6-1-2009

Development of Test of Academic Readiness

Tanachit Kasintorn
University of Denver

DEVELOPMENT OF TEST OF ACADEMIC READINESS

A Dissertation

Presented to

the Morgridge College Of Education

University of Denver

In Partial Fulfillment

of the Requirement for the Degree

Doctor of Philosophy

---

by

Tanachit Kasintorn

June 2009

Advisor: Professor Kathy E. Green

Author: Tanachit Kasintorn
Title: Development of Test of Academic Readiness
Advisor: Professor Kathy E. Green
Degree Date: June 2009

ABSTRACT

The purpose of this study was to develop a test of academic readiness for first grade instruction in Thailand. Test of Academic Readiness (TAR) consists of six domains: verbal, visual, memory, math, logical, and general knowledge. Two pilot studies were carried out and a main study tested items in those domains. Rasch model was used to assess the scale's level of reliability and item discrimination. Content validity was claimed through extensive review of literature and similar readiness tests both in the U.S. and Thailand.

TAR achieved a range of reliability between 0.73-0.93 with the exception of the visual subtest, with a reliability of 0.43. With the exception of the visual subtest, TAR has sufficient reliability to be worthwhile for future research to improve TAR as a measure of academic readiness for first grade instruction in Thailand. Further validation studies are recommended.

ACKNOWLEDGEMENT

This dissertation is dedicated to my mother (Mrs. Kwanyoen Kasintorn), my father (Mr.Charoenchai Kasintorn), and my wife (Mrs.Phuangpaka Kasintorn).

This dissertation is not possible without the following extraordinary persons: Professor Kathy E. Green agreed to serve as a chairperson to give a much need direction for this dissertation. She has also made it possible for this dissertation to utilize one of the most advanced item response theory for scale development. She has provided excellent and individualized trainings. She has consistently lent her industry leading expertise, administrative support, and moral support for almost 10 years of this dissertation endeavor. Professor Martin L. Tombari has inspired the idea of creating a scale as a topic for this dissertation. He has also provided consistent assistance and moral support throughout the entire dissertation process. Dean Emerita Ellinor L. Katz has been very supportive and accommodating. She is ready to help a student to fulfill the dissertation requirements even at her own difficulty. Professor Karin Dittrick-Nathan has been a warm addition to the dissertation committee. Her expertise has helped to inspire future research topics. Professor Enid O. Cox has been very supportive and accommodating. Her participation in the committee has helped this dissertation completion a reality.

There are many other people who have contributed in a way or another toward the successful completion of this dissertation. They are recognized here for their involvement and support.

Table Of Contents

List of Tables

List of Figures

CHAPTER 1

Statement of the Problem

The concept of school readiness has been debated for more than a century (Kagan, 1990). Different conceptualizations of school readiness may have been the cause of this long debate (Kagan, Moore, & Bredekamp, 1995). Take, for example, the two prevalent perspectives of school readiness: ready schools and ready students (SECA Public Policy Institute Report, 1993). The first perspective reflects the schools being ready to provide a learning environment that caters to children's needs, and hence ready schools (SECA Public Policy Institute Report, 1993; Lewit & Baker, 1995). The second perspective concerns the children being ready for their formal schooling and hence ready students. Since this dissertation deals with a child's level of readiness for school, any further usage of the term "school readiness" is based on the second perspective.

Besides ready school versus students, readiness can include readiness for first grade or for kindergarten. The literature uses the term "school readiness" interchangeably when referring to first grade and kindergarten readiness. The reason for such usage of the term may be due to a close similarity between kindergarten readiness and first grade readiness (SECA Public Policy Institute Report, 1993; Kagan et al., 1995; Nurss, 1987; Nurss & Hodges, 1982). On the one hand, the level of readiness required for kindergarten is much lower than that for first grade. On the other hand, the skill types

for readiness are surprisingly similar for both. There are three explanations for such close similarity.

First, beginning in the latter part of the 19th century, kindergartens became part of elementary schools (De Cos, 1997). Prior to that period, kindergarten was philosophically distinct from the primary grades. Kindergarten was meant to foster the natural development of children. Play was an important means of self-development. During most of the 20th century, kindergarten's curriculum has focused on discipline, neatness, structured lessons, and recitations. Kindergarten was viewed as a transition for children from home to elementary school. Now kindergarten is an integral part of the elementary school's curriculum (Nurss & Hodges, 1982). The content of the kindergarten curriculum is now more tightly coordinated with that of the primary grades (De Cos, 1997; Nurss, 1987). This is one reason why the concept of readiness for kindergarten has become very similar to that of readiness for first grade.

Second, the focus of the kindergarten curriculum has shifted from social to a cognitive developmental emphasis. Since the cognitive emphasis has always been central to the elementary curriculum (Nurss & Hodges, 1982), the line between kindergarten and first grade readiness has been blurred.

Third, the concept of school readiness is equally applicable to either kindergarten readiness or first grade readiness. The conceptualization of school readiness depends more on the context in which the term is applied. For example, experts defined school readiness as a child's level of readiness to meet the demands of schooling (Nurss, 1987; Kagan, 1992). Accordingly, kindergarten readiness means a child's level of readiness to

meet the demands of kindergarten schooling. First grade readiness means a child's level of readiness to meet the demands of first grade schooling.

Since this dissertation deals with the readiness for first grade instruction, the context to which the term refers in this dissertation is first grade. Any further usage of the term "school readiness" means readiness for first grade instruction.

Need for an Academic Readiness Measure

Success in school depends on many factors. Academic readiness is one such factor. Some students enter school more ready than others. The gap in the students' readiness level has many ramifications on both the students and the schools. For example, more ready students often get bored and become disinterested when the teachers are forced to teach more slowly to the less ready students. And schools face issues such as ability grouping, after-school tutoring, disinterested students, failing students, and enrollment screening.

Enrollment screening is one of many strategies schools use to combat the problem. As opposed to a selection tool, a screening tool provides a direction for appropriate intervention rather than a decision about promotion (Axford, 1992). A screening tool can help schools to identify the less ready students prior to admission. A major benefit of a screening tool is to allow schools to institute an appropriate intervention program for the less ready students. The intervention program can help to improve the readiness level of the less ready students. When the readiness gap disappears, many of the problems schools have been facing are eliminated.

Schools in Thailand use a wide variety of screening tools to screen student applicants, possibly because Thai parents are as concerned about their child's level of academic readiness for first grade as are U.S. parents. The screening tools used by Thai schools range from entrance examination, to parent interviews, to observation. The most commonly used tools are the entrance examination, which consists of items in a variety of self-response question formats (e.g., multiple choices, fill in the blank, and matching). Many schools in Thailand develop their own entrance exam questions. The internally developed tests often require group administration. Other schools use a translated version of achievement tests which were developed in other countries.

There are many problems with both types of tests the schools have been using to screen students. For example, the translated tests were not developed with Thai children in mind. Although translated, the items have not been adjusted for the differences in cultural and socioeconomic conditions. Consequently, the extensive reliability and validity that supports the original tests are not carried over to the translated version. The internally developed tests are equally, if not more, problematic. There have been no attempts by school officials, who developed the tests, to ensure the reliability and the validity of the tests. Consequently, the scores of such tests do not necessarily provide a reliable and valid inference of the students' readiness level for first grade instruction. As opposed to those of a readiness test, the scores of a translated achievement test do not necessarily provide an indicator of readiness. The skills sampled by the tests such as achievement tests do not necessarily reflect the readiness areas found in the literature

review. As will be described later in Chapter 2, there is a great difference between an achievement and a readiness test.

Thai government officials are calling for a true readiness test that can assess the readiness for first grade instruction. A true measure of academic readiness represents the first grade content areas for which the students should be ready. It is reliably measuring the level of readiness. And it provides an acceptable level of indication of first grade success.

Since there is no such measure available in Thailand, a new measure of academic readiness is needed. The purpose of this dissertation was to develop a new scale, called Tests of Academic Readiness (TAR), which measures different components of academic readiness for first grade instruction in Thailand. If reliable and valid, TAR will be an improvement over the entrance examinations used by many schools in Thailand. TAR will be a true academic readiness measure, as opposed to the achievement tests used by many schools in Thailand. TAR is designed specifically for children aged 5-6. This age range was selected because it is the age range of most kindergarten graduates, who continue on to first grade.

To ensure that the new scale contains the characteristics described above, two questions were raised. These questions helped to guide the development of the new scale. The research questions were as follow:

Specific Research Questions

1. Does the Test of Academic Readiness possess content validity?
2. Is the Test of Academic Readiness a reliable measure of academic readiness for first grade instruction in Thailand?

Definitions

Academic readiness for this dissertation was defined as the level of readiness in academic skill areas that help a child in Thailand to successfully complete first grade instruction.

The skill areas, which were contained in TAR, included verbal, visual, memory, math, logical, and general knowledge. Verbal skills were defined as a child's ability to read vocabulary, to write vocabulary, and to read and comprehend a short story. Visual skills were defined as a child's ability to recognize an everyday object, to identify everyday objects, to matching objects, to recognize objects in a different spatial arrangement. Memory skills were defined as a child's ability to immediately recall pictures, alphabets, words, or a series of events and a child's ability to recall pictures, alphabets, words, or a series of events after an extended period of time. Logical skills were defined as a child's ability to find conceptual relationship in objects, to find a pattern of a series of events, and to be able to identify a correct order of a series of events. General knowledge was defined as a child's level of common knowledge, which is acquired through grade appropriate training, everyday experiences, and an age appropriate supplementary self study.

Delimitations

This study was limited to children aged between 5 and 6 from urban schools in Bangkok, Thailand. As such, results from this study may not be generalizable to children of the same age from other types of schools in Thailand, who are entering first grade. In addition, this study was limited to developing a test containing six academic readiness areas that were thought to measure corresponding academic readiness in matching subject areas taught in first grade in Thailand. This study may only be inferred that a child, who scores highly in this study, would do well in first grade in corresponding subject matters the six academic readiness components found in TAR purported to measure.

CHAPTER 2

Literature Review

This chapter begins with a review of the literature on the concept of school readiness. The literature review highlights variations in the conceptualization of school readiness. The explanation of the conceptual variations leads to a discussion of readiness types. Then, school readiness as defined by Thai experts and organizations is discussed. A comparison among the experts' definitions of school readiness is made. Another comparison is also made between types of school readiness for American students and for Thai students.

Next, available measures of school readiness are evaluated. The shortcomings of these measures indicate the need for a new measure of first grade readiness for Thai students. The strengths of these measures serve as a list of desirable test characteristics the new measure should incorporate. The desirable test characteristics as well as the experts' definitions of school readiness serve as a framework, within which school readiness was defined for this dissertation.

Debates on Concept of School Readiness

There is great variation in the conceptualization of school readiness. For instance, Nurss (1987) defined readiness as the preparedness for what comes next. Kagan (1992) defined school readiness as the ability to meet the task demands of schooling and to

successfully acquire the curriculum content. Cronbach (1970) defined school readiness as the learner's capacity and maturity in different areas that allow learners to perform or respond to requirements. He also outlined readiness as two types: physical maturity and intellectual maturity. Others defined readiness more specifically in terms of the child's characteristics in such areas as social, emotional, and/or intellectual abilities (Lewit & Baker, 1995). These experts disagree, however, about what constitutes an ideal list of readiness characteristics. Many experts believed the list should include language capacity, intellectual and perceptual functioning, and gross and fine motor coordination (Kagan et al., 1995; Katz, 1991; Kunesh & Farley, 1993; Seefeldt & Barbour, 1994).

Lamberty and Crnic (1994) included physical, intellectual, and social development standards in their list. Some experts consider age as another legitimate standard since it dictates the level of certain intellectual development (e.g., language skills) (Downing & Thackrey, 1971; Katz, 1991; Lewit & Baker, 1995). Other experts go beyond intellectual and physical capacities. They argue that children with behavioral problems often cannot participate fully during class activities. Such behaviors cost not only the child but also his or her classmates opportunities to learn. Therefore, these experts include the social-emotional domain as part of the list as well (Gracey, Carey, & Reinherz, 1984).

Doherty (1997) proposed five components of school readiness: physical well being and motor development; social knowledge and competence; emotional health and a positive approach to new experiences; language skills; and general knowledge and cognitive skills. Her five readiness components encompass many of the experts' ideal list

of readiness indicators (Gracey et al., 1984; Kagan et al., 1995; Katz, 1991; Kunesh & Farley, 1993; Lamberty & Crnic, 1994; Lewit & Baker, 1995). Nonetheless, the five readiness components identified were not accepted as an ideal list by every expert.

The National Task Force on School Readiness was created to redefine school readiness. The task force acknowledged that school readiness concerns not only academic skills but also good health, self-confidence, and social competence (Kunesh & Farley, 1993).

Also in an attempt to redefine the term, Lamberty and Crnic (1994) proposed that school readiness, as a "static" concept, be replaced by the notion of "continuing" readiness to learn. They believed that there are multiple component states of readiness (e.g., cognitive, social, and psychological) and that children at different ages achieved these states of readiness differently (Katz, 1991). Accordingly, readiness should imply a continuing process of adaptations to cognitive and social challenges.

Kagan et al. (1995) found similarly that children not only develop differently but they also develop in "spurts." A child may achieve with ease what was once difficult for him or her. The argument follows therefore that school readiness should not be viewed in static terms (i.e., ready versus not ready) but in a continuing fashion (e.g., more ready or less ready).

Nonetheless, Nurss (1987) and the task force (Kunesh & Farley, 1993) pointed out that readiness should depend on which type of program the child is entering. They believed that the types of readiness a child possesses rarely matter if the readiness types do not match what the child needs to do well in the program.

Types of School Readiness

Even though the experts' definitions of school readiness are more different than they are similar, the definitions provide a starting place to study the concept of school readiness. The debates on school readiness have ignited the interest of the experts at the national levels to understand what school readiness really means. The National Education Goals Panel drew together the best-informed experts on the subject to figure out what it means to be ready to learn (Kagan et al., 1995). The panel articulated a broad concept of readiness, with at least five major areas that together form the notion of school readiness. The areas include health and physical development, emotional well-being and social competence, approaches-to-learning, communication skills, and cognition and general knowledge. Each of these areas is described below.

Health and Physical Development

A growing body of research shows a strong link between a child's health and his or her school performance (Kagan et al, 1995). Experts believed that a healthy child is more able to engage actively in class activities. Physical readiness involves such skills as gross and fine motor coordination (e.g., walking up and down the stairs, turning pages, and printing), eye-hand coordination (e.g., use of a pencil or scissors), visual discrimination of objects (i.e., by colors, shapes, sizes, names, and types), and auditory discrimination (Morrongiello, 1997).

Emotional Well-Being and Social Competence

Experts generally agree that children who were reared in a stable and caring relationship tend to be more productive in school (Morrongiello, 1997). Children with emotional maturity persevere with difficult tasks and are able to regulate emotions in difficult situations. Emotionally mature children are able to function as members of a group (Panpum, 2004). They can work within the time constraints of the school program. They know the difference between work and play and when and where each is appropriate (Bradley, 1984; LeCompte, 1980; Panpum, 2004). They are aware of what is socially acceptable and know appropriate ways of relating to others.

Emotionally unstable children (i.e., unhappy, fearful, or angry) tend to be preoccupied and unable to participate effectively in class activities. Lack of emotional maturity has been found to be a cause of peer rejection, exclusion, and disengagement in learning activities (Doherty, 1997). Conversely, children with a sense of social competence are more likely to form good relationships with teachers and peers. Key skills leading to a sense of social competence include respecting the rights of others, relating to others without being too submissive or overbearing, and being willing to give and receive support (Panpum, 2004).

Approaches-to-Learning

This readiness type concerns not only academic skills but also motivation, learning styles, habits, and attitudes. Children approach learning experiences differently. Some are more adventurous, playful, and open to new learning experiences. Others are more deliberate, less willing to experiment, and/or hesitant to take on new challenges

(Good, 1973). Although there is no clear preference for a particular approach to learning, measurement experts often look upon these factors to determine the child's level of readiness (Morrongiello, 1997).

Communication Skills

Through communication, children learn new ideas, acquire meaningful knowledge, and construct relationships between new and the existing knowledge. Since learning occurs through intellectual exchanges of ideas, communication skills are key predictors of a child's academic success (Morrongiello, 1997). Children with appropriate communication skills are able to express themselves not only orally but also in a written format. They feel more competent in school when they can understand and use the language of various academic subject matters. They are also more confident in their own ability when they can relate to ideas and topics introduced by the teachers and peers during class discussion and activities (Katz, 1991).

Communication skills are also part of social skills. Good communication skills help children to establish and promote meaningful relationships with teachers and peers. Children use their communication skills to express their feelings, wants, and needs in a socially acceptable way. Good listening skills help children to understand others' feelings, wants, and needs (Morrogiello, 1997). Appropriate communication skills help children to behave more appropriately toward other children at school. Good communication skills allow children to coexist in a meaningful manner.

Cognition and General Knowledge

The acquired knowledge helps a child to make sense of new concepts. Knowledge may include facts in such subject areas as science, social studies, and ethics, or information about significant people, places, things, and events that are relevant to the child's life. It is important that a child is able to organize learned information and assimilate it into a new set of knowledge (Morrogiello, 1997).

Similarities between Various Definitions of School Readiness

The panel's (Kagan et al., 1995) five components of school readiness closely resemble parts or all of those proposed by many experts mentioned earlier. For example, some experts proposed language, intellectual and perceptual functioning, and gross and fine motor coordination as components of school readiness (Lewit & Baker, 1995). Their proposal parallels the panel's first (i.e., Health and Physical Development), fourth (i.e., Communicative Skills), and fifth (i.e., Cognition and General Knowledge) component of school readiness (Kagan et al., 1995). Similarly, Gracey et al. (1984) recommended adding a social-emotion domain to the list. The social-emotional domain matches the panel's second component of school readiness.

The components proposed by Nurss (1987) match at least four of the panel's five school readiness components (Kagan et al., 1995). Nurss (1987) included social-behavioral, sensory-motor, cognitive-language, and age as the components of school readiness. These components match the panel's first (i.e., Health and Physical Development), second (i.e., Emotional Well-Being and Social Competence), fourth (i.e.,

Communication Skills), and fifth (i.e., Cognition and General Knowledge) component of school readiness (Kagan et al., 1995).

The components of school readiness proposed by Katz (1991) also bear a close resemblance to those of the panel (Kagan et al., 1995). Katz (1991) proposed three components of school readiness, which include physical well-being, emotional well-being, and cognitive readiness. These components match the first (i.e., Health and Physical Development), the second (i.e., Emotional Well-Being and Social Competence), and the fifth (i.e., Cognition and General Knowledge) component of the panel.

Concept of School Readiness in Thailand

The history of kindergarten education dates back to the era of King Rama V of Thailand, who went on an unofficial visit to Europe. After the return, King Rama V decreed that a group of national-level educational administrators visit Europe and study the prevalent educational models. The decree resulted in a long-term national education project beginning in 1898. The project started a three-year education program for 7-9 year-old children. The three-year education program is what we now know of as Thailand's first kindergarten education program (Tongdee & Kanjanakij, 1994).

An aspect of the original kindergarten education in Thailand received an influence from two schools of thought in Europe. The first school of thought was from Friedrich Froebel, who was a German educator and best known as the originator of kindergarten system. The second school of thought was from Maria Montessori, who was an Italian educator and best known for her child-centered alternative educational method for children (Tongdee & Kanjanakij, 1994).

There were three main types of kindergarten at the early stage of kindergarten education in Thailand. The different types do not necessarily dictate differences in the underlying educational philosophy the school followed. Rather, the differences were in the location of the school and/or school affiliations. The first type was meant to teach children all necessary knowledge before entering an elementary school. Some of these schools were separate institutions and some were part of an elementary school. The first type of kindergarten enrolled children at seven years of age. The second type was meant to teach children to think, read, write, and do some easy math. These schools were usually part of a temple or a household. There was open access to this type of school. Anybody could enroll and there was no age limit. When a student from these schools was able to read, write, and do some easy math, they would be allowed to enter elementary school. The third type of kindergarten, aptly known by Thais as "Kindergarten", was similar to the first two types. There was open enrollment and no age limit. The children are taught to read, write, and do math in the "old" way. This type of kindergarten could usually be found at a temple or at a household (Tongdee & Kanjanakij, 1994).

From the period of three loosely defined types of kindergarten, Thailand saw five versions of the kindergarten curriculum, where each version was named after the year of publication. The five versions were 1940, 1953, 1960, 1975, and 1979. Contrary to earlier conceptions of kindergarten education in Thailand in 1898, which involved a three-year program, all five versions of these kindergarten curriculae involved only two-year programs. All five versions stressed a heavy emphasis on academic achievement.

Learning was divided into different subjects: Social, Thai, Math, Nature, Arts, Health, Singing, and Music. Objectives of each subject were outlined in a very broad sense and there was no emphasis on assessment. Instruction focused mainly on making sure children could read and write, not on cultures and traditions or on social norms. This was because Thai social conditions, unlike in the present, involved simplistic ways of life and uncomplicated social rules, norms, and problems (Witantam, 1990).

Presently, the kindergarten curriculum in Thailand is implemented inter-disciplinarily in a three-year program. That is, there is not a clear separation of which subject is being taught at a particular moment. But in the end, children are meant to learn four subject areas: pre-Thai skills, pre-math skills, pre-social skills, and pre-health skills. One might notice the word "pre" in front of each subject area (Witantam, 1990). This is to imply that kindergarten education is to prepare children for a more "intense" learning of each subject area in first grade (Boonsawat, Issarangkul Na Ayudhaya, Laungsuwan, & Tosupan, 1980; Office of the National Education Commission [ONEC], 1994; Prawalpruk, 1975; Tangjitsomkit, 1996; Witantam, 1990). In addition, the current kindergarten education also allows for ways to assess children so that teachers can evaluate the children's level of readiness for first grade (Witantam, 1990).

