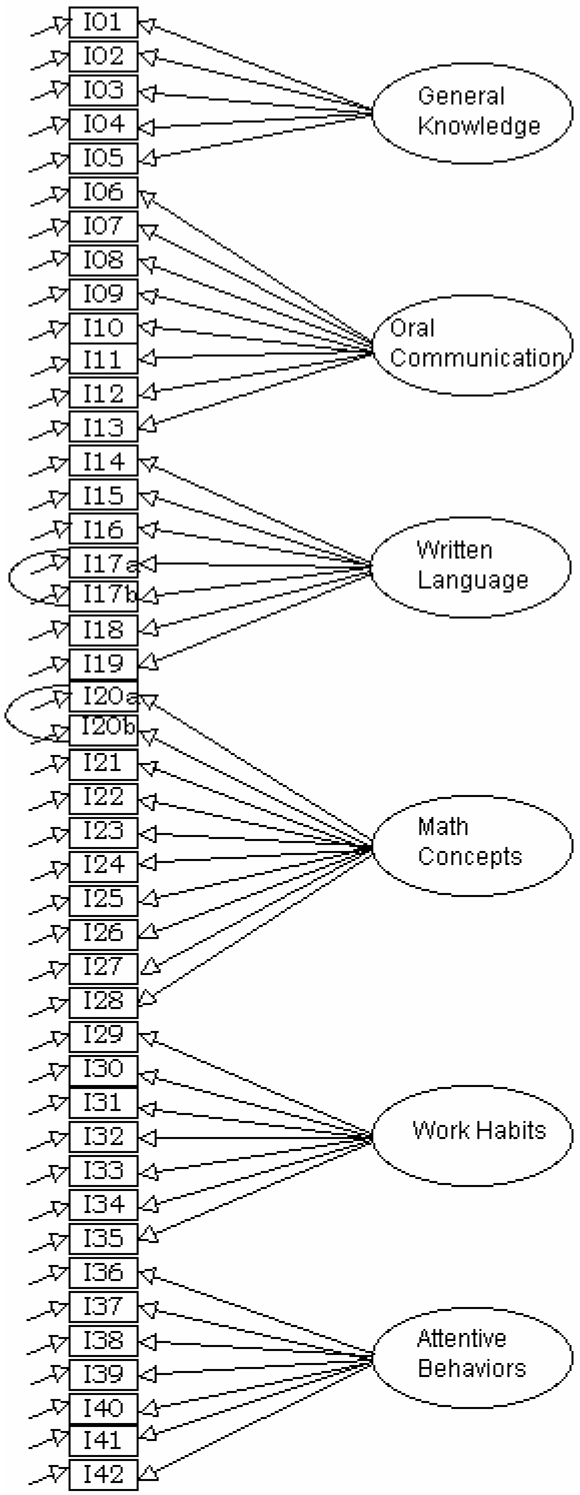

Figure 2. Model B, six uncorrelated factors.

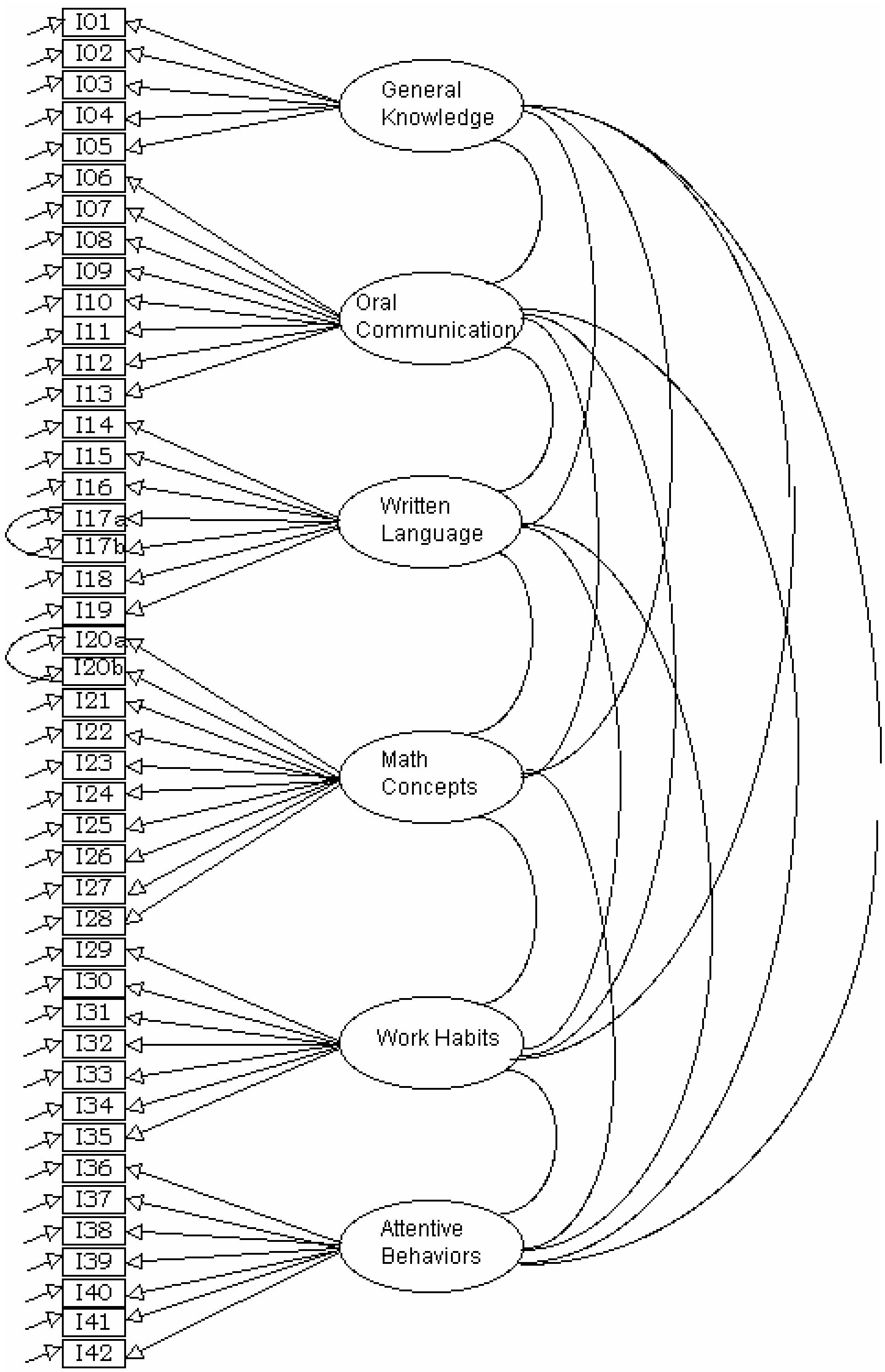

Figure 3. Model C, six correlated factors.

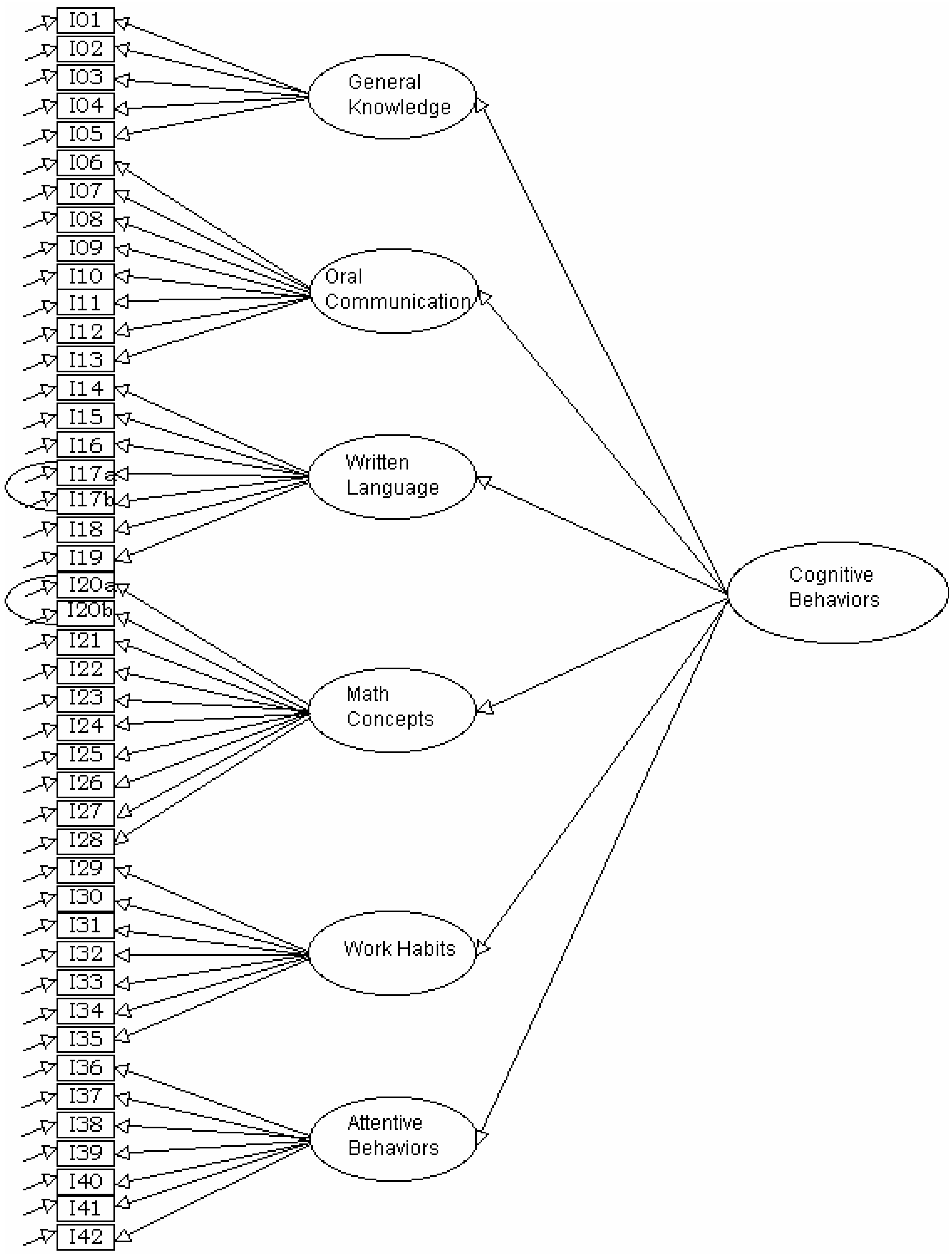

Figure 4. Model D, six uncorrelated factors and one general factor (second order factor).

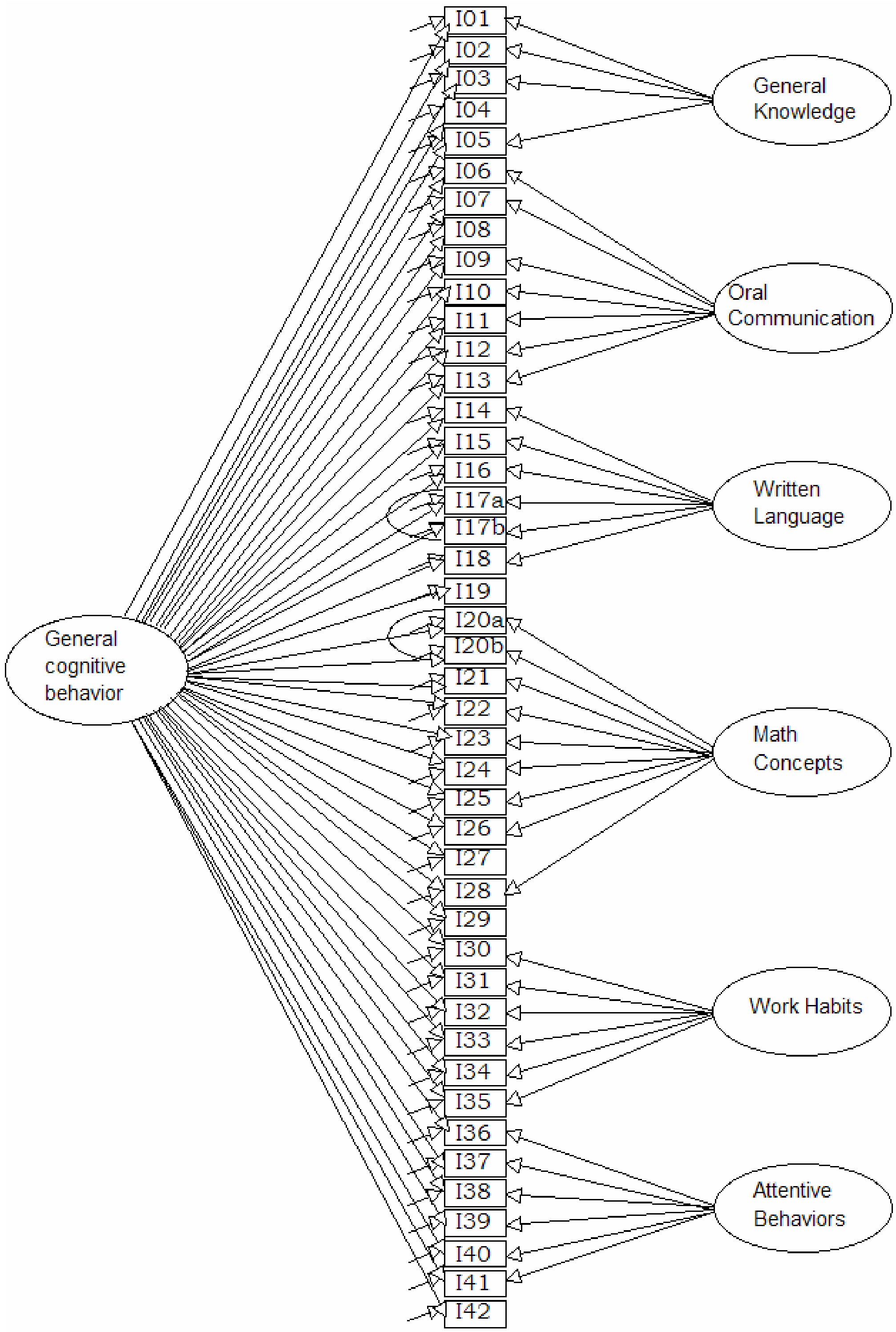

Figure 5. Model E, six uncorrelated factors one general factor.

from each behavior area was removed from the respective factor and used to identify the general factor, apart from that, the loadings on the specific factors were as described previously. These six items were chosen based on investigations of correlations of the items to all six behavior areas and on modification indices from confirmatory factor analysis of the six related factors model. Under this model the general factor accounts for all that is common among the behavior areas, but the specific factors can vary in any way without being related to each other. The model is shown in figure 5.

Consequences of test use:
Sex and ethnic bias

It is important to consider bias regarding the validity of test scores. From the perspective of validity theory, bias exists whenever student characteristics other than the construct measured are represented in the test score in such a way that one group of students is shown in more favorable or unfavorable light than another group (Cole & Moss, 1989; Messick. 1989). Differential item functioning (DIF) is the statistical evaluation of potential bias. Irrelevant variables, such as students' sex or ethnic origin, may pose a potential threat to score interpretations if they relate to the underlying construct. Teacher ratings of student cognitive behavior may be susceptible to the influence of demographic variables due to cultural differences in student behavior. For example, some behavior patterns that may be more frequent among either girls or boys may be confounded with cognitive behaviors. Patterns of behaviors frequent among African American, or Hispanic or White students may be similarly confounded with cognitive behaviors. Such differences may influence teacher ratings, and might produce IELI scores that over- or underrepresent the students' actual cognitive behaviors. Any such bias potentially diminishes the validity and usefulness of the IELI scores for instructional planning, and may place certain groups of students at a disadvantage.

In the most common testing situation the source of influences potentially resulting in DIF operates only through one medium, the student himself/herself. In the current situation, as well as with any other rating instrument, the rater can be a second medium through which effects associated with characteristics of groups of students can affect the test score. The raters perceptions or previous ideas regarding the members of the groups that influence their ratings of individuals from each group can become a second source of errors. However, this second source of potential influence does not alter the basis of DIF analysis. The focus of a DIF analysis is the consistency of the interpretation of a score. It does not matter whether a student receives disproportionally low score due to characteristics associated with his/her group membership status or whether the raters perceptions or ideas of the groups are at effect. Both situations result in a challenge to the consistency of one and the same score interpretation. For this reason the DIF effects are viewed as a single source of error in the analyses conducted as part of this study.

Methods used for DIF analyses

A great number of statistical methods intended to identify DIF have been suggested in the psychometric literature. The majority are applicable for dichotomous items (Camilli & Shepard, 1994), while several methods for polychotomous items have also been developed (Penfield & Lam, 2000; Potenza & Dorans, 1995). Potenza and Dorans (1995) adopted a two dimensional classification scheme for DIF methods that takes into account two important methodological issues. The first of these considers the nature of the measure of ability used to match individuals of comparable ability; the latter has to do with the use of a statistical model relating ability and performance. The matching variable can either be observed or latent. An observed matching variable is a sum over the individual item scores, while a latent matching variable is an estimate of an underlying ability or trait. The second dimension has to do with the amount of structure the statistical methods impose on the data. Parametric methods assume that a mathematical model describes the relationship among the ability, the matching variable, and performance on individual items. Nonparametric methods do not employ an explicit mathematical model for this purpose (Penfield & Lam, 2000; Potenza & Dorans, 1995). Methods that fall under different classifications of the above scheme have different strengths and weaknesses and make different demands in terms of data. Therefore, Potnza & Dorans (1995) and Penfield & Lam (2000) recommended that an investigator can strengthen her/his study by using more than one method, and selecting methods that belong to different categories of the scheme. For example parametric methods tend to require more data and demand more complex computations than do the nonparametric methods.

Several factors affect the performance of DIF detection statistics. The most important ones are sample size and differences in the location, or mean ability, of the focal and reference groups on the relevant ability distribution. All the available methods require larger sample sizes for desired performance than are available for the present study. In some cases it is advised that at least 500 examinees would be required for acceptable performance of the statistics while in other cases sample size should be more than 2000 examinees. The probability of type I error increases as sample size gets smaller. (Penfield & Lam, 2000; Zwick, Thayer, & Mazzero, 1997) The other factor of primary relevance is the relative location of the focal and reference groups on the matching ability variable. The performance of most methods to correctly identify DIF items without flagging non-DIF items decreases as the means of the two groups are located farther apart on the ability scale (Penfield & Lam, 2000).

<u>Methods used for sex DIF</u>

Following the advice of Penfield & Lam (2000), four statistical methods were selected to investigate the presence of sex DIF, SIBTEST (Shealy & Stout, 1993; Stout & Roussos, 1996), Poly-SIBTEST (Chang, Mazzeo & Roussos, 1996), MHχ^2 (Mantel & Hanzel, 1959; Holland & Thayer, 1988), HW3 (Welch & Hoover, 1993). Two of them make use of the original polychotomous response scale of the items, poly-SIBTEST and HW3, while the other two use a dichotomised form of the item response scale, SIBTEST and MHχ^2. SIBTEST and Poly-SIBTEST use a latent ability variable but MHχ^2 and HW3 use an observed ability variable. All four methods are nonparametric. Due to insufficient data, the use of parametric methods was abandoned.

SIBTEST and Poly-SIBTEST (Shealy & Stout, 1993; Roussos & Stout, 1996; Chang, Mazzeo & Roussos, 1996) are nonparametric methods that use a latent ability variable to match individuals from the two groups under investigation. The method essentially tests whether an item measures a second dimension in addition to the primary unidimensional ability or trait measured by the test.

The MH method is designed for analysing the relationship between two dichotomous variables, while controlling for the influence of a third variable that has many levels. The method is an analysis of a series of 2 by 2 contingency tables that represent the group membership and the correct/incorrect responses to the studied items, over the levels on the ability variable. The method gives the MHχ^2 statistic, a test of the null hypothesis that the odds ratio is constant across all levels of the levels of the ability variable (Holland & Thayer, 1988).

The Hoover-Welch 3 (HW 3) statistic is a nonparametric method that uses an observed ability variable for matching designed for polychotomous response variables (Welch & Hoover, 1993). This statistic tests for differences between groups over levels of a matching variable, for example levels of test scores (Welch & Hoover, 1993). The method has been applied using an external ability variable in situations where the number of responses is limited (Welch & Hoover, 1993), but the number of items in the present study allows the use of the composite score as an observed ability variable.