The current kindergarten curriculum aims at producing a warm and loving person, who is physically capable, emotionally stable, socially competent, and intellectually ready for first grade instruction (Maikaew, 2000; Department of Curriculum and Instruction Development [DCID], 1997; ONEC, 1999; ONPEC, 1982; Sangmali, 1986, Tangjitsomkit, 1996; Tongsawat, 1994; Yimyong, 2001). The curriculum addresses three

broad school readiness components, in which each subject resides. They include physical, emotional/social, and academic readiness (DCID, 1997). Each of these readiness components is described later in this section.

In addition to how readiness components are defined by the various versions of the kindergarten curriculum, the concept of readiness is understood by Thai experts in a similar way. Panpum (2004) defines readiness as developmental flourishing in four areas: physical, emotional, social, and intellectual. Other Thai experts have similar definition, and areas, of readiness (Charoensuk, 1986; Chonchop, 1982; Kangpenkae & Tongnui, 1986; Kisawatkon, 2000; Moopung, 1982; Ratana, 1992; Setsukko, 1981). Many experts, including Thais, agree that other factors such as age maturity, prior experiences like academic training, adaptability, and interests contribute greatly to the level of readiness in school entering children (Aonjumras, 1985; Bupawes, 1984; Chuthai, 1982, Downing & Thackrey, 1971; Good, 1973; Kaosim, 1994; Nilarun, 1987; Nilwichian, 1989; Panich, 1988; Panpum, 2004). These factors help children to learn new things quickly, effectively, and with ease and satisfaction (Downing & Thackrey, 1971; Good, 1973; Panpum, 2004; Pengsawat, 2001).

Physical Readiness

A physically ready child shows appropriate progress for his or her physical development. A healthy child is physically active and energetic (Suwanatat, 1996). He or she is able to meet the demands of daily school activities. Other examples of a healthy child include the abilities to (a) alternate feet while walking up and down the stairs, (b) button his or her shirt and tie his or her shoes, (c) to use pencils and spoons, and (d) to

practice healthy habits (DCID, 1997; Office of the National Primary Education Commission [ONPEC], 1983).

Emotional/Social Readiness

An emotionally and socially ready child seems happy, playful, and confident around other people. A socially ready child is eager to make new friends. He or she knows appropriate ways to interact with peers and adults. An emotionally ready child is confident in his or her own abilities to meet the demands of daily school activities. The child relies mainly on his or her own emotional strengths to function well in the new environments. An emotionally mature child is ready to take risks in learning new skills. Other examples of an emotionally and socially ready child include the child who (a) enjoys playing with others rather than in isolation, (b) cares as much about his or her needs as others', (c) is outgoing and loves to interact with his peers and the adults, and (d) loves to listen to music and to see beautiful things (DCID, 1997; ONPEC, 1983).

Academic Readiness

An academically ready child possesses appropriate academic skills. The child shows imagination and creativity in solving challenging problems (ONPEC, 1983). Examples of an academically ready child include the child who (a) begins to use a more complex sentence structure, (b) possesses more vocabulary knowledge, (c) knows the information relevant to himself or herself and his or her environments, (d) shows appropriate math skills, (e) is able to discriminate between different colors, sounds, shapes, pictures, numbers, and objects (DCID, 1997).

Within the academic readiness area, many Thai experts and organizations define domains that encompass the academic readiness area. For example, ONPEC (1991) and ONPEC (1998) list language, math, surrounding, concept formation, and problem-solving domains. The Office of Private Education Commission [OPEC] (1990) lists language, math, spatial relation, sensory, memory, and creativity and imagination. Panich (1988) lists counting and number value, language, differentiation, comparison, and situational recall. And Pluksawan (1975) lists the ability to re-order events and pictures, to memorize, to communicate ideas, to listen attentively, and to use language well. Sintuwej (1986) lists computation and writing. Malumpong (1982) lists logical classification, logical ordering, numerical comparison, space, and language. Other psychologists in Thailand proposed memory, perception, concept, reasoning, problem solving, imagination, creativity, interest, and judgment (Pinjinda, Jongpayuha, & Charoensuk, 1973).

Similarities between Readiness Components in the U.S. and Thailand

The components of school readiness in Thailand are similar to those in the U.S. The three components (i.e., physical, emotional/social, and academic) fall almost perfectly into three of the five school readiness components in the U.S. (i.e., Health and Physical Development, Emotional Well-Being and Social Competence, and Cognition and General Knowledge). Although Approaches-to-Learning and Communication Skills are not specified as a distinct component of school readiness in Thailand, they are already embedded in the three readiness components. For example, communication skills are part of academic readiness. Students must be able to comprehend the new concepts

through listening, reading, and writing and to apply them in new situations. Similarly, Approaches-to-Learning readiness falls under emotional and social readiness. Students who are playful and active are more likely to be motivated and curious to learn new concepts than those who are unhappy and withdrawn.

Readiness Tests

Distinction between Achievement and Readiness Tests

A measure of academic readiness is the instrument that assesses a child's level of readiness for the next grade instruction. The term "readiness" implies a satisfactory level of preparedness in content areas (Kagan, 1992). It may also imply a specified level of traits shown to be necessary prerequisites for the next grade assignment (Nurss, 1987).

A fundamental difference between a measure of academic achievement and that of academic readiness lies in the purpose. A measure of academic achievement evaluates the extent to which a child has mastered the academic content areas. A measure of academic readiness determines how ready a child is for the next grade instruction based on his or her level of content mastery. The focus of a measure of academic achievement lies more in the past. Such a measure is concerned with the mastery of what has been taught. On the contrary, the focus of a measure of academic readiness lies more in the future. Such a measure predicts the level of a child's future academic success based on the current level of content mastery.

Summary of Available Readiness Tests

Twenty-three readiness tests were found in the Mental Measurements Yearbooks (Conoley & Impara, 1995; Conoley & Kramer, 1989; Impara & Plake, 1998; Kramer & Conoley, 1992; Mitchell, 1985). Literature review uncovered two readiness tests created for a Thai population. They are "Language Readiness Test for Continuing Education in Prathom Suksa 1" [LRTCEPS1] and "Intellectual Readiness Test for Pre-school Children" [IRTPC].

Both tests were found in an on-line search through "ThaiLit" (Boonruang, 1991; Ineay, 2004). While the term "school readiness" was used in the search, many of the tests found do not use the term school readiness. Examples include the Battelle Developmental Inventory (Paget; 1989; Stinnett, 1989), the Boehm Test of Basic Concepts-Revised (Fitzmaurice & Witt, 1989; Linn, 1989), and the Developmental Indicators for the Assessment of Learning (3rd Edition) (Cizek, 2001; Fairbank, 1998).

Regardless of the naming convention, these tests are considered readiness tests for two primary reasons. First, at least one of its stated purposes is to assess a child's knowledge and skills that are critical for "future" school success. Second, inclusion of the domains reflects the attempts to measure the abilities necessary for success in the first grade. Such abilities include language, math, and visual discrimination, which together provide an indicator of a child's level of readiness for first grade instruction (DCID, 1997; Tangjitsomkit, 1996).

Except for LRTCEPS1 and IRTPC, which were developed as part of a graduate thesis, each measure referenced in Mental Measurements Yearbooks was reviewed by

two experts, each of whom specialize in psychology, psychometrics, education, or test administration. Based on the recommendation of the reviewers, only seven tests warrant a detailed discussion. Each of the five tests in the U.S. received a "full recommendation of use" (i.e., without any reservation) from both reviewers. The two tests in Thailand claimed sufficient reliability and validity. Other tests received a full recommendation from one reviewer and a "partial" recommendation from the other. Some tests received no recommendation at all from both reviewers. The experts' review of the seven tests is summarized in detail below as these tests form the framework for TAR development.

Battelle Development Inventory (BDI).

Purpose. The BDI (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984) was designed with three primary purposes: assessment and identification of the handicapped child, assessment of the nonhandicapped child, and planning and providing instruction. The BDI was intended for the determination of individual strengths and weaknesses, the development of Individual Education Plan (IEP), tracking individual progress, and planning and evaluation of instructional programs.

Administration. The BDI is divided into five domains: Personal-Social, Adaptive, Motor, Communication, and Cognitive. Each domain has its own manual, allowing specific diagnosticians (e.g., speech pathologists) to administer the appropriate section of the test. Separate domains can be administered independently (Stinnett, 1989). Administration and scoring guidelines are provided in the manuals. The BDI has a short version, called "Screening Tests," which take between 10 and 30 minutes to complete. The results from the screening tests will dictate further administration of certain BDI

domains. The administration of "full" BDI takes between 1 and 2 hours. Since the administration of the entire battery may be too long for preschool children, the administration may take place over different sessions (Paget, 1984).

Reliabilities and Validities. The standardization sample included 800 children (Paget, 1984). Excellent reliability data were reported (Stinnett, 1989). For example, test-retest reliability coefficients ranged from 0.71 to 0.99, most of which were above 0.80. Interrater reliability coefficients were between 0.70 to 1.0, again with most coefficients above 0.80 (Stinnett, 1989). The evidence from test-retest and interrater reliability suggests that BDI is a stable instrument (Paget, 1984). The manual also reported small standard errors of measurements (SEM), which suggest little variability between the "observed" score and the "true" scores.

Excellent validity data were also reported. Initial validity information confirmed the BDI as a measure of development (Stinnett, 1989). Content validity was ensured by the rigorous item selection and test development procedures. Item-total score and domain-total score correlations were very good, suggesting the BDI as a homogeneous measure of development. The validity of BDI's developmental domains was also supported by significant t-test comparisons between adjacent age groups on BDI components (Stinnett, 1989). Several studies supported the construct and criterion-related validity of BDI and its domains. Factor analyses of the pilot data provided considerable support for the BDI's domain organization as well.

The evidence of concurrent validity was also provided by correlating BDI with several measures such as the Peabody Picture Vocabulary Test-Revised (PPVT-R), the

Developmental Test of Visual Motor Integration (VMI), the Reaction Time (RT), the Bayley Scales of Infant Development, the Minnesota Child Development Inventory, the Stanford-Binet (S-B), the WISC-R, the Preschool Language Scale-Revised (PLS-R), the Arizona Articulation Proficiency Scale-Revised (AAPS-R), and the first grade Wide Range Achievement test (WRAT). It was found that preschool children's performance on the BDI accurately predicts their performance on the PPVT-R, VMI, and RT (Stinnett, 1984). In addition, the BDI was found to moderately correlate with S-B, WISC-R, and PPVT-R, indicating the BDI measures some similar skills but is still sufficiently distinct from being an intelligent test. Evidence for predictive validity was also supported when the BDI was found to predict achievement of first grade WRAT better than the well-known Metropolitan Readiness Test (MRT). The Personal-Social, Communication, and Cognitive domains of the BDI were found to be the best predictors of academic achievement (Stinnett, 1989).

Cautions. No normative data for handicapped children were reported in the manual. Only normal children were included in the standardization data (Stinnett, 1989). Therefore, cautions must be exercised in the interpretation of the scores (Paget, 1984). Regarding predictive validity, the BDI's domains were correlated with the domains of tests measuring similar constructs. All domains except Adaptive had evidence of predictive validity only on the tests with which they were correlated. Until further validation of each domain is conducted, it was recommended the BDI be used as an overall measure of development rather than used separately by the domains.

Another caution concerns the materials provided with the administration kit. The test kit does not include all materials necessary for test administration. Some missing materials have to be manufactured by the administrator. Therefore, there are possibilities of the materials not conforming strictly to the specifications (i.e., size, shape, color). Different materials raise the problems of reliabilities for BDI.

Conclusion. Stinnett (1989) believed the BDI to be a well-developed assessment instrument in early childhood development. Despite some of the BDI's limitations, many practitioners have adopted the BDI as the instrument of choice for assessing current developmental status and documenting children's progress (Paget, 1984).

Boehm Test of Basic Concepts-Revised (BTBCR).

Purpose. The BTBCR (Fitzmaurice & Witt, 1989; Linn, 1989) was designed to assess a child's mastery of basic concepts that are both fundamental to understanding verbal instruction and essential for early school achievement. It is also intended to provide classroom teachers with a means of identifying children whose overall level of concept mastery is low and who therefore may need special attention. The measure can also help to identify the "concepts" with which large numbers of children in a class may be unfamiliar.

Administration. The BTBCR consists of a set of 50 relational (e.g., front, below, fewest) and standard (e.g., left-right) concepts. The measure also has an Applications test, which consists of 26 items assessing the mastery of basic concepts that are used frequently in combinations with other basic concepts. While the BTBCR taps a child's knowledge of the concepts, the Application tests ask a child to apply the knowledge

within the context of multiple-step directions. The BTBCR can be administered in groups. The administration of a complete test (i.e., Form C and D and the Applications test) takes approximately one hour.

A typical BTBCR item consists of line drawings of three objects or sets of objects (e.g., a lamp, a shirt, and a shoe). The test administrator reads a sentence instructing the child to mark a particular picture (e.g., Mark the thing that a child should never wear). The instructional sentence is repeated and the administrator is instructed to emphasize the key word. Scoring of the test is simple and straightforward. The test manual is easy to follow.

Reliabilities and Validities. Split-half reliabilities were reported to be between 0.55 (Form C, 2nd grade) and 0.87 (Form D, kindergarten). Content validity was claimed because the test items were chosen from school curricular materials and teachers' verbal instructions. The correlations with other tests of achievement were reported to be in the range of 0.24 to 0.64.

Cautions. One of the reviewers stated that the BTBCR has a reasonable level of reliability for the end of kindergarten and the end of first grade (e.g., low to mid 0.80s) but poor reliability for the end of second grade (e.g., alternate form correlation of 0.65 and split-half coefficients of 0.64 to 0.73 for Forms C and D). In other words, the BTBCR is much too easy for students at the end of grade 2. The manual confirms the norms to be nationally representative of school district, geographic region, and socioeconomic status of students based on 1980 U.S. census. However, only 15 states were represented in the norming process. In addition, the correlations between total

scores for Form C and D (between 0.65 and 0.82) are low for alternate forms (Fitzmaurice & Witt, 1989). Therefore, the use of common conversion tables to obtain percentile equivalents of raw scores for both forms is questionable (Fitzmaurice & Witt, 1989; Linn, 1989).

The discussion in the manual about construct validity lacks any clear conception or direction. The evidence of criterion-related validity through correlations with other tests does not provide any conceptual framework from which the domains were identified (Fitzmaurice & Witt, 1989; Linn, 1989). The claims for content validity have reasonable support. However, the evidence that targeted instruction based on the test results is beneficial is limited (Linn, 1989). Fitzmaurice and Witt (1989) questioned if students who do poorly on the test would actually do more poorly in school or preschool.

Conclusion. Linn (1989) recommended the use of the test only at kindergarten or the beginning of first grade. Fitzmaurice and Witt (1989) supported the author's claims of the test as a "primary" screening device. They recommended the Brigance Preschool Screen (Fitzmaurice & Witt, 1989; Linn, 1989) as a more appropriate screening instrument of academic tasks.

Clymer-Barett Readiness Test-Revised (CBRT-R).

Purpose. The CBRT-R (McCarthy, 1985; Proger, 1985) was designed to measure the important skills necessary for success in beginning instruction (especially reading).

Administration. The CBRT-R contains three components: Visual Discrimination, Auditory Discrimination, and Visual-Motor coordination. Each of these components comprises two subtests. The visual discrimination subtests consist of Recognizing

Letters (35 items) and Matching Words (20 items). The auditory discrimination subtests consist of Beginning Sounds (20 items) and Ending Sounds (20 items). And the visual-motor coordination subtests consist of Completing Shapes (20 items) and Copy-A-Sentence (7 possible points). There are Form A and B, which are equivalent in terms of items from the six subtests (Proger, 1985). The CBRT-R also has the Short Form, which contains only two subtests (Recognizing Letters and Beginning Sounds). The manual is very well written and easy to follow (Proger, 1985).

Reliabilities and Validities. Norming was based on 5,565 first-grade students in 188 classrooms. Reliability coefficients (i.e., split-half, corrected) for Visual Discrimination, Auditory Discrimination, Visual-Motor Coordination, Total Score-Short Form, and Total Score-Full Form are all in the 0.90s (Proger, 1985). Five internal consistency studies were performed on Form A on five "atypical groups" (i.e., first graders in a bilingual, rural, southwestern school system). Again, except one reliability coefficient of 0.89, all coefficients were in the 0.90s (McCarthy, 1985; Proger, 1985). The raw score standard errors of measurement based upon internal consistency reliabilities were 3, 3, 2, 3, and 4 for Visual Discrimination, Auditory Discrimination, Visual-Motor Discrimination, Total Score-Short Form, and Total Score-Full Form respectively (McCarthy, 1985).

Concurrent validity was claimed through the correlations between the CBRT-R and other readiness tests. The correlations ranged from 0.55 to 0.80, indicating the CBRT-R fitting as a test of school readiness. The evidence for construct validity was found through the low intercorrelations (0.02 to 0.45) among Form A's six subtests. The

low intercorrelations indicate that the CBRT-R's subtests are each tapping relatively independent aspects of the school readiness (Proger, 1985). In addition, there was evidence that the students' performance on the CBRT-R tasks was not entirely linked to the students' intelligence. This was done by correlating the score totals from the CBRT-R Form A subtests, Short Form, and Full Form with those of various measures of first grade intelligence (i.e., Stanford Binet (Form L-M), Pinter-Cunningham, California Test of Mental Maturity, and Kuhlman-Anderson). The correlations with these tests were not high (i.e., generally in the 0.30s, 0.40s, and 0.50s) (Proger, 1985).

Predictive validity studies show corelational data of 0.30s to 0.70s between first grade CBRT-R scores in the Fall and various reading achievement test scores (i.e., Stanford Achievement Test, Gates Primary Reading Test, Gates-MacGinite Reading Test, and MRT) in the Spring (Proger, 1985).

Cautions. No details were given of the construct validity although the manual claimed construct validity studies were performed. Furthermore, the CBRT-R Full Form scores were not any better at predicting first grade success than were the CBRT-R Short Form scores. Despite the claimed equivalency between Form A and B, the correlations between both forms were only moderate (i.e., 0.57 to 0.79).

All of the technical data (i.e., reliability and validity) were generated only for Form A. By stating that Form B is really the predictor of school readiness, the usefulness of the technical data is questionable. No information was supplied with regard to the selection of content and items (McCarthy, 1985). This may be the reason for the lack of content validity evidence.

Conclusion. McCarthy (1985) perceived the CBRT-R to be a very strong test of school readiness with only a few minor weaknesses. Proger (1985) believed that the CBRT-R would be a useful readiness instrument for all practitioners.

DABERON-2 Screening for School Readiness (DABERON-2).

Purpose. The DABERON-2 (Axford, 1992; Hughes, 1992) was designed to identify students who may not be ready for formal academic instruction. It was intended as a "selection" tool for entrance into educational programs (Axford, 1992).

Administration. The DABERON-2 takes approximately 20-40 minutes to complete. The test kit contains everything needed for the administration. The materials are easy to handle. There are 122 items, which are scored "right" (R), "Wrong" (W), "no response" (N), or "inappropriate" (I). The DABERON-2 assesses areas such as body parts, color concepts, number concepts, prepositions, following directions, plurals, general knowledge, visual perception, gross motor development, and categories (Axford, 1992).

Reliabilities and Validities. The standardization sample of the DABERON-2 included 1,647 children, whose demographic representation was similar to that of the target population. There was evidence of high internal consistency (i.e., four of the five coefficients exceeding 0.90) (Axford, 1992; Hughes, 1992). The standard errors of measurement by age are considered adequate (Axford, 1992).

Concurrent validity was established by correlating the scale with the MRT. The total battery's Pearson product moment correlation was 0.83 ($p<0.05$). Predictive validity was established by correlating the kindergarten-aged subjects' scores with the follow-up

behavior checklist ratings involving fifteen experts. The correlation was 0.84 ($p<0.001$). Construct validity was established through research regarding the relationship between the subtest scores with chronological age and with aptitude, through cluster intercorrelations, and through item validity. All research findings support the validity of the DABERON-2 (Axford, 1992).

Cautions. Hughes (1992) pointed out that evidence of the relationship between the subtest scores with aptitude was too weak (e.g., 0.60). Although the items may satisfy statistical criteria, the items still represent a very limited range of skills. Huges (1992) believed that there is still little emphasis on language and cognition skills.

Axford (1992) cautioned that the DABERON-2 be used as "selection" tool for entrance into educational programs, not as a screening instrument. According to Axford (1992), a screening tool provides a direction for further assessment rather than a decision of promotion. The instrument does not measure more advanced classification and quantitative reasoning skills—characteristics of highly developed kindergarten-aged students. Therefore, the range of application is limited to measuring early-to-late "preoperational" skills of kindergarten students. Finally, the dichotomous (i.e., right-wrong) scoring system is inconsistent with the measurement of development, where gradation is psychometrically preferred.

Conclusion. Axford (1992) believed that the DABERON-2 is among the better measure for school readiness. Hughes (1992) maintained that the DABERON-2 is a useful instrument for the practitioners who need the information to assist in individualized curriculum development for a child.

Developmental Indicators for the Assessment of Learning (3rd Edition) (DIAL-3).

Purpose. The DIAL-3 (Cizek, 2001; Fairbank, 2001) is a screening test designed to identify potential difficulties for children at school-entering age. The test outcomes may indicate the need for further assessment of the child, who is identified as having difficulties with motoric, conceptual, and language areas.