Method used for ethnic DIF

The data at hand contained less than sufficient samples of students from two of the groups of interest for ethnic DIF, 91 African Americans students and 71 Hispanic students. Research on the sample size required for adequate performance of DIF procedures for dichotomous items indicates that sample sizes of about 200 for each examinee group are required for adequate detection rates (Mazor, Clauser, & Hambleton, 1992; Parshall & Miller, 1995). These studies have shown that with samples of 100 for each group, the procedure routinely fails to detect items with considerable DIF. For this reason Muňiz, Hambleton and Xing (2001) have suggested the use of procedures that do not rely on statistical significance tests. One such method is the Delta Plot. This method was one of the earliest methods used for DIF analysis (Angoff, 1982). It was more or less abandoned because it fails to control for differences in group mean ability level or impact (Angoff, 1982; Camilli & Shephard, 1994). The Delta plot method basically evaluates whether an items the difference in difficulty of an item differs from ‘a typical group difference in item difficulty’ (Camilli & Shepard, 1994, p.25). The item difficulties are transformed to Delta values and then plotted. The difficulty values are first transformed to normalized z-scores, then a linear transformation to a scale with a mean of 13 and standard deviation of 4 is conducted. In the present case the polytomous item difficulties had to be transformed to the range of item difficulties of dichotomous items before the method could be applied (Angoff, 1982; Camilli & Shepard, 1994).

The Delta plot method has the advantage of being applicable despite very small samples. Some authors advise against its use in general (Camilli & Shephard, 1994), while others recommend its use in situations where reasonable samples are unattainable (Scheuneman & Bleistein, 1989; Muňiz, Hambleton and Xing, 2001). The most recent reference recommending its use, Muňiz, Hambleton and Xing, (2001), has been in the area of translating or adapting tests to different languages or cultures. In this field, DIF analyses are useful at an early stage of the adaptation process when data are sparse, and the Delta method can be useful if its flaws are taken into account. This method was selected to conduct a 'quick-and-dirty' DIF analysis in the present study to provide an indication of potential DIF using the available data.

The reliability of the IELI scores

The fifth and final purpose of this study is an investigation of the reliability of the IELI. Reliability represents the stability of the test scores. It is defined as the ratio of true score variance to observed score variance (Feldt & Brennan, 1989). Reliability estimates can be obtained by several data collection designs, a single test administration, repeated testing using the same test form, and repeated testing using different test forms. The present study has a single administration design. The appropriate estimate of reliability was an internal consistency estimate, for example Cronbach's alpha. Internal consistency estimates of reliability address primarily sources of errors arising from sampling of items and carelessness or disturbances during administration.

Reliability of the six IELI scales was estimated, as well as the reliability of the total IELI scores. Estimates for the six cognitive behavior areas were obtained using Cronbach's alpha. Estimating the reliability of the IELI total score requires that the structure of the instrument as a composite of multiple cognitive behavior areas be taken into account. A formula for the reliability of composites of several subtests is the following:

$$_{comp.\alpha}r_{xx'} = \left(1 - \frac{S_{Ez}^2}{S_Z^2}\right) = \frac{n}{n-1}\left(1 - \frac{\sum w_i^2 S_{Xi}^2 \left(1 - {}_{\alpha}r_{xx'}\right)}{\sum w_i^2 S_{Xi}^2 + \sum\sum w_i w_j S_{XiXj}}\right),$$

where $_{comp.\alpha}r_{xx'}$ is a alpha reliability coefficient for composites scores, S_Z^2 is the variance of the composite score, $S_{E_Z}^2$ is the error variance for the composite score,

$S^2_{X_i}$ is the variance of a subtest or cognitive behavior area, $S_{X_i X_j}$ is the covariance among the subtests, and w_i are weights that depend on the number of items or score points for each subtest (Feldt & Brennan, 1989).

Research hypotheses

The hypotheses that are of interest in this study can be placed into five groups, four regarding aspects of the validity of the interpretations of the IELI and the fifth being estimation of the reliability of the IELI. These research questions are described below.

Several research hypotheses regarding the relations of IELI cognitive behavior areas with ITBS test scores were derived from the construct theories underlying each instrument and investigated:

1. The IELI total score would correlate significantly with the ITBS Composite score. Sample B was used to investigate this hypothesis.
2. The oral communication behavior area would correlate significantly with the ITBS language and listening tests. Also it would show a modest correlation with the 1st grade reading test. Sample B was used.
3. The written language behavior area would correlate significantly with the ITBS word analysis. A weak, but significant, relation with the 1st grade reading test was also expected. Sample B was used.
4. The IELI math concepts behavior area would correlate significantly with the ITBS math score. Sample B was used.
5. The IELI general knowledge, work habits and attentive behaviors areas would correlate significantly with the ITBS Composite score. Sample B was used.
6. Work habits and attentive behaviors were expected to be modestly but significantly correlated with each of the ITBS subtest scores. Sample B was used.

7. Each of the content matter related behavior areas would account for a significant portion of variance in the theoretically relevant ITBS scores in addition to what the other behavior areas account for. The general areas would account for specific variance in the ITBS scores in addition to that accounted for by the relevant subject matter areas. Multiple regression analysis was used. Sample B was used.
8. The general behavior areas would account for variance in the ITBS composite score in addition to variation accounted for by other behavior areas. Multiple regression analysis was used. Sample B was used.

The research hypothesis regarding the structure of the IELI is fairly complex, but was stated as:

9. The IELI has six factors corresponding to the cognitive behavior areas, each defined by a set of items written to measure that area. Items 1 through 5 define general behavior, items 6 through 13 define oral communication, items 14 through 19 define written language, items 20a through 28 define math concepts, items 29 through 35 define work habits, and items 36 through 42 define attentive behaviors. All IELI items would have high loadings on their respective factors, and lower loadings on other factors. Confirmatory factor analysis was used to evaluate several models of the above relations among the 44 items and six factors. Sample A was used for the confirmatory factor analysis.

The research hypotheses regarding sex and racial/ethnic bias were the following:

10. Each item on the IELI would be free of sex DIF. Poly-SIBTEST and HW3 methods were used to investigate the hypothesis. For reasons addressed

later, these methods were supplemented using SIBTEST and MHχ^2. Also, Kindergarten and 1st grade samples were combined for the investigation of this hypothesis. Sample C was used.

11. Each item on the IELI would be free of African American - White students DIF. The Delta plot method was used to investigate this hypothesis. Also, Kindergarten and 1st grade samples were combined for the investigation of this hypothesis. Sample C was used.
12. Each item on the IELI would be free of Hispanic -White DIF. The Delta method was used to investigate this hypothesis. Also, Kindergarten and 1st grade samples were combined for the investigation of this hypothesis. Sample C was used.
13. Each item on the IELI would be free of African American - Hispanic DIF. The Delta plot method was used to investigate this hypothesis. Also, Kindergarten and 1st grade samples were combined for the investigation of this hypothesis. Sample C was used.

The final purpose of this research is to obtain estimates of the reliability of the IELI cognitive behavioral areas and the total score.

14. The reliability of the total IELI score and each cognitive behavior area would be adequate. Sample A was used to estimate reliability.

CHAPTER IV
RESULTS

The results discussed in this chapter all pertain to the validity of the IELI. The chapter is divided into four parts. The first consisted of descriptive statistics and information on the reliability for the areas of cognitive behaviors identified. The second part discusses the relations of IELI scales with certain measures of academic achievement, the ITBS. These results are important for the evaluation of the concurrent validity of the IELI. The third part discusses of the structure of the IELI using confirmatory factor analysis. This is relevant for the evaluation of the construct validity of IELI. A final part of the results contains an investigation of the differential item functioning (DIF) with respect to sex and ethnic background for the IELI. This is an investigation of important potential sources of bias where gender and, to a limited extent, ethnic differences are investigated.

Descriptive statistics

Descriptive statistics for the cognitive behavior areas are shown in tables 6 and 7 for the kindergarten students and 1st grade students, respectively. In addition to the mean and standard deviation, the tables report the median and maximum scores. It should be noted that the medians were consistently greater than the means. This indicates that the data were negatively skewed as result of ceiling effects in all areas as well as the composite. In three of twelve instances, the medians were equal to the maximum score, indicating that 50% or more of the subjects received the maximum score on these scales. The skewness and kurtosis reported also reflected these properties of the score distributions. Due to this ceiling effect the data were problematic and were likely to lead to problems in subsequent analyses.

Girls were rated higher on all cognitive behavior areas than boys at both grade levels (see tables 8 and 9). At the kindergarten level the greatest difference

Table 6. Descriptive statistics for the IELI scores in the kindergarten sample.

Area	N	Mean	Median	St.dev.	Skewness	Kurtosis	Max
General knowledge	476	16.0	17	1.6	-2.0	3.8	17
Oral communication	480	29.1	31	4.0	-1.7	2.5	32
Written language	469	23.4	25	3.2	-1.5	1.7	26
Math concepts	468	31.0	32	4.1	-1.5	2.2	35
Work habits	483	24.9	26	3.4	-1.2	1.0	28
Attentive behaviors	481	25.2	27	3.6	-1.3	1.2	28
Composite	422	149.8	155	16.9	-1.4	1.5	166

Table 7. Descriptive statistics for the IELI scores in the 1st grade sample.

Area	N	Mean	Median	St.dev.	Skewness	Kurtosis	Max score
General knowledge	580	16.3	17	1.3	-2.4	6.5	17
Oral communication	578	29.1	31	3.8	-1.7	2.7	32
Written language	577	25.3	26	1.7	-3.4	13.7	26
Math concepts	555	32.3	33	3.1	-1.6	3.0	35
Work habits	577	24.8	26	3.4	-1.1	0.5	28
Attentive behaviors	577	24.8	26	3.7	-1.1	0.5	28
Composite	522	152.6	156	13.7	-1.5	2.7	166

occurred for attentive behaviors (girls 25.8 and boys 24.4), written language (girls 24.0 and boys 22.7), and work habits (girls 25.5 and boys 24.3). The greatest differences at the 1st grade level occurred for attentive behaviors (girls 25.3 and boys 24.2), work habits (girls 25.2 and boys 24.3), and oral communication (girls 29.3 and boys 28.9). The differences appear small, but are stistically significant. The differences accumulate for the composite score, resulting in score difference of 5.9 points (girls 152.4 and boys 146.4) at kindergarten and 2.9 points (girls 154.0 and boys 151.1) in 1st grade. An investigation focusing on the severity of the sex differences is reported later in this chapter.

Among the information collected in addition to the IELI scores were student scores on the ITBS. These data served as criteria in the evaluation of the validity of the IELI cognitive behavior areas that were part of this study.

Table 8. Descriptive statistics for kindergarten boys and girls on the IELI.

	Boys			Girls		
Area	N	Mean	St.dev.	N	Mean	St.dev.
General knowledge	222	15.8	1.8	254	16.2	1.4
Oral communication	221	28.7	4.3	259	29.5	3.7
Written language	216	22.7	3.5	253	24.0	2.7
Math concepts	219	30.5	4.4	249	31.4	3.7
Work habits	227	24.3	3.5	256	25.5	3.1
Attentive behaviors	226	24.4	3.9	255	25.8	3.2
Composite	189	146.5	17.8	233	152.4	15.6

Table 9. Descriptive statistics for 1st grade boys and girls on the IELI.

	Boys			Girls		
Area	N	Mean	St.dev.	N	Mean	St.dev.
General knowledge	284	16.2	1.4	296	16.4	1.2
Oral communication	286	28.9	4.1	292	29.3	3.6
Written language	283	25.2	1.8	294	25.4	1.5
Math concepts	272	32.4	3.2	283	32.3	3.1
Work habits	286	24.3	3.5	291	25.2	3.1
Attentive behaviors	285	24.2	3.9	292	25.3	3.6
Composite	256	151.1	14.5	266	154.0	12.7

Descriptive statistics for the ITBS raw scores can be found in tables 10 and 11 for the Kindergarten and 1st grade levels, respectively.

Correlations among the IELI cognitive behavior areas

All six IELI cognitive behavior areas related with each other to a certain extent. In Kindergarten, the average correlation was $r = .66$ (median $r = .64$); in 1st grade the average correlation was $r = .59$ (median $r = .56$) (see tables 12 and 13). Two important observations should be made here. First, the correlations were strong, and second, the correlations were weaker in the 1st grade sample than the kindergarten sample. Several factors, or a combination of them, may account for the strength of the correlations. One methodological reason was the

Table 10. Descriptive statistics for ITBS raw scores for the kindergarten sample.

	N	Mean	St.dev.
Vocabulary	167	21.3	3.8
Listening	137	20.7	4.5
Language	122	19.6	5.2
Math total	132	23.1	4.7
Word analysis	162	23.8	6.0
Composite	122	136.4	11.0

Table 11. Descriptive statistics for ITBS raw scores for the 1st grade sample.

	N	Mean	St.dev.
Vocabulary	190	20.9	5.9
Reading	74	24.5	9.3
Reading total	110	40.2	7.9
Listening	190	22.6	3.9
Language	175	23.8	6.3
Math total	111	29.8	4.0
Word analysis	190	27.9	6.0
Composite	175	155.8	17.5

Note: Level 6 and level 7 data are combined.

fact that all areas were measured as ratings by the same teachers based on their perceptions of the student. Some aspects of the possible structure underlying cognitive behaviors may also have accounted for the magnitude of the correlations. For example the behavior areas may not be differentiated in the young students, or a general trend in student behavior is reflected to some extent in the six cognitive behavior areas. The correlations obtained from the current data offered no support for one rather than another of these possibilities. The most likely explanation for lower correlations in the 1st grade sample is that it had lesser variability than the kindergarten sample. But both student behavior and rating issues could also offer explanations regarding the difference in overall level of correlations in Kindergarten and in 1st grade.

Correlations of items with the cognitive behavior areas.

The correlations of individual items with the IELI cognitive behavior areas in the kindergarten and 1st grade samples are reported in tables 14 and 15. In general, the greatest correlations for each item corresponded with the behavior area it was intended to measure. Exceptions were items 20b and 21 in the first grade sample. The problem was not replicated in the kindergarten sample. In addition, item 35 had almost as great a correlation with a second behavior area.

On average, the correlations obtained from the kindergarten sample were greater than those from the 1st grade sample, average correlations were $r = 0.53$ (SD = 0.13) for the kindergarten sample and $r = 0.47$ (SD = 0.15) for the 1st grade sample. The respective medians were $r = 0.52$ and $r = 0.44$. The most likely explanation for lower correlations in the 1st grade sample is that it had lesser variability than the kindergarten sample.

There was considerable variation in the general magnitude of the correlations of the behaviors within one area with those of a second area. For example, the correlations of items within work habits and attentive behaviors were greater than those in other areas. To a lesser extent, a similar situation occurred for items within the general behavior and math concepts areas. This may have reflected the similarity of the constructs measured by these areas.

One last observation is that several items related strongly to all behavior areas, e.g. items 4, 5, 12, 11, 35 and 42. This may have reflected the fact that some items captured general aspects of cognitive behaviors, e.g. item 42, or behaviors strongly related to cognitive development in general.

Reliability of the IELI

An important psychometric property of scores is their reliability. It is important to estimate the extent to which the IELI scores reflect the student attribute of interest and to what extent they are contaminated by error. The reliability coefficients estimated from the current sample are reported in tables 16

Table 12. Correlations among the IELI behavior areas at the Kindergarten level.

Area	General Knowledge	Oral Communication	Written Language	Math Concepts	Work Habits
Oral communication	0.72				
Written language	0.70	0.67			
Math concepts	0.77	0.72	0.73		
Work habits	0.54	0.64	0.62	0.63	
Attentive behaviors	0.57	0.58	0.61	0.60	0.86

N rangesfrom 452 to 483.

All table entries are significantly greater than zero, $p < .01$.

Table 13. Correlations among the IELI behavior areas at the 1st grade level.

Area	General Knowledge	Oral Communication	Written Language	Math Concepts	Work Habits
Oral communication	0.67				
Written language	0.67	0.55			
Math concepts	0.68	0.64	0.62		
Work habits	0.48	0.57	0.50	0.50	
Attentive behaviors	0.48	0.56	0.52	0.53	0.86

N ranesfrom 546 to 580.

All table entries are significantly greater than zero, $p < .01$.

and 17 along with the number of items in each cognitive behavior area, the potential score range and the number of individuals responding to all items within that area. The greatest reliability coefficients for both Kindergarten and 1st grade samples were found for oral communication and attentive behaviors, being greater than 0.90. The reliability coefficients for written language, math concepts, and work habits ranged from 0.83 to 0.87. Finally, the reliability of general knowledge was found to be 0.76 in Kindergarten and 0.74 in 1st grade. One factor that may contribute the relatively low reliability of general knowledge is that this area has the fewest items.

Table 14. Correlations of IELI items with the behavior area scores in kindergarten.

Area	Item	GK	OC	WL	MC	WH	AB
General knowledge	I01	**.53**	.49	.51	.54	.39	.39
General knowledge	I02	**.32**	.17	.30	.29	.11	.11
General knowledge	I03	**.60**	.50	.47	.50	.32	.37
General knowledge	I04	**.72**	.64	.59	.69	.48	.50
General knowledge	I05	**.68**	.64	.63	.69	.54	.53
Oral communication	I06	.58	**.67**	.49	.56	.47	.43
Oral communication	I07	.61	**.77**	.55	.62	.56	.52
Oral communication	I08	.53	**.72**	.46	.49	.49	.39
Oral communication	I09	.61	**.82**	.56	.59	.55	.49
Oral communication	I10	.65	**.81**	.65	.63	.56	.55
Oral communication	I11	.62	**.82**	.63	.64	.52	.49
Oral communication	I12	.64	**.75**	.65	.66	.60	.57
Oral communication	I13	.61	**.50**	.50	.61	.52	.44
Written language	I14	.41	.31	**.42**	.35	.29	.32
Written language	I15	.40	.32	**.50**	.34	.25	.27
Written language	I16	.58	.59	**.69**	.58	.51	.48
Written language	I17A	.58	.52	**.74**	.59	.51	.52
Written language	I17B	.54	.49	**.71**	.59	.50	.49
Written language	I18	.57	.56	**.71**	.63	.52	.49
Written language	I19	.56	.56	**.71**	.60	.53	.53
Math concepts	I20A	.38	.32	.34	**.37**	.21	.15
Math concepts	I20B	.53	.49	.55	**.56**	.36	.38
Math concepts	I21	.52	.47	.56	**.49**	.44	.44
Math concepts	I22	.61	.48	.58	**.56**	.36	.38
Math concepts	I23	.51	.49	.52	**.58**	.43	.42
Math concepts	I24	.67	.63	.58	**.74**	.48	.45
Math concepts	I25	.66	.56	.60	**.73**	.51	.50
Math concepts	I26	.50	.48	.51	**.68**	.49	.43
Math concepts	I27	.58	.60	.56	**.78**	.53	.48
Math concepts	I28	.64	.57	.54	**.71**	.49	.47
Work habits	I29	.35	.50	.34	.36	**.35**	.31
Work habits	I30	.43	.44	.49	.47	**.67**	.74
Work habits	I31	.47	.56	.52	.52	**.71**	.64
Work habits	I32	.45	.50	.55	.54	**.72**	.74
Work habits	I33	.29	.32	.33	.37	**.55**	.59
Work habits	I34	.34	.46	.43	.48	**.71**	.67
Work habits	I35	.48	.53	.54	.52	**.76**	.82
Attentive behaviors	I36	.48	.52	.48	.51	.74	**.75**
Attentive behaviors	I37	.43	.44	.47	.46	.76	**.86**
Attentive behaviors	I38	.48	.47	.49	.49	.70	**.74**
Attentive behaviors	I39	.47	.49	.51	.47	.72	**.84**
Attentive behaviors	I40	.44	.43	.43	.46	.65	**.71**
Attentive behaviors	I41	.43	.47	.53	.48	.73	**.81**
Attentive behaviors	I42	.63	.62	.66	.65	.78	**.76**

Note: N ranges from 461 to 483. For all table entries, $p < .01$.

Table 15. Correlations of IELI items with the behavior area scores in 1st grade.