Administration. The DIAL-3 consists of five developmental areas: Motor, Concepts, Language, Self-Help, and Social Development. A short form, called Speed DIAL, assesses only three basic areas, the details of which were not provided in the reviews. The administration of the DIAL-3 takes approximately 30 minutes. The Speed Dial takes only 15 minutes. The DIAL-3 is administered in a group of three children, who are observed by several administration team members in a specially designed area. Three operators staff each of the three stations. Each child moves through each station after responding to the questions or performing the tasks related to one of the areas assessed (i.e., jumping for the motor station). The test materials contained in the test kit are comprehensive, easy to use, and of high quality. Instructions are clear and easy to follow.

Reliabilities and Validities. The standardization of the DIAL-3 included a nationally representative sample of 1,560 children. The alpha coefficients for the total test and subtests are greatest in the preschool age range (i.e., 5 years old) and lower as the age of the child is at either extreme (i.e., 3 or 7 years old). For the 5-year-old range, internal consistency estimates are 0.90 for total score and between 0.71 and 0.85 for the subtest totals. The test-retest coefficients were in the 0.80s for the Total test and never

lower than 0.69 for the subtests (Cizek, 2001). The Rasch model was utilized with the DIAL-3 to identify any items that were not consistent with others in the test (Fairbank, 2001).

Development of the DIAL-3 involved a review of the literature on child development, task tryouts and refinements, and bias reviews. Extensive evidence of content validity for Motor, Concepts, and Language areas was provided. On the contrary, very little evidence of content validity was presented for the Self-Help and Social areas. Evidence of convergent and discriminant validity includes low intercorrelations between the DIAL-3 domains. Intercorrelations between the DIAL-3 and other tests were also reported to be between 0.48 and 0.79 (Cizek, 2001). Evidence of content validity was claimed through Rasch analysis results and expert reviews of items (Fairbank, 2001).

Cautions. Cizek (2001) pointed out that the scores in some areas (i.e., Motor) do not have acceptable dependability for decision making. The selection of five cutoff points appeared arbitrary. In addition, information regarding standard errors of measurement at the critical cutoff points is missing. Such a lack of information supporting cutoff points leads to more questions about the validity of the cutoff points. Evidence of predictive validity is absent (Cizek, 2001). Although there was evidence that the DIAL-3 measures the intended construct, the level of confidence to use the DIAL-3 scores for a refer-do-not-refer decision is uncertain. Finally, the standard errors of measurement are unacceptably large for very young children.

Conclusion. Fairbank (2001) believed that the DIAL-3 should only be used as a selection instrument with cautions taken in the interpretation of the results. Cizek (2001)

was confident in the dependability of the DIAL-3 scores. The reviewer stated that the DIAL-3 serves as an instrument that is comparatively superior to other alternative tests and provides a defensible way to help educators identify children at risk for school failure resulting from developmental delays.

Language Readiness Test for Continuing Education in Prathom Suksa 1 (LRTCEPS1).

Purpose. The purpose of the LRTCEPS1 (Ineay, 2004) to measure the level of language readiness in children entering first grade. The LRTCEPS1 was intended to determine the level of prior experience and development in each child. The information gained from the administration of the test helps teachers to create appropriate intervention programs to remedy any weakness in language capacity of the child.

Administration. The LRTCEPS1 is divided into three sections: vocabulary, sentences, and stories. The vocabulary section contains twenty five items and takes approximately twenty five minutes. The sentences section contains twenty items and takes approximately twenty five minutes. The stories section contains ten items and takes approximately twenty five minutes.

Reliabilities and Validities. The population included 3,909 kindergarteners entering first grade in 2003. Two groups of samples were drawn. The first group consisted of 600 children, who were recruited through multi-stage random sampling method. The results from the first group were used to determine the quality of the measure. The second group consisted of 581 children. The sampling method for the second group was not discussed. The results from the second group were used to

construct norms. The test developer claimed content validity through an index of consistency of 1.00 for each test item. The reported item difficulty level in the vocabulary section ranged from 0.37 to 0.75 with an average item difficulty level of 0.62. The reported item discrimination ranged from 0.37 to 0.89 with an average discrimination level of 0.79. The reported reliability level for the vocabulary section was 0.926 with the standard deviation of 1.9102.

The reported item difficulty level in the sentences section ranged from 0.32 to 0.78 with an average difficulty level of 0.59. The reported item discrimination ranged from 0.69 to 0.86 with an average discrimination level of 0.79. The reported reliability level for the sentences section was 0.93. The reported item difficulty level in the stories section ranged from 0.50 to 0.75 with an average difficulty level of 0.70. The reported item discrimination ranged from 0.52 to 0.93 with an average discrimination of 0.80. The reported reliability level for the stories section was 0.83.

Cautions. The test developer does not claim predictive validity. The developer stated that the measured level of language readiness through the LRCEPS1 does not predict the level of success in children in higher grades. It merely indicated the level of language capacity of children at the time of testing.

Conclusion. The LRCEPS1 contains easy to moderately difficult items. The test has the ability to discriminate children with less degrees of language capability from the more able children. The test has a high level of reliability.

Intellectual Readiness Test for Pre-school Children (IRTPC).

Purpose. The purpose of the IRTPC (Boonruang, 1991) was to measure the level of intellectual readiness in children entering first grade. The IRTPC was intended to determine the level of intellectual readiness for pre-elementary education, and to later create norms for children in the whole northern region of Thailand. The information gained from the administration of the test helps to further develop tests of intellectual readiness using different designs and methodologies and to further develop tests of other types of readiness (Boonruang, 1991).

Administration. The IRTPC is divided into nine sections: general knowledge, event ordering, categorization, listening, following orders, visual discrimination, auditory discrimination, matching, and counting and number value. The general knowledge section was designed to measure level of personal experiences in different areas. The event ordering section contains items with three pictures arranged in a logical order. The categorization section measures the discrimination skills of objects with different categories. The listening section measures the children's ability to remember a story told verbally and to be able to answer questions pertaining to the story. The following order section measures the children's ability to follow order. The visual discrimination section measures the children's ability to visually match the picture choices with the picture given. The auditory discrimination measures the children's ability to discriminate against different sounds. The counting and number value section measures the children's ability to count numbers, understand number value, compare numbers, add and subtract, order numbers ascendingly (Boonruang, 1991).

Reliabilities and Validities. Three samples were drawn for test administrations. The first pilot group consisted of 27 children from Chaingmai, Thailand. The second pilot group consisted of 100 children from Chaingmai, Thailand. The main study group consisted of 100 children from Tak, Thailand. The test developer used multi-stage random sampling method. The average item difficulty level was 0.52 while the item separation was 0.61 for general knowledge (Boonruang, 1991).

The average item difficulty level was 0.56 and the item separation index was 0.65 for event ordering. The average item difficulty level was 0.58 while the item separation was 0.65 for categorization. The average item difficulty level was 0.67 while the item separation was 0.57 for listening section. The average item difficulty level was 0.54 while the item separation was 0.59 for following order section. The average item difficulty level was 0.57 while the item separation was 0.68 for visual discrimination. The average item difficulty level was 0.56 while the item separation was 0.60 for auditory discrimination. The average item difficulty level was 0.58 while the item separation was 0.70 counting and number value section (Boonruang, 1991).

The test developer claimed a range of reliability levels for IRTPC between 0.74-0.91 and 33 to 59 percent predictive ability for successful first grade instruction. The test developer claimed content validity through extensive review of literature and expert panel reviews of items (Boonruang, 1991).

Cautions. The test developer claimed high predictive ability of the IRTPC for successful first grade instruction. Due to the limited geographical areas used for

sampling, the test developer recommended a wider administration and normalization process.

Conclusion. IRTPC contains moderately difficult items. The test has the ability to discriminate children with less ready from the more ready ones. The test has a moderate to high level of reliability and content validity, albeit having low predictive validity.

Summary of Strengths and Shortcomings of Reviewed Readiness Tests

The reviewed tests have many strengths. The common strength is the content validity of the domains. There was evidence that the domains were measuring part or all of school readiness. Another strength unique to the DIAL-3 is the use of Rasch model, the literature review, and expert review for item generation and revision (Cizek, 2001; Fairbank, 2001).

The common shortcoming of the tests is that five of them were not specifically designed for Thai students. One test from Thailand only measures one readiness component (i.e., verbal) while the other used small sample sizes. Other important shortcomings from at least one of the tests include poor predictive validity studies (McCarthy, 1985; Paget, 1984; Proger, 1985; Stinnett, 1989), a poor norming process (Fitzmaurice & Witt, 1989; Linn, 1989; Paget, 1984; Stinnett, 1989), a limited range of skills (Axford, 1992; Hughes, 1992; Ineay, 2004), and arbitrary cutoff points (Cizek, 2001; Fairbank, 2001). As the new measure was developed, the strengths and shortcomings of these tests were taken into consideration.

CHAPTER 3

Method

This chapter describes how this dissertation addressed the two research questions listed in Chapter 1. The first question concerning content validity is addressed by a discussion of how the findings from the literature review in Chapter 2 informed item construction. The second question concerning reliability is addressed by a discussion about test theory and the selection of the Rasch model as the underlying theory, the scale development process, and the method used in the first pilot, the second pilot, and the main study.

Content Validity

Content validity refers to "the degree to which the scores yielded by a test adequately represents the content, or conceptual domain, that these scores purport to measure" (Gall, Borg, & Gall, 1996, p. 250). There are many ways to establish evidence of content validity. A review of the literatures on content areas is one way to contribute such evidence. The literature review identifies the content and the domains that authors think best represent the content. A domain may be reflected by responses to questions, tasks, or behaviors representing the content the test purports to measure. The more a measure contains the domains found in the literature review, the stronger the evidence of content validity.

A review by the experts regarding the representativeness of the item contents to the targeted content universe is another way to establish evidence of content validity. After a measure is initially assembled, a group of experts may be asked to review the items. The content experts have expertise in the content areas the test is trying to measure. Their expertise helps to define in precise terms the domains of specific content that the test is assumed to represent. The experts then determine how well the test items reflect the content. The more valid items a measure has, the stronger the evidence of content validity.

Literature Review

Literature review identifies the content universe representing school readiness and academic readiness for first grade instruction in Thailand. The content universe is circumscribed by the school readiness definitions found in the literature. The content universe is limited further by the definitions of academic readiness. The literature review finds the domains that purport to measure the content representing academic readiness for first grade in Thailand. The following paragraphs describe the findings from the literature review.

School Readiness.

One finding from the literature review is the different conceptualizations of school readiness. To avoid a possible misconception, it is important to define what school readiness means for this dissertation. The definition provides a direction for the

scale development process. It is also important to point out the context in which the definition is used. The context provides an additional support for the definition.

School Readiness Defined.

School readiness is defined as the level of skills and knowledge, which are necessary for a successful completion of first grade requirements in Thailand. The skills and knowledge can be classified into three readiness components: physical, emotional/social, and academic readiness.

Interpretation of School Readiness Definition.

The word "level" in the definition signifies that school readiness varies in degree. A student can be described as "more ready" or "less ready" than other students. The definition is not intended to describe a student simply as "ready for school" or "not ready for school." The definition indicates that a more ready student possesses more skills and knowledge than does a less ready student. This does not mean that a less ready student will fail first grade. The definition simply implies a less ready student may possess a lesser degree, yet at least with a minimum level, of ability to survive the demands of first grade education. That is, a less ready student may very well successfully complete, at least at a minimum level or higher of, the first grade requirements. A less ready student is less likely to achieve as high a level of academic performance in first grade as are his or her more ready peers.

In addition to the degree of school readiness, it is also important to point out the context in which the definition is applied. This definition concerns kindergarten

graduates in Thailand. The scale being developed for this dissertation is intended for use exclusively with Thai kindergarten graduates. Therefore, the definition is aligned more closely with the concept of school readiness in Thailand than the ones in the U.S. The readiness components in the definition are reflective of the three readiness components in Thailand. Nonetheless, the definition does incorporate the knowledge gained from the literature review on the concept of school readiness in the U.S. to make it more complete.

Academic Readiness.

Although the definition includes three readiness components, this dissertation only used the academic readiness component. The reasons for using only one component are as follows. First, the inclusion of every component would result in an incredibly long scale. As will be explained later in this chapter, it is important to include a large enough number of items during the pilot administration. This guarantees that every aspect of the construct is measured by at least a few items. To exhaust all of the possible item contents and formats, the number of items can become too large even for a single construct. The size of the scale multiplies with the number of constructs involved. The inclusion of three components would lead to a very long scale. Too long a scale has many undesirable ramifications. For example, the administration of such a scale requires hours. A lengthy administration leads to examinee's fatigue, which greatly reduces the examinee's performance.

Second, the other two readiness components require a much more involved kind of administration. For example, the physical readiness component requires students to perform a lot of physical activities. Some of the activities (i.e., climbing stairs or

jumping ropes) can potentially lead to an injury. The parents and the research sites are most likely not willing to take such a risk. Since their consent is mandatory, the study would likely experience a drop in participation by the sample. Similar to physical readiness, the social/emotional readiness component is difficult to administer. Observation is usually the preferred method of administration for social/emotional constructs. In addition to time and involvement from the administrators, observation is more prone to the observers' subjectivity than are other methods. The subjectivity may result from an unclear scoring policy or a biased personal judgment by the observers. Subjectivity could lead to an incorrect interpretation of the examinee's responses and hence inaccurate reflection of the examinee's true score. Unless the administrator is experienced, it is hard to guarantee an accurate measurement. With the limited test administration experiences of the prospective administrators, the researcher is not confident that the social/emotional readiness construct would be accurately measured.

Third, schools in Thailand value academic readiness more than other readiness components. The inclusion of only the academic readiness component will result in a measure that directly addresses their needs for a measure of academic readiness. The application of an academic readiness measure is more direct than that of a school readiness scale measuring additional two readiness components.

The three rationales have led the researcher to select academic readiness as the only construct to develop a measure for this dissertation. The following paragraphs contain a discussion about the definition of academic readiness, the meaning of the definition, and the domains of academic readiness.

Academic Readiness Defined.

Academic readiness is defined as the level of academic skills and knowledge, which are necessary to fulfill the academic requirements of first grade in Thailand. The literature review has uncovered six academic readiness components: verbal skills, math skills, visual abilities, logical skills, memory capacities, and general knowledge. For example, DCID (1997) proposed vocabulary knowledge, relevant information, math skills, and visual abilities as part of academic readiness components. ONPEC (1991) and ONPEC (1998) proposed language, math, and concept formation. OPEC (1990) proposed language, math, spatial relation, and memory. Panich (1988) proposed language, math, visual, and verbal. Pluksawan (1975) proposed sequential ordering, memory, and language. Sintuwej (1986) proposed math and language. Malumpong (1982) proposed visual matching, ordering, math, spatial relation, and language. Pinjinda, Jongpayuha, and Charoensuk (1973) proposed memory and concept. Table 1 below summarizes the domains with the sources from the literature review of Thai research.

Table 1
Domains and Sources from Literature Review in Thailand

Domains	Sources
Verbal	Malumpong (1982), ONPEC (1991), ONPEC (1998), OPEC (1990), Panich (1988), and Pluksawan (1975).
Visual	Malumpong (1982) and Panich (1988).
Memory	OPEC (1990), Pluksawan (1975), and Pinjinda, Jongpayuha, & Charoensuk (1973)
Math	Malumpong (1982), ONPEC (1991), ONPEC (1998), OPEC (1990), Panich (1988), Sintuwej (1984)
Logical	Malumpong (1982), Panich (1988), and Pinjinda, Jongpayuha, & Charoensuk (1973).
General Knowledge	ONPEC (1991) and ONPEC (1998).

Interpretation of Academic Readiness Definition.

Similar to school readiness, academic readiness varies in degree. A student can be described as "more academically ready" or "less academically ready." A more academically ready student possesses more of the skills and knowledge in the six academic readiness components than does a less academically ready student. A more academically ready student is more likely to achieve a higher academic performance in first grade than is a less academically ready student.

The context in which the definition of academic readiness is applied is also first grade in Thailand. The first grade curriculum consists mainly of ten subjects: math, language, household skills, music, arts, science, health, ethics, English as a second language, boy/girl scout, and physical education. Six academic readiness components

address a big part of the ten subjects in first grade. The five subjects that will not be measured by this measure of academic readiness are music, arts, boy/girl scout, English as a second language (ESL), and physical education.

A reason for not including ESL was due to the fact that English is considered, as the name implies, a second language. To start learning a second language in first grade (at the age of six) is considered a luxury rather than a necessity. The school puts more emphasis on acquiring the first language (i.e., Thai), which is the "language" subject matter. The success of students in first grade and in higher grades depends on their abilities and skills in Thai language for several reasons. First, teachers speak in Thai while teaching all subjects (including ESL). To not have acquired necessary Thai language skills, students will not be able to comprehend what is being taught in all subjects and hence they will likely fail in all subjects. Second, all textbooks are in Thai. Students will not be able to comprehend the content of the subject matters provided in the textbooks if they do not have necessary Thai language skills. Third, all tests are written and answered in Thai. Students who do not have necessary reading skills in Thai language will not be able to understand the instructions, the questions, and the answer choices. Similarly, students who do not have necessary writing skills in Thai language will not be able to provide written answers in, for example, a fill-in-the-blank or an open-ended question. As a result, they would likely fail all subjects.

The reasons for not including physical readiness were provided earlier. The reason for not including music and arts was similar to the subjectivity rationale given for social/emotional readiness. Music and arts are subject to the audience's interpretation.

Certain audiences may be able to appreciate certain genres of music and not others. Similarly, certain audiences may prefer a certain type of arts to the others. As opposed to math, music and art are not exact sciences. It depends on who is evaluating the piece of art or music. One evaluator may not perceive doodling by a student as a piece of art while another evaluator may be able to appreciate the student's artful expression. The former may give the student a low score while the latter may give the student a high score for the performance. Similar to arts, an evaluator may give a low score to a piece of music while another may give a high score to the very same piece. It is therefore difficult to ensure objective and accurate measurement of musical or artistic components of school readiness.

Although the memory capacity component is not supported explicitly by the subjects in first grade, memory is found as one of the main academic readiness components in the literature review. Memory is a main factor in other academic readiness components. For example, memory is essential for remembering numbers and mathematical procedures. Memory is also essential for language studies. Vocabularies are learned through memory. Different syntax stored in the memory permits students to form intelligible sentences. General knowledge and home living information are also learned through memory. Students rely heavily on their memorization skills to learn critical and relevant information. Therefore, reasonable memory capability is a necessary prerequisite of success in first grade in Thailand. Similarly, although there is not a separate subtest for household skills, science, health, and ethics, these subject matters are included as part of general knowledge readiness component.

In addition to being part of the subjects in first grade in Thailand, the six academic readiness components can be found in several readiness tests reviewed in Chapter 2. Examples include verbal (Clymer & Barrett, 1966; Danzer, Gerber, Lyons, & Voress, 1967; Ineay, 2004; Mardell & Goldenberg, 1983; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), visual (Clymer & Barrett, 1966; Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), memory (Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), math (Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), logical (Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), and general knowledge (Danzer, Gerber, Lyons, & Voress, 1967).

Operational Definitions of Academic Readiness Domains.

An operational definition refers to a series of tasks devised to elicit behaviors that are indicative of the constructs (Thorndike, 1997). Operational definitions of the academic readiness components describe a series of tasks devised to elicit behaviors that are indicative of the academic readiness constructs. The number and the types of task devised to elicit behaviors that are indicative of six academic readiness domains are determined based on the findings from the literature review and the feedback from the expert panel. The literature review has helped to identify the tasks for each domain. The tasks were selected from similar measures (Boonruang, 1991; Ineay, 2004; Roid & Miller, 1997), resource books (First Grade Exams, 1997, Mati Silapin, 1987; Nontapuk, 2009; Pangwiruitrak, 2007; Pinyo Anantapong, 1993; Pinyo Anantapong, 1996; Sang Asanee, Boon Urapeepinyo, Wong Wijit Sin, Ruji Rek, & Apichartimanon, 1990; Trium

Sop Por 1, 2009; Wai Prip Trium Sop, 2009), and textbooks from Thailand (Helair, 1996). Based on the types of task identified, as many items for each type as possible were created.

Similarly, the content experts helped to identify the major areas of the content universe for each of the domains. Content areas signify the types of task to be created for each construct. Depending on the construct, as many items of each type as possible were created. For example, the literature review and the expert panel identified three types of tasks for the math readiness construct. They include math concepts and vocabulary, arithmetic, and word problems. For each of these types, as many items as possible were created so that there are at least a few items to capture each aspect of the math construct.

For other readiness constructs, a similar process was followed. Content areas are identified and many items are created. The operational definitions of each academic readiness construct serve as a boundary within which the items are created. The operational definitions are provided in the following paragraphs which list the nature of the item and the number of items created in the second pilot and final administration.

Verbal.

The literature review and the expert panel recommendations have identified three types of tasks to elicit the behaviors that are indicative of verbal readiness. They are reading vocabulary, writing dictation, and reading comprehension. For the second pilot study, reading vocabulary contained a list of forty words, which ranged from easy to difficult. For the main study, the researcher was able to use expert reviewer comments and pilot study analysis results to create double the number of items

found in the second pilot study. For the main study, reading vocabulary contained a list of eighty words. This section tests vocabulary knowledge, the ability to recognize the vocabulary by sight, and the ability to pronounce words aloud. Each examinee was required to correctly pronounce each word aloud to the examiner. Table 2 below provides sample vocabulary found in the reading vocabulary section.

Table 2
Vocabulary for Reading Vocabulary Section of Verbal Subtest

Vocabulary	
RV-03	รับประทาน
RV-14	ล่องแก่ง

In both the second pilot and the main study, writing dictation contained another list of forty words, which ranged from easy to difficult. This section also tests vocabulary knowledge, the ability to recognize vocabulary by sound, and the ability to spell words by writing them down on a piece of paper. The examiner pronounced the first word aloud. Each examinee was required to correctly write the spelling of the word down on an answer sheet. The examiner then pronounced the second word aloud. The examinee was required to correctly write the spelling of the second word down on the same answer sheet. The process repeats until the examiner has gone through every word in the list. Table 3 provides a list of sample vocabulary for writing dictation section.

Table 3
Vocabulary for Writing Dictation Section of Verbal Subtest

Vocabulary	
WP-04	ดนตรี
WP-08	สมาธิ

In the second pilot study, reading comprehension contained a short story and ten multiple-choice questions. In the main study, the researcher was able to use expert reviewer comments and analysis results to create double the number of items found in the second pilot study. Reading comprehension in the main study contained four short stories and twenty multiple-choice questions. This section tests the ability to read and comprehend a short story. Each examinee was provided with the same short stories. Each examinee was required to read the stories and try to remember as much of its details as possible. Then, the examinee was presented with questions. Each examinee was required to identify and select the correct answer to each question. Table 4 provides a sample story and question found in reading comprehension section.

Table 4
Sample Story from Reading Comprehension Section of Verbal Subtest

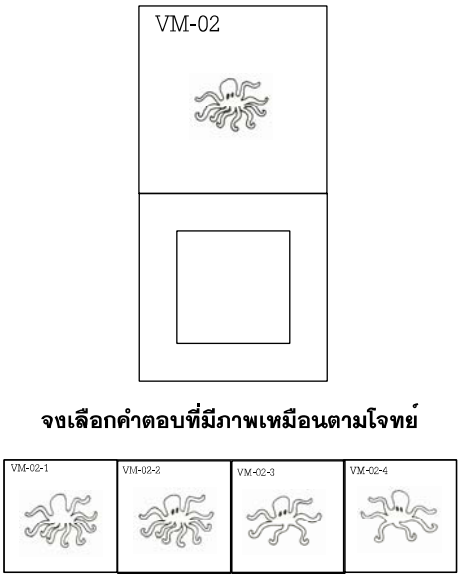
Story
เรื่องที่ 2 เช้าวันหนึ่งขณะที่แดงกำลังเล่นฟุตบอลอยู่ แม่ออกมาเรี่ยกให้แดงไปรับประทานอาหาร เมื่อรับประทานอาหารเสร็จแล้ว พ่อก็พาแดงขึ้นเกวียนไปที่แปลงผัก พ่อปลูกผักไว้หลายชนิด เช่น ตะไคร้ มะกรูด พริกขี้หนู โหระพา แตงกวา เป็นต้น แดงช่วยพ่อรดน้ำผักและใส่ปุ๋ย 1. ตอนเช้าแดงกำลังเล่นกีฬาอะไร ก. ตะกร้อ ข. ฟุตบอล ค. ว่ายน้ำ

Visual.

The literature review and the expert panel recommendations identified five types of tasks to elicit the behaviors that are indicative of visual readiness. They are visual matching, visual recognition, visual identification, mental folding, and mental rotation. For the second pilot study, visual matching contained thirty-three items that test the ability to visually discriminate different pictorial objects by size, color, shape, and location. For the main study, the researcher was able to use expert review comments and analysis results to create three more items, which brought the total number of items for visual matching in the main study to thirty six. Each item contains a pictorial object or a series of pictorial objects located at the top of the page. The bottom of the page contains a series of pictorial objects, one of which looks exactly like the one(s) at the top. When there was only one pictorial object shown at the top, the examinees were required to correctly select the pictorial object (from the choices at the bottom) that perfectly

matches the top one. When there was a series of pictorial objects at the top, the examinees were required to select the same number of pictorial objects at the bottom and correctly identify which selected pictorial objects match which pictorial objects at the top. Figure 1 shows a sample visual matching item found in Visual subtest.

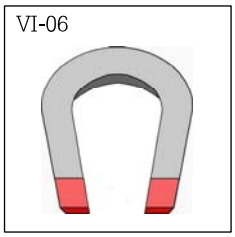
Figure 1
Sample Visual Matching Item from Visual Subtest



For the second pilot study, visual identification contained forty-four items testing the ability to identify a pictorial object, which is intermingled with other pictorial objects in a rectangular frame at the top of the page. For the main study, the researcher deleted some items from the visual identification in the second pilot study and added some items based on the expert reviewer comments and analysis results. For the main study, visual identification contained forty items. Each item contains a picture of an

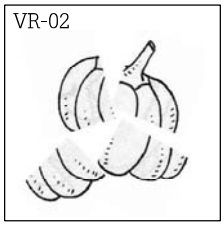
object. The examinees were required to identify the correct answer that best describes the object. Figure 2 shows a sample visual identification item found in Visual subtest.

Figure 2
Sample Visual Identification Item from Visual Subtest



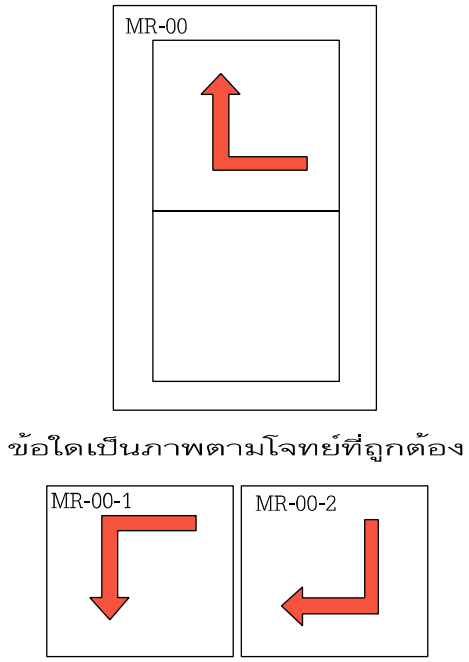
For the second pilot study, visual recognition contained ten items that test the ability to recognize objects based on the pictures provided. For the main study, the researcher was able to use expert reviewer comments and analysis results to create triple the numbers of items found in the second pilot study. For the main study, visual recognition contained thirty items. Each item contains a picture depicting object fragments arranged randomly. The examinees were required to look at the pictures and correctly identify the name of the thirty objects. Figure 3 shows a sample visual recognition item found in Visual subtest.

Figure 3
Sample Visual Recognition Item from Visual Subtest



For the second pilot study, mental rotation contained eleven items that test the ability to recognize pictorial objects, which are simply being displayed differently. For the main study, the researcher deleted some items from the mental rotation in the second pilot study and added some items based on the expert reviewer comments and analysis results. For the main study, mental rotation contained eight items. Each item contains a pictorial object at the top of the page. The bottom of the page contains a series of pictorial objects, one of which is the exact copy of the top pictorial object. Except being displayed at a different angle, the correct pictorial object at the bottom looks exactly like the one at the top. For each item, the examinees were required to select the correct pictorial object (at the bottom) that perfectly matches the top one. Figure 4 shows a sample mental rotation item found in Visual subtest.

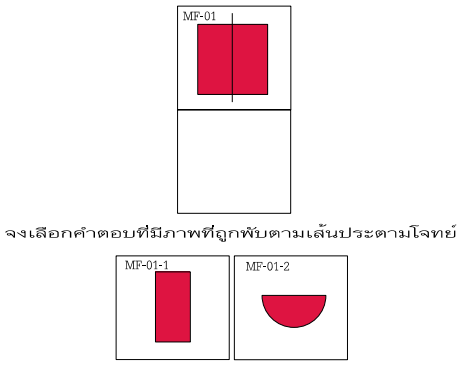
Figure 4
Sample Mental Rotation Item from Visual Subtest



For the second pilot study, mental folding contained ten items that test the ability to recognize the pictorial objects after being folded. For the main study, the researcher was able to use expert review comments and analysis results to create five additional items, which brought the total number of mental folding in the main study to seventeen. Each item contained a pictorial object at the top of the page. There was a dashed line cutting across the pictorial object. The dashed line signifies where the pictorial object is folded. The bottom of the page contained a series of pictorial objects, one of which looked exactly like the top pictorial object once folded as specified. The examinees were required to select the correct pictorial object (at the bottom) that

perfectly matches the top one once it is folded. Figure 5 shows a sample mental folding item found in Visual subtest.

Figure 5
Sample Mental Folding Item from Visual Subtest

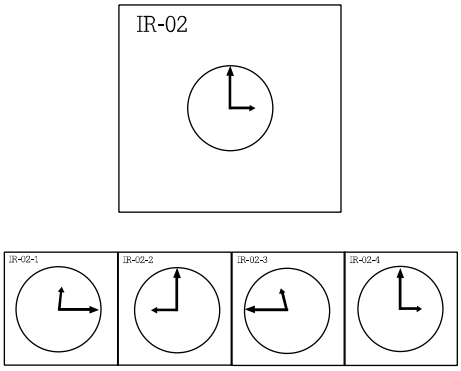


Memory.

The literature review identified three types of tasks to elicit the behaviors that are indicative of memory readiness. They are immediate recognition, spatial memory, and delayed recognition. For the second pilot study, immediate recognition contained eleven items that test short-term memory. For the main study, the researcher was able to use expert reviewer comments and analysis results to create almost triple the number of items found in the second pilot study. For the main study, immediate recognition contained thirty items. Each item contained a pictorial object on the first page. The second page contained a series of pictorial objects, one of which looked exactly like the one on the first page. The examiner showed the object on the first page to the examinees for five seconds. After the time limit, the examiner put away the first

page and provided the examinees with the second page. For each item, the examinees were required to select the correct pictorial object (on the second page) that perfectly matched the one on the first page. Figure 6 shows a sample immediate recognition item found in the Memory subtest.

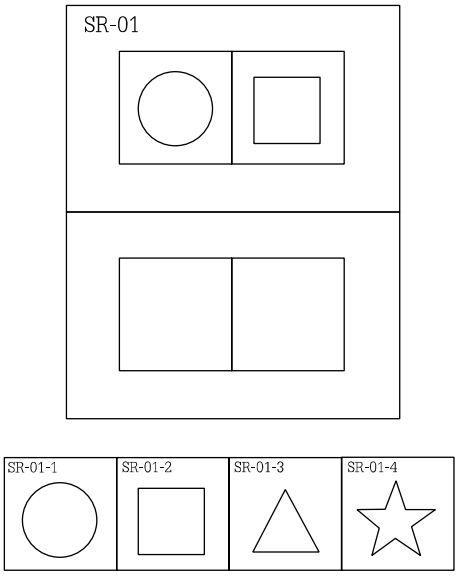
Figure 6
Sample Immediate Recognition Item from Memory Subtest



For the second pilot study, spatial memory contained twelve items that test short-term spatial memory. For the main study, the researcher created almost triple the number of items found in the second pilot study. For the main study, spatial memory contained thirty items. Each item contained a series of pictorial objects on the first page. The second page contained the same number of spaces as the number of the objects on the first page. The examiner showed the series of objects on the first page to the examinees for five seconds. After the time limit, the examiner put away the first page and provided the examinees with the second page. For each item, the examinees were

required to arrange the pictorial objects in the order shown on the first page. Figure 7 shows a sample spatial memory item found in Memory subtest.

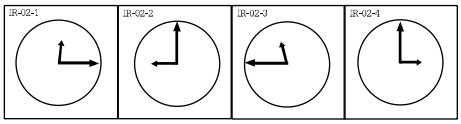
Figure 7
Sample Spatial Memory Item from Memory Subtest



For the second pilot study, delayed recognition contained eleven items that test long-term memory. For the main study, the researcher created almost triple the number of items found in the second pilot study. For the main study, delayed recognition contained thirty items. Delayed recognition items used the very same items in the very same order as the immediate recognition section. For example, the first delayed recognition item used the very same pictorial objects from the first immediate recognition item. Delayed recognition items used the pictorial objects on the second page of the immediate recognition items. The examiner showed each item containing the pictorial

objects from the corresponding immediate recognition item. After each item was shown, the examinee was required to select the correct pictorial objects that matched the ones on the first page of each corresponding immediate recognition item. Figure 8 shows a sample delayed recognition item found in Memory subtest.

Figure 8
Sample Delayed Recognition Item from Memory Subtest



Math.

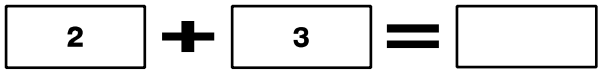
The literature review identified three types of tasks to elicit the behaviors that are indicative of math readiness. They are math concept and vocabulary, arithmetic, and word problems. For the second pilot study, math concept and vocabulary contained twenty statements, each of which described the mathematical relationship between two or more numbers. For the main study, math concept and vocabulary contained fifteen statements, each of which described the mathematical relationship between two or more numbers. After the examiner read the first statement, the examinee was required to write it down in a mathematical operation format. For example, the examiner read "one plus one." The examinees were expected to write "1+1." The examinees were required to write each of the statements in mathematical operation format correctly. Table 5 provides a sample math concept item found in Math subtest.

Table 5
Sample Math Concept and Vocabulary Item from Math Subtest

Items	
MCV-1	เขียน "ห้าลบสิบ" เป็นเลขอารบิก.

In the second pilot study, arithmetic contained thirty items that test computational skills. In the main study, the researcher created almost double the number of items found in the second pilot study. In the main study, arithmetic contained fifty items that test computational skills. Each item contained an equation with a missing number signified by a space. Each examinee was required to perform arithmetic computation to find the correct number for the space. The examinees were required to find the correct answer for each item. Table 6 provides a sample arithmetic item found in Math subtest.

Table 6
Sample Arithmetic Item from Math Subtest

Items	
MC-01	

For the second pilot study, word problems contained twenty math-related word problems. For the main study, the researcher created almost double the number of items found in the second pilot study. For the main study, word problems contained thirty two math-related word problems. This section tests the ability to interpret math-related word questions and to perform the necessary arithmetic computation to find the correct

answers. Each examinee was required to interpret the questions, use the relevant information provided with the questions, and perform the arithmetic computation to find the correct answers for each of the twenty two questions. Table 7 provides a sample word problem item found in Math subtest.

Table 7
Sample Word Problem Item from Math Subtest

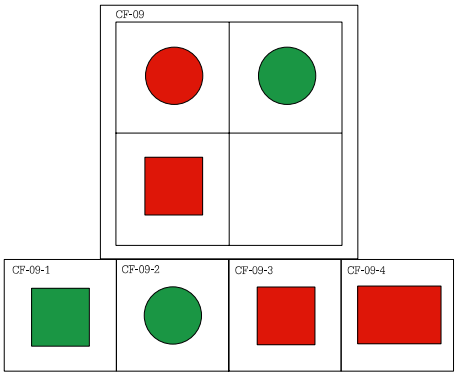
Items
WP-01 แม่มีไข่อยู่ 3 ฟอง แม่ซื้อไข่มาอีก 2 ฟอง แม่มีไข่ทั้งหมดกี่ฟอง ?

Logical.

The literature review identified three types of tasks to elicit the behaviors that are indicative of logical readiness. They are concept formation, sequential order, and pattern finding. For the second pilot study, concept formation contained twenty two items that test the deductive and inductive reasoning skills and the ability to conceptualize the relationship between the shown pictorial objects. For the main study, the researcher deleted some items from the concept formation in the second pilot study and added some items based on the expert reviewer comments and analysis results. For the main study, concept formation contained twenty five items. Each item contains a square divided into four quadrants. Except for the fourth, every quadrant contains a pictorial object. There is a relationship between the pictorial objects in the first and the second quadrant. There is a relationship between the pictorial objects in the first and the third quadrant. There is also a relationship between the pictorial object in the second

quadrant and the missing pictorial object in the fourth quadrant. There is no relationship between either the pictorial objects in the second and the third quadrant or the ones in the first and the fourth quadrant. The four quadrants are located at the top of the page. The bottom of the page contains a series of pictorial objects, one of which will correctly conform to the relationship between itself and other objects at the top. For each of the twenty-two items, the examinees were required to select the correct pictorial object that conforms to the relationship as described above. Figure 9 shows a sample concept formation item found in Logical subtest.

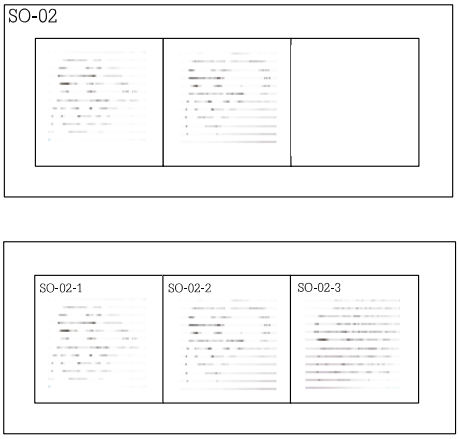
Figure 9
Sample Concept Formation Item from Logical Subtest



For the second pilot study, sequential order contained twenty items that test the ability to recognize the logical progressions of pictorial objects or events. For the main study, the researcher deleted some items from the sequential order in the second pilot study and added some items based on the expert review comments and analysis results. For the main study, sequential order contained fifteen items. Each item contains

a series of pictorial objects or events at the top of the page. Located in the middle or at the end of the series, the blank space(s) signifies the missing pictorial object(s) or event(s). The bottom of the page contains another series of objects, the right one(s) of which correctly conforms to the rules that govern the relationship in the series. The examinees were required to select the correct pictorial object(s) or event(s) that conform(s) to the rules, which govern the relationship of the series at the top. Figure 10 shows a sample sequential order item found in Logical subtest.

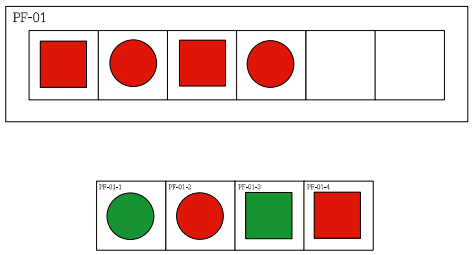
Figure 10
Sample Sequential Order Item from Logical Subtest



For the second pilot study, pattern finding contained twenty items that test the skills in deductive and inductive reasoning and conceptualization of the shown patterns. For the main study, the researcher deleted some items from the pattern finding in the second pilot study and added some items based on the expert reviewer comments and analysis results. For the main study, pattern finding contained twenty two items.

Each item contains a series of pictorial objects, figures, or characters. Located in the middle or at the end of the series, the blank space(s) signifies the missing pictorial object(s), figure(s), or character(s). The bottom of the page contains another series of object(s), figure(s), or character(s), the right one(s) of which conform(s) to the rules that govern the relationship of the series pattern at the top. The examinees were required to select the correct pictorial object(s), figure(s), or event(s) that conform(s) to the rules, which govern the relationship of the series pattern at the top. Figure 11 shows a sample pattern finding item found in Logical subtest.

Figure 11
Sample Pattern Finding Item from Logical Subtest



General Knowledge.

The literature review identified one type of task to elicit the behaviors that are indicative of general knowledge readiness. For the second pilot study, there were ten items assessing knowledge of important information, which is critical for academic success in first grade. For the main study, the researcher created eight times more than the number of items found in the second pilot study. For the main study, general

knowledge contained seventy five items. Each item contained a question and multiple-choice answers. The examinees were required to choose the correct answer for each of the eighty questions. Table 8 provides a sample word problem item found in Math subtest.

Table 8
Sample General Knowledge Item from General Knowledge Subtest

Items
1. ข้อใดไม่ใช่อวัยวะของร่างกาย ก. ตา ข. แขน ค. แหวน
10. สิ่งใดไม่ได้ช่วยให้ร่างกายของเราเจริญเติบโตและแข็งแรง ก. อาหาร ข. การออกกำลังกาย ค. ยาเสพติด

Expert Review

Expert review is one way to establish content validity. Content experts define in precise terms the universe of specific content that the test is assumed to represent. The content universe may come from school curricular materials used in kindergarten and first grade instruction (Linn, 1989). Then, the experts determine how well that content universe is sampled by the test items (Gall et al., 1996). After the review, content experts may suggest removing certain items that do not represent the content. Being content experts, they are good resources for additional valid items that are absent from the test.

In late 1999, the first pilot items were given to a group of content experts for the review (See Appendix 1 for titles and expertise). They were asked to comment on the

representativeness of item content to the content universe (See Appendix 2 for a sample list of comments). Their suggestions included addition of new items with a higher level of difficulty (See Appendix 3 for a list of sample additional items provided by the content experts). Their comments were incorporated in the development of the second pilot measure.

The second pilot items were given to another group of content experts in early 2001 (See Appendix 4 for titles and expertise). They were asked to evaluate the representativeness of the content of the second pilot items to the content universe. Specifically, they were asked to define the content universe for academic readiness for first grade in Thailand. Then, they evaluated how well the second pilot items sampled the content universe. The review resulted in the identification of the content areas that were not yet covered by the items. In addition, the experts were asked to provide a sample of items covering those areas and to identify the items that did not represent any areas of the content universe. The review resulted in suggestions to remove those items (See Appendix 5 for the list of content review questions). Another group of experts in the field of measurement and evaluation, child development, psychometric theory, curriculum and instruction, and test administration was asked to comment on the items and the test formats. (See Appendix 6-7 for titles and expertise and their comments.) Items were further revised based on those comments.

Internal Consistency

Internal consistency is one way of estimating reliability. As will be explained in more detail later, reliability is a quantity derived from classical test theory. The

psychometric model used in this project, which is the Rasch measurement model, relies more on a somewhat different formulation of reliability. Although the interpretation is consonant with that of classical test theory, the Rasch model uses the reliability of the "person separation index" to evaluate the internal consistency of a scale. Reliability of person separation is the ability of a set of items to reliably discriminate among people based on their trait level.