Area	Item	GK	OC	WL	MC	WH	AB
General knowledge	I01	**.48**	.45	.44	.43	.34	.33
General knowledge	I02	**.41**	.35	.41	.24	.29	.27
General knowledge	I03	**.61**	.50	.44	.47	.30	.28
General knowledge	I04	**.72**	.63	.56	.62	.38	.40
General knowledge	I05	**.54**	.47	.56	.55	.41	.43
Oral communication	I06	.53	**.54**	.44	.40	.40	.37
Oral communication	I07	.49	**.67**	.39	.44	.48	.48
Oral communication	I08	.42	**.69**	.32	.40	.37	.34
Oral communication	I09	.56	**.76**	.42	.49	.40	.42
Oral communication	I10	.60	**.78**	.51	.57	.48	.52
Oral communication	I11	.62	**.78**	.56	.58	.47	.51
Oral communication	I12	.55	**.78**	.48	.64	.57	.56
Oral communication	I13	.57	**.42**	.42	.54	.45	.40
Written language	I14	.17	.13	**.37**	.27	.18	.19
Written language	I15	.20	.21	**.35**	.28	.26	.25
Written language	I16	.60	.49	**.71**	.56	.42	.42
Written language	I17A	.54	.37	**.69**	.43	.38	.42
Written language	I17B	.50	.37	**.67**	.43	.39	.41
Written language	I18	.57	.41	**.72**	.47	.39	.41
Written language	I19	.61	.52	**.67**	.58	.47	.48
Math concepts	I20A	.23	.09	.27	**.28**	.22	.19
Math concepts	I20B	.34	.20	.46	**.37**	.24	.24
Math concepts	I21	.32	.17	.43	**.35**	.25	.26
Math concepts	I22	.51	.37	.47	**.48**	.35	.36
Math concepts	I23	.43	.43	.38	**.56**	.31	.32
Math concepts	I24	.60	.57	.48	**.72**	.43	.41
Math concepts	I25	.59	.53	.48	**.69**	.41	.43
Math concepts	I26	.47	.49	.45	**.70**	.41	.41
Math concepts	I27	.58	.60	.51	**.76**	.40	.45
Math concepts	I28	.53	.48	.51	**.70**	.44	.47
Work habits	I29	.29	.44	.24	.24	**.36**	.33
Work habits	I30	.39	.46	.42	.38	**.61**	.67
Work habits	I31	.35	.37	.37	.35	**.71**	.66
Work habits	I32	.41	.50	.43	.47	**.75**	.73
Work habits	I33	.19	.23	.26	.26	**.51**	.52
Work habits	I34	.32	.36	.34	.38	**.70**	.70
Work habits	I35	.43	.48	.47	.45	**.70**	.79
Attentive behaviors	I36	.46	.47	.44	.48	.71	**.72**
Attentive behaviors	I37	.39	.47	.42	.44	.79	**.84**
Attentive behaviors	I38	.42	.49	.47	.43	.70	**.74**
Attentive behaviors	I39	.37	.46	.39	.41	.75	**.84**
Attentive behaviors	I40	.31	.37	.37	.34	.68	**.75**
Attentive behaviors	I41	.33	.39	.36	.38	.73	**.80**
Attentive behaviors	I42	.56	.64	.58	.59	.71	**.73**

Note: N ranges from 549 to 580. For all table entries, $p < .01$.

Table 16. Reliability of the IELI scores for the kindergarten sample.

Area	N	Reliability	Number of items	Number of score points
General knowledge	476	0.76	5	13
Oral communication	480	0.93	8	25
Written language	469	0.85	7	20
Math concepts	468	0.87	10	26
Work habits	483	0.86	7	22
Attentive behaviors	481	0.93	7	22
Composite	422	0.97	6	123

Table 17. Reliability of the IELI scores for the 1st sample.

Area	N	Reliability	Number of items	Number of score points
General knowledge	580	0.74	5	13
Oral communication	578	0.91	8	25
Written language	577	0.83	7	20
Math concepts	555	0.84	10	26
Work habits	577	0.85	7	22
Attentive behaviors	577	0.92	7	22
Composite	522	0.97	6	123

Criterion related validity:
IELI and academic achievement

In this part of the chapter, the results of investigations pertaining to the criterion related validity of the IELI scores are reported. First, there are results indicating that each area correlated to a certain extent with academic achievement scores. Secondly, the results of investigations of the extent to which the variance accounted for by each IELI behavior area relative to its theoretically relevant academic achievement measures overlapped with the variance accounted for by the other behavior areas are presented in the following section.

The correlations among the IELI behavior areas and the ITBS scores are reported in tables 18 and 19. All correlations were significantly greater than zero at the .01 level. Their magnitudes ranged from $r = .19$ to $r = .66$ in the kindergarten sample and from $r = .16$ to $r = .54$ in the 1st grade sample (see individual correlation coefficients and *p*-values in tables 18 and 19). It is worth pointing out that only in six out of 13 table columns did the greatest correlation occur where the IELI construct theory postulated a strong relationship between a behavior area and ITBS score. Those were word analysis and math concepts at the kindergarten level and reading vocabulary, reading total, listening and word analysis at the 1st grade level. The first order correlations among IELI and ITBS are not appropriate to investigate the theoretically relevant relationships among scales on the two measures. To do so it is necessary to investigate their relationship independently of the other cognitive behavior areas. Such analysis is reported later in this chapter.

Finally, the correlations of the IELI composite score, as well as the behavior areas, with the ITBS composite are given for each test level of the ITBS in table 20. These correlations ranged from $r = .21$ to $r = .61$, with median correlations of $r = .43$ at level 5, $r = .31$ at level 6, and $r = .42$ at level 7. The first observation about the table is that the correlations were systematically lower for level 6 than levels 5 and 7. Secondly, at all test levels, the correlation of at least one behavior area with the ITBS composite exceeded the correlation of the IELI composite with the ITBS composite. Those behavior areas were written language at level 5, attentive behaviors and oral communication at level 6, and general knowledge at level 7. Finally the rankings of the correlations of individual behavior areas with the ITBS composite were unstable, with the exception of math concepts that ranked close to the median correlation for all three test levels.

Table 18. Correlations among IELI behavior areas and ITBS scores in the kindergarten sample.

Area	Reading vocabulary			Listening			Word analysis		
	r	*p*-value	N	*R*	*p*-value	N	*r*	*p*-value	N
General knowledge	.35	.000	159	.39	.000	130	.55	.000	154
Oral communication	.38	.000	163	.35	.000	134	.60	.000	158
Written language	.39	.000	158	.39	.000	129	.61	.000	153
Math concepts	.40	.000	160	.45	.000	131	.60	.000	155
Work habits	.22	.005	164	.27	.001	137	.48	.000	159
Attentive behaviors	.26	.001	165	.25	.004	136	.42	.000	160
IELI composite	.40	.000	139	.41	.000	114	.66	.000	134

Table 18. Continued.

Area	Math total			Language			IELI composite		
	r	*p*-value	N	*R*	*p*-value	N	*r*	*p*-value	N
General knowledge	.51	.000	127	.35	.000	117	.43	.000	94
Oral communication	.51	.000	129	.35	.000	119	.39	.000	94
Written language	.55	.000	125	.41	.000	115	.41	.000	94
Math concepts	.59	.000	127	.42	.000	117	.44	.000	94
Work habits	.43	.000	132	.19	.038	122	.32	.002	94
Attentive behaviors	.41	.000	132	.24	.007	122	.32	.002	94
IELI composite	.61	.000	114	.41	.000	122	.45	.000	94

Table 19. Correlations among IELI behavior areas and ITBS scores in the 1st grade sample.

	Reading vocabulary			Reading Comprehension			Reading Total		
Area	*r*	*p*-value	N	*r*	*p*-value	N	*r*	*p*-value	N
General knowledge	.40	.000	187	.59	.000	73	.27	.005	108
Oral communication	.41	.000	188	.45	.000	72	.42	.000	110
Written language	.28	.000	187	.52	.000	73	.49	.000	108
Math concepts	.22	.002	184	.54	.000	70	.34	.000	108
Work habits	.27	.000	188	.56	.000	73	.30	.001	109
Attentive behaviors	.31	.000	188	.58	.000	74	.43	.000	108
IELI composite	.40	.000	190	.66	.000	74	.45	.000	110

Table 19. Continued.

	Listening			Word Analysis			Math Total		
Area	*r*	*p*-value	N	*r*	*p*-value	N	*r*	*p*-value	N
General knowledge	.39	.000	187	.44	.000	187	.48	.000	109
Oral communication	.40	.000	188	.38	.000	188	.44	.000	111
Written language	.21	.005	187	.46	.000	187	.27	.005	109
Math concepts	.37	.000	184	.34	.000	184	.47	.000	109
Work habits	.31	.000	188	.40	.000	188	.37	.000	110
Attentive behaviors	.35	.000	188	.39	.000	188	.44	.000	109
IELI composite	.45	.000	190	.49	.000	190	.54	.000	111

Table 19. Continued.

	Language			ITBS composite		
Area	*r*	*p*-value	N	*r*	*p*-value	N
General knowledge	.32	.000	172	.38	.000	157
Oral communication	.24	.001	173	.40	.000	157
Written language	.16	.037	172	.20	.011	157
Math concepts	.33	.000	169	.30	.000	157
Work habits	.27	.000	173	.25	.002	157
Attentive behaviors	.30	.000	173	.36	.000	157
IELI composite	.35	.000	175	.42	.000	157

Note: Both levels 6 and 7 of the ITBS were used.

Table 20. Correlations of the IELI behavior areas and the IELI composite with the ITBS composite at test levels 5, 6 and 7.

Area	Level 5[1]			Level 6[2]			Level 7[2]		
	r	*p*-value	N	*R*	*p*-value	N	*r*	*p*-value	N
IELI Composite	.47	.000	73	.34	.001	102	.53	.000	45
General Knowledge	.45	.000	78	.29	.002	109	.61	.000	53
Oral Communication	.43	.000	82	.43	.000	111	.40	.003	52
Written Language	.51	.000	81	.21	.029	109	.35	.010	53
Math Concepts	.42	.000	80	.31	.001	109	.44	.001	50
Work Habits	.33	.002	83	.29	.002	110	.47	.000	53
Attentive Behaviors	.35	.001	83	.39	.000	109	.44	.001	54

[1] All students tested with level 5 were in kindergarten.

[2] Correlations for level 6 and 7 are reported only for students in 1st grade.

Construct validity: Relation of IELI areas with academic achievement

The evidence of construct validity discussed next consists of an investigation of the variation in achievement scores accounted for by each IELI behavior area and not shared with the other behavior areas. This provided information regarding the extent to which a given behavior area supplied information relevant to a given facet of academic achievement over and above what was supplied by the other areas. These questions were investigated using regression methods. On the one hand, the hypotheses predicted that the subject matter oriented behavior areas provided specific information on selected achievement measures. On the other hand, it was also predicted that the general behavior areas provided specific information on the same academic achievement measures in addition to that provided by the relevant subject matter behavior area. Separate regression equations were constructed as described in Chapter III. The objective of these analyses was to evaluate the specific relationship of the subject matter oriented cognitive behavior areas with their theoretically relevant achievement scores. In order to achieve this the cognitive behavior area of interest was entered as a second predictor variable in the regression equations, using each of the other cognitive behavior areas, in turn, as the first predictor variable. As explained above, the results of interest were whether the cognitive behavior area of interest accounted for a significant amount of variance in the achievement scores in addition to that accounted for by each of the other behavior areas. In the following sections results for the subject matter oriented behavior areas are presented first, oral communication, written language and math concepts. Then, the results for the three general academic behavior areas are reported, general knowledge, work habits, and attentive behaviors.

Oral communication

The theoretically relevant relation of oral communication was determined, based on the construct theory, to be with listening at both grade levels and with reading measures at the 1st grade level. Among the ITBS scores obtained for the present sample were a listening score and two measures of reading.

Oral communication and listening

ITBS listening scores were available for 137 students at the kindergarten level and 190 students the 1st grade level. A regression analysis was conducted at both grade levels with the ITBS Listening score as a dependent variable. Tables 21 and 22 report the results of these analyses. At the 1st grade level, oral communication accounted for a significant portion of the variance in listening in addition to that accounted for by any of the other behavior areas (see *F*-values and *p*-values in table 22). At the kindergarten level, oral communication accounted for a significant portion of the ITBS listening variance only in addition to two of the general behavior areas, work habits and attentive behaviors (see *F*-values and *p*-values in table 21). The magnitudes of the independent contributions of oral communication were, in general, also greater at the 1st grade level than at the kindergarten level. The significant R^2 changes at the kindergarten level ranged from $R^2 = .04$ to $R^2 = .05$. At the 1st grade level these

Table 21. Regression of oral communication on kindergarten ITBS listening with each of the other IELI behavior areas as the first predictor.

Area	*r*	*R*	R^2 change	N	*F* change	df_1	df_2	*p*-value.
General Knowledge	0.414	0.415	0.001	126	0.204	1	124	0.652
Written Language	0.374	0.391	0.013	126	1.834	1	124	0.178
Math Concepts	0.439	0.440	0.001	127	0.236	1	125	0.628
Work Habits	0.298	0.356	0.038	133	5.197	1	131	0.019
Attentive Behaviors	0.269	0.354	0.053	132	7.841	1	130	0.006

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 22. Regression of oral communication on 1st grade ITBS listening with each of the other IELI behavior areas as the first predictor.

Area	*r*	*R*	R^2 change	N	*F* change	df_1	df_2	*p*-value.
General Knowledge	0.370	0.446	0.062	184	14.021	1	182	0.000
Written Language	0.192	0.418	0.138	184	30.351	1	182	0.000
Math Concepts	0.360	0.440	0.064	181	14.192	1	179	0.000
Work Habits	0.285	0.419	0.094	185	21.024	1	183	0.000
Attentive Behaviors	0.337	0.426	0.068	185	15.152	1	183	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

values were in the $R^2 = .06$ to $R^2 = .14$ range. Recall that the correlations of the IELI subtests were generally higher in the kindergarten sample

<u>Oral communication and reading</u>

The construct theory of the IELI supports predictions of specific relations of oral communication with reading measures at the 1st grade level. Two measures of reading among the ITBS scores obtained may provide information regarding this hypothesis, reading vocabulary and reading total. Reading vocabulary scores were available for 190 students and reading total scores for 74 students. Thus, two sets of regression analyses were conducted to investigate the specific relationship of oral communication with reading achievement. These are reported in tables 23 and 24. A significant increase in R^2 was obtained for both reading vocabulary and reading total when oral communication was added to the regression equations for each of the five other cognitive behavior areas (see *F*-values and *p*-values in tables 23 and 24). The significant R^2 change values ranged from $R^2 = .07$ to $R^2 = .14$ for reading vocabulary and for reading total from $R^2 = .06$ to $R^2 = .12$.

Written language

The analysis of the construct of written language was the basis of hypotheses that written language would account for a significant portion of the variance of the ITBS word analysis scores at both grade levels and for reading at the 1st grade level. These hypotheses were tested using a series of regression analyses.

Written language and word analysis

The data contained word analysis scores for 162 kindergarten students and 190 1st grade students. Results of the regression analysis for the kindergarten sample are reported in table 25, and table 26 contains results for the 1st grade sample. The R^2 change attributed to written language, after entering each of the other IELI behavior areas as first predictors, was significant beyond the .01 level in all five regressions at both grade levels (see *F*-values and *p*-values in tables 25 and 26). The magnitude of the R^2 change was in the $R^2 = .06$ to $R^2 = .08$ range for two equations, in the $R^2 = .11$ to $R^2 = .14$ for six regressions, and close to $R^2 = .20$ for the remaining two. At the kindergarten level the R^2 change was considerably larger when work habits and attentive behaviors were first predictors than when the other three were first predictors, while the magnitudes of the R^2 change were similar over all regressions at the 1st grade level.

Table 23. Regression of oral communication on ITBS 1st grade reading vocabulary with each of the other IELI areas as the first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.377	0.465	0.074	184	17.086	1	182	0.000
Written language	0.269	0.457	0.136	184	31.458	1	182	0.000
Math concepts	0.217	0.424	0.133	181	29.019	1	179	0.000
Work habits	0.234	0.406	0.110	185	24.070	1	183	0.000
Attentive behaviors	0.291	0.419	0.091	185	20.061	1	183	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 24. Regression of oral communication on ITBS 1st grade reading total with each of the other IELI behavior areas as the first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.267	0.442	0.124	107	16.223	1	105	0.000
Written language	0.487	0.549	0.064	107	9.640	1	105	0.002
Math concepts	0.336	0.446	0.086	107	11.296	1	105	0.001
Work habits	0.303	0.426	0.090	108	11.647	1	106	0.001
Attentive behaviors	0.425	0.487	0.057	107	7.688	1	105	0.007

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 25. Regression of written language on kindergarten ITBS word analysis with each of the other IELI behavior areas as the first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.564	0.657	0.114	144	28.150	1	142	0.000
Oral communication	0.612	0.661	0.062	150	16.442	1	148	0.000
Math concepts	0.600	0.662	0.078	147	20.046	1	145	0.000
Work habits	0.491	0.662	0.197	149	35.041	1	147	0.000
Attentive behaviors	0.436	0.615	0.188	150	44.995	1	148	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 26. Regression of written language on 1st grade ITBS word analysis with each of the other IELI areas as the first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.423	0.550	0.124	183	32.017	1	181	0.000
Oral communication	0.369	0.519	0.133	184	33.178	1	182	0.000
Math concepts	0.310	0.483	0.137	180	31.674	1	178	0.000
Work habits	0.379	0.508	0.114	184	27.938	1	182	0.000
Attentive behaviors	0.368	0.490	0.105	184	25.188	1	182	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 27. Regression of written language on kindergarten ITBS reading vocabulary with each of the other IELI areas as first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.389	0.427	0.031	149	5.504	1	147	0.020
Oral Communication	0.391	0.417	0.021	154	3.891	1	152	0.050
Math concepts	0.425	0.453	0.025	152	4.636	1	150	0.033
Work Habits	0.236	0.386	0.093	154	16.635	1	152	0.000
Attentive Behaviors	0.272	0.395	0.082	155	14.847	1	153	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 28. Regression of written language on 1st grade ITBS reading vocabulary with each of the other IELI areas as first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.419	0.450	0.027	183	6.178	1	181	0.014
Oral Communication	0.430	0.457	0.024	184	5.633	1	182	0.019
Math concepts	0.229	0.305	0.041	180	7.976	1	178	0.005
Work Habits	0.274	0.334	0.036	184	7.431	1	182	0.007
Attentive Behaviors	0.325	0.356	0.021	184	4.412	1	182	0.037

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Written language and reading

Results of the regressions of written language on ITBS reading are reported in table 27 for the kindergarten sample, and the 1st grade results are reported in table 28. The R^2 change values in reading vocabulary attributed to written language, after entering each of the other IELI behavior areas as a first predictor, were significant in all five regressions in both the kindergarten (see *F*-values and *p*-values in table 27) and 1st grade samples (see *F*-values and p-values in table 28). The magnitudes of the increase in explained variances were generally in the $R^2 = .02$ to $R^2 = .04$ range, but two of them were in the $R^2 = .08$ to $R^2 = .09$ range. The two larger values were obtained when work habits and attentive behaviors were first predictors at the kindergarten level.

Math Concepts

Predictions were drawn from the IELI construct theory regarding the relationship of math concepts with academic achievement in math. These predictions were tested using the ITBS math total score as a dependent variable. Results are reported in table 29 for the kindergarten sample and in table 30 for the 1st grade sample. Math total scores were obtained for a sample of 132 students at the kindergarten level, and for 111 students at the 1st grade level. The R^2 change values in ITBS math total attributed to math concepts, after the other IELI behavior areas were entered as first predictors, were significant at least at the .01 level in all five regressions at the kindergarten level (see *F*-values and *p*-values in table 29). At the 1st grade level, four of the five R^2 increases were significant at the .01 level while the remaining one was significant at the .05 level (see *F*-values and *p*-values in table 30). The magnitudes of the R^2 change were generally in the .08 to .11 range, but at the kindergarten level two of them were close to $R^2 = .20$. Finally, one R^2 at the 1st grade level was = .03. The two largest values were obtained when work habits and attentive behaviors were first predictors at the kindergarten level.

The general behavior areas

Predictions were derived from the IELI construct theory that the general cognitive behavior areas added specific variance in academic achievement, not accounted for by the subject matter oriented cognitive behavior areas. These hypotheses were tested in the same way as the hypotheses regarding the subject matter oriented areas and discussed in the previous sections. In the analyses discussed next, each of the general cognitive behavior areas was added as a second predictor variable in a regression equation, after the theoretically relevant subject matter behavior area was added as first predictor.

Table 29. Regression of math concepts on kindergarten ITBS math total with each of the other IELI behavior areas as the first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.514	0.587	0.080	121	14.510	1	119	0.000
Oral communication	0.514	0.593	0.087	123	16.309	1	121	0.000
Written language	0.559	0.640	0.097	120	19.456	1	118	0.000
Work habits	0.435	0.605	0.177	126	34.609	1	124	0.000
Attentive behaviors	0.417	0.606	0.193	126	38.022	1	124	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 30. Regression of math concepts on ITBS 1st grade math total with each of the other IELI behavior areas as the first predictor.

Area	r	R	R^2 change	N	F change	df_1	df_2	p-value.
General knowledge	0.479	0.512	0.033	106	4.558	1	104	0.035
Oral communication	0.434	0.524	0.086	108	12.641	1	106	0.001
Written language	0.266	0.394	0.084	106	10.376	1	104	0.002
Work habits	0.366	0.491	0.107	107	14.751	1	105	0.000
Attentive behaviors	0.433	0.528	0.091	106	13.190	1	104	0.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

General knowledge and academic achievement

The results of the regression analysis for general knowledge are reported in table 31 for the kindergarten sample and table 32 for the 1st grade sample. At the kindergarten level, the R^2 change attributable to general knowledge was statistically significant for three regression equations; reading total with oral communication as the first predictor, and word analysis and reading vocabulary with written language as the first predictor (see *F*-values and *p*-values in table 31). The magnitudes of the R^2 change were $R^2 = .07$ for the first and $R^2 = .03$ for the latter two. Five of the R^2 changes at the 1st grade level were statistically significant (see *F*-values and *p*-values in table 32). At the 1st grade level, the significant R^2 change values ranged from.04 or .08.

Work habits and academic achievement

The results of the regression analysis for work habits are reported in table 33 for the kindergarten sample and table 34 for the 1st grade sample. At the kindergarten level, the R^2 change values attributable to work habits were significant at the .05 level for two regression equations, word analysis with written language as the first predictor and math total with math concepts as the first predictor (see *F*-values and *p*-values in table 33). The magnitudes of the R^2 change values were $R^2 = .02$ in both cases. At the 1st grade level three R^2 change values involving work habits were statistically significant, oral communication with listening as first predictor, written language with word analysis as first predictor, and math concepts with math total as first predictor (see *F*-values and *p*-values in table 34). The significant R^2 values ranged from .02 to .05.

Attentive behaviors and academic achievement

The results of the regression analysis for attentive behaviors are reported in table 35 for the kindergarten sample and table 36 for the 1st grade sample. At the kindergarten level, the increase in R^2 for ITBS math total with math concepts

Table 31. Regression of general knowledge on kindergarten ITBS achievement scores with the subject matter oriented IELI areas as first predictors.

Area and ITBS score	r	R	R^2 change	N	F change	Df_1	df_2	p-value.
OC - Reading vocabulary	0.366	0.384	0.014	154	2.357	1	152	0.127
OC – Reading total	0.327	0.418	0.068	126	9.834	1	124	0.002
OC – Listening	0.343	0.384	0.030	113	3.894	1	111	0.051
WL – Word analysis	0.632	0.657	0.032	144	7.875	1	142	0.006
WL – Reading vocabulary	0.394	0.427	0.027	149	4.808	1	147	0.030
MC – Math total	0.580	0.587	0.008	121	1.422	1	119	0.236

OC = Oral communication; WL = Written language; MC = Math concepts.

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 32. Regression of general knowledge on ITBS 1st grade achievement scores with the subject matter oriented IELI areas as first predictors.

Area and ITBS score	r	R	R^2 change	N	F change	df_1	df_2	p-value.
OC - Reading vocabulary	0.423	0.465	0.037	184	8.668	1	182	0.004
OC – Reading total	0.399	0.446	0.040	184	8.891	1	182	0.003
OC – Listening	0.237	0.315	0.043	169	7.977	1	167	0.005
WL – Word analysis	0.468	0.550	0.083	183	21.786	1	181	0.000
WL – Reading total	0.491	0.501	0.010	105	1.256	1	103	0.265
MC – Math total	0.457	0.512	0.053	106	7.439	1	104	0.007

OC = Oral communication; WL = Written language; MC = Math concepts.

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 33. Regression of work habits on kindergarten ITBS achievement scores with the subject matter oriented IELI areas as first predictors.

Area and ITBS score	r	R	R^2 change	N	F change	df_1	df_2	p-value.
OC – Reading vocabulary	0.370	0.371	0.001	159	0.088	1	157	0.767
OC – Reading total	0.346	0.356	0.007	133	1.052	1	131	0.307
OC –Listening	0.347	0.350	0.002	118	0.218	1	116	0.641
WL - Word analysis	0.606	0.622	0.020	149	4.979	1	147	0.027
WL - Reading vocabulary	0.386	0.386	0.000	154	0.000	1	152	0.997
MC – Math total	0.587	0.605	0.021	126	4.293	1	124	0.040

OC = Oral communication; WL = Written language; MC = Math concepts.

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 34. Regression of work habits on 1st grade ITBS achievement scores with the subject matter oriented IELI areas as first predictors.

Area and ITBS score	r	R	R^2 change	N	F change	df_1	df_2	p-value.
OC – Reading vocabulary	0.404	0.406	0.002	184	0.287	1	182	0.593
OC – Reading total	0.409	0.419	0.008	185	1.857	1	183	0.175
OC –Listening	0.243	0.285	0.022	170	4.066	1	168	0.045
WL - Word analysis	0.458	0.508	0.048	184	11.715	1	182	0.001
WL - Reading total	0.485	0.493	0.008	106	1.091	1	104	0.299
MC – Math total	0.465	0.491	0.025	107	3.355	1	105	0.070

OC = Oral communication; WL = Written language; MC = Math concepts.

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

as the first predictor and attentive behaviors as the second predictor was significant at the .05 level, (see *F*-values and *p*-values in table 35). The magnitude of the R^2 change value was .02. None of the R^2 change values for oral communication or written language was statistically significant (see *F*-values and *p*-values in table 35). At the 1st grade level five R^2 change values were statistically significant (see *F*-values and *p*-values in table 36). Only reading vocabulary with oral communication as a first predictor was non-significant. The significant R^2 change for the ITBS scores were in the .02 to .05 range.

Cognitive behavior areas and the ITBS composite

The results of the regression analyses for general behavior areas on the ITBS composite after entering the subject matter specific behavior areas are reported in table 37 for the kindergarten sample and table 38 for the 1st grade sample. At the kindergarten level, the general behavior areas did not, with one exception, add a significant portion of variance additional to that accounted for by the subject matter areas (see *F*-values and *p*-values in table 37). The only significant R^2 change was the contribution of work habits after oral communication was entered as first predictor. Eight of the nine R^2 change values at the 1st grade level were statistically significant (see *F*-values and *p*-values in table 38). The R^2 change values ranged from .04 to .10. The greater ones being associated with work habits and attentive behaviors.

The results of the regression analysis for subject matter specific behavior areas on ITBS composite after entering the general behavior areas are reported in table 39 for the kindergarten sample and table 40 for 1st grade. These tables show a different pattern of results than those in tables 37 and 38. The effects observed at the kindergarten level, were somewhat stronger than those at the 1st grade level. At the kindergarten level eight out of nine R^2 change values were significant; with R^2 change values ranging from .04 to .13 (see *F*-values and *p*-values in table 39). Six of the nine R^2 changes values at the 1st grade level were

Table 35. Regression of attentive behaviors on kindergarten ITBS achievement scores with the subject matter oriented IELI areas as first predictors.

Areas	*r*	*R*	R^2 change	N	*F* change	df_1	df_2	*p*-value.
OC - Reading vocabulary	0.373	0.375	0.001	160	0.322	1	158	0.571
OC – Reading total	0.347	0.354	0.005	132	0.719	1	130	0.398
OC –Listening	0.347	0.351	0.003	118	0.362	1	116	0.549
WL - Word analysis	0.608	0.615	0.009	150	2.135	1	148	0.146
WL - Reading vocabulary	0.392	0.395	0.002	155	0.501	1	153	0.480
MC – Math total	0.587	0.606	0.023	126	4.573	1	124	0.034

OC = Oral communication; WL = Written language; MC = Math concepts.

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 36. Regression of attentive behaviors on 1st grade ITBS achievement scores with the subject matter oriented IELI areas as first predictors.

Areas	*r*	*R*	R^2 change	N	*F* change	df_1	df_2	*p*-value.
OC - Reading vocabulary	0.409	0.419	0.008	185	1.683	1	183	0.196
OC – Reading total	0.400	0.426	0.021	185	4.874	1	183	0.029
OC –Listening	0.238	0.301	0.034	170	6.249	1	168	0.013
WL – Word analysis	0.460	0.490	0.029	184	6.904	1	182	0.009
WL - Reading vocabulary	0.282	0.356	0.047	184	9.796	1	182	0.002
MC – Math total	0.474	0.528	0.054	106	7.795	1	104	0.006

OC = Oral communication; WL = Written language; MC = Math concepts.

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

statistically significant (see *F*-values and *p*-values in table 40). The significant R^2 change values ranged from .03 to .07, the largest obtained for all general cognitive behavior areas when oral communication was the first predictor.

A summary: Construct relevant variance

In this section an important question related to the validity of the IELI scores has been investigated. In general, the results reported indicated that all the cognitive behavior areas accounted for specific variance in the academic achievement measures in addition to the variance shared with at least some of the other cognitive behavior areas. This supports the hypothesis that the cognitive behavior areas provided what has been labeled construct relevant variance to the IELI total score, variance specifically related to the construct being measured or a relevant aspect of it (Messick, 1989). Overall, the magnitude of the specific variation accounted for was small, but the shared variation accounted for by any pair of cognitive behavior areas was in most cases of medium strength. A substantial increase in variance accounted for should therefore not have been expected. The analysis did support the interpretation of the IELI cognitive behavior areas as relevant predictors for academic achievement in the language and math related subjects. This type of validity evidence is among the validity issues most relevant for potential users of the IELI. Another validity issue of similar importance for potential users concerns the structure of the IELI areas. Results of investigations into this aspect of validity will be reported next.

Investigation of the structure of the IELI: Construct validity

As noted in Chapter 3, a series of confirmatory factor analyses (CFA) were conducted to investigate the structure of the IELI areas. The results of these analyses will be reported next. Each of the five potential models of the IELI theory was evaluated separately for kindergarten and 1st grade students using CFA. Three estimation methods were used due to potential problems with the

Table 37. Regression of the general behavior areas on the ITBS composite score with the subject matter oriented IELI behavior areas as the first predictor in the kindergarten sample.

Areas	*r*	*R*	R^2 change	N	*F* change	df_1	Df_2	*p*-value.
Oral communication								
General knowledge	.395	.426	.026	103	3.155	1	101	.079
Work habits	.417	.460	.038	99	4.635	1	97	.034
Attentive behaviors	.454	.461	.007	101	.908	1	99	.343
Written language								
General knowledge	.395	.401	.005	108	.630	1	106	.429
Work habits	.444	.445	.001	104	.128	1	102	.722
Attentive behaviors	.461	.464	.002	106	.305	1	104	.582
Math concepts								
General knowledge	.395	.408	.010	108	1.313	1	106	.255
Work habits	.444	.448	.003	104	.386	1	102	.536
Attentive behaviors	.461	.470	.008	106	1.127	1	104	.291

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 38. Regression of the general behavior areas on the ITBS composite score with the subject matter oriented IELI behavior areas as the first predictor in the 1st grade sample.

Areas	*r*	*R*	R^2 change	N	*F* change	df_1	Df_2	*p*-value.
Oral communication								
General knowledge	.409	.470	.053	159	1.664	1	157	.001
Work habits	.276	.418	.099	158	18.646	1	156	.000
Attentive behaviors	.354	.432	.061	155	11.513	1	153	.001
Written language								
General knowledge	.408	.429	.018	160	3.556	1	158	.061
Work habits	.275	.365	.058	159	1.461	1	157	.001
Attentive behaviors	.361	.405	.034	156	6.196	1	154	.014
Math concepts								
General knowledge	.408	.459	.044	160	8.794	1	158	.003
Work habits	.264	.393	.084	159	15.621	1	157	.000
Attentive behaviors	.360	.435	.059	156	11.232	1	154	.001

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 39 Regression of the content matter specific behavior areas on the ITBS composite score with the general behavior areas as the first predictor in the kindergarten sample.

Area	*r*	*R*	R^2 change	N	*F* change	df_1	df_2	*p*-value.
Oral communication								
General knowledge	.415	.426	.010	103	1.182	1	101	.280
Work habits	.322	.401	.057	108	7.187	1	106	.009
Attentive behaviors	.331	.408	.057	108	7.194	1	106	.008
Written language								
General knowledge	.416	.460	.038	99	4.680	1	97	.033
Work habits	.302	.445	.107	104	13.657	1	102	.000
Attentive behaviors	.310	.448	.104	104	13.307	1	102	.000
Math concepts								
General knowledge	.408	.461	.046	101	5.826	1	99	.018
Work habits	.292	.464	.130	106	17.210	1	104	.000
Attentive behaviors	.296	.470	.133	106	17.797	1	104	.000

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

Table 40. Regression of the content matter specific behavior areas on the ITBS composite score with the general behavior areas as the first predictor in the 1st grade sample.

Areas	*r*	*R*	R^2 change	N	*F* change	df_1	df_2	*p*-value.
Oral communication								
General knowledge	.402	.470	.059	159	11.820	1	157	.001
Work habits	.336	.429	.072	160	13.910	1	158	.000
Attentive behaviors	.386	.459	.061	160	12.296	1	158	.001
Written language								
General knowledge	.393	.418	.020	158	3.803	1	156	.053
Work habits	.324	.365	.028	159	5.098	1	157	.025
Attentive behaviors	.373	.393	.015	159	2.716	1	157	.101
Math concepts								
General knowledge	.420	.432	.011	155	2.008	1	153	.158
Work habits	.355	.405	.038	156	6.921	1	154	.009
Attentive behaviors	.404	.435	.026	156	4.860	1	154	.029

r = Correlation of the first predictor with the dependent variable.

R = Correlation of first and second predictors with the dependent variable.

R^2 change = Additional variance explained by the second predictor.

data. Results for each estimation method are reported separately for the kindergarten and 1st grade samples. Due to a great number of results, the discussion will focus on the range of results for each fit index as well as the rank of the models according to the indices, indicating the models obtaining the best and second best fit. The criteria used for the evaluation of fit were discussed in Chapter III. The remaining values are, of course, included in the relevant tables. First reported are the results obtained using the most commonly employed estimation method, maximum likelihood analysis (ML) of complete data records. This method assumes the items to be measured as continuous variables. These are followed by results using a weighted least-squares estimation method, WLSMV, which assumes the items to be ordered categorical variables. Finally, results of estimation using maximum likelihood estimation that allows inclusion of missing values are reported. This method assumes the items to be measured as continuous variables but allows for inclusion of missing values.

Results of model evaluation at the kindergarten level

The five models under investigation were estimated for the kindergarten sample using a ML estimation method, and assuming the items to be measured on a continuous level. This analysis is limited to cases with no missing values, 422 subjects. The resulting statistics relevant for evaluating the fit of the five models are reported in table 41. The chi-square tests of model fit ranged from $\chi^2 = 5440.437$ (*d.f.* = 900; $p = 0.000$) to $\chi^2 = 2997.157$ (*d.f.* = 885; $p = 0.000$) (see table 41 for remaining values). The largest χ^2 value was obtained for the one general factor model, while the smallest value was obtained for the six related factors model. These results indicate that none of these models fit the data well, meaning that the estimate of the population covariance matrix assumed by each theoretical model differed significantly from the sample covariance matrix. A different perspective on this interpretation is that a different and a more complex theoretical model is required to capture the relationships implied by the data

Table 41. Fit statistics for the estimated models in kindergarten using ML estimation and assuming items to be continuous.

Model	Chi-Square		CFI		RMSEA		SRMR	
	value	rank	value	rank	Value	rank	value	rank
Baseline	15037.813 *d.f.*=946 *p*=.000							
One general factor	5440.437 *d.f.*=900 *p*=.000	5	0.678	5	0.109	5	0.087	4
Six unrelated factors	5090.273 *d.f.*=900 *p*=.000	4	0.703	4	0.105	4	0.364	5
Six related factors	2997.157 *d.f.*=885 *p*=.000	1	0.850	1	0.075	1	0.064	1
Second order factor	3387.948 *d.f.*=894 *p*=.000	3	0.823	3	0.081	2	0.080	3
General & specific factors	3062.481 *d.f.*=782 *p*=.000	2	0.832	2	0.083	3	0.072	2

N = 422; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

(Bollen, 1989). Considering the other fit indices obtained from the ML estimation method; the standardized root mean square residual (SRMR) values ranged from 0.364 to 0.064; the root mean square error of approximation (RMSEA) for the models ranged from 0.109 to 0.075; and the CFI ranged from 0.678 to .850 (see table 41). The ranking of the models was fairly consistent over the fit indices. For all three indices, the values most indicative of acceptable fit were obtained for the six related factors model. The second order factor model and the general and specific model ranked second and third, the unrelated factors model ranked fourth, and the one factor model ranked fifth. The values of the CFI and RMSEA were similar for these three best fitting models. The SRMR indicated acceptable fit for the six related factors, second order factor and general and specific models, the while the CFI and RMSEA indicated unacceptable fit for all models.

The results at the kindergarten level using the WLSMV estimation method are reported in table 42. This method assumes the items to be ordered categorical measures. The chi-square tests of model fit ranged from χ^2 = 4189.670 (*d.f.* = 31;

Table 42. Fit indices for the estimated models in kindergarten using WLSMV and assuming categorical items.

Model	Chi-Square			CFI		RMSEA		SRMR	
	value		rank	value	rank	value	rank	value	rank
Baseline	7603.719	*d.f.*=31 *p*=.000							
One general factor	583.121	*d.f.*=63 *p*=.000	5	0.931	4	0.140	4	0.121	4
Six unrelated factors	4189.67	*d.f.*=31 *p*=.000	4	0.451	5	0.564	5	0.524	5
Six related factors	252.155	*d.f.*=81 *p*=.000	1	0.977	1	0.071	1	0.077	1
Second order factor	398.730	*d.f.*=71 *p*=.000	2	0.957	2	0.105	2	0.102	3
General & specific factors	506.041	*d.f.*=73 *p*=.000	3	0.943	3	0.119	3	0.098	2

n = 422; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

p = 0.000) to χ^2 = 252.155 (*d.f.* = 81, *p* = 0.000) (see table 42 for remaining values). The smallest value was again obtained for the six related factors model, and the second smallest for the second order factor model. The χ^2 tests indicated unacceptable fit for all models. The ranges of the other fit indices were as follows; SRMR ranged from 0.524 to 0.077; the RMSEA ranged from 0.564 to 0.071; and the CFI ranged from 0.451 to 0.977 (see table 42). The values for the six related factors model best approximated the criterion values indicative of acceptable fit on the three indices. The second order factor model and the general and specific factors model ranked second and third, while the one factor and six unrelated factor models ranked fourth and fifth. Both the SRMR and CFI indicated that the six correlated factors model had acceptable fit to the data. The CFI further supported acceptable fit for the second order factor model. None of the models had acceptable fit according to the RMSEA.

The final set of results for the kindergarten sample was obtained using ML estimation method that allows the inclusion of cases where responses are missing for some items. The analyses used responses of 487 students, including those with ten or fewer missing responses, or 13.3% more subjects than the other two estimation methods. The χ^2 values ranged from χ^2 = 5909.581 (*d.f.* = 900, *p* = 0.000) to χ^2 = 3116.167 (*d.f.* = 885 *p* = 0.000) (see table 43 for remaining values). The smallest value was obtained for six related factors model. Again, the χ^2 failed to indicate that the theoretically implied covariance matrixes adequately reproduced the sample covariance matrix. The SRMR values ranged from 0.353 to 0.061; the RMSEA values ranged from 0.107 to 0.072; and the CFI ranged from 0.684 to 0.859 (see table 43). Similar to the results obtained in the analyses using the other two estimation methods, the values indicating best approximation to acceptable fit were obtained for the six related factors model, with the second order factor and general and specific factor models ranking second or third. The values of the SRMR index obtained for the six related factors model and the general and specific model support the adequacy of fit for these models to the data. The CFI and RMSEA fail to support the fit of even the best fitting models.

Three observations regarding the results of the CFA for the kindergarten s ample must be noted. First, the rank order of the fit indices exhibited some fairly constant patterns for all sets of results. Secondly, the values most indicative of goodness of fit were consistently obtained for the six related factors model. The second order factor model and the general and specific factors models ranked second and third according to all indices. Results least indicative of goodness of fit were in all situations obtained either for six unrelated factors model and the one factor model. Thus, the rank order of the fit values along a dimension from worse to better fit was stable for the three sets of results, indicating the six related factors model as the one closest to adequate fit in all instances. The three sets of results were reported for the sake of comparison, with the results from the WLSMV estimation as potentially the most reliable because its assumptions are in best accord with certain known properties of the data. Finally, only the SRMR index consistently indicated adequate fit for one of the models, the six correlated

Table 43. Fit indices for the estimated models in kindergarten using ML estimation that allows missing values and assumes items to have continuous distributions.

Model[1]	Chi-Square		CFI		RMSEA		SRMR	
	value	Rank	value	rank	value	rank	value	rank
Baseline	16812.203 *d.f.*=946 *p*=.000							
One general factor	5909.581 *d.f.*=900 *p*=.000	4	0.684	4	0.107	4	0.085	2
Six unrelated factors	5505.767 *d.f.*=900 *p*=.000	3	0.710	3	0.103	3	0.353	4
Six related factors	3116.167 *d.f.*=885 *p*=.000	1	0.859	1	0.072	1	0.061	1
Second order factor	3802.902 *d.f.*=895 *p*=.000	2	0.817	2	0.082	2	0.106	3

n = 487; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

[1] Due to convergence problems the general & specific factors was excluded from this analysis.

factors model. Note that this index was evaluated as the most sensitive to misspecified factor covariances of all the indexes evaluated by Hu & Bentler (1998). The only results indicating adequate fit by the indices sensitive to factor loadings, CFI and RMSEA, were obtained when using the WLSMV method. Overall, the fit of the model must be considered questionable. There was partial support regarding the fit of the six related factors model, but other sources of evidence spoke against model fit.