Internal consistency of a measure can be ensured when the development of a measure follows a proven scale development approach. The following paragraphs describe the steps that were taken to ensure acceptable reliability levels of tested domains for the main study. The discussion begins with the scale development process, which outlines the necessary steps to be taken during scale development. The discussion continues with the statistical theories used for item analysis of the test results. Then, the developmental process of, and the results from, the first pilot and the second pilot study are summarized. The discussion ends with the development process of the main study measure.

Scale Developmental Process

There are a number of books and articles on measure development (Bode & Wright, 1999; Benson & Clark, 1982; DeVellis, 1991; Wright & Stone, 1979). Benson and Clark (1982) recommended the steps can be classified into planning, construction, quantitative evaluation, and validation phase. DeVellis (1991) proposed seven steps, which were construct identification, item generation, format selection, expert review, inclusion of validation items, pilot administration, item evaluation, and scale length

optimization. Wright and Stone (1979), as well as Bode and Wright (1999), provided several suggestions for variable construction, test design, measure development, and scale selection.

The cited authors recommended a very similar measure development process. For instance, many similarities of the recommendations by DeVellis (1991) and Benson and Clark (1982) lie in the careful determination of construct, item generation, format selection, expert review, pilot administration, and item evaluation. The recommendations from Wright and Stone (1979) and Bode and Wright (1999) also fall quite perfectly with those of DeVellis (1991) and Benson and Clark (1982).

The development of the main study took advantage of the recommendations from these authors. However, the steps resembled most closely those from Benson and Clark (1982). The recommendations from other authors were incorporated into those steps when appropriate.

Planning Phase.

This phase involved careful planning prior to actual item generation. The steps in the planning phase include statement of purpose, domain identification and definition, and literature review. The statement of purpose defines the intended purpose of the measure. Determination of the purpose may be the most critical step in the development of a measure. An unclear purpose may lead to measuring a wrong construct and hence invalid measure.

After a statement of purpose is written, the next step is to identify the domains. Once the domains are identified, each of them is given a specific definition. Next, a

review of the literature is conducted. The literature review provides two major benefits. First, a literature review indicates if a valid measure with the same purpose already exists. Second, it confirms the number and types of domains specified earlier. In some cases, the literature review will also identify additional domains for the constructs.

Construction Phase.

The first step in construction phase is the generation of item pool. This involves much more than creation of item content. Item generation also concerns the selection of item formats. The item generation for verbal readiness, for example, involves not only the selection of vocabularies (item content) but also how (item formats) to test verbal readiness through vocabularies. DeVellis (1991) suggested that every item format be considered. The author believed that the number of items, and of formats, should be more than what will be included in the final scale. Bode and Wright (1999) also provided a very good consideration for item creation. They maintained that good items always aim at measuring different amount of the trait.

The next step in the construction phase involves expert review of the item pool. Experts are asked to review several aspects of the items. Depending on the type of measure, different types of experts are recruited. For Tests of Academic Readiness, two groups of experts were needed. The first group consisted of content experts. These experts specialize in the content areas the test purports to measure. Feedback from the content expert is necessary if content validity needs to be established.

The second group of experts included those whose expertise fell in such areas as child development, curricular and instructional theories, test administration, and

psychometric theories. The feedback from this group helped to improve the item quality based on the respective area of expertise. For example, a child development expert can help to identify items that may not be developmentally appropriate. A curriculum and instructional specialist can help to verify if the items address the curricular objectives of first grade subject matters. A test administration expert can pinpoint the items that can become problematic during the administration. A psychometric expert can help to ensure that the collection of items conforms to the requirements of the chosen psychometric theory.

Although the purpose of expert review is to solicit suggestions regarding item quality, the decision to follow the suggestions lies mostly with test developers. It is appropriate if the test developers choose to follow only the suggestions that are applicable to the scale. The suggestions may lead to revisions of existing items and inclusion of additional items to form a scale for the main study.

Quantitative Evaluation Phase.

After the pilot administrations, the raw scores were analyzed using item response theory (IRT). The analysis yields several statistics such as reliability coefficients, person separation, item discrimination, and fit statistics. These statistics help to improve the quality of items. For example, poorly discriminating items warrant either deletion or revision. The analysis results suggest which items should be dropped or revised and where new items are needed for the final measure. The results from the main study administration were used to further refine the instrument to form a final scale.

Validation Phase.

To establish validity of the scale, at least one or more of the following validation approaches must be undertaken: content, criterion-related, or construct validation. After the validation is performed, the scale is finalized. The final step is to provide test norms, publish the test results, the instrument, and the manual. After the first validation of the scale, additional validation studies may be undertaken to assess how the scale continues to function over time.

Statistical Theories for Test Development

There are two current statistical theories for test development: classical test theory and item response theory (IRT) (Hambleton & Jones, 1993; Snyder & Sheehan, 1992). Until a few decades ago, test developers used classical test theory as the primary test development tool. Now they choose between classical test theory and IRT. Each of the theories will be briefly discussed below. The differences between the two theories will also be explained.

Classical Test Theory.

Classical test theory concerns three types of scores: test (observed) score, true score, and error score. The observed score (X) is linked in a simple linear fashion by the true score (T) and the error score (E), hence $X = T + E$. From this formulation, classical theory assumes that the observed score is the result of the true score and some error due to factors unrelated to the ability of the examinees or the difficulty of the test items. In order for this formulation to work, certain assumptions are made. First, true scores and

error scores are uncorrelated. That is, classical test theory assumes that error scores are constant. As a result, the error score will neither decrease nor increase due to a change in the true score. Second, in the long run, the average error score from the examinee population is zero. Third, error scores on parallel tests are uncorrelated (Hambleton & Jones, 1993). Classical test theory assumes that parallel tests are two tests that yield the same true score and the same variances of error score (Lord & Novick, 1968). The existence of a correlation between errors in both forms implies that some systematic traits are simply not yet accounted for by test items. Fourth, repeated administrations of a test yield a value of the observed score exactly equal to that of the true score.

Item Response Theory.

Item response theory (IRT) links item scores to trait level by showing how test performance is determined by the abilities of examinees and the difficulty level of items (Hambleton & Jones, 1993). IRT provides a mathematical statement of the probability that an examinee with a particular level of ability will experience success on a particular item measuring that trait. IRT assumes that there are one or more underlying trait(s) which determine(s) an examinee's observed responses to test items. The trait can be defined in a quantitative (natural log) unit, called "logit" (Elliot, 1983; Snyder & Sheehan, 1992). The values of item difficulty and person ability (in logits) can be used to locate their position along a latent trait continuum. The difference between the position of a person and an item determines the probability of the person responding correctly to the item (Snyder & Sheehan, 1992).

There are two assumptions of IRT models. The first assumption concerns the dimensional structure of the model. Two major variations include unidimensional models and multidimensional models. The unidimensional model (e.g., Rasch model) assumes that a single latent trait accounts for differences in person performance. The Rasch model assumes that one parameter—the difference between person position and item difficulty—can measure the trait level. Therefore, the Rasch model is recognized as a single parameter model.

Another variation of a unidimensional model is the "two-parameter" model, which assumes that two parameters—item difficulty and item discrimination—are needed to model the data. The last variation of the unidimensional model is the "three-parameter" model. It assumes that three parameters—item difficulty, item discrimination, and guessing—are needed.

Unlike unidimensional models, multidimensional models assume that more than one latent trait accounts for the differences in the person performance (Hambleton & Jones, 1993).

Differences between Classical Test Theory and IRT.

There are a number of differences between classical test theory and IRT. As each difference is discussed, it will become obvious that certain differences are benefits of IRT and limitations of classical test theory.

The first difference between classical test theory and IRT lies in the long history of classical test theory. More familiarity with the statistics has led test developers to prefer classical test theory to IRT (Dun & Dun, 1981). Having been used in the

development of numerous measurement instruments (Snyder & Sheehan, 1992), classical test theory has been recognized as the *de facto* statistical theory of test development.

The second difference concerns the complexity of the theories. Classical test theory is more straightforward and requires simpler mathematical analyses. IRT, on the other hand, is more complex and difficult to comprehend. Consequently, some test developers are more inclined to use classical test theory.

The third difference between classical test theory and IRT concerns sample characteristics. The item difficulty and item discrimination statistics that form the cornerstones of classical test theory are sample dependent (Hambleton & Jones, 1993; Wright & Masters, 1982). Sample dependency means that item difficulty and item discrimination statistics are based on the ability of the specific sample to which a test is being administered. Consequently, estimates of item difficulty and item discrimination are mathematically confounded with specific characteristics of the examinees in the sample. Since the sample always differs in some way from the population, the item statistics are applicable only to the particular sample (Hambleton & Jones, 1993). Unlike classical test theory, IRT's item difficulty estimate is independent of the sample characteristics (Snyder & Sheehan, 1992). Therefore, the application of the IRT's item statistic is not limited to a particular sample.

The fourth difference concerns test score dependency. This is another important limitation of classical test theory. Test score dependency means an examinee's scores depend on the particular difficulty level of the items to which the examinee responses. There are two reasons that dependency makes it difficult to predict how an examinee may

perform on a different test. The first reason is that the scores on the two measures are on different scales. The second reason is that no functional relationship exists between those scales. Unlike classical test theory, IRT determines the probability of a particular examinee correctly answering any given item. Therefore, the measurement of an examinee's ability is independent of the administered items (Snyder & Sheehan, 1992). Being test dependent, the classical test theory is often described as "test-based." IRT in contrast is described as "item-based" (Rasch, 1980).

The fifth difference between classical test theory and IRT concerns the sample size. Because IRT's estimates of item difficulty and person ability are sample independent, the sample size required for a meaningful standardization is lower for IRT than for the classical test theory. Since the statistics in classical test theory are sample dependent, a big enough sample is needed to achieve a reasonable representation of the population (Snyder & Sheehan, 1992; Hambleton & Jones, 1993). While IRT needs a smaller sample for the standardization, its complex analysis actually requires a larger sample during the administration stage.

The sixth difference lies in the additional amount of information obtained from IRT. First, the information indicates precisely where an item is doing its best measurement on the ability scale. Second, the information helps to determine the exact relationship between item performance and person ability (Hambleton & Jones, 1993). Third, the information allows a broader range of interpretation at the item level. Fourth, the information permits a prediction of persons' scores at any given ability level. Fifth, IRT provides information regarding the contribution of particular items to the ability

assessment. The more the information provided by a test at a particular ability level, the lower the errors are for the ability estimation (Hambleton & Jones, 1993).

The seventh difference concerns the technical side of scale development. IRT does not require strict parallel tests for assessing reliability. Unlike IRT, classical test theory does not require strict goodness-of-fit tests to ensure a good fit of model to the test data.

The eighth difference between the classical test theory and IRT lies in the property of "model-parameter" invariance. IRT incorporates information about the examinees' ability into the item-parameter-estimation process (Hambleton, Swaminathan, & Jane, 1991). There are three benefits of invariance. First, it allows for the investigation of possible item bias. Second, it permits equating of tests through linking common items with known difficulties. Third, it allows for similar estimates of person ability and of item difficulty regardless of which items are being administered and of the ability of the persons taking the test.

In summary, the benefits of the classical test theory and the IRT are as follow:

Classical Test Theory

- Long track record
- Straightforward and simple mathematical analyses
- Smaller sample size for administration
- Assumptions are easier to meet

IRT

- Sample-Independence
- Test-Independence
- Linear Measure
- Broader range of interpretation possible at item level

- Investigation of possible item bias
- Linking tests through common items
- Smaller sample size for standardization.

The advantages of IRT over classical test theory have led to the selection of IRT as the primary development tool. In addition, the Rasch model is chosen from among other IRT models for the following reasons. The Rasch model is a single parameter model. It is an inevitable choice if one wishes to measure singular constructs (Green, 1996). The Rasch statistics provide a means to evaluate if the data fit the model. The other IRT models, in contrast, add parameters to enhance the fit of data to the model. The Rasch model is a 'stochastic' realization of Guttman scaling. Other IRT models do not follow this joint transitivity property recognized by Guttman as a necessity for the construction of a measure.

Rasch Measurement Model.

Trait Continuum.

A ruler is a useful example of how a measure must behave. A good measure must be able to define a trait continuum in the same manner as the ruler defining the people's height. Jones (1971) defined measurement as a determination of the magnitude of an object's attribute, which must be observable and can be counted in equal unit of like meaning. Accordingly, a ruler is a measurement system that determines the magnitude of a person's height, which is observable by standing the ruler on the ground and parallel to the person while observing the mark that is closest to the top of the person's head. A ruler type of measure implies equal intervals. It also implies some

standard process for use. Further, the trait can be measured by any ruler in the class of rulers.

The use of a ruler as the representation of a measure has many implications. First, a construct varies in degree. It is not a matter of all or nothing. For example, there is never a person with zero height. Some people may be shorter while others may be taller. Second, there is always a direction from the lower end of the construct to the higher end. The lower end signifies a lesser degree of the construct while the higher end signifies for a greater degree. Progressing from the lower end toward the higher end leads to increases in degree of construct attributes. Third, a measure, like a ruler, is universal. A standard ruler is the accepted measurement instrument of height around the world. If a person is measured at 6' 2", he or she is undeniably 6' 2". It does not matter where the person is being measured as long as a standard ruler is being used. Fourth, a measure is sample free. In other words, the 6' 2" person is still 6' 2" regardless of who else is being measured along with the person. There is never a case where a 6' 2" person will be any shorter or taller than 6' 2" when he is being measured along with many others.

Statistical Formulation.

In the Rasch measurement model, person responses are determined by person ability and item difficulty. Person ability b_v and item difficulty d_i interact to produce the responses. The difference between b_v and d_i defines the probability of correct or incorrect response when a person uses his or her ability to respond to an item of a given difficulty.

Since the difference between b_v and d_i varies from minus infinity to plus infinity, two steps are taken to bring the probability of response to within zero and one. First an exponent of the natural constant "e" (i.e., e=2.71828) is first applied to limit the difference to within zero and plus infinity and hence the exponential function $\exp(b_v - d_i)$. Second, a ratio $[\exp(b_v - d_I)/(1 + \exp(b_v - d_I))]$ is formed to bring the interval to within zero and one. The probability of a successful response is therefore $p\{x_{vi} = 1|b_v,d_I\} = p_{vi} = \exp(b_v - d_I)/[1 + \exp(b_v - d_I)]$. This Rasch model is used with a dichotomous $x_{vi} = 0,1$. The logarithmic version of the Rasch model is $\log(p_{vi1}/ p_{vi0}) = b_v - d_i$.

Rasch Analysis Process.

The first step in the Rasch analyses involves item calibration and person ability estimation. After persons and items with extreme scores are set aside, the data are summarized into person and item scores by summing each row and each column in the data matrix. These scores are then transferred into the proportions of their maximum possible values in order to free person and item scores from sample size and test length. To linearize these proportions, the log odds, or "logits," are calculated by taking the natural log of the proportion incorrect (for item) or success (for persons) divided by the proportion correct (for items) or failures (for persons). Means and variances for the person and item logit distributions are also computed. The mean for item logits is used to remove the effects of the ability level of the sample on the items. This centers the item calibrations at zero. The variances are used to calculate two expansion factors, by which each person and item estimate is multiplied. The expansion factors are necessary to widen the distance between any two persons of similar ability and any two items of

similar difficulty. Finally, standard errors of the person and item estimates are calculated to assess the precision of item-free person ability measures and person-free item difficulty measures.

The second step involves the analyses of fit. The Rasch model requires a person to respond to an item in a certain way. The analyses of fit evaluate how well the data fit this expectation. The evaluation of fit examines if the response of each person to an item is consistent with the general pattern of the responses observed. This is accomplished by defining the expected value of the variable realized in any response in terms of the probability of that response occurring. The expected value is used to calculate standardized residuals, which indicates the unexpectedness of any observed response. The chi-square's degree of freedom is used to evaluate whether the estimated standardized residuals deviate significantly from their model expectations.

There are two types of misfit estimates: infit and outfit. The misfit statistics are useful indicators of "noise." Infit indicates irregular patterns of responses for items close to a person's ability level. A large infit implies a central pattern of response incoherence. Outfit indicates unexpected responses to items far from the person's ability level. A large outfit implies the presence in the data of unexpected off-target responses.

The third step in the Rasch analysis involves the determination of variable existence and usefulness. A variable exists only when it measures different amounts of the trait. The item separation is an indicator of the spread in item difficulties. The larger the item separation, the wider the range of the attribute defined by the set of items. A variable is useful only if persons differ in the extent to which they possess the trait.

Person separation is an indicator of the spread in person measures. This index indicates the number of distinct levels into which the sample of persons can be classified. This degree of separation indicates that the difference must be due to the differences in the magnitudes of the person's underlying attribute (Bode & Wright, 1999).

Pilot Studies

A pilot study refers to a small-scale testing of the items and administration procedures that a researcher plans to use in the main study (Gall et al., 1996). In some cases, the pilot study is carried out after the research proposal has been approved by the dissertation committee. In other cases like in the first pilot situation of this dissertation, a pilot study may be carried out prior to the proposal approval. Such cases happen when the research problem involves trying out a new procedure, the use of which has no precedent in the literature. The findings from the pilot study can be used to prove the merit of the new procedure or to justify for conducting a formal, full-scale study.

The development of Test of Academic Readiness involved trying out new items that measure the academic readiness for first grade in Thailand. The literature review has found no precedent in the use of such measure in Thailand. It was therefore necessary to conduct a small-scale test of the items to prove the merit of a more formal, full-scale study. The first pilot study was the result of such small-scale testing. The knowledge gained from conducting the first pilot study provided a proof of merit for a full-scale study. Additionally, the knowledge proved useful for the development of the second pilot measure. Specifically, the item analysis results of the first pilot study were used to avoid mistakes committed during the development of the first pilot items.

First Pilot Study.

Purpose.

The researcher conducted a pilot test of items that represented six school readiness domains. The first version of Test of Academic Readiness was piloted in late 1999. The purpose of the first pilot was to evaluate the appropriateness of the items measuring various school readiness domains. The measure developed for the first pilot administration was designed to assess kindergarten graduates of their readiness for first grade instruction. The instrument provided a standardized measure of four skills. The measure assessed the child's visual, verbal, logical-mathematical, and spatial development.

Development.

The developmental process of the first pilot measure followed a similar process as that described earlier. The development of the first pilot measure began with a statement of purpose. A clearly defined purpose helped with the domain identification. Several readiness tests provided a list of readiness domains, which were potential candidates for the first pilot domains. The available readiness tests were not the only sources for domains. Knowledge of such psychological development theory as multiple intelligences provided a framework for the formulation of the readiness domains (Gardner, 1993). Additionally, the subject matters, which are taught to first grade students in Thailand, provided a support for the domains.

After the domains were selected, items were created for each domain (See Appendix 8). Some items were created anew while others were adopted from other readiness tests measuring similar constructs. The first pilot items were then given to two groups of experts for a review. The first group consisted of content experts discussed earlier. The second group consisted of experts in the field of child development, measurement and evaluation, test administration, and psychometric theory. The comments from both groups of experts are summarized in Appendices 2 and 7.

Administration.

The first pilot administration took place in late 1999. The sample was drawn from the population of graduating kindergarten students at a private school in Bangkok, Thailand. The sample included 32 boys and 29 girls. The test administration took approximately two hours for each student to complete. Permission to conduct the pilot study was obtained from the University of Denver Institutional Review Board and from the principal of the private school in Thailand.

The administration of the subtests normally followed the same sequence (e.g., motor, visual, verbal, logical-mathematical, music, and spatial). Some students were allowed to follow a different sequence if they so stated their preference. There were five sections of motor, seven sections of visual, seven sections of verbal, four sections of logical-mathematical, three sections of musical, and one section of spatial subtest. Each subtest contained a varying number of items. There were 392 items for the whole battery (See Appendix 9 for the number of items for each subtest for the first pilot study). Unless expressing his or her wish otherwise, every student was asked to complete every item.

The scores of the student who asked to discontinue the administration were not included in the analysis.

Results.

The data obtained from the first pilot administration were tabulated and entered into a computer spreadsheet. The scores from each subtest were entered into a separate spreadsheet. The data in each spreadsheet were then visually inspected for invalid entries. After the data were "cleaned," each spreadsheet was transformed into the correct file format (i.e., .dat) for "Winsteps"—a computer program for Rasch analysis (See Appendix 10 for an example of an input file). To perform a Rasch analysis for each subtest, Winsteps asked for the appropriate file.

After performing the analysis for a subtest, Winsteps generated a number of tables, which were saved into a computer output file. The tables contained information regarding the item quality of those particular subtests. If the information suggested a removal of one or more children or items, a change was made to the input file for another iteration of analysis. This process repeated itself until there was no more children or item scores that misfit the model.

Certain indicators signify if another iteration should take place. One of the indicators is the "infit" and "outfit" statistics. Any items having either an infit or an outfit statistic outside of the accepted 0.7-1.3 range should be removed. Another indicator is the reliability coefficient. There are two types of reliability coefficients: person reliability and item reliability. These reliability coefficients exhibit a sign of possible improvement to the model if more kid(s) and/or item(s) are removed. However,

the coefficients do not indicate which kid(s) or item(s) should be removed. The removal candidates could only be identified by their misfit statistics. The following paragraphs summarize the item analysis process performed for each subtest.

Visual Discrimination. The researcher ran eighteen iterations and found seventeen items with poor fit (i.e., item 1- 4, 12-13, 20-24, 26, 28, and 43- 47,), which were deleted. Table 9 below provides information regarding the person and item separation and reliability.