Results of model evaluation in 1st grade

The resulting statistics relevant for evaluating the fit of the five models in the 1st grade sample, using the ML estimation method are reported in table 44. This analysis was based on all cases with no missing values over the 44 items, 522 subjects. The chi-square tests of model fit ranged from χ^2 = 6902.232 (*d.f.* = 900; *p* = 0.000) to χ^2 = 3707.278 (*d.f.* = 885; *p* = 0.000) (see table 44 for remaining values).

Table 44. Fit indices for the estimated models in 1st grade using ML estimation and assuming items to have continuous distributions.

Model	Chi-Square		CFI		RMSEA		SRMR	
	value	rank	value	rank	value	rank	Value	rank
Baseline	16413.48 *d.f.*=946 *p*=.000							
One general factor	6902.232 *d.f.*=900 *p*=.000	5	0.612	4	0.113	5	0.100	4
Six unrelated factors	5708.604 *d.f.*=900 *p*=.000	4	0.689	5	0.101	4	0.301	5
Six related factors	3707.278 *d.f.*=885 *p*=.000	1	0.818	1	0.078	1	0.071	1
Second order factor	4202.332 *d.f.*=894 *p*=.000	3	0.786	2	0.084	2	0.089	3
General & specific factors[1]	4177.013 *d.f.*=822 *p*=.000	2	0.778	3	0.088	3	0.084	2

n = 522; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

[1] Due to convergence problems item 20a was omitted from the present analyses.

The smallest value was obtained for the six related factors model, but the second smallest value was obtained for the general and specific factors model. These results indicate that none of these models fit the data adequately. Looking at the other fit indices, the SRMR values ranged from 0.301 to 0.071; the RMSEA for the models ranged from 0.113 to 0.078; and the CFI ranged from 0.612 to 0.818. All statistics relevant for evaluation of model fit indicated that the six related factors model best approached acceptable fit, but only the SRMR indicated acceptable fit. The second best fitting models were the second order factor or the general and specific factors models.

The second set of statistics relevant for evaluating the fit of the five models in 1st grade are reported in table 45. These are based on estimation using the WLSMV method. The chi-square tests of model fit ranged from $\chi^2 = 3176.216$ (*d.f.* = 32; *p* = 0.000) to $\chi^2 = 312.878$ (*d.f.* = 79, *p* = 0.000) (see table 45 for remaining

Table 45. Fit indices for the estimated models in 1st grade using WLSMV and assuming categorical items.

Model	Chi-Square			CFI		RMSEA		SRMR	
	value		rank	value	rank	value	rank	Value	rank
Baseline	6295.42	*d.f.*=32 *p*=.000							
One general factor	848.628	*d.f.*=72 *p*=.000	4	0.876	4	0.144	4	0.143	4
Six unrelated factors	3176.216	*d.f.*=32 *p*=.000	5	0.498	5	0.434	5	0.484	5
Six related factors	312.878	*d.f.*=79 *p*=.000	1	0.963	1	0.075	1	0.090	1
Second order factor	523.711	*d.f.*=77 *p*=.000	2	0.929	2	0.105	2	0.114	2
General & specific factors	621.942	*d.f.*=75 *p*=.000	3	0.913	3	0.118	3	0.118	3

n = 522; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

values). The smallest χ^2 value was obtained for the six related factors model, the second smallest for the second order factor model. As before, the χ^2 failed to support the fit of the models. The SRMR values for the estimated models ranged from 0.484 to 0.090; the RMSEA values ranged from 0.434 to 0.075; and the CFI ranged from 0.498 to 0.963 (see table 45). All statistics relevant for evaluation of model fit produced identical rank orders of values for the five models; the six related factors model always obtained the values most indicative of acceptable fit. The second order factor model ranked second and general and specific factor model third. The only value that fell within the range of acceptable fit was the CFI for the six related factors model.

The final set of results for the 1st grade sample was obtained using the ML estimation method that allows the inclusion of cases where responses are missing for some items. The analyses used responses for 583 students; thus 10.5% more subjects were available for this estimation method than for the other two methods. The χ^2 values ranged from χ^2 = 7534.092 (*d.f.* = 900, *p* = 0.000) to

Table 46. Fit indices for the estimated models in 1st grade using ML estimation that allows missing values and assumes items to have continuous distributions.

Model[1]	Chi-Square			CFI		RMSEA		SRMR	
	value		rank	Value	rank	value	rank	value	rank
Baseline	18681.493	*d.f.*=946 *p*=.000							
One general factor	7534.092	*d.f.*=900 *p*=.000	4	0.626	4	0.112	4	0.096	2
Six unrelated factors	6271.120	*d.f.*=900 *p*=.000	3	0.697	3	0.101	3	0.305	4
Six related factors	3905.084	*d.f.*=885 *p*=.000	1	0.830	1	0.077	1	0.068	1
Second order factor	4881.868	*d.f.*=895 *p*=.000	2	0.775	2	0.087	2	0.115	3

n = 583; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

[1] Due to convergence problems the general & specific factors was excluded from this analysis.

χ^2 = 3905.084 (*d.f.* = 885 *p* = 0.000) (see table 46 for remaining values). The smallest value was once again obtained for the six related factors model. The χ^2 indicate that the theoretically implied covariance matrixes failed to adequately reproduce the sample covariance matrix. The SRMR values ranged from 0.305 to 0.068; the RMSEA values ranged from 0.112 to 0.077; and the CFI ranged from 0.626 to 0.830 (see table 46). The values best indicative of acceptable fit were once again obtained for the six related factors model, the second order factor model ranked second and the general and specific factor model ranked third. The obtained value of the SRMR index indicate acceptable fit for the six related factors model. All other values fell outside the range of acceptable fit for the respective indices. Estimation problems plagued the general and specific factors model under this method of estimating model parameters. The model was therefore excluded from the analysis.

Similar observations can be made regarding the results from the 1st grade sample as were made regarding the kindergarten sample. The rank order of the obtained values of the fit indices is fairly consistent across estimation methods. The six correlated factors model was consistently the one with obtained fit values most favorably indicative of goodness of fit; the second order factor model ranked second in all but one case, and the general and specific factors model ranked third. Only SRMR under the ML estimation method and the CFI under WLSMV method indicated acceptable fit. Finally, values of fit indices for the first grade sample were generally further removed from the direction indicating goodness of fit than those obtained for the kindergarten sample.

A modified six related factors model:
Additional correlated errors

When evaluating the results obtained from both samples and reported above, it can confidently be stated that the six related factors model is the most promising of the models evaluated. On the other hand, the results offered only limited support regarding acceptance of this model as adequately fitting the data. One observation from the above results is that the index that most constantly indicated adequate fit for the six related factors model was the SRMR, an index reported by Hu and Bentler (1999) to be most sensitive of the fit indexes they studied to misspecifications of the latent factors. Less support was, however, received from indices sensitive to minor misspecifications, such as of factor loadings. This observation leads to the possibility that minor modifications within the modeling framework adopted above may lead to important improvements of fit. Therefore, a slightly modified version of the six related factors model will be investigated next.

One aspect of the constraints imposed during the model construction may contribute to the lack of fit observed, the absence of correlated errors among items within each behavior area. Several items wree constructed from behaviors that share characteristics or appear similar from the observer's perspective. Correlated errors among them were not considered in the model construction. An exception was made in the case of two item pairs, 17a and 17b and 20a and 20b, which explicitly aimed at different complexity of the same behavior.

Investigation of modification indices obtained from estimation of the six related factors model in both kindergarten and 1st grade samples revealed that in several cases model fit would improve if correlations among error terms for pairs of items were allowed. Investigations of those item pairs that were flagged in both samples using either the ML or WLSMV methods for estimation are listed in table 47, along with descriptions of the item content. Only item pairs belonging to the same behavior area were considered for modification and investigated. Several other pairs of items have similar content, e.g. 4 and 27, 17 and 21, and 35 and 37, but allowing correlations among items across behavior areas was viewed as violating the underlying construct.

A second type of relationship that contributed to the lack of fit was correlations of individual items to a second behavior area, above and beyond the relation implied via the correlation of the two respective behavior areas. As an example, item 42 had a stronger relation with all five 'other' areas than that implied by the correlation with the attentive behavior area. In order to maintain that the model investigated remained consistent with the construct theory outlined above, modifications of relations of this type were not considered.

A modified six related factors model was evaluated allowing error terms for selected pairs of items to correlate. These modifications are reported in table 47. The results obtained are reported in tables 48 and 49 for the kindergarten and 1st grade samples, respectively. Generally speaking, the obtained values of the fit indices were closer to the range of values indicating acceptable fit than the original six related factors model, without yielding any firmer evidence of acceptable fit. No further modifications of the CFA model were investigated in this study.

Table 47. Item pairs with correlated error terms in the modified six related factors model.

Item pair	Area	First item	Second item
3 & 4	gk	opposites of common characteristics	likeness & difference in pictures
6 & 7	oc	speak in complete sentences	answer a direct question
10 & 11	oc	recall facts from a story	retell simple story
08 & 13	oc	share ideas and information	use personal experiences when speaking
14 & 15	wl	recognize own first/last name in print	Print own first & last name
16 & 18	wl	match simple words and pictures	Copy simple words
16 & 19	wl	match simple words and pictures	write simple words from memory
20B & 21	mc	count from 1 to 20	Copy simple numerals
26 & 27	mc	measure objects with self defined units	classify objects according to numeral properties
33 & 34	wh	wait for directions before beginning an assignment	follow teacher suggestions
33 & 35	wh	wait for directions before beginning an assignment	Work independently
39 & 40	at	pay attention in class	listen while others are talking
40 & 42	at	listen while others are talking	Keep up with pace of instruction

Table 48. Fit statistics for the modified six correlated factors model in the kindergarten sample.

	Estimation method		
	ML	WLSMV	ML-missing
Chi-Square	2501.815	215.310	2602.904
d.f.	872	80	872
P-value	0.000	0.000	0.000
CFI	0.884	0.982	0.892
RMSEA	0.067	0.063	0.064
SRMR	0.060	0.074	0.058

CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

Note: Correlations among the following item pairs were added:
3 & 4; 6 & 7; 10 & 11; 08 & 13; 14 & 15; 16 & 18; 16 & 19; 20B & 21; 26 & 27; 33 & 34; 33 & 35; 39 & 40; 40 & 42.

Table 49. Fit statistics for the modified six correlated factors model in the 1st grade sample

	Estimation method		
	ML	WLSMV	ML-missing[1]
Chi-Square	3095.939	268.253	3105.689
d.f.	872	79	831
P-value	0.000	0.000	0.000
CFI	0.856	0.970	0.868
RMSEA	0.070	0.068	0.069
SRMR	0.068	0.087	0.065

CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

Note: Correlations among the following item pairs were added:
3 & 4; 6 & 7; 10 & 11; 08 & 13; 14 & 15; 16 & 18; 16 & 19; 20B & 21; 26 & 27; 33 & 34; 33 & 35; 39 & 40; 40 & 42.

[1] Due to convergence problems item 20a was omitted from this estimation.

Investigation of potential bias: Differential item functioning

Investigation of sex DIF

Two sets of DIF analyses were conducted in this study and will be reported next. First, an investigation of sex DIF will be reported followed by an investigation of ethnic DIF. A convention that has been established in the field of DIF analysis is to use two or more statistical methods to flag items potentially exhibiting DIF. The items that are consistently flagged by more than two methods are then evaluated as potential biased items (Camilli and Shephard, 1994). The present study follows this convention regarding sex DIF. Therefore results of four analyses are reported for sex DIF, but due to the limited sample, only one for ethnic DIF.

The original plan of the present study included analyses of DIF at each grade level. Problems with the data at hand, limited distributions on most of the items, precluded this. For this reason it was decided to combine students at both grade levels for the analyses, and the choice of methods also had to be reconsidered. The four methods finally selected for use were SIBTEST, Poly-SIBTEST, MHχ^2 and HW3. Furthermore, all remaining IELI items were used for the criterion variable in the DIF analysis. This method was adopted for practical reasons. In reality it assumes unidimensionality across the areas, an assumption that is unlikely to hold.

SIBTEST and poly-SIBTEST were computed using the SIBTEST software. For the analysis of each item, all remaining items were included in the estimation of the matching variable. The proportion of data discarded from the analysis of each item, as reported by the program, ranged from 15 to 22% for each item. SIBTEST and poly-SIBTEST statistics were not obtained for one item.

Microsoft EXCEL spreadsheets were used to compute MHχ^2 and HW3, but SPSS was used to prepare frequency tables necessary for this computation. The IELI composite could not be used directly as a matching variable because of the considerable number of score levels with zero variance for either boys or girls. Therefore the number of score levels on the matching variable was reduced for each analysis. In a few instances the number of score levels had to be reduced even further due to limited variability on the matching variable. Due to limited data, the MHχ^2 and HW3 statistics were not obtained for four items.

Two of the methods used a dichotomized item response scale, SIBTEST and MHχ^2. The data were dichotomized by assigning the value '1' to the maximum score for each item while all other response values were assigned a '0'. The analysis is in a sense limited to investigating whether a different proportion of boys and girls obtained the maximum score rather than a lower score on each item. The rationale supporting the use a dichotomized item response scale was the limited number of responses at the lower score levels for many of the items. Given this limitation of the data, it was believed to be valuable to supplement the analysis of the polytomous data by analyses of the dichotomized data. The results of the analysis of sex DIF will be reported next, followed by the results of the ethnic DIF analysis.

Results of DIF analyses for sex

Four methods were used for the sex DIF analyses. Two of these made full use of the polytomous data, poly-SIBTEST and HW3. The results from these analyses are reported in table 50. Poly-SIBTEST flagged 11 items as potentially exhibiting DIF at the $p < 0.01$ level. The flagged items belonged to four of the cognitive behavior areas, six of the items favored boys and five favored girls (see table 50 for item identity and numerical values). All items that favored girls belonged to work habits and attentive behaviors. The items that favored boys, on the other hand, all belonged to oral communication and math concepts. The magnitudes of the DIF effects were greater for math concepts and attentive behaviors than for items from the other two areas. The HW3 test flagged six items as potentially exhibiting DIF at the $p < 0.01$ level (see table 50). Two of these items belonged to math concepts; both items favored boys. The remaining four items all favored girls and belonged to work habits or attentive behaviors (see table 50 for item identity and numerical values).

Results for the methods that used the dichotomized item responses are reported in table 51. The SIBTEST method flagged thirteen items as potentially exhibiting DIF at the $p < 0.01$ level (see table 51 for item identity and numerical values). Three of these items favored girls; one belonged to each of the written language, work habits and attentive behaviors areas. The ten items that favored boys belonged to general knowledge, math concepts and oral communication.

The MH method flagged seven items; five items favored girls and two items favored boys (see table 51 for item identity and numerical values). All items that favored girls belonged to work habits and attentive behaviors while the items that favored boys belonged to math concepts and general knowledge.

Overview of results for the investigation of sex DIF

An overview of the DIF analysis is given in table 52. The items flagged as potentially exhibiting DIF by each method are identified by the letters B or G, B indicating that an item favored boys and G an item that favored girls. The last column, headed consistency, indicates items that were identified by three or four methods and should therefore be considered as DIF items in the evaluation of these results. As can be seen in table 52, five items were constantly flagged by three or four of the methods; items 23, 24, 34, 37, and 39. These are considered to have exhibited DIF in the present study. These items belonged to three cognitive behavior areas, math concepts, work habits and attentive behaviors. Further, two items favored boys and three favored girls. The items that favored boys both belonged to the math concepts area, items 23 and 24. The items that favored girls belonged to work habits, item 34, and attentive behaviors, items 37 and 39. The attentive behavior area was most problematic, with two out of six items flagged as DIF items. A second area of concern was math concepts, with two out of ten items that were flagged.

It is noteworthy that the HW3 and MH methods flagged fewer items than SIBTEST and poly-SIBTEST, and those flagged by the former method were in all but one case also flagged by the latter. These were not considered as DIF items in the present study, but the fact that three additional items that belonged to math concepts and oral communication were flagged by one or two methods calls for a further study using a larger dataset to investigate the issue.

Table 50. Results of sex DIF analysis using polytomous item responses.

	Poly-SIBTEST				HW3		
Item	Beta	St. error	*p*-value	Direction[2]	Test statistic	*p*-value	Direction[2]
I01	-0.021	0.026	0.425		-1.636	0.103	
I02	-0.001	0.004	0.857		***[1]		
I03	0.047	0.021	0.028		1.296	0.194	
I04	0.059	0.024	0.013		1.142	0.255	
I05	-0.001	0.021	0.946		-1.077	0.283	
I06	-0.002	0.021	0.938		-0.999	0.317	
I07	-0.022	0.025	0.377		-0.867	0.384	
I08	0.074	0.035	0.036		1.430	0.153	
I09	0.042	0.027	0.126		0.082	0.936	
I10	0.082	0.027	0.003	* B	1.351	0.187	
I11	0.071	0.029	0.014		1.398	0.162	
I12	0.116	0.032	0.000	* B	2.192	0.028	
I13	0.070	0.035	0.047		0.927	0.352	
I14	-0.012	0.010	0.195		***[1]		
I15	-0.007	0.020	0.739		-0.639	0.522	
I16	0.039	0.033	0.234		0.306	0.757	
I17A	-0.003	0.022	0.885		-0.699	0.484	
I17B	-0.006	0.022	0.789		-0.700	0.484	
I18	-0.027	0.024	0.272		-1.622	0.105	
I19	-0.001	0.045	0.979		-0.293	0.772	
I20A	****[1]				***[1]		
I20B	0.014	0.010	0.170		***[1]		
I21	-0.001	0.011	0.937		-1.023	0.308	
I22	0.023	0.016	0.139		0.589	0.555	
I23	0.132	0.033	0.000	* B	2.512	0.012	* B
I24	0.119	0.031	0.000	* B	2.594	0.010	* B
I25	0.065	0.027	0.017		0.948	0.342	
I26	0.125	0.042	0.003	* B	1.179	0.238	
I27	0.099	0.037	0.008	* B	1.003	0.317	
I28	0.037	0.028	0.184		1.151	0.250	
I29	-0.067	0.043	0.119		-1.211	0.301	
I30	-0.019	0.035	0.580		-1.395	0.163	
I31	-0.084	0.031	0.007	* G	-1.312	0.187	
I32	-0.033	0.032	0.302		-1.236	0.219	
I33	-0.067	0.035	0.054		-2.088	0.037	
I34	-0.110	0.029	0.000	* G	-3.151	0.002	* G
I35	-0.018	0.031	0.558		0.026	0.976	
I36	-0.059	0.024	0.014		-2.055	0.040	
I37	-0.131	0.033	0.000	* G	-3.556	0.001	* G
I38	-0.021	0.021	0.321		-1.833	0.074	
I39	-0.131	0.031	0.000	* G	-3.579	0.000	* G
I40	-0.080	0.034	0.018		-2.936	0.003	* G
I41	-0.095	0.037	0.009	* G	-2.276	0.023	
I42	0.063	0.028	0.022		1.650	0.099	

[1] Statistic not obtained; [2] * $p < .01$; B = favors boys; G = favors girls.

Table 51 Results of sex DIF analysis using dichotomized item responses.

	SIBTEST		Direction[2]	MH	Direction[2]
Item	Beta	SE beta	*p*-value	MH χ^2	*p*-value
I01	-0.024	0.026	0.372	5.614	0.018
I02	-0.011	0.005	0.043	***[1]	
I03	0.045	0.021	0.031	1.344	0.246
I04	0.066	0.024	0.006* B	10.140	0.001 * B
I05	0.012	0.020	0.563	2.200	0.138
I06	-0.009	0.022	0.671	2.599	0.107
I07	-0.010	0.023	0.663	***[1]	
I08	0.072	0.029	0.014	2.668	0.102
I09	0.047	0.026	0.067	2.529	0.112
I10	0.072	0.025	0.003* B	2.511	0.113
I11	0.073	0.026	0.005* B	2.031	0.154
I12	0.089	0.026	0.001* B	4.854	0.028
I13	0.072	0.028	0.009* B	4.701	0.030
I14	***[1]			***[1]	
I15	-0.895	0.013	0.000* G	0.816	0.366
I16	0.035	0.025	0.162	2.271	0.132
I17A	0.001	0.018	0.935	0.202	0.653
I17B	-0.007	0.019	0.707	0.487	0.485
I18	-0.022	0.021	0.297	5.691	0.017
I19	0.026	0.030	0.389	0.882	0.348
I20A	-0.006	0.004	0.165	***[1]	
I20B	0.021	0.012	0.086	2.252	0.133
I21	-0.011	0.013	0.408	0.268	0.604
I22	0.018	0.018	0.313	2.711	0.100
I23	0.133	0.031	0.000* B	6.308	0.012
I24	0.111	0.027	0.000* B	10.473	0.001 * B
I25	0.086	0.025	0.001* B	5.491	0.019
I26	0.113	0.030	0.000* B	2.908	0.088
I27	0.096	0.028	0.001* B	2.371	0.124
I28	0.060	0.024	0.013	3.616	0.057
I29	-0.007	0.032	0.822	2.051	0.152
I30	0.031	0.028	0.264	1.317	0.251
I31	-0.054	0.027	0.042	13.857	0.000 * G
I32	-0.012	0.027	0.658	4.483	0.034
I33	-0.039	0.031	0.199	1.723	0.189
I34	-0.090	0.026	0.001* G	11.495	0.001 * G
I35	0.027	0.026	0.291	0.880	0.348
I36	-0.045	0.023	0.054	3.831	0.050
I37	-0.064	0.026	0.014	8.713	0.003 * G
I38	0.001	0.022	0.947	0.535	0.464
I39	-0.072	0.027	0.007* G	12.026	0.001 * G
I40	-0.057	0.029	0.048	12.925	0.000 * G
I41	-0.033	0.026	0.215	3.604	0.058
I42	0.067	0.023	0.003	2.584	0.108

[1] Statistic not obtained; [2] * $p < .01$; B = favors boys; G = favors girls.

Table 52 Overview of significant sex DIF.

		Polytomous		Dichotomous		
Area	Item	Poly-SIBTEST	HW-3	SIBTEST	MHχ^2	Consistency
General knowledge	I01					
General knowledge	I02					
General knowledge	I03					
General knowledge	I04			B	B	
General knowledge	I05					
Oral communication	I06					
Oral communication	I07					
Oral communication	I08					
Oral communication	I09					
Oral communication	I10	B		B		
Oral communication	I11			B		
Oral communication	I12	B		B		
Oral communication	I13			B		
Written language	I14					
Written language	I15			G		
Written language	I16					
Written language	I17A					
Written language	I17B					
Written language	I18					
Written language	I19					
Math concepts	I20A					
Math concepts	I20B					
Math concepts	I21					
Math concepts	I22					
Math concepts	I23	B	B	B		X
Math concepts	I24	B	B	B	B	X
Math concepts	I25			B		
Math concepts	I26	B		B		
Math concepts	I27	B		B		
Math concepts	I28					
Work habits	I29					
Work habits	I30					
Work habits	I31	G			G	
Work habits	I32					
Work habits	I33					
Work habits	I34	G	G	G	G	X
Work habits	I35					
Attentive behaviors	I36					
Attentive behaviors	I37	G	G		G	X
Attentive behaviors	I38					
Attentive behaviors	I39	G	G	G	G	X
Attentive behaviors	I40		G		G	
Attentive behaviors	I41	G				
Attentive behaviors	I42					

B = item favoring boys; G = item favoring girls.

Investigation of ethnic DIF

Part of the original study plan was to investigate ethnic DIF, at least for African American – White and Hispanic – White groups. The number of students with the different ethnic backgrounds in the data at hand was too small for such analyses. As can be seen in table 53 there were 91 students with African American background and 71 with Hispanic background in the data. These samples were too small to conduct a DIF study using the intended methods designed for polytomus data or other currently common methods. In order to gain preliminary information regarding ethnic DIF, the Delta plot method was used (Angoff, 1982; Muňiz, Hambleton and Xing, 2001).

The Delta plot method assumes item difficulties in the zero to one range, but the IELI items have three or four score points. In order to place the item means on a difficulty scale for dichotomous items, the mean for each item was divided by its maximum score. This re-expressed the item mean on the item difficulty scale for dichotomous items. These difficulty values were transformed into delta values. Figures 6 to 8 display the Delta plots for the African American – White, Hispanic – White, and African American –Hispanic comparisons respectively. The item difficulties and their differences can be found in tables 54 to 56. Traditionally, a deviation of 1.5 deltas has indicated that an item was potentially exhibiting DIF

Table 53. Ethnic background of students by grade level.

	Kindergarten		1st grade		Combined grades	
Ethnic origin	Total N	Complete data records	Total N	Complete data records	Total N	Complete data records
African American	55	46	54	45	109	91
Hispanic	41	35	42	36	83	71
White	360	313	450	400	810	713
Asian/Pacific	6	6	12	12	18	18
Native America	11	9	14	13	25	22
Mixed	11	10	7	7	18	17
Total	484	419	579	513	1063	932

(Angoff, 1982; Camilli & Shephard, 1994). These values were drawn as broken lines on figures 6 to 8. A solid line representing equal difficulty of an item in both groups was also drawn on the figures.

Considering the large and very large differences, 10 items are flagged as potential DIF items for the African American – White comparison, two for the African American – Hispanic comparison but 25 items for the Hispanic – White comparisons. The sheer number of items flagged as potentially exhibiting DIF must be of concern. Given that the present study relies on a small sample and a weak method flags may be interpreted as indicating a need for a study using more adequate sample sizes for the respective ethnic groups.

An investigation of figure 6 reveals that 10 items fell on or just outside the dotted line in the lower-right half of the figure (see table 54 for numerical values). No item fell above the solid line. All potential DIF items therefore favored white students. The two items with largest delta differences were located near the top of the difficulty scale, an area where the Delta plot method is known to have a tendency to falsely flag items (Angoff, 1982; Camilli& Shephard, 1994). The flagging of these two items, one from math concepts and one from general knowledge, was therefore suspect. The remaining items, all of which fell on or close to the dotted line, belonged to all of the behavior areas. The greatest number of potential DIF items within a given area was three items, oral communication and math concepts. The results for the African American – White comparisons indicated a need for further investigation. The sheer number of items flagged was counterbalanced by the fact that most of them fell on or just outside the cut-point for flagging an item or were suspicious for methodological reasons. But the limited sample size used limits the results.

The delta method flagged 25 items as potentially exhibiting Hispanic-White DIF (see figure 7 and table 55). It is noteworthy that all but one item from the work habits and attentive behaviors areas were flagged. Further, seven out of ten math concepts items were flagged. Again, two items fell in the range of extreme difficulty, where the Delta method is unreliable. The magnitude of the remaining flagged items was often of considerable magnitude. The heavy emphasis of potential DIF items in three of the cognitive behavior areas was a reason for concern despite the small sample size. Considerations of the implications for the behavior areas have to wait until studies using lager samples and stronger methods have been conducted.

Finally the Delta method was also used to investigate potential Hispanic - African American DIF. Three items, one from general knowledge and two from math concepts were flagged as potential DIF items in this analysis. The Delta plot is in figure 8 and item difficulties and delta values are reported in table 56.

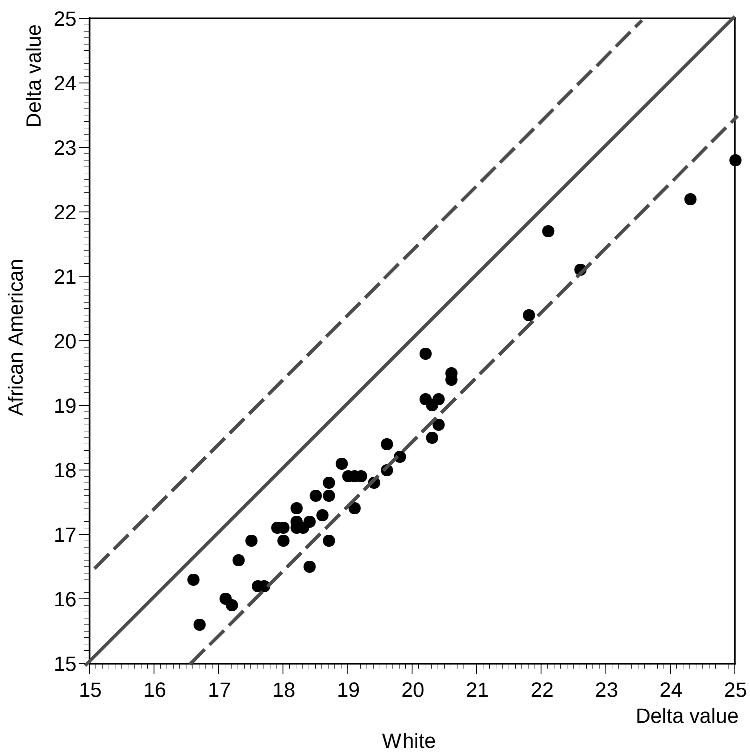

Figure 6. Delta plot of item difficulties of African American and White students at both grade levels.

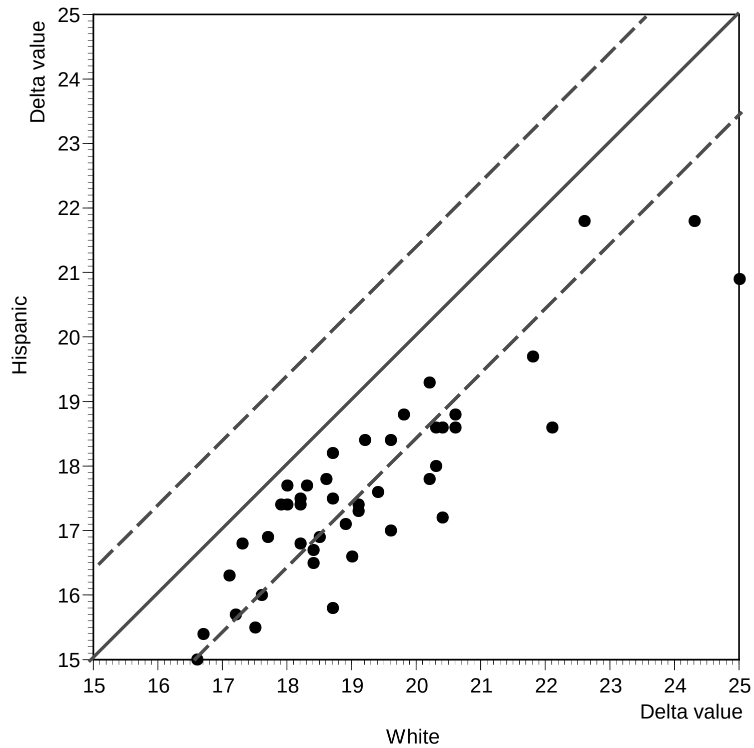

Figure 7. Delta plot of item difficulties of Hispanic and White students at both grade levels.

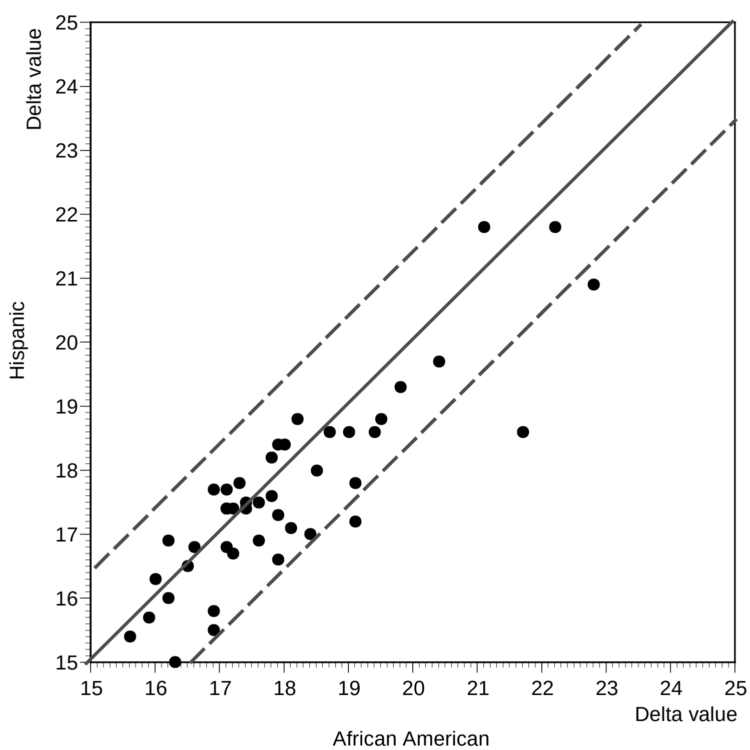

Figure 8. Delta plot of item difficulties of African American and Hispanic students at both grade levels.

Table 54. Statistics for the Delta plot of the White - African American comparison

Aerea	Item	Item mean		Rescaled difficulty		Delta value		Difference
		White	Afr.Am.	White	Afr.Am.	White	Afr.Am.	
GK	I01	2.8	2.7	0.93	0.89	19.0	17.9	1.1
	I02	3.0	3.0	1.00	0.99	24.3	22.2	**2.2**
	I03	2.9	2.8	0.97	0.94	20.4	19.1	1.3
	I04	3.8	3.7	0.95	0.91	19.6	18.4	1.2
	I05	3.9	3.7	0.97	0.92	20.4	18.7	**1.7**
OC	I06	3.9	3.8	0.96	0.94	20.2	19.1	1.1
	I07	3.8	3.6	0.95	0.90	19.6	18.0	**1.6**
	I08	3.7	3.5	0.91	0.88	18.5	17.6	0.8
	I09	3.7	3.5	0.94	0.87	19.1	17.4	**1.7**
	I10	3.7	3.5	0.92	0.87	18.7	17.6	1.1
	I11	3.7	3.4	0.91	0.85	18.4	17.2	1.2
	I12	3.5	3.2	0.88	0.79	17.6	16.2	1.4
	I13	3.7	3.2	0.91	0.81	18.4	16.5	**2.0**
WL	I14	3.0	2.9	0.99	0.98	22.6	21.1	**1.5**
	I15	2.9	2.9	0.96	0.96	20.2	19.8	0.3
	I16	3.7	3.6	0.93	0.90	18.9	18.1	0.8
	I17A	3.9	3.8	0.97	0.95	20.6	19.4	1.2
	I17B	3.9	3.7	0.97	0.93	20.3	19.0	1.3
	I18	3.9	3.7	0.97	0.91	20.3	18.5	**1.8**
	I19	3.5	3.4	0.87	0.84	17.5	16.9	0.6
MC	I20A	3.0	3.0	1.00	0.99	25.0	22.8	**2.2**
	I20B	3.0	3.0	0.99	0.99	22.1	21.7	0.4
	I21	3.0	2.9	0.99	0.97	21.8	20.4	**1.5**
	I22	2.9	2.9	0.97	0.95	20.6	19.5	1.1
	I23	2.5	2.4	0.82	0.79	16.6	16.3	0.3
	I24	3.7	3.4	0.92	0.84	18.7	16.9	**1.7**
	I25	3.7	3.6	0.94	0.89	19.1	17.9	1.2
	I26	3.3	3.0	0.82	0.74	16.7	15.6	1.1
	I27	3.4	3.1	0.85	0.77	17.2	15.9	1.3
	I28	3.8	3.5	0.94	0.88	19.4	17.8	**1.6**
WH	I29	3.4	3.3	0.86	0.81	17.3	16.6	0.7
	I30	3.6	3.4	0.90	0.85	18.2	17.2	1.0
	I31	3.7	3.4	0.92	0.86	18.6	17.3	1.4
	I32	3.5	3.2	0.88	0.79	17.7	16.2	**1.5**
	I33	3.6	3.4	0.91	0.85	18.3	17.1	1.1
	I34	3.7	3.5	0.92	0.88	18.7	17.8	0.9
	I35	3.6	3.5	0.90	0.87	18.2	17.4	0.7
AT	I36	3.8	3.6	0.94	0.89	19.2	17.9	1.3
	I37	3.6	3.4	0.89	0.85	17.9	17.1	0.8
	I38	3.8	3.6	0.96	0.90	19.8	18.2	**1.6**
	I39	3.6	3.4	0.89	0.84	18.0	16.9	1.0
	I40	3.6	3.4	0.89	0.85	18.0	17.1	0.9
	I41	3.4	3.1	0.85	0.77	17.1	16.0	1.1
	I42	3.6	3.4	0.90	0.85	18.2	17.1	1.1

Table 55. Statistics for the Delta plot of the White - Hispanic comparison

Aerea	Item	Item mean		Rescaled difficulty		Delta value		Difference
		White	Hispanic	White	Hispanic	White	Hispanic	Difference
GK	I01	2.8	2.5	0.93	0.82	22.6	21.8	0.8
	I02	3.0	3.0	1.00	0.99	24.3	21.8	**2.5**
	I03	2.9	2.6	0.97	0.85	18.0	17.7	0.3
	I04	3.8	3.4	0.95	0.84	20.2	19.3	0.8
	I05	3.9	3.7	0.97	0.92	18.7	18.2	0.5
OC	I06	3.9	3.6	0.96	0.89	25.0	20.9	**4.0**
	I07	3.8	3.7	0.95	0.91	18.3	17.7	0.5
	I08	3.7	3.4	0.91	0.84	19.2	18.4	0.8
	I09	3.7	3.5	0.94	0.86	18.0	17.4	0.6
	I10	3.7	3.5	0.92	0.87	17.9	17.4	0.6
	I11	3.7	3.3	0.91	0.82	19.8	18.8	1.0
	I12	3.5	3.1	0.88	0.77	18.2	17.5	0.7
	I13	3.7	3.2	0.91	0.81	17.3	16.8	0.5
WL	I14	3.0	3.0	0.99	0.99	21.8	19.7	**2.1**
	I15	2.9	2.8	0.96	0.94	18.6	17.8	0.8
	I16	3.7	3.4	0.93	0.85	19.6	18.4	1.2
	I17A	3.9	3.7	0.97	0.92	18.2	17.4	0.8
	I17B	3.9	3.7	0.97	0.92	17.7	16.9	0.7
	I18	3.9	3.6	0.97	0.89	20.6	18.8	**1.8**
	I19	3.5	2.9	0.87	0.74	20.3	18.6	**1.7**
MC	I20A	3.0	2.9	1.00	0.98	20.4	18.6	**1.8**
	I20B	3.0	2.8	0.99	0.92	20.6	18.6	**2.0**
	I21	3.0	2.9	0.99	0.95	18.7	17.5	1.2
	I22	2.9	2.8	0.97	0.92	17.1	16.3	0.8
	I23	2.5	2.1	0.82	0.69	19.4	17.6	**1.7**
	I24	3.7	3.0	0.92	0.76	22.1	18.6	**3.5**
	I25	3.7	3.4	0.94	0.86	18.2	16.8	1.4
	I26	3.3	2.9	0.82	0.73	20.3	18.0	**2.3**
	I27	3.4	3.0	0.85	0.75	19.1	17.4	**1.7**
	I28	3.8	3.5	0.94	0.88	18.5	16.9	**1.5**
WH	I29	3.4	3.3	0.86	0.83	19.1	17.3	**1.8**
	I30	3.6	3.5	0.90	0.86	20.2	17.8	**2.4**
	I31	3.7	3.5	0.92	0.88	18.9	17.1	**1.7**
	I32	3.5	3.4	0.88	0.84	18.4	16.7	**1.7**
	I33	3.6	3.5	0.91	0.88	16.7	15.4	1.3
	I34	3.7	3.6	0.92	0.90	17.6	16.0	**1.6**
	I35	3.6	3.5	0.90	0.87	17.2	15.7	**1.5**
AT	I36	3.8	3.7	0.94	0.91	18.4	16.5	**2.0**
	I37	3.6	3.5	0.89	0.86	19.6	17.0	**2.6**
	I38	3.8	3.7	0.96	0.93	20.4	17.2	**3.2**
	I39	3.6	3.5	0.89	0.88	19.0	16.6	**2.4**
	I40	3.6	3.5	0.89	0.87	16.6	15.0	**1.6**
	I41	3.4	3.2	0.85	0.79	17.5	15.5	**2.0**
	I42	3.6	3.3	0.90	0.83	18.7	15.8	**2.9**

Table 56. Statistics for the Delta plot of the African American - Hispanic comparison

Aerea	Item	Item mean		Rescaled difficulty		Delta value		Difference
		Afr.Am.	Hispanic	Afr.Am.	Hispanic	Afr.Am.	Hispanic	Difference
GK	I01	2.7	2.5	0.89	0.82	17.9	16.6	1.3
	I02	3.0	3.0	0.99	0.99	22.2	21.8	0.4
	I03	2.8	2.6	0.94	0.85	19.1	17.2	**1.9**
	I04	3.7	3.4	0.91	0.84	18.4	17.0	1.4
	I05	3.7	3.7	0.92	0.92	18.7	18.6	0.1
OC	I06	3.8	3.6	0.94	0.89	19.1	17.8	1.3
	I07	3.6	3.7	0.9	0.91	18.0	18.4	-0.4
	I08	3.5	3.4	0.88	0.84	17.6	16.9	0.7
	I09	3.5	3.5	0.87	0.86	17.4	17.4	0.0
	I10	3.5	3.5	0.87	0.87	17.6	17.5	0.1
	I11	3.4	3.3	0.85	0.82	17.2	16.7	0.5
	I12	3.2	3.1	0.79	0.77	16.2	16.0	0.2
	I13	3.2	3.2	0.81	0.81	16.5	16.5	0.0
WL	I14	2.9	3.0	0.98	0.99	21.1	21.8	-0.7
	I15	2.9	2.8	0.96	0.94	19.8	19.3	0.5
	I16	3.6	3.4	0.9	0.85	18.1	17.1	1.0
	I17A	3.8	3.7	0.95	0.92	19.4	18.6	0.8
	I17B	3.7	3.7	0.93	0.92	19.0	18.6	0.4
	I18	3.7	3.6	0.91	0.89	18.5	18.0	0.5
	I19	3.4	2.9	0.84	0.74	16.9	15.5	1.4
MC	I20A	3.0	2.9	0.99	0.98	22.8	20.9	**1.8**
	I20B	3.0	2.8	0.99	0.92	21.7	18.6	**3.1**
	I21	2.9	2.9	0.97	0.95	20.4	19.7	0.7
	I22	2.9	2.8	0.95	0.92	19.5	18.8	0.8
	I23	2.4	2.1	0.79	0.69	16.3	15.0	1.3
	I24	3.4	3.0	0.84	0.76	16.9	15.8	1.2
	I25	3.6	3.4	0.89	0.86	17.9	17.3	0.6
	I26	3.0	2.9	0.74	0.73	15.6	15.4	0.2
	I27	3.1	3.0	0.77	0.75	15.9	15.7	0.2
	I28	3.5	3.5	0.88	0.88	17.8	17.6	0.2
WH	I29	3.3	3.3	0.81	0.83	16.6	16.8	-0.2
	I30	3.4	3.5	0.85	0.86	17.2	17.4	-0.2
	I31	3.4	3.5	0.86	0.88	17.3	17.8	-0.5
	I32	3.2	3.4	0.79	0.84	16.2	16.9	-0.7
	I33	3.4	3.5	0.85	0.88	17.1	17.7	-0.6
	I34	3.5	3.6	0.88	0.90	17.8	18.2	-0.4
	I35	3.5	3.5	0.87	0.87	17.4	17.5	-0.1
AT	I36	3.6	3.7	0.89	0.91	17.9	18.4	-0.6
	I37	3.4	3.5	0.85	0.86	17.1	17.4	-0.2
	I38	3.6	3.7	0.9	0.93	18.2	18.8	-0.6
	I39	3.4	3.5	0.84	0.88	16.9	17.7	-0.8
	I40	3.4	3.5	0.85	0.87	17.1	17.4	-0.3
	I41	3.1	3.2	0.77	0.79	16.0	16.3	-0.3
	I42	3.4	3.3	0.85	0.83	17.1	16.8	0.3

CHAPTER V
DISCUSSION

The construction of a new teacher rating instrument assessing six areas of cognitive behavior, the IELI, is timely. Twenty-nine years ago Spivack and Swift (1973) lamented that the needs of teachers for useful rating instruments had not been served. Still today, the need for a rating instrument providing instructionally relevant information regarding student behavior has not been fulfilled. The IELI is intended to fulfill such a need for the primary grades. The IELI scores are intended to reflect student cognitive behaviors and provide information helpful for instructional planning. The six areas of cognitive behavior provide the underlying theoretical construct for the IELI. Specific interpretations and score properties were targeted in the test construction process. But the psychometric properties of this instrument were still unknown. Empirical evidence regarding the score properties were required to support or refute the validity of the intended score interpretations. The purpose of the present study was an empirical investigation of these properties of the IELI scores.