Table 9
Person and Item Reliability and Separation—Visual-Discrimination Subtest

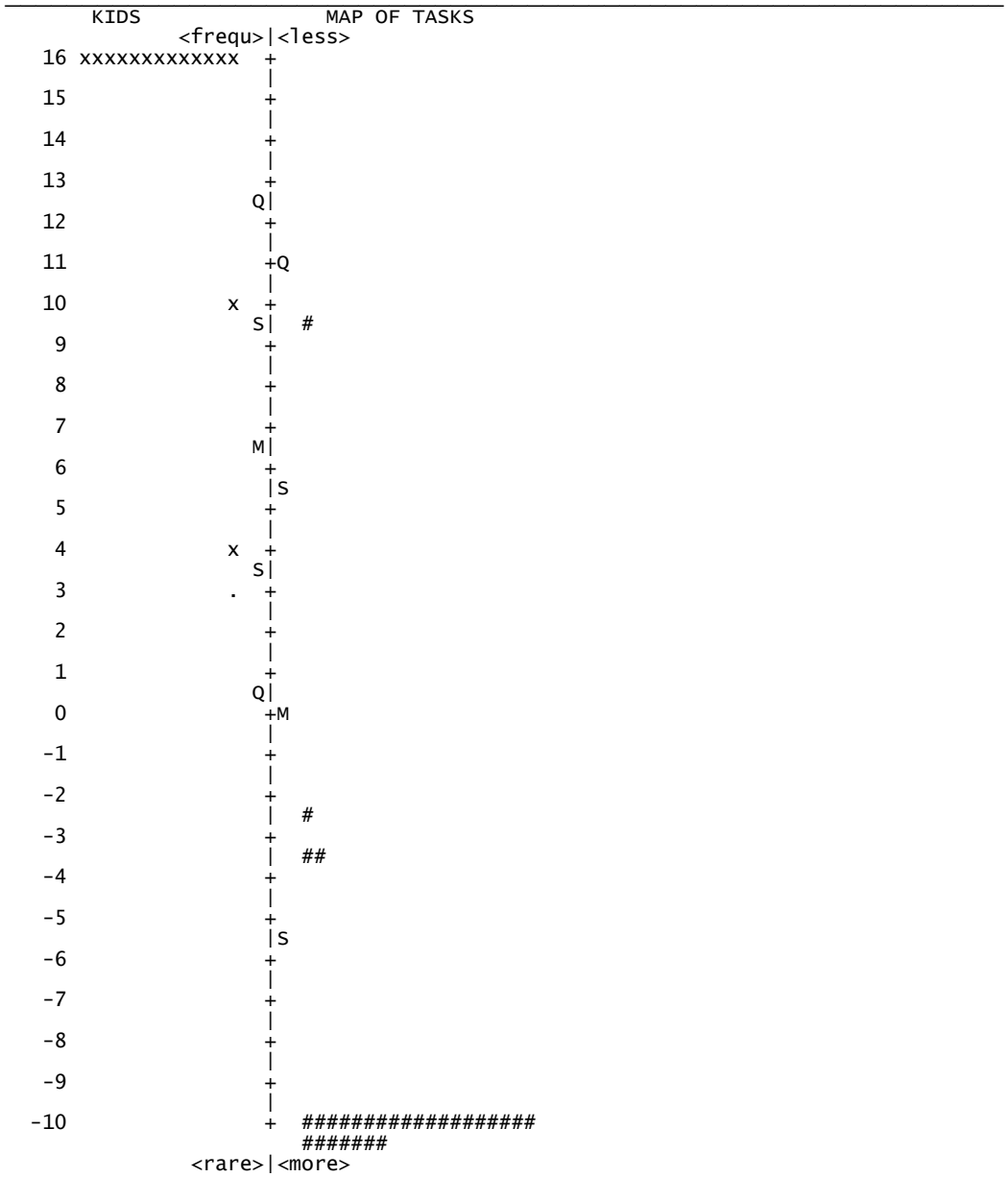
Statistics	Value
Reliability of person separation	.00
Reliability of item separation	.96
Item Separation (Real RMSE Separation)	.00
Person Separation (Real RMSE Separation)	4.83

The achieved level of reliability for person separation was nonexistent and it was high for item separation. The items were not able to separate children into groups (with a Real RMSE Separation of 0.00) while the children were able to separate items into almost five difficulty levels (with a Real RMSE Separation of 4.83).

Figure 12 below maps or plots the positions of the children ability (to the right of the vertical line) relative to those of the item difficulty (to the left of the line). The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 6.57 compared to the arbitrary item mean position of 0.0. In

other words, children on average found the items to be very easy. Thus, this item set was poorly targeted for these children, which led to low reliability.

Figure 12
Item/Child Position Map for Visual Discrimination Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Logical-Mathematical. The researcher ran ten iterations and found nine items with poor fit (i.e., item 1, 4, 8, 13, 14, 17, 27, 33, and 34), which were deleted. Table 10 below provides information regarding the person and item separation and reliability.

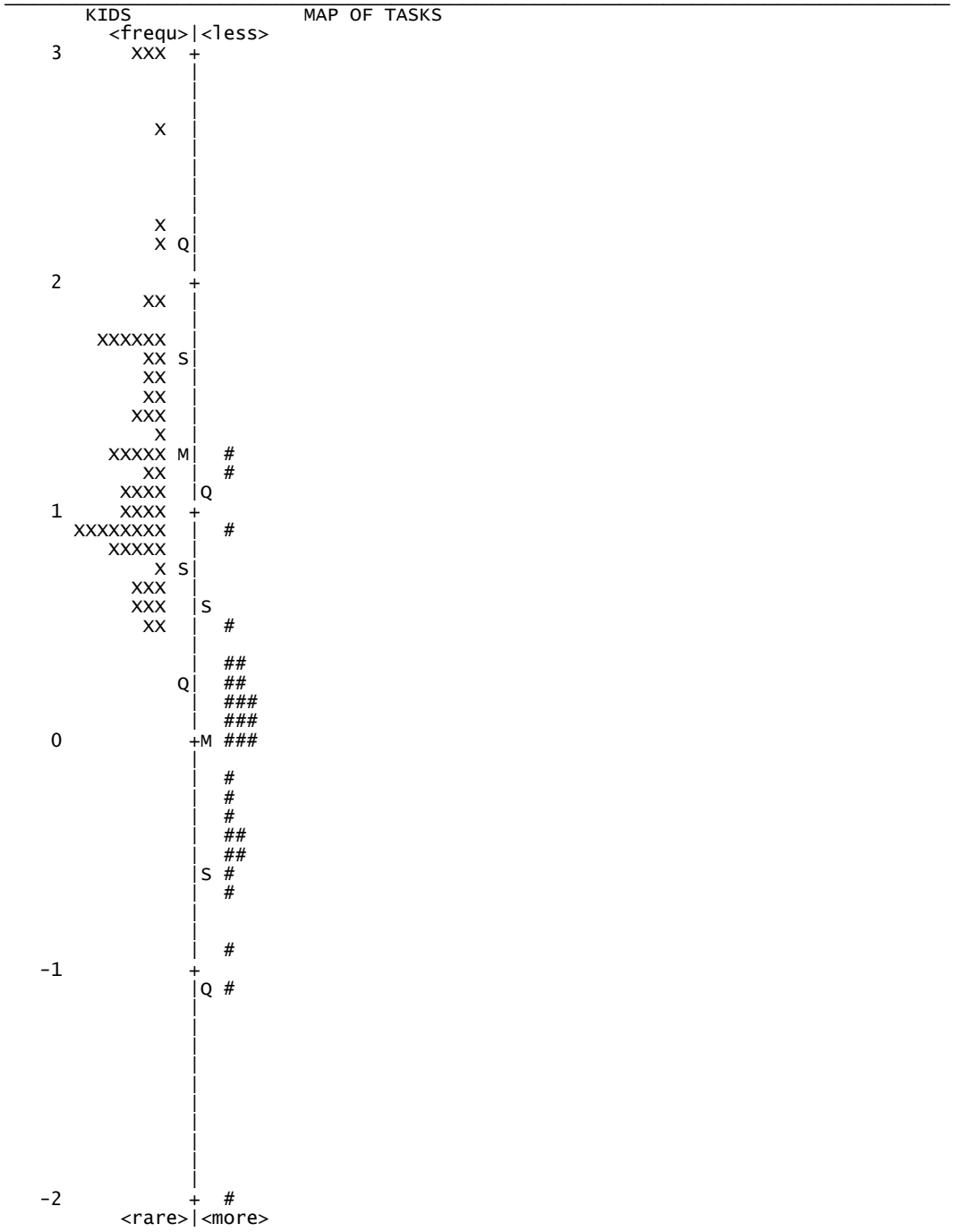
Table 10
Person and Item Reliability and Separation—Logical-Mathematical Subtest

Statistics	Value
Reliability of person separation	.39
Reliability of item separation	.77
Item Separation (Real RMSE Separation)	.80
Person Separation (Real RMSE Separation)	1.83

The achieved level of reliability for person separation was low and it was moderate for the item separation. The items were not able to separate children into groups (with a Real RMSE Separation of 0.80) while the children were able to separate items into almost two difficulty levels (with a Real RMSE Separation of 1.83).

Figure 13 below shows a map of item/child position for the Logical-Mathematical Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 1.21. In other words, children on average found the items to be very easy. Thus, this item set was poorly targeted for these children, which led to low reliability.

Figure 13
Item/Child Position Map for Logical-Mathematical Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Verbal. The researcher ran sixty-eight iterations but used the results from the twentieth iteration as it yielded the best level of reliability. The researcher found nineteen items with poor fit (i.e., item 10, 14, 18, 28, 26, 31, 33, 50, 93, 113, 116, 117, 158, 174, 179, 176, 191, 192, and 196), which were deleted. Table 11 below provides information regarding the person and item reliability.

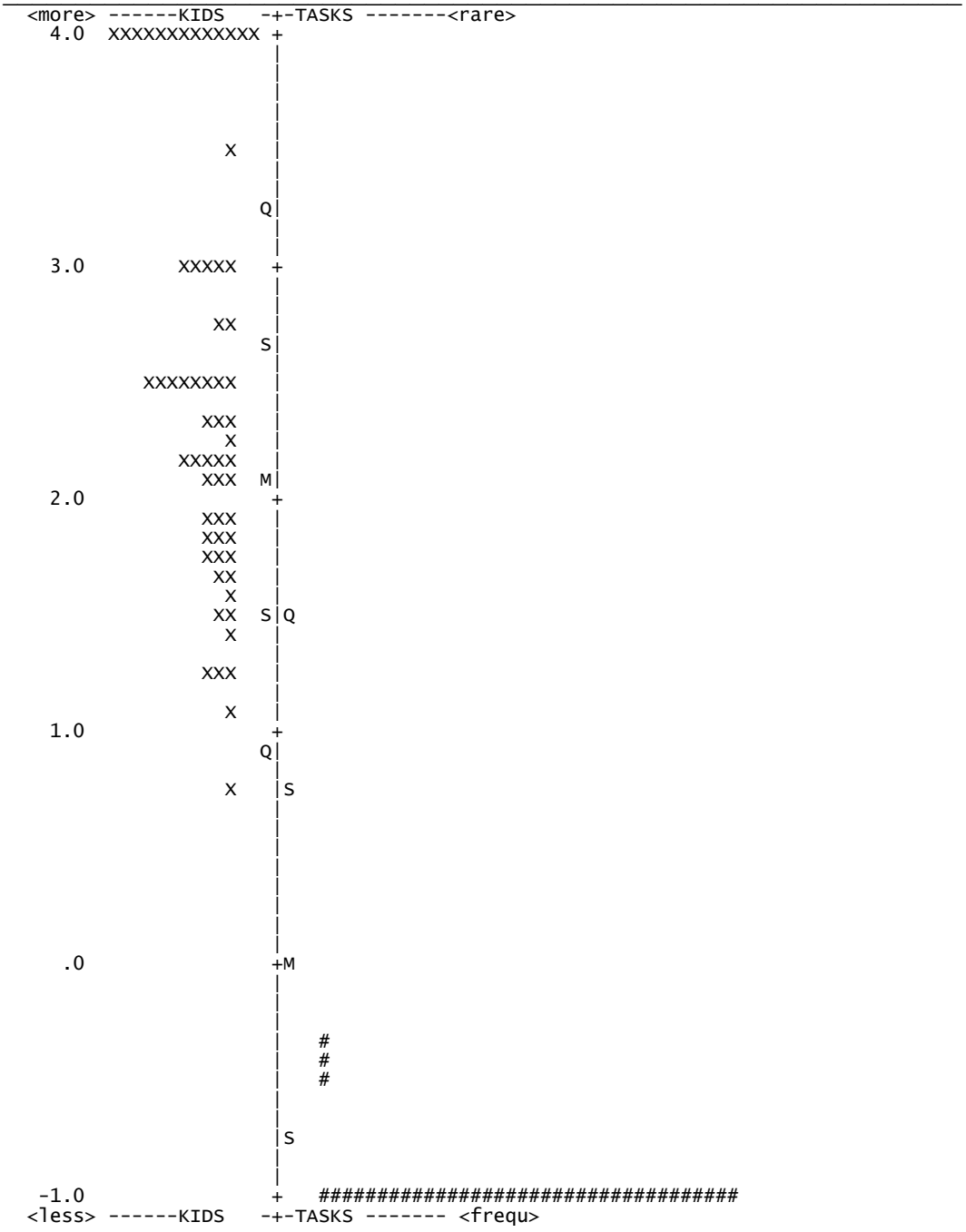
Table 11
Person and Item Reliability and Separation—Verbal Subtest

Statistics	Value
Reliability of person separation	.44
Reliability of item separation	.43
Item Separation (Real RMSE Separation)	.89
Person Separation (Real RMSE Separation)	.00

The achieved level of reliability for both person separation and item separation was low. The items were not able to separate children into groups (with a Real RMSE Separation of 0.89). Similarly, the children were not able to separate items into difficulty levels (with a Real RMSE Separation of 0.00).

Figure 14 below shows a map of item/child position for the Verbal Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 2.12. In other words, children on average found the items to be very easy. Thus, this item set was poorly targeted for these children, which led to low reliability.

Figure 14
Item/Child Position Map for Verbal Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Spatial. The researcher ran four iterations found two items with poor fit (i.e., item 2 and 4), which were deleted. Table 12 below provides information regarding the person and item reliability.

Table 12
Person and Item Reliability and Separation—Spatial Subtest

Statistics	Value
Reliability of person separation	.00
Reliability of item separation	.92
Item Separation (Real RMSE Separation)	.00
Person Separation (Real RMSE Separation)	3.36

The achieved level of reliability for person separation was low and it was high for item separation. The items were not able to separate children into groups (with a Real RMSE Separation of 0.00) while the children were able to separate items into more than three difficulty levels (with a Real RMSE Separation of 3.36).

Figure 15 below shows a map of item/child position for the Spatial Subtest. The map indicates that, on average, children were as able as items were difficult, with a child logit mean position of 0.00. In other words, children on average could respond correctly to items with approximately a .50 likelihood.

Figure 15

Item/Child Position Map for Spatial Subtest

```
                                    KIDS    -MAP- TASKS
                                        <more>|<rare>
    6                              X X X X X X  +
                                                |
                                                |
                                                |
                                                |
    5                                           +
                                                |
                                                |
                                                |
                                                |
    4                                           +
                                                |
                                                |S #
                                                |
    3                                           +
                                                |
                                                |
                                                |
                                                |
    2                                           +
                                                |
                                                |
                                                |
                                                |
    1                                           +
                                                |
                                                |
                                                |
    0                      X X X X X X X X X X M+M
                                                |
                                                |
                                                |
   -1                                           +
                                                |
                                                |
                                                |
   -2                                           +
                                                |
                                                |
                                                |
   -3                                           +
                                                |
                                                |
                                                |S #
                                                |
   -4                                           +
                                                |
                                                |
                                                |
   -5                                           +
                                                |
                                                |
                                                |
   -6 X X X X X X X X X X X X X X X X X X X X X X X X  +  # #
                                        <less>|<frequ>
```

Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Discussion.

The item analysis results of every subtest indicate that the first pilot items were too easy for the children. There are two alternative courses of actions that can be taken to address the difficulty level. First, items with higher difficulty levels can be generated for the second pilot subtests. Second, a whole new set of items can be created for the second pilot measure. The rationale for selecting the second alternative is explained below.

Following the first pilot study, the literature review uncovered two additional readiness domains, which were memory and general knowledge. In addition, a decision was also made to exclude certain readiness domains of the first pilot. The domains were motor and music. The rationales for such exclusion are given above (pp. 32-34). The decision resulted in a somewhat different set of readiness components. For the first pilot measure, the domains included motor, visual, verbal, logical-mathematical, music, and spatial. The domains of the second pilot measure include verbal, visual, math, memory, logical, and general knowledge. If more items of the original subtests were to be created, they would not fit some of the second pilot domains. The different set of second pilot domains, along with the lack of a match between item difficulty and person ability in the first pilot study, dictates the need to start anew.

Second Pilot Study.

Purpose.

The purpose of the second pilot study was to test the items that represent the six academic readiness domains as described earlier.

Development.

The development process for the second pilot measure is described in the following paragraphs.

Planning.

The development process of the second pilot measure began with the statement of purpose. Unlike that for the first pilot study, the purpose of the measure for the second pilot study was much narrower and more refined. The purpose of the instrument was to measure the *academic* readiness of the kindergarten graduates for their first grade instruction.

After the determination of the purpose, the identification of domains takes place. Based on the reviews of readiness tests (Boonruang, 1991; Ineay, 2004; Roid & Millers, 1997), resource books (First Grade Exams, 1997, Mati Silapin, 1987; Nontapuk, 2009; Pangwiruitrak, 2007; Pinyo Anantapong, 1993; Pinyo Anantapong, 1996; Sang Asanee, Boon Urapeepinyo, Wong Wijit Sin, Ruji Rek, & Apichartimanon, 1990; Trium Sop Por 1, 2009; Wai Prip Trium Sop, 2009), and first grade curriculum in Thailand (Helair, 1996), six domains were identified. They were verbal (Clymer & Barrett, 1966; Danzer, Gerber, Lyons, & Voress, 1967; Mardell & Goldenberg, 1983; Newborg, Stock, Wnek,

Guidubaldi, & Swinicki, 1984), visual (Clymer & Barrett, 1966; Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), memory (Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), math (Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), logical (Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), and general knowledge (Danzer, Gerber, Lyons, & Voress, 1967).

After the domains were identified, a review of the literature was conducted. The literature review suggested five components of school readiness for the U.S. and for Thailand. The literature review found academic readiness as a common component of school readiness for the U.S. and Thailand. The literature review also found the domains that can best measure academic readiness in Thailand (Malunpong, 1982; ONPEC, 1991; ONPEC, 1998; OPEC, 1990; Panich, 1988; Pluksawan, 1975; Pinjinda, Jongpayuha, & Charoensuk, 1973; Sintuwej, 1984). The domains were discussed in the previous paragraph. The literature review did not uncover any tests that served the same purpose as that of this measure. This was evidence to prove the merit of the development of a measure of academic readiness for first grade in Thailand.

Construction.

After the domains are identified and operationalized, the generation of item pool takes place. The process involves generation of item content and selection of item formats. The item content was derived from the findings of the first pilot study, the suggestions of the content experts, and the literature review. The items contained in the second pilot measure included new items adapted from several resources such as the

Leiter International Performance Scale (Roid & Miller, 1997), resource books (First Grade Exams, 1997, Mati Silapin, 1987; Nontapuk, 2009; Pangwiruitrak, 2007; Pinyo Anantapong, 1993; Pinyo Anantapong, 1996; Sang Asanee, Boon Urapeepinyo, Wong Wijit Sin, Ruji Rek, & Apichartimanon, 1990; Trium Sop Por 1, 2009; Wai Prip Trium Sop, 2009), and first grade textbooks from Thailand (Helair, 1996) (See Appendix 11 for an example of second pilot items). Operational definitions of the second pilot domains are provided earlier in the content validity section.

As previously mentioned, the second pilot items were given to a group of content experts and another group of experts in the field of child development, educational measurement and evaluation, test administration, and psychometric theories. The comments from both groups of experts were incorporated into the development of the second pilot measure. The revised measure was used for the second pilot administration.

Quantitative Evaluation.

The second pilot administration took place midyear, 2001. The sample was drawn from the population of graduating kindergarten and beginning first grade students at a private school in Bangkok, Thailand. The sample included 45 boys and 56 girls. The test administration involved three separate sessions, during each of which the two subtests were administered and each lasted approximately two hours. The administration followed the same sequence (e.g., verbal, logical, visual, math, memory, and general knowledge). Some students were allowed to follow a different sequence if they so stated their preference. There were three sections of verbal, five sections of visual, three sections of math, three sections of memory, three sections of logical, and one

section of general knowledge subtest. There were 361 items for the whole battery (See Appendix 12 for the number of items for each subtest in the second pilot study). Unless expressing his or her wish otherwise, every student was asked to complete every item. The responses of students who asked to discontinue the administration were not included in the analysis.

Results.

The researcher followed the same procedure for data preparation, Winsteps analyses, and consideration for removal of children or items.

Verbal. The researcher ran 43 iterations but used the results from the 37th iteration as it yielded the best level of reliability. The researcher found 42 items with poor fit (i.e., items 2, 5, 12-13, 21, 25, 26-34, 37, 45, 50, 53, 56-57, 59-61, 63, 66-68, 71, 73-74, 78, 80, and 90-99),which were deleted. Table 13 below provides information regarding the person and item separation and reliability.

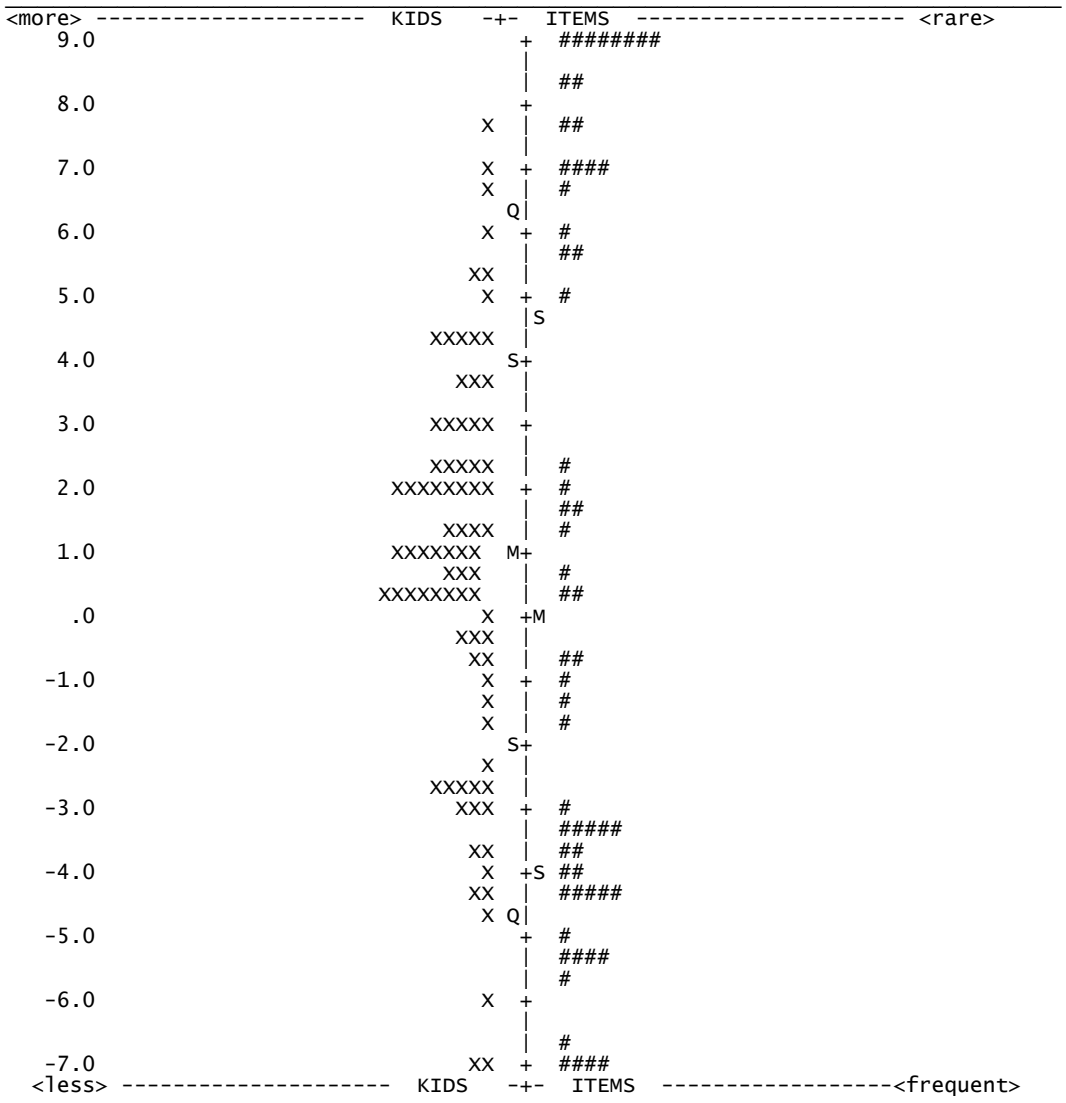
Table 13
Person and Item Reliability and Separation—Verbal Subtest

Statistics	Value
Reliability of person separation	.94
Reliability of item separation	.98
Item Separation (Real RMSE Separation)	3.81
Person Separation (Real RMSE Separation)	7.44

The achieved level of reliability for both person separation and item separation was high. The items were able to separate children into almost four groups (with a Real RMSE Separation of 3.81) while the children were able to separate items into more than seven difficulty levels (with a Real RMSE Separation of 7.44).

Figure 16 below shows a map of item/child position for the Verbal Subtest. The map indicates that, on average, children were a little more able than items were difficult, with a child logit mean position of 1.22. In other words, children on average found items to be a little easy.

Figure 16
Item/Child Position Map for Verbal Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Visual. The researcher ran 24 iterations and found 23 items with poor fit (i.e., items 2, 46, 49, 51-53, 55-57, 65-66, 70, 74, 79, 80, 84-85, 91-93, and 95-97), which were deleted. Table 14 below provides information regarding the person and item separation and reliability.

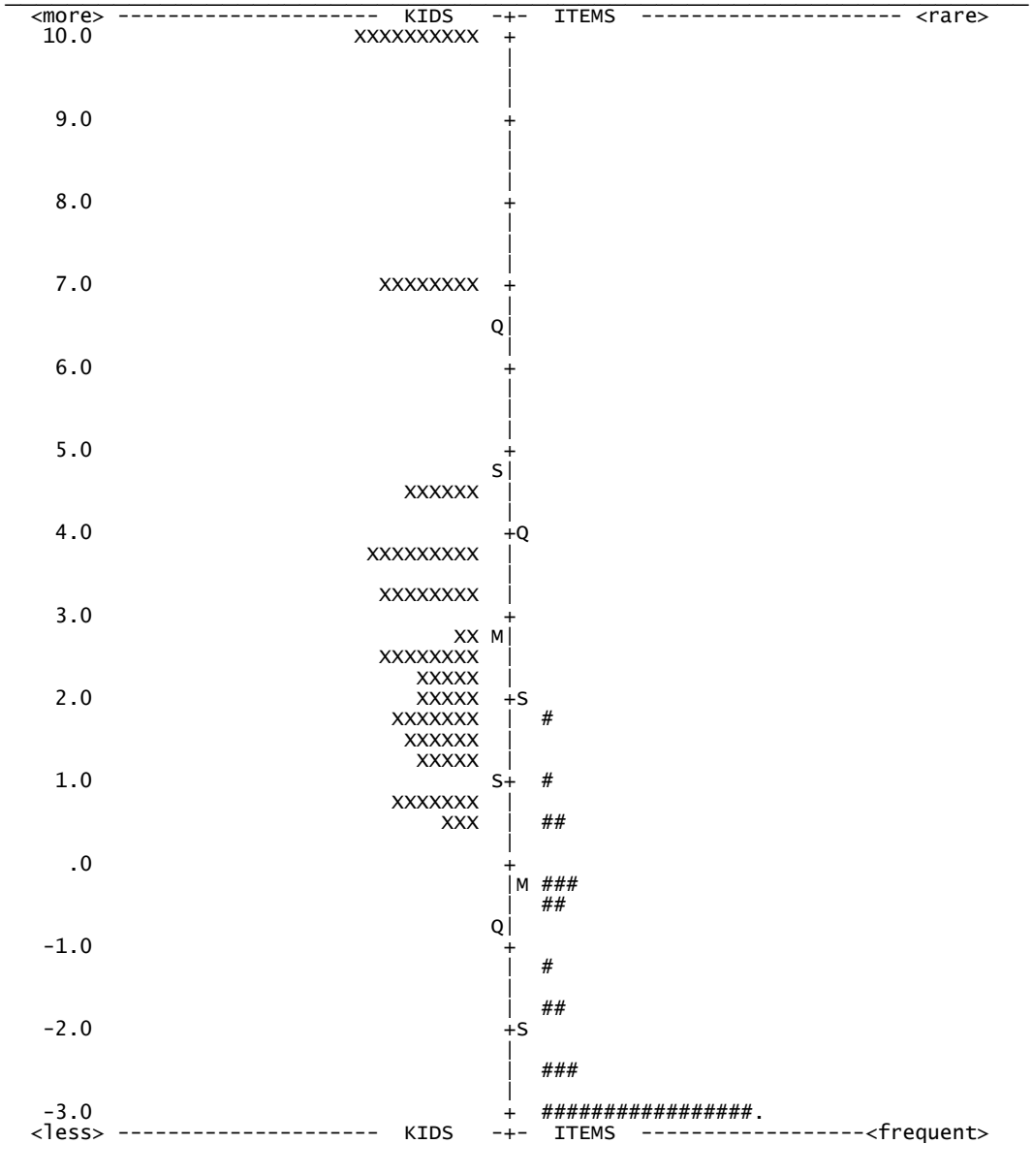
Table 14
Person and Item Reliability and Separation—Visual Subtest

Statistics	Value
Reliability of person separation	.72
Reliability of item separation	.92
Item Separation (Real RMSE Separation)	1.59
Person Separation (Real RMSE Separation)	3.36

The achieved level of reliability for both person separation and item separation was high. The items were not able to separate children into groups (with a Real RMSE Separation of 1.59) while the children were able to separate items into more than three difficulty levels (with a Real RMSE Separation of 3.36).

Figure 17 below shows a map of item/child position for the Visual Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 2.85. In other words, children on average found the items to be very easy.

Figure 17
Item/Child Position Map for Visual Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Memory. The researcher ran 15 iterations and found 10 items with poor fit (i.e., items 3, 5, 6, 11, 15, 16, 21, 22, 26, and 27), which were deleted. Table 15 below provides information regarding the person and item separation and reliability.

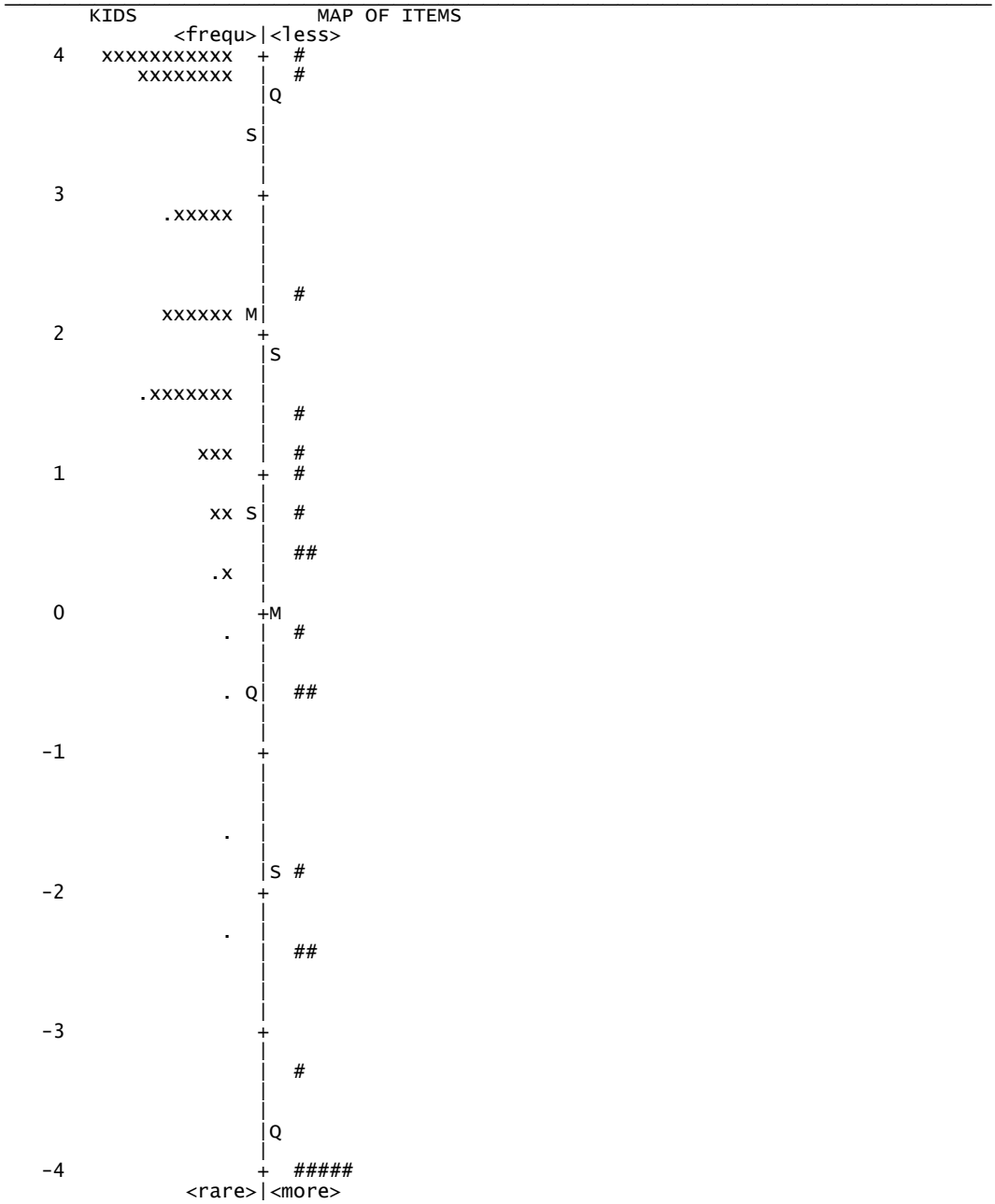
Table 15
Person and Item Reliability and Separation—Memory Subtest

Statistics	Value
Reliability of person separation	.51
Reliability of item separation	.92
Item Separation (Real RMSE Separation)	1.02
Person Separation (Real RMSE Separation)	3.43

The achieved level of reliability for person separation was low and it was high for item separation. The items were not able to separate children into groups (with a Real RMSE Separation of 1.02) while the children were able to separate items into more than three difficulty levels (with a Real RMSE Separation of 3.43).

Figure 18 below shows a map of item/child position for the Memory Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 2.09. In other words, children on average found the items to be very easy.

Figure 18
Item/Child Position Map for Memory Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Math. The researcher ran 27 iterations but used the results from the 36th iteration as it yielded the best level of reliability. The researcher found 24 items with poor fit (i.e., items 2-4, 6, 16, 22, 26, 30, 32-34, 43, 46, 48-49, 50, 52-54, 58, 60-62, and 65),which were deleted. Table 16 below provides information regarding the person and item separation and reliability.

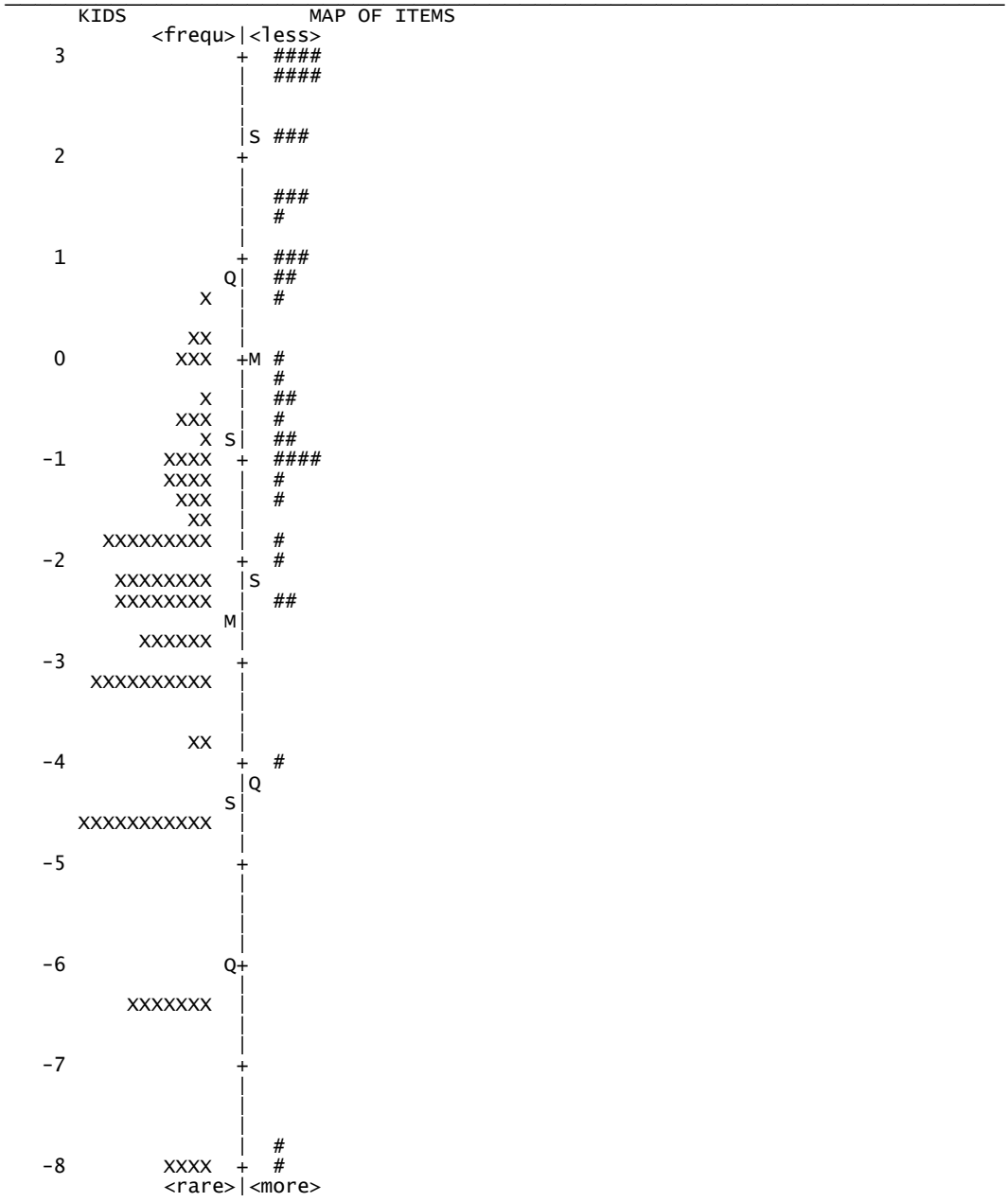
Table 16
Person and Item Reliability and Separation—Math Subtest

Statistics	Value
Reliability of person separation	.76
Reliability of item separation	.93
Item Separation (Real RMSE Separation)	1.77
Person Separation (Real RMSE Separation)	3.75

The achieved level of reliability for both person separation and item separation was high. The items were almost able to separate children into two groups (with a Real RMSE Separation of 1.77) while the children were able to separate items into more than three difficulty levels (with a Real RMSE Separation of 3.75).

Figure 19 below shows a map of item/child position for the Math Subtest. The map indicates that, on average, children were much less able than items were difficult, with a child logit mean position of -2.60. In other words, children on average found the items to be very difficult.

Figure 19
Item/Child Position Map for Math Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Logical. The researcher ran 15 iterations and found 13 items with poor fit (i.e., items 16, 18, 19, 27, 29-32, 44-45, 48, 52, and 56), which were deleted. Table 17 below provides information regarding the person and item separation and reliability.

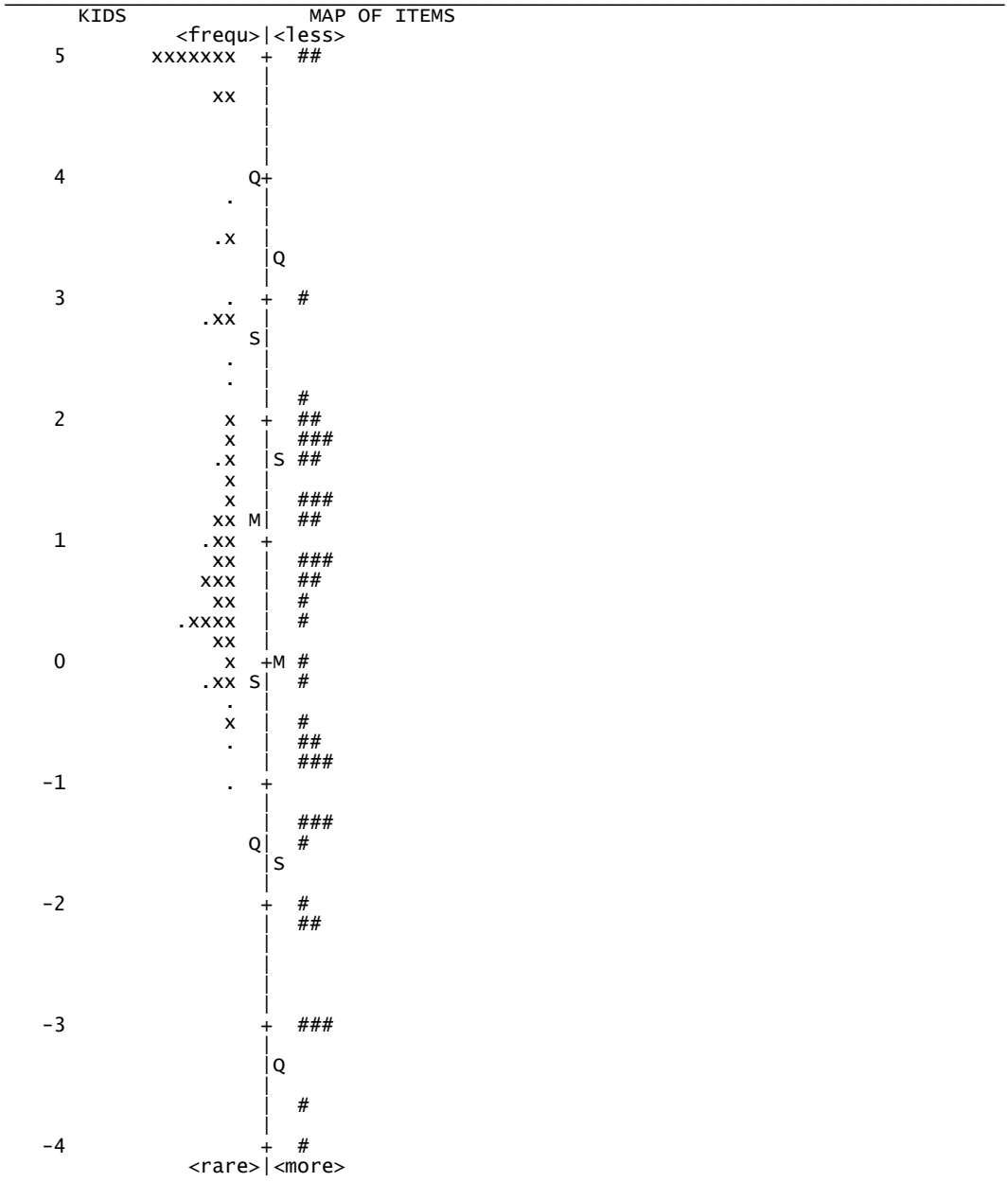
Table 17
Person and Item Reliability and Separation—Logical Subtest

Statistics	Value
Reliability of person separation	.86
Reliability of item separation	.94
Item Separation (Real RMSE Separation)	2.50
Person Separation (Real RMSE Separation)	4.05

The achieved level of reliability for both person separation and item separation was high. The items were able to separate children into more than two groups (with a Real RMSE Separation of 2.50) while the children were able to separate items into four difficulty levels (with a Real RMSE Separation of 4.05).