The present study focused on four aspects of the validity of the IELI; its content validity, its pattern of relations with academic measures of achievement, its internal structure, and the consistency of its interpretation across defined groups. This chapter contains a discussion of the results pertaining to each of these four aspects of the validity of the IELI in turn, beginning with a discussion of the IELI construct theory and possible interpretations.

Content validity: Construct theory and the domain of behaviors measured

The IELI was intended to measure observable student behaviors that relate to academic achievement at the primary grade levels. Such behaviors are part of the learning process, indicative of student learning, or are supportive of the learning process. Such behaviors were labeled cognitive behaviors for the purpose of the construction of the IELI. The IELI construct theory was then derived empirically, identifying a collection of statements describing school related student behaviors using document analysis and searching for distinct areas of these behaviors that share similar characteristics. The document analysis used a sample of state curriculum guides, various existing measurement instruments and empirical research. Six areas of cognitive behaviors were then derived through a qualitative analysis of the resulting collection of behavior statements. They all shared the defining characteristic of relating to the learning process and being related to academic achievement. These cognitive behavior areas are: general knowledge, oral communication, written language, math concepts, work habits, and attentive behaviors. Two broad areas of behaviors, social behaviors and affective behaviors or conduct disorders, were excluded from the IELI on the grounds that these behavior areas did not fall under the definition of cognitive behaviors used. A short description of the behavior areas follows.

The general knowledge area can be described as behaviors or skills that are useful in multiple situations and are relevant to the independence of the student. An example is recognizing common opposites. Oral communication represents the students' mastery of the spoken language, capturing speech and listening as well as the communicated message. Examples are describing a thing or an event and choosing a relevant topic. Written language represents behaviors and skills involved in the manipulation of the letters and words as tools for symbolic representation of language. An example is recognizing letters.

Math concepts represents behaviors relevant for dealing with quantitative aspects of phenomena. An example is recognizing the relative value of coins. Work habits represents students' approach to the task at hand, both the disciplined execution of skills as well as persistence in the process of working on or solving a task. An example is how a student organizes her/his work. Attentive behaviors include items assessing the students' attendance to the instruction, learning tasks, or ongoing group work. An example is attending to or working on the assigned task.

The method of test construction provided an important piece of evidence relevant for the support of the content related aspect of the IELI. The content theory of the IELI was empirically derived from analysis of relevant literature and using qualitative analysis of the resulting pool of behavior statements. Its content was therefore systematically derived from the views and empirical results in the relevant literature. The method systematically identified separate areas with a common content in the pool of statements. The statements within an area shared an important theme that sets them apart from the other statements. The IELI items were then written using selected behaviors from within each area. Its content thus reflects the theme of each area identified in the qualitative analysis.

Statistical properties of the IELI

Some problems with the statistical properties of the IELI were evident at both grade levels. In particular the data suffered from skewness at both the individual item level and at the scale level. The seriousness of the problem can best be described by the fact that over half of the students obtained the highest score on a number of items as well as on a few scales. This led to problems with data analysis and called for changes in the plan for the analysis on two occasions. The fault may lie with the sample obtained for the present study, but it may also be that the behaviors described by the items have already been mastered by a majority of students at the age levels the IELI is intended to serve. These two possible underlying reasons have different implications. An explanation of how the first possibility may have occurred is that the literatures sampled for the document analysis may present an inaccurate view of student development. This would then be reflected in the statistical properties of the items. In the first case, a second sample would not necessarily share the skeweness. In the latter case, a second sample would probably also suffer from similar problems. Only further research could verify whether the source of the problems lies more with the items and the scale or with the data at hand. In either case this is not a very surprising outcome. Many measures of young children's characteristics have limited variability (Salvia & Ysseldyke, 1998).

The correlations of the IELI items with their respective cognitive behavior areas were generally quite high. The correlations of each item with behavior areas other than the one they were intended to measure were also fairly high. This indicates that there is, in general, something common among all the cognitive behavior areas as well as the behaviors chosen to construct the items representing them. This is also addressed directly in the results of the confirmatory factor analysis and indirectly in the regression analyses. Both supported, to some extent, the validity of separate but related constructs. The results are discussed further below.

IELI and academic achievement: Construct relevant variance

The present study investigated questions regarding the criterion related validity of the IELI cognitive behavior areas. Evidence regarding the relations of the areas with theoretically relevant external measures of academic achievement, or the absence of such relations, was important to indicate whether the behavior areas actually measured behaviors relevant for academic progress. The relationships an instrument such as the IELI is involved in are complex. The results of the present study indicate that all six cognitive behavior areas relate to each other as well as with all academic measures from the ITBS that were used as criterion variables. In fact, in four out of ten instances, the greatest correlation of the behavior areas with an ITBS score was not with one of the areas implied by the content theory as being the main relationship hypothesized. A plausible explanation for this is that all facets of achievement are interrelated to some extent. In order to establish or refute the hypotheses regarding cognitive behavior area to academic domain relationships a method that directly targets these relationships was used. These results established that the IELI and the ITBS shared variance that was relevant for the interpretation of both the IELI and the ITBS, or construct relevant variance in the language of modern validity theory (Messick, 1989). The results of these analyses are extremely important for the validity of the IELI interpretations. They indicate that a behavior area, say math concepts, yields information regarding a child that is related to academic achievement in mathematics. Similar statements can be made regarding the other IELI behavior areas.

The evidence regarding individual cognitive behavior areas varies. For example, oral communication added a significant portion of variance in the prediction of ITBS listening over and beyond all of the other behavior areas at the 1st grade level, but only with respect to two of the general areas at the kindergarten level. When its relation to reading was considered oral communication added an important portion of variance in ITBS reading at both grade levels. Written language had a strong specific relation with the ITBS word analysis test, but a somewhat weaker relationship with ITBS reading. Math concepts had a significant relationship with ITBS math over and above all other behavior areas. The results regarding the general behavior areas were somewhat different. General knowledge provided noteworthy information over and above some of the subject matter behavior areas at both grade levels, but work habits and attentive behaviors only added weak information with respect to a few achievement measures at the 1st grade level only.

The pattern of relationships of IELI behavior areas with ITBS achievement measures that are reported in the present study give rise to some questions that can not be investigated from the present data. The relationship of some behavior areas, for example, the general areas as well as some of the subject matter oriented areas, appeared to be stronger at the 1st grade level than at the kindergarten level. A potential explanation that is both consistent with background theory and may potentially be supported by further research is that absence of the behaviors measured by these areas hinders the child's acquisition of academic skills, but they have not accumulated to the extent that they are detectable by the instruments used at the kindergarten level. This interpretation assumes that the theoretically relevant variation accumulates over the years of schooling, and thus would predict stronger relationships of early IELI area scores with later achievement. For example, predicting 1st or 2nd grade achievement from kindergarten IELI might yield stronger relationships than those observed in the concurrent results that are reported in the present study. This is an important validity question that needs to be addressed in further studies of the IELI.

Overall, the results of the regression analyses indicated that the behavior areas are related to relevant measures of achievement. The behaviors used for the IELI items therefore relate to the measures of academic achievement that the IELI construct theory pinpoints as theoretically relevant. This supports the purported interpretation of the IELI area and composite scores as indicating the presence or absence of behaviors that enhance academic achievement.

Construct validity

When investigating the validity of an instrument such as the IELI, it is important to consider the extent to which empirical data regarding the internal structure of the instrument are consistent with its theoretical structure. The IELI theory implies six related but distinct behavior areas. In general, the series of confirmatory factor analyses conducted indicated weak support for a model consisting of six related constructs represented by the six cognitive behavior areas.

The model that best approached acceptable fit received only marginal support according to the evaluation criteria adopted for the present study. Two observations must be made on this point. First, the data obtained for the present study suffered from skewed distributions and limited ranges of responses to all of the items. The results of the analysis that explicitly took the problematic data into account yielded the strongest support for the goodness of fit for some of the models under study. Also, the most consistent evidence indicating adequate fit was obtained using an index sensitive to major misspecifications, the underlying factor structure, while indexes sensitive to misspecifications of individual loadings failed to support the models investigated. The model ranking second on most accounts also consisted of the six behavior areas, but their relationship was represented by a general factor rather than correlations among behavior areas. Other representations of models involving six behavior areas were shown to be in less accord with the empirical data.

Secondly, a modified model allowing direct correlations among items within behavior areas, but not across areas, improved the measures of fit without yielding firm evidence of fit of the model to the data. Given the similarity of item content for a number of item pairs belonging to different behavior areas, it is not unreasonable to expect that a respectable fit may be achieved by further modification, but that would violate the IELI construct theory. Also, improved fit might be obtained by allowing items to relate to a second behavior area. Such modifications would also violate the construct theory and were therefore not attempted in the present study.

Two facts indicate that the problems with the data referred to earlier served a role in the limited fit for the theoretical models obtained in the current CFA. One is the fact that the CFA estimation method that best takes such problems into account gave a better indication of model fit than the other estimation methods. Secondly, a better fit was obtained in the sample that suffered to a lesser extent from such problems, i.e. the kindergarten sample.

A post hoc study revealed that combining the cognitive behavior areas with the strongest correlations among them did not significantly improve model fit with respect to the results reported above. A model with a combined area consisting of work habits and attentive behavior was estimated, but failed to achieve acceptable fit.

The current study concludes that a model that best represents the IELI construct is one with six related factors, corresponding with the six behavior areas. This model is consistent with the IELI construct theory derived in the qualitative study of behavior statements described above. Despite the fact that the results failed to achieve levels of model fit recommended by current statistical research, it must still be considered the most promising model.

Differential item functioning analyses: Potential bias

The differential item functioning analyses conducted revealed distinct patterns of potential DIF for sex but results regarding DIF for ethnic groups were not interpretable. In the sex based analyses, about equally many DIF items favored boys as girls. Further, items from subject matter cognitive behavior areas tended to favor boys while items from the general areas tended to favor girls. DIF analyses for ethnic comparisons were based on insufficient sample sizes to yield meaningful results.

Methodological problems led to a change in the original plan regarding DIF analyses. This involved both a change of methods of analyses and of adopting a different perspective on the data. First, due to heavy skewness of responses at the item level, it was believed that a method that had originally been part of the plan of analyses should not be used. The method assumed a parametric relationship of item performance and the construct measured. Secondly, the number of respondents available for the ethnic DIF study was too limited to allow the use of anything but a simple 'quick and dirty' method. For these reasons the choice of methods was altered.

A further development due to the skewness of the data was the adoption of DIF methods designed to evaluate DIF for dichotomous items. Essentially these consisted of redefining the data and regarding them as a maximum score or a lower score. The fact that this led to results fairly consistent with methods utilizing the polytomous nature of the data must be regarded as supporting the robustness of the methods despite the problems with the data.

A DIF identification rule for the sex DIF was that an item must be flagged by at least three methods from different categories of methods. Based on this, relatively few items, seven out of 44 items, were flagged as DIF items.

The results of the sex DIF investigations draw attention primarily to three behavior areas. Two of the general behavior areas, work habits and attentive behaviors, appear to be particularly sensitive to differences in behavior of the sexes. Five of seven items consistently flagged as potentially exhibiting DIF belonged to these two areas. In addition, all items favored girls. Two items from math concepts favored boys.

This leads to a result that may potentially be one of the strengths of the IELI. It is well known that boys are more likely to be referred to special education or special needs programs on the basis of potential learning difficulties. How does this relate to the results regarding the IELI? Teachers may rely more heavily on recollections of behavior from work habits and attentive behaviors when referring a student to a diagnostic process. Or teachers may evaluate similar behaviors differently depending on the gender of student. Then an instrument such as the IELI that systematically captures teacher knowledge of various areas of student behavior may be useful in deriving a more balanced judgment. Then the question can be raised whether the addition of IELI profile to the information used when such referrals are made leads to a more balanced judgment. Given this situation, that three of the behavior areas are mostly free of DIF while the other three suffer from sex specific DIF, knowledge of the patterns among these areas could be useful to teachers and diagnosticians.

Overall the present results indicate that the IELI is only moderately affected by sex DIF. Relatively few items were flagged as exhibiting DIF, and they were concentrated in three behavior areas. The implications of this are that further studies of DIF on the IELI are need.

These results regarding sex and ethnic DIF share methodological problems. All analyses were conducted on combined samples of kindergarten and 1st grade students due to the limited data. Larger sample sizes would have been desirable for the analyses. Finally, the distributions of responses tended to be heavily skewed towards the upper end of the response scales. All these drawbacks on the methodological side tend to make the results regarding DIF weak. Further study must be conducted, and the present results can only be regarded tentatively. But they do serve as a warning sign that these issues should be investigated further in the future, if or when sufficient data are available.

Limitations

The present study is based solely on measures obtained during a single school year. Its focus is therefore limited in terms to an aspect of validity traditionally labeled concurrent validity. Data of this type are unsuitable for answering many very important questions related to the validity of the IELI; questions having to do with the relations of kindergarten or 1st grade IELI scores with measures of academic achievement at later grade levels. These questions are particularly important for an instrument such as the IELI that is intended for early warnings regarding potential academic difficulties later in a student's academic career. It is important to keep this limitation in mind because at its heart the IELI is a measurement tool intended to assess constructs that are of interest because of their relationship with students' future academic development. The present study was initiated as the construction phase of the IELI was ending, and practical considerations at the time precluded the collection of longitudinal data for this study. The investigation of the relationship of the IELI with students' later academic standing had therefore to be set aside in the present study. Despite this limitation, the present study provides important results pertaining to the validity of the IELI.

Another limitation that affected the methods of analysis used was that the obtained sample suffered from distributional characteristics that made certain analyses problematic. A probable reason is that relatively many students have mastered the measured behaviors. In that case, an alternative random sample may exhibit different distributional characteristics. The first results from the tryouts of an Icelandic translation and adaptation of the IELI have similar distributional characteristics. But other underlying processes may also be at work. For example, this may be an artifact of the large proportion of the sample that was obtained in a region where educational standing ranks relatively high. In that case, an alternative random sample may exhibit different distributional characteristics.

Further, the ethnic combination of the sample prohibited all but the simplest investigation of ethnic DIF. This, again, had the result that a large proportion of the sample was obtained in a state with a low proportion of the largest ethnic groups in the US population. Sample sizes for the relevant groups were too small for anything but a study that utilized a simplistic method.

This study was initiated when the construction of the IELI was being completed. The inventory has been modified since then and the published version currently available is slightly different from the version used for this study. The results of these analyses have been used in making these modifications.

Conclusion

The current study investigated several aspects of the validity of a new teacher rating instrument which has considerable potential utility for teachers working with kindergarten and 1st grade students. It was developed to serve the needs of teachers for early warning of students' potential learning difficulties. The IELI measures cognitive behaviors that support the learning process. An instrument that focuses on these behaviors can be useful for teachers needing to identify such students early in their educational career. Students, who at the beginning of their academic career lack these behaviors, are likely to experience learning difficulties later during their academic progress. It is more likely that a student's teacher and parents will be able to help her/him if aid is targeted at her/his needs early in her/his educational career. The IELI can help teachers identify these students early on and thus provide valuable information for the teachers in early grades.

REFERENCES

Achenbach, T. M. (1978). The Child Behavior Profile: I. Boys aged 6-11. *Journal of Consulting and Clinical Psychology, 46*(3), 478-488.

Achenbach, T. M. (1992). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington, VT: University of Vermont Department of Psychology.

Adelman, H., and Feshbach, S. (1971). Predicting reading failure: Beyond the readiness model. *Exceptional Children*, 37, 349-354.

Alabama State Department of Education (1992). *Alabama course of study: Social studies.* Montgomery, AL: Author

Alexander, K. L., Entwisle, D. R., and Dauber, S. L. (1993). First-Grade classroom behavior: Its short- and long-term consequences for school performance. *Child Development, 64*, 801-814.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A., and Urbana, S. (1997). *Psychological testing* (7th. ed.). Upper Saddle River, NJ: Prentice-Hall.

Anderson, J. R. (1983). *Architecture of cognition*. Cambridge, MA: Harvard University press.

Anderson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). New York: Freeman.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook for methods of detecting test bias.* (pp. 96-116). Baltimore, MD: Johns Hopkins University Press.

Badian, N. A. (1982). The prediction of good and poor reading before kindergarten entry: A 4-year follow-up. *Journal of Special Education,16*, 309-18.

Badian, N. A. (1988). The prediction of good and poor reading before kindergarten entry: A nine-year follow-up. *Journal of Learning Disabilities, 21,* 98-103, 123.

Bentler, P. M., and Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bergkamp, S., and Rosegrant, T. (1992). Reaching potentials through appropriate curriculum. In S. Bergkamp and T. Rosegrant (Eds.), *Reaching potentials: Appropriate curriculum and assessment for young children*. Washington, DC: National Association for the Education of Young Children.

Biehler, R. F., and Snowman, J. (1993). *Psychology applied to teaching* (7th ed.). Boston: Houghton Mifflin.

Boling, J. and Fletcher, G. (1973). The MRT vs. ratings of kindergarten children as predictors of success in first grade. *Educational Leadership, 30,* 637-39.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Bond, G. L., and Dykstra, R. (1997/1967) The cooperative research program in first-grade reading instruction. *Research in Reading Quarterly, 32*, 345-427.

Bracken, B. A., Keith, L. K., and Walker, K. C. (1998). Assessment of preschool behavior and social-emotional functioning: A review of thirteen third-party instruments. *Journal of Psychoeducational Assessment, 16* (2), 153-169.

Brookhardt, S. (1998). Review of the Iowa Tests of Basic Skills. In J. C. Impara, and B. S. Plake, (Eds.), *Thirteenth mental measurement yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Bryan, T. S. (1974). An observational analysis of classroom behaviors of children with learning disabilities. *Journal of Learning Disabilities, 7*, 34-43.

Bryan, T. S. and McGrady, H. J. (1972). Use of a teacher rating scale. *Journal of Learning Disabilities, 5*, 199-206.

Bryan, T. S., and Wheeler, R. (1972). Perceptions of learning disabled children: The eye of the observer. *Journal of Learning Disabilities, 5*, 484-488.

Bush, R. F. (1980). Prediction of first-grade reading achievement. *Learning Disabilities Quarterly, 3*, 38-38.

Butler, S. R., Marsh, H. W., Sheppard, M. J., and Sheppard, J. L. (1982). Early prediction of reading achievement with the Sheppard School Entry Screening Test: A four year longitudinal study. *Journal of Educational Psychology, 74*, 280-290.

Butler, S. R., Marsh, H. W., Sheppard, M. J., and Sheppard, J. L. (1985). Seven-year longitudinal study of early prediction of reading achievement. *Journal of Educational Psychology, 77*, 349-361.

California Department of Education (1988). *History-social science framework for California public schools Kindergarten through grade twelve*. Sacramento, CA: Author.

Camilli, G., and Shephard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Camp, B. W., and Zimet, S. G. (1974). The relationship of teacher rating scales to behavior observations and reading achievement of first-grade children. *Journal of Special Education, 8*, 353-359.

Cartledge, G., and Milburn, J. (1978). The case for teaching social skills in the classroom: A review. *Review of Educational Research, 48*, 133-56.

Chang, H., Mazzeo, J., and Roussos, , L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Chiu, L.-H. (1997). Development and validation of the School Achievement Motivation Rating Scale. *Educational and Psychological Measurement, 57*, 292-305.

Cobb, J. A. (1972). The relationship of discrete classroom behaviors to fourth-grade academic achievement. *Journal of Educational Psychology, 63*, 74-80.

Cole, N., and Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 201-219). Phoenix, AZ: American Council on Education, Oryx Press.

Collingen, R. C. (1979). Predictive utility of the Mykleburst Pupil Rating Scale: A two year follow up. *Journal of Learning Disabilities, 12,* 264-267.

Cone, J. (1982). Validity of direct observation assessment procedures. *New Directions for Methodology of Social and Behavioral Science, 14*, 67-79.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry, 126*, 884-888.

Conoley, J. C., and Impara, J. C. (Eds.)(1995). *The twelfth mental measurement yearbook*. Lincoln, NE: Buros Institute of Mental Measurement.

Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace.

Cross, S. (1998). Review of the Iowa Tests of Basic Skills. In J. C. Impara, and B. S. Plake (Eds.), *The thirteenth mental measurement yearbook*. Lincoln, NE: Buros Institute of Mental Measurement.

Daly, E. J., Wright, J. A., Kelly, S. Q., and Martens, B. K. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. *School Psychology Quarterly, 12* (3), 268-280.

Day, A. M. L., and Peters, R. D. (1989). Assessment of attentional difficulties in underachieving children, *Journal of Educational Research, 82* (6), 356-361.

Derevensky, J. L., Hart, S. Farrell, M. (1983). An examination of achievement-related behavior of high- and low achieving inner city pupils. *Psychology in the Schools, 20*, 328-334.

DuPaul, G. J., Rapport, M. D., and Perroello, L. (1991). Teacher ratings of academic skills: The development of the academic performance rating scale. S*chool Psychology Review, 20*, 284-300.

Eggen, P., and Kauchak, D. (1999). *Educational psychology: Windows on classrooms* (4th ed.). Upper Saddle River, NJ: Merrill.

Fad, K., and Ryser, G. (1993). Social/Behavioral variables related to success in general education. *Remedial and Special Education, 14*, 25-35.

Feldt, L. S., and Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 105-146). Phoenix, AZ: American Council on Education, Oryx Press.

Ferster, C. B., Culbertson, S., and Boren, M. C. P. (1975). *Behavior principles* (2nd ed.). Englewood Cliffs, N.J: Prentice-Hall.

Feschbach, S., Adelman, H., and Fuller, W. W. (1974). Early identification of children with high risk of reading failure. *Journal of Learning Disabilities, 7*, 639-644.

Flynn, J. M., and Rahbar, M. H. (1998a). Improving teacher prediction of children at risk for reading failure. *Psychology in the Schools, 35*, 163-172.

Flynn, J. M., and Rahbar, M. H. (1998b). Kindergarten screening for risk of reading failure. *Journal of Psychoeducational Assessment, 16,* 15-35.

Foster, S. L. and Cone, J. D. (1980). Current issues in direct observation. *Behavior Assessment, 2*, 313-338.

Gambell, L. B., Wilson, R. M., and Gnatt, W. N. (1981). Classroom observation of task-attending behaviors of good and poor readers. *Journal of Educational Research, 74* (6), 400-404.

Gene, C. and Peshkin, A. (1992). *Becoming a qualitative researcher: An introduction.* White Plains, NY: Longman.

Glazzard, P. (1980). Teacher rating and reading readiness as predictors of vocabulary and comprehension achievement in first, second, third and fourth grade. *Learning Disability Quarterly, 5*, 35-45.

Gleitman, H. (1992). *Basic* p*sychology* (3rd Ed.). NY: Norton.

Gorsuch, R. L. (1983). *Factor analysis* (2nd. ed.).Hillsdale, NJ: Lawrence Earlbaum Associates.

Goyette, C. H. Conners, C. K. and Ulrich, R. F. (1978). Normative data on Revised Connors Parent and Teacher Rating Scales. *Journal of Abnormal Child Psychology, 6*. 221-236.

Greeno, J. G., Collins, A. M., and Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner and R. C. Calfee, (Eds.), *Handbook of educational psychology* (pp. 15-46). NY: Macmillan.

Gresham, F. M. (1982). A model for the behavioral assessment of behavior disorders in children: Measurement considerations and practical application. *Journal of School Psychology, 20*, 131-144.

Grogan, S. C. (1995). Which cognitive abilities at age four are the best predictors of reading ability at age seven? *Journal of Research in Reading, 18*, 24-31.

Gronlund, N. E. (1993). *How to make achievement tests and assessments* (5th ed.). Boston: Allyn & Bacon.

Harrison, P. L. (1985). *Classroom edition manual: Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Service.

Hightower, A. D., Work, W. C., Cowen, E., Lotsyczewski, B., Spinell, A., Guare, J. G., and Rhohrbeck, C. A. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary problem behaviors and competencies.. *School Psychology Review, 15*, 393-409.

Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *Journal of Special Education, 17*, 401-429.

Hoge, R. D. (1985). Validity of direct observation measures of pupil classroom behavior. *Review of Educational Research, 55*, 469-483.

Hoge, R. D. and Luce, S. (1979). Predicting academic achievement from classroom behavior. *Review of Educational Research, 49*, 479-496.

Holland, P. W., and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 35-66). Hillsdale, NJ: Lawrence Earlbaum Associates.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., and Qualls, A. L. (2001). *Iowa Early Learning Inventory*. Chicago, IL: Riverside.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberly, K. R., Cantor, N. K., Bray, G. B., Lewis, J. C., and Qualls-Payne, A. L. (1993). *Iowa Tests of Basic Skills: Interpretative guide for teachers and counselors: Levels 5-8*. Chicago, IL: Riverside.

Hu, L-T., and Bentler, P. M. (1998). Fit indexes in covariance structure modeling: Sensitivity to underparameterized misspecification. *Psychological Methods, 3*, 424-453.

Hu, L-T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Impara, J. C., and Plake, B. S. (Eds.) (1998). *The thirteenth mental measurement yearbook*. Lincoln, NE: Buros Institute of Mental Measurement.

Indiana Department of Education (1987). *Social studies proficiency guide*. Indianapolis, IN: Author.

Indiana Department of Education (1992). *English/Language arts proficiency guide: Essential skills for Indiana students*. Indianapolis, IN: Author.

Jorm, A. F., Share, D. L., McClean, R., and Matthews, R. (1986). Cognitive factors at school entry predictive of specific reading retardation and general reading backwardness: A research note. J*ournal of Child Psychology and Psychiatry, 27*, 45-54.

Kelly, M. S., and Peverly, S. T. (1992). Identifying bright kindergartners at risk for learning difficulties: Predictive validity of a kindergarten screening tool. *Journal of School Psychology, 30*, 245-258.

Kenny, D., and Chekaluk, E. (1993), Early reading performance: A comparison of teacher-based and test-based assessments. *Journal of Learning Disabilities, 26*, 227-236.

Kentucky Department of Education (1993a). *State regulations and recommended best practices for Kentucky's primary program*. Frankfort, KY: Author.

Kentucky Department of Education (1993b). *Transformation: Kentucky's curriculum framework (Vol. I)*. Frankfort, KY: Author.

Kim, Y., Anderson, H. E., and Bashaw, W. L. (1968). Social maturity, achievement, and basic ability. *Educational and Psychological Measurement, 28*, (2), 535-543.

Kohn, M. and Roshman, B. L. (1974). Social-emotional, cognitive, and demographic determinants of poor school achievement: Implications for a strategy of intervention. *Journal of Educational Psychology, 66*, 266-276.

Lahaderne, H. M. (1968). Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. *Journal of Educational Psychology*, 59 (5), 320-324.

Lahaderne, H. M., and Jackson, P. W. (1970). Withdrawal in the classroom: A note on some educational correlates of social desirability among school children. *Journal of Educational Psychology, 61*, 97-101.

Landy, F. J. and Farr, J. L. (1980). Performance ratings. *Psychological Bulletin, 87*, 72-107.

Lam, C. M., and Beale, I. L. (1991). Relations among sustained attention, reading performance, and teachers' rating of behavior problems. *Remedial and Special Education, 12*, 40-47.

Lane, S. (1995). Review of the Iowa Tests of Basic Skills. In J. C. Conoley and J. C. Impara, (Eds.) *The twelfth mental measurement yearbook*. Lincoln, NE: Buros Institute of Mental Measurement.

LeCompte, M. C., and Goetz, J. P. (1982). Problems of reliability and validity in ethnographic research. *Review of Educational Research, 52*, 31-60 .

Lindholm, B., Touliatos, J., and Rich, A. (1976). A canonical correlation analysis of behavior problems and school achievement for different grades, sexes and races. *Journal of Educational Research, 70*, 340-42.

Lloyd, J. W. and Loper, A. B. (1986). Measurement and evaluation of task-related behaviors: Attention to task and metacognition. *School Psychology Review, 15*, 336-345.

Lowell, R. E. (1971). Reading readiness factors as predictors of success in first grade reading. *Journal of Learning Disabilities, 4*(10), 24-28.

Mantel, H., and Haenzel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Marsh, H. W., and Hau, K. T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64*, 364-390.

Mazor, K. M., Clauser, B. E. and Hambleton, R. K. (1992). The effect of sample size on the functioning of the Maentel-Haenzel DIF statistic. *Educational and Psychological Measurement, 52(2)*, 443-451.

McDermont, P. A., and Beitman, B. S. (1984). Standardization of a scale for the study of children's learning styles: Structure, stability, and criterion validity. *Psychology in the Schools, 21*, 5-14.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, N.J. : Lawrence Erlbaum Associates.

McDonald, R. P. and Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64-82.

McKinney, J. D. (1989). Longitudinal research on the behavior characteristics of children with learning disabilities. *Journal of Learning Disabilities*, *22*, 141-150.

McKinney, J. D., Mason, J., Perkerson, K., and Clifford, M. (1975). Relationship between classroom behavior and academic achievement. *Journal of Educational Psychology*, *67*, 198-203.

Mercer, C. D., Algozzine, B., and Trifiletti, J. (1988). Early identification- an analysis of the research. *Learning Disabilities Quarterly, 11*, 176-188.

Messick, S. (1984) The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 13-103). Phoenix, AZ: American Council on Education, Oryx Press.

Meyers, C. E., Atwell, A. A., and Orpit, R. E. (1968). Prediction of fifth-grade achievement from kindergarten test and rating data. *Educational and Psychological Measurement, 28*, 457-463.

Millman, J., and Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 335-366). Phoenix, AZ: American Council on Education, Oryx Press.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Muňiz, J., Hambleton, R. K., and Xing D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115-135.

Muthén, L. K., and Muthén, B. O., (2002). *Mplus 2.12.* (Computer software) Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., and Muthén, B. O., (2001). *Mplus user's guide* (2^{nd} Ed.) Los Angeles, CA: Muthén and Muthén.

Naglieri, J. A., Bardos, A. N., and LeBuffe, P. A. (1995). Discriminant validity of the Devereux Behavior Rating Scale-School Form for students with serious emotional disturbance. *School Psychology Review, 24*, 104-111.

Naglieri, J. A. and Flanagan, D. P. (1993). Psychometric characteristics of commonly used behavior rating scales. *Comprehensive Mental Health, 2*, 225-239.

National Association for the Education of Young Children (1988). Position statement on standardized testing of young children 3 through 8 years of age. *Young Children, 43* (3), 42-47.

Nebraska Department of Education and Iowa Department of Education (1993). *The primary program: Growing and learning in the heartland*. Lincoln, NE: Nebraska Department of Education.

Neeper, R., and Lahey, B. B. (1984). Identification of two dimensions of cognitive deficits through the factor analysis of teacher ratings. *School Psychology Review, 13* , 485-490.

Neeper, R., and Lahey, B. B. (1986). The Children's Behavior Rating Scale: A factor analytic developmental study. *School Psychology Review, 15,* 277-288.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J, and Svinicki, J. (1984a). *Battelle Developmental Inventory: Cognitive domain*. Chicago, IL: Riverside Publishing Company.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J, and Svinicki, J. (1984b). *Battelle Developmental Inventory: Communication domain*. Chicago, IL: Riverside Publishing Company.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J, and Svinicki, J. (1984c). *Battelle Developmental Inventory: Personal-social domain*. Chicago, IL: Riverside Publishing Company.

Nitko, A. J. (1996). *Educational assessment of students* (2nd. Ed.). Englewood Cliffs, NJ: Merrill.

Oregon State Department of Education (1986). *Essential learning skills*. Salem, OR: Author.

Oregon State Department of Education (1988). *Oregon Kindergarten guide*. Salem, OR: Author.

Oregon State Department of Education (1990). *Social studies: Common curriculum goals*. Salem, OR: Author.

Parshall, C.G., and Miller, T. R. (1995). Exact versus asymptotic Mantel-Henzel DIF statistic: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32,* 302-316.

Patterson, D. S. (1997). *An exploration of the reliability and validity of a rating scale of primary school children's classroom-academic-related behaviors*. Unpublished doctoral dissertation: Pennsylvania State University.

Penfield, R. D., and Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19* (3), 5-15.

Perkins, H. V. (1965). Classroom behavior and underachievement. *American Educational Research Journal*, 2, 1-12.

Potenza, M. T., and Dorans, N. J. (1995). DIF Assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Quay, L. C., and Steele, D. C. (1998). Predicting children's achievement from teacher judgments: An alternative to standardized testing. *Early Education and Development, 9*, 207-218.

Resnick, L. B., and Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford and M. C. O'Connor (Eds.), *Changing assessments: Alternate views of aptitude, achievement, and instruction*. Boston: Kluwer.

Reynolds, C. R., and Kamphaus, R. W. (1992). *Behavior assessment systems for children: Manual*. Circle Pines, MN: American Guidance Service.

Reynolds, W. M. (1978). *Development of a scale to measure learning-related behaviors*. Paper presented at the annual meeting, American Psychological Association, Toronto, Canada, August, 1978. (ERIC ED173366).

Reynolds, W. M. (1979). Development and validation of a scale to measure learning related classroom behaviors. *Educational and Psychological Measurement, 39,* 1011-1018.

Reynolds, W. M., and Bernstein, S. M. (1982). A factorial validity study of the Devereux Elementary School Behavior Rating Scale. *Journal of Abnormal Child Psychology, 10*, 113-122.

Ross, A. O., Lacey, H. M., and Parton, D. A. (1965). The development of a behavior checklist for boys. *Child Development*, *36*, 1013-1027.

Roussos, L. A. and Stout, W. F. (1996). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantél-Haenszel Type I error performance. *Journal of Educational Measurement, 33* (2), 215-230.

Saal, F., Downey, R., and Lahey, M. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413-428.

Salvia, J., and Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston, MA: Houghton Mifflin.

Samuels, S. J., and Turnure, J. E. (1974). Attention and reading achievement in first-grade boys and girls. *Journal of Educational Psychology, 66*, 29-32.

Satz, P., and Fletcher, J. (1988). Early identification of learning disabled children: An old problem revisited. *Journal of Consulting and Clinical Psychology, 56*, 824-829.

Schaefer, E. S. (1975). *Major replicated dimensions of adjustment and achievement: Cross-cultural, cross-sectional and longitudinal research*. Paper presented at the annual meeting, American Educational Research Association, Washington, D.C., April 3, 1975. (ERIC ED115390).

Schaefer, E. S., Baker, E., and Zawel, D. (1975). A factor analysis and reliability study of the Devereux Elementary School Behavior Rating Scale. *Psychology in the Schools, 12*. 295-300.

Scheuneman, J. D., and Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2*, 255-275.

Seidman, E., Rapport, J., Kramer, J., Liney, J. A., Herzeberger, S., and Alden, L. (1979). Assessment of classroom behavior: A multiattribute, multisource approach to instrument development and validation. *Journal of Educational Psychology, 71*, 451-464.

Selders, J. (1981). *The predictive effectiveness of the Kindergarten Rating Scale as measured by standardized readiness tests and total reading achievement*. Unpublished doctoral dissertation, The University of Kansas.

Sharpley, C. F. and Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*, 106-111.

Shealy, R. and Stout, W. F. (1993). A moel-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58* (2), 159-194.

Shealy, R. T., and Stout, W. T. (1993). An item response model for test bias and differential test functioning. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 197-239). Hillsdale, NJ: Lawrence Earlbaum Associates.

Shepard, L. A. (1993). Evaluating test validity. In L. Darlington-Hammond, (Ed.) *Review of Research in Education*, *19*, 405-450. Washington, DC: AERA.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Snow, R. E. (1980). Aptitude and achievement. In W. Schrade (Ed.), *New Directions for Testing and Measurement, no.19: Measuring Achievement: Progress over a Decade* (pp. 39-59). San Francisco, CA: Jossy-Bass.

Snow, R. E., and Lohman, D. F. (1984). Towards a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*, 347-76.

Snow, R. E., and Lohman, D. F. (1989). The implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 263-331). Phoenix, AZ: American Council on Education, Oryx Press.

Soli S. D., and Devine, V. T. (1976). Behavioral correlates of achievements: A look at high and low achievers. *Journal of Educational Psychology*, *68,* 335-341.

Sparrow, S. S., Balla, D. A., and Cicchetti, D. V. (1985). *Vineland Adaptive Behavior Scales: Classroom edition*. Circle Pines, MN: American Guidance Service.

Spivack, G., and Swift, M. S. (1967). *Devereux Elementary School Behavior Rating Scale: manual*. Devon, PA: Devereux Foundation.

Spivack, G., and Swift, M. S. (1973). The classroom behavior of children: A critical review of teacher-administered rating scales. *Journal of Special Education 7*, 55-89.

Stiggins, R. J., and Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, *22,* 271-286.

Stiggins, R. J., Conklin, N. F., and Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice*, *5* (2), 5-17.

Stout, W., and Roussos, L. (1996). *SIBTEST manual.* Urbana-Champaign, Il: The William Stout Institute for Measurement.

Swift, M., and Spivack, G. (1968). The assessment of achievement-related classroom behavior. *Journal of Special Education, 2*, 137-153.

Swift, M. S., and Spivack, G. (1969). Clarifying the relationship between academic success and overt classroom behavior. *Exceptional Children, 36,* 99-104

Telgredy, G. A. (1975). The effectiveness of four readiness tests as predictors of first grade academic achievement. *Psychology in the Schools, 12*. 4-11.

Tennessee Department of Education (1992). *Tennessee comprehensive curriculum guide: Grades K - 8* Nashville, TN: Tennessee Department of Education.

Tollefson, M., Rodriquez, R., and Glazzard, P. (1985). Predicting reading achievement for kindergarten boys and girls. *Psychology in the Schools, 22*, 34-39.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., and Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. NY: Macmillan.

Tramontana, M. G., Hooper, S. R., and Selzer, S. C. (1988). Research on the preschool prediction of later achievement: A review. *Developmental Review, 8*, 89-146.

Turnure, J. E. and Samuels, S. J. (1972). *Attention and reading achievement in first grade boys and girls.* Research Report No. 43. Minneapolis, MN: Center in Education of Handicapped Children, Minnesota University (ERIC ED074447).

Wallbrown, J. D., Wallbrown, F. H., and Engin, A. (1977). The prediction of first grade achievement with behavioral rating taken during kindergarten. *Journal of Experimental Education, 45* (4), 16-20.

Welch, C. (1991). *Procedures for extending item bias detection techniques to polytomously scored items.* Unpublished Ph.D. dissertation, University of Iowa.

Welch, C., and Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6*, 1–19.

White, K. R. (1986). Efficacy of early intervention. *Journal of Special Education, 19*, 401-416.

Wisconsin Department of Public Instruction (1986). *A guide to curricular planning in social studies*. Madison, WI: Author.

Wisconsin Department of Public Instruction (1991). *A guide to curricular planning in English language arts*. Madison, WI: Author.

Yu, C-Y., and Muthén, B. O. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Technical report,

Zwick, R., Thayer, D. T., and Mazzero, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321-344.