Figure 20 below shows a map of item/child position for the Logical Subtest. The map indicates that, on average, children were more able than items were difficult, with a child logit mean position of 1.23. In other words, children on average found the items to be easy.

Figure 20
Item/Child Position Map for Logical Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

General Knowledge. The researcher ran one iteration and found one item with poor fit (i.e., item 1), which was deleted. Table 18 below provides information regarding the person and item separation and reliability.

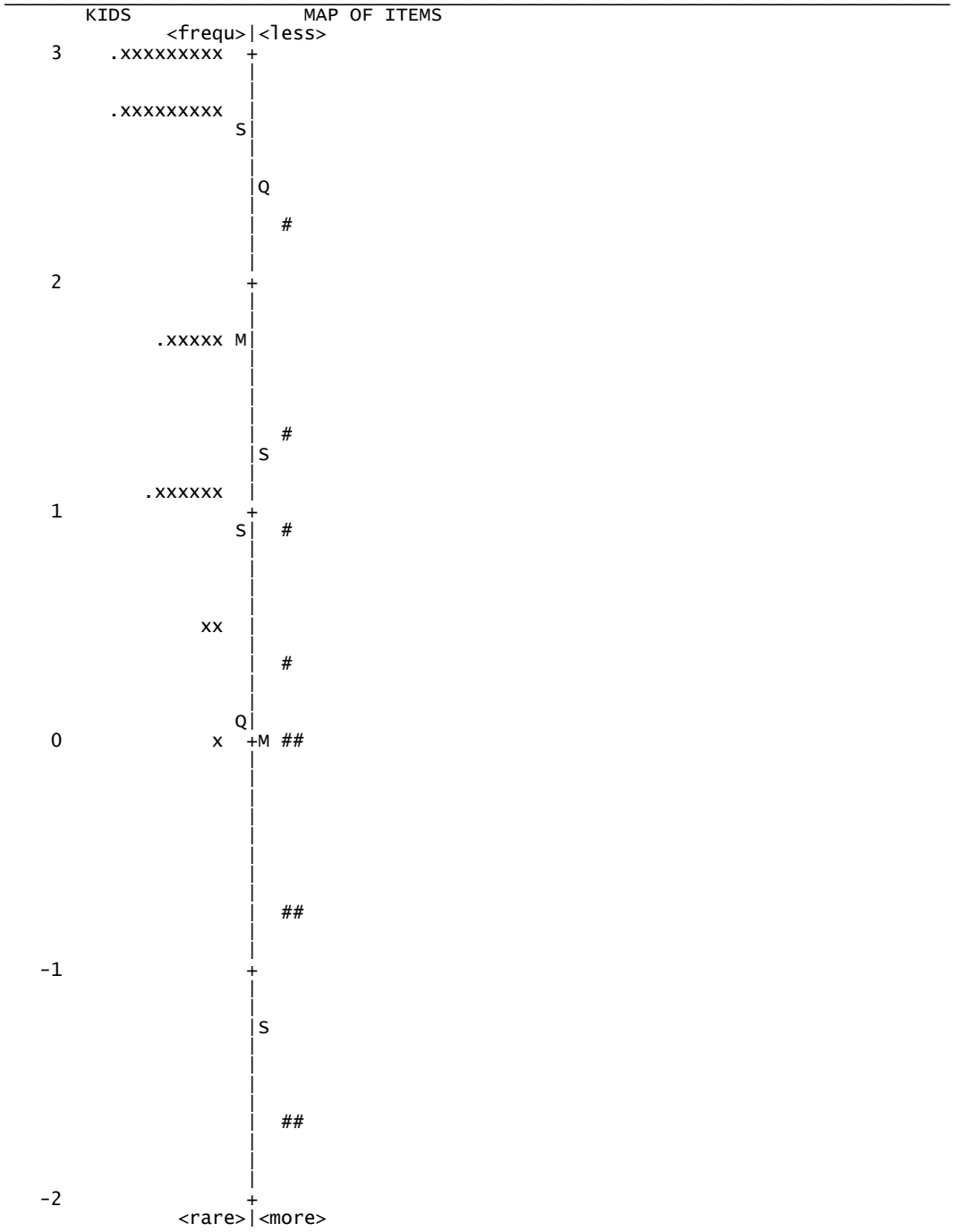
Table 18
Person and Item Reliability and Separation—General Knowledge Subtest

Statistics	Value
Reliability of person separation	.00
Reliability of item separation	.89
Item Separation (Real RMSE Separation)	0.00
Person Separation (Real RMSE Separation)	2.86

The achieved level of reliability for person separation was low and it was high for item separation. The items were not able to separate children into groups (with a Real RMSE Separation of 0.00) while the children were able to separate items into four difficulty levels (with a Real RMSE Separation of 2.86).

Figure 21 below shows a map of item/child position for the General Knowledge Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 1.79. In other words, children on average found the items to be easy.

Figure 21
Item/Child Position Map for General Knowledge Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Discussion.

The item analysis results for every subtest, except for Math Subtest, indicate that the second pilot items were still easy for the children. Therefore, items with higher difficulty levels were created for the main study subtests. Additionally, items whose difficulty levels were considered to fill the gaps in the difficulty continuum were created. The creation of items was based on the literature review and expert reviewer comments. As discussed in the Operational Definitions of Academic Readiness Domains section, the researcher was able to create, in most cases, double or triple the number of items for the main study as that of the second pilot study.

Main Study.

Purpose.

The purpose of the main study was to finalize the set of items that represent the six academic readiness domains as described earlier.

Development.

The developmental process of the main study measure is described in the following paragraphs.

Planning.

The development process of the main study measure began with the statement of purpose. The purpose of the main study measure was very much similar to

that of the second pilot study. That is, the purpose of the instrument was to measure the *academic* readiness of the kindergarten graduates for their first grade instruction.

After the determination of the purpose, the identification of domains takes place. Based on the reviews of additional resource books (Nontapuk, 2009; Pangwiruitrak, 2007; Trium Sop Por 1, 2009; Wai Prip Trium Sop, 2009) and first grade curriculum in Thailand (Helair, 1996), six domains were confirmed. They were verbal (Clymer & Barrett, 1966; Danzer, Gerber, Lyons, & Voress, 1967; Mardell & Goldenberg, 1983; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), math (Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), memory (Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), visual (Clymer & Barrett, 1966; Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), logical (Danzer, Gerber, Lyons, & Voress, 1967; Newborg, Stock, Wnek, Guidubaldi, & Swinicki, 1984), and general knowledge (Danzer, Gerber, Lyons, & Voress, 1967).

After the domains were confirmed, additional review of the literature was conducted. The literature review did not uncover additional components of school readiness for the U.S. and for Thailand. The literature review still confirmed academic readiness as a common component of school readiness for the U.S. and Thailand. The literature review also confirmed the domains that can best measure academic readiness. The domains were discussed in the previous paragraph. The literature review did not uncover any tests that serve the same purpose as that of this measure. This serves as

added evidence to support the development of a measure of the academic readiness for first grade in Thailand.

Construction.

After the domains are confirmed and operationalized, the generation of additional items takes place. The process involves generation of item content and selection of item formats. The item content was derived from the findings of the second pilot study, the suggestions of the content experts, and the literature review. The items contained in the main study measure include new items adapted from several resources such as resource books First Grade Exams (Nontapuk, 2009; Pangwiruitrak, 2007; Trium Sop Por 1, 2009; Wai Prip Trium Sop, 2009) and first grade textbooks from Thailand (Helair, 1996) (see an example of main study items in the discussion of Operational Definitions). Operational definitions of the main study domains are provided earlier in the content validity section.

As previously mentioned, the main study items were given to a group of content experts for suggestions. The revised measure was used for the main study administration.

Quantitative Evaluation.

The main study administration took place in four separate time periods: May, 2007 through July, 2007, November, 2007 through March, 2008, May, 2008 to July 2008, and November, 2008 through February, 2009. University of Denver Institutional Research Board approval as well as approval of the school principals was granted for the study. The test administration involved three separate sessions, during each of which two

subtests were administered and each lasted approximately two hours. The administration followed the same sequence (i.e., verbal, logical, visual, math, memory, and general knowledge). Some students were allowed to follow a different sequence if they so stated their preference. There were three sections of verbal, five sections of visual, three sections of memory, three sections of math, three sections of logical, and one section of general knowledge subtest. There were 595 items in the whole battery (See Appendix 13 for the number of items for each subtest in the main study). Unless expressing his or her wish otherwise, every student was asked to complete every item. The scores of the students who asked to discontinue the administration were not included in the analysis.

Sample. The sample was drawn from the population of graduating kindergarteners and beginning first grade students at two private schools in West Bangkok, Thailand. The sample came from middle- to upper-class families. All students were Thai nationals. The sample consisted of 237 boys and 198 girls.

Setting. The administration was carried out in an empty classroom, which was located far from noisy areas. Four small kindergarten desks were used and placed together during the administration to have room to place the items and the answer sheets. There were pencils, erasers, and blank pieces of paper available for student use. The student was seated in a kindergarten chair across from the proctor. The room was air-conditioned with the temperature at 24 Celsius degree. All the lights were turned on to ensure sufficient lighting.

Training. The researcher recruited four proctors, who were classroom teachers or teacher's aids. These individuals have a minimum of a bachelor degree with a

minimum classroom experience of one year. The researcher provided an orientation for each proctor to familiarize proctors with the purpose of the study, types of subtests, and types of items. The researcher performed a role play with each proctor by having the proctor be the student and the researcher be the proctor. The researcher (i.e., the proctor) then had each proctor (i.e., the student) take each subtest. The researcher then asked that the proctor allowed enough time to establish rapport with the student and to allow the student to become accustomed to the proctor and the testing environment (i.e., the test room) before the test began. The researcher explained to the proctor that the researcher asked the parent(s) to talk to the student before the test date to give the student enough advance notice and time to mentally prepare for the test. The researcher asked the parents to tell the student that the proctor would show, for example, some pictures and asked the student to tell what the picture was. The researcher asked the proctor to be friendly and use a soft voice as opposed to a loud and authoritative tone to avoid tension during the testing. The researcher discussed with the proctor that the proctor could move to the next item when the student was able to provide a response or after two attempts without a response. The researcher showed the proctor how to properly record the responses by writing down the responses in the appropriate spaces in the proctor's answer sheet and to make sure the student wrote down the responses properly by writing down the responses in the appropriate spaces in the student's answer sheet. If there were subtests with a proctor's answer sheet, the proctors were asked to record the responses on the answer sheet. If there were subtests with a student's answer sheet, the student was asked to respond by writing down the answer on the student's answer sheet. The proctors

were asked not to translate responses into scores in the proctor's or the student's answer sheet. The researcher discussed with the proctors to look for signs when a student became uncomfortable continuing taking the test. The proctor was trained to ask the student, if there was such a sign, if the student wanted to continue taking the test. If the student said yes, the proctor continued the test. If the student said no, the proctor asked if the student wanted a five-minute break. If the student said yes, the proctor gave a five minute break and then continued the test. If the student said he or she did not want to continue the test, the proctor asked if the student was sure. If the student said yes, the proctor said ok and gave praise. The proctor then escorted the student back to the classroom and reported to the researcher. The researcher then contacted the parents and asked the parents to talk to the student if he or she wanted to come back to the test. If the student said yes, the researcher rescheduled the test. If the student said no, the researcher removed the student from the participant list.

The researcher then switched roles. The researcher played a student and the proctor played a proctor. The researcher then observed and provided specific recommendations with regard to different situations during the role-playing of the test administration. After the orientation, if the proctors showed that they could use the items, follow directions, and record responses correctly, the researcher would allow the proctors to perform a few administrations with participants within the researcher's presence. The researcher attended these first few administration sessions for each proctor. During such sessions, the researcher helped the proctors in situations where the proctor performed the procedure incorrectly. The researcher then had a brief meeting

with each proctor, talking about what the proctor did well and what needed improvement. Afterwards, the researcher attended another administration session to make sure if all recommendations were followed. And if the proctors performed all administration procedures correctly, the researcher allowed the proctors to continue with the rest of the participants. The researcher periodically attended additional administration sessions with each proctor to make sure of consistent test administration practices.

Procedure. The researcher made an appointment with the classroom teacher for the date and time of administration for each student. The proctor was informed of the name, date, and time of the administration for each student one day prior to the test date. The proctor went to the classroom to escort the student to the test room on the test date. The proctor asked the student to take a seat. The proctor introduced herself in case the student did not know the proctor beforehand. The proctor then asked if the student knew why he or she was there. If the student did not know, the proctor explained the reason and asked if the student wanted to take the test. If the student said yes, the proctor started the test. If the student said no, the proctor then asked if the student was sure. If the student still said yes, the proctor took the student back to his or her classroom. The proctor then reported this event to the researcher. The researcher then contacted the parent(s) if the student understood he or she would take the test and if he or she agreed. If the parent(s) confirmed that the student understood and had agreed, the parent(s) asked the student if he or she still wanted to take the test. If the answer was yes, the researcher rescheduled the test for the student. If the answer was no, the researcher removed the student from the participant list.

The administration followed the same sequence (i.e., verbal, logical, visual, math, memory, and general knowledge). Some students were allowed to follow a different sequence if they so stated their preference. There were three sections of verbal, five sections of visual, three sections of memory, three sections of math, three sections of logical, and one section of general knowledge subtest. There were 595 items in the whole battery (See Appendix 13 for the number of items for each subtest in the main study). Unless expressing his or her wish otherwise, every student was asked to complete every item. The scores of the students who asked to discontinue the administration were not included in the analysis.

Verbal Subtest. The first session started with verbal subtest. The reading vocabulary section was given to the participant first. The proctor showed the first vocabulary card (i.e., RV-01) which contains a vocabulary word to the student and asked the student to pronounce the vocabulary word aloud. After the student pronounced the word, the proctor recorded the response and then showed the next vocabulary card (i.e., RV-02). The proctor recorded the pronunciation phonetically into the proctor's answer sheet. If, after 30 seconds of silence, the student did not pronounce the word, the proctor would ask the student if he or she could pronounce the word. If the student said he or she did not know how to pronounce the word, the proctor would record "Did Not Respond" and move on to the next word. If the student said he or she knew how to pronounce the word, the proctor would ask the student to pronounce the word. If the student still did not pronounce the word after the second trial, the proctor would record "Did Not Respond" and move on to the next word. This process repeated until the proctor had showed all of

the vocabulary cards. The process of response recording, of first response timing, and of a second opportunity to try an item was similar for the other subtests.

After the end of the reading vocabulary section, the proctor continued with writing dictation section. The proctor gave a blank answer sheet with a list of item number and a blank space next to it to the student. The proctor used a sheet containing a list of writing dictation items to give writing queue to the student. The proctor read the first vocabulary word (i.e., WD-01) on the list and asked the student to write down the answer on the space for item number one (i.e., WD-01). After the student had written the vocabulary word, the proctor then read the second vocabulary word (i.e., WD-02) to the student and asked the student to write down the answer on the space for item number two. This process repeated until the proctor had read all of the vocabulary cards.

After the end of the writing dictation section, the proctor gave a sheet containing the first short story and a list of multiple choice questions. The proctor asked the student to read the story. After finishing reading the story, the proctor asked the student to read the first question (i.e., RC-01) and select an answer from the choices given. The proctor then recorded the response into the proctor's answer sheet. This process was repeated until the proctor had asked student to read all the stories and respond to all questions.

Logical Subtest. After the proctor finished the verbal subtest, the proctor gave the student a five-minute break. Then the proctor asked the student to return to his or her seat. The proctor started the logical subtest with the concept formation section. The proctor showed the first item (i.e., CF-01) to the student and explained to the student that there is a relationship between the pictures in the top left and the top right square.

There is a relationship between the pictures in the top left and the bottom left square. There is also a relationship between the picture in the top right and the missing picture in the bottom right square. There is no relationship between either the pictures in the top right and the bottom left or the ones in the top left and the bottom right square. The proctor gave a set of pictures, one of which correctly conforms to the relationship between itself and other pictures in the squares. After the proctor showed the first item, the proctor asked the student to choose an answer. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had showed the student all concept formation items and recorded the response for all items.

After the concept formation section, the proctor continued with the sequential order section. The proctor showed the first sequential order item (i.e., SO-01) and explained that there is a series of pictures. Located in the middle or at the end of the series, the blank space(s) is/are for the missing picture(s). The proctor then gave a set of picture answers and told the student that one of the given pictures correctly give the correct meaning to the relationship of the pictures in the order. Then the proctor asked the student to select an answer. The proctor then recorded the response on the proctor's answer sheet. This process was repeated until the proctor had showed the student all sequential ordering items and recorded the response for all items.

After the sequential ordering section, the proctor continued with the pattern finding section. The proctor showed the first pattern finding item (i.e., PF-01) and explained that there is a series of pictures and blank spaces located in the middle or at the end of the series. The proctor gave a set of picture answers and told the student that one

of the given pictures correctly give the correct meaning to the relationship of the pictures in the order. Then the proctor asked the student to select an answer. The proctor then recorded the response on the proctor's answer sheet. This process was repeated until the proctor had showed the student all pattern finding items and recorded the response for all items. At the end of pattern finding, the first session of administration ended. The proctor praised the student. The proctor escorted the student back to the classroom. The proctor returned the answer sheets to the researcher.

Visual Subtest. The second session of the main study administration began with the visual subtest. The proctor started with the visual matching section. The proctor showed the first visual matching item (i.e., VM-01), which is a picture card, to the student and also showed a set of answer cards. The proctor asked the student to look at the picture card and choose a matching picture from one of the answer cards. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had showed the student all visual matching items and recorded the response for all items.

After the visual matching section, the proctor continued with the visual identification section. The proctor showed the first visual identification item (i.e., VI-01), which is a picture card, to the student. The proctor asked the student to look at the picture card and ask which of the following answers best described the picture. The proctor then read answer choices. The proctor asked the student to choose from one of the answer choice and then recorded the response on the proctor's answer sheet. This

process repeated until the proctor had showed the student all visual identification items and recorded the response for all items.

After the visual identification section, the proctor continued with the visual recognition section. The proctor showed the first visual recognition item (i.e., VR-01), which is a picture card, to the student. The proctor then explained that the picture was cut into pieces and rearranged and the proctor wanted the student to name the object shown in the picture. The proctor then read possible answers to student and asked the student to choose an answer. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had showed the student all visual recognition items and recorded the response for all items.

After the visual recognition section, the proctor continued with mental folding section. The proctor showed the first mental folding item (i.e., MF-01), which is a picture card, and also showed a set of answer cards to the student. The proctor explained to the student that one of the answer cards contained the same picture shown as the question except the picture is folded by the dotted line. The proctor then asked the student to choose an answer. The proctor then recorded the response into the proctor's answer sheet. This process repeated until the proctor had showed the student all mental folding items and recorded the response for all items.

After the mental folding section, the proctor continued with mental rotation. The proctor showed the first mental rotation item (i.e., MR-01), which is a picture card, and also showed a set of answer cards to the student. The proctor explained to the student that one of the answer cards contained the same picture shown as the question except the

picture is rotated. The proctor then asked the student to choose an answer. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had showed the student all mental rotation items and recorded the response for all items.

Math Subtest. After the mental rotation, the proctor gave the student a five-minute break. The proctor then continued with the math subtest. The proctor started with the math concepts and vocabulary section. The proctor explained that the student would listen to the proctor reading each math concept and vocabulary item and that the student would write it down in a mathematical format. The proctor asked the student to listen to the proctor carefully. The proctor then read the first math concept and vocabulary item (i.e., MCV-01) to the student and asked that the student write down what he or she heard on the student's answer sheet. This process repeated until the proctor had read to the student all math concept and vocabulary items and until the proctor had all responses recorded.

After the math concept and vocabulary section, the proctor continued with the arithmetic (or math computation) section. The proctor then provided the student with a sheet of arithmetic questions and asked that the student find an answer for each question. The proctor asked the student to start from the first item (i.e., MC-01) and then continue on to the second item (i.e., MC-02). The student was asked to put the answer on the student's answer sheet. This process repeated until the student had gone through all arithmetic items and until the proctor had all responses recorded.

After the arithmetic section, the proctor continued with word problems. The proctor provided a sheet of word problem questions to the students and asked the student to find an answer for each question and put the answer on the student's answer sheet. The proctor asked the student to start from the first item (i.e., WP-01) and then continue on to the second item (i.e., WP-02). This process repeated until the student had gone through all word problem items and until the proctor had all responses recorded. At the end of word problem section, the second session of the administration ended. The proctor praised the student. The proctor escorted the student back to the classroom. The proctor returned the answer sheets to the researcher.

Memory Subtest. The third session of the main study administration began with the memory subtest. The proctor started with the immediate recognition section. The proctor showed the first immediate recognition item (i.e., IR-01), which is a picture card, to the student for five seconds. After the time limit, the proctor put away the picture card and showed a set of answer cards, one of which contained the same picture shown earlier. The proctor asked the student to choose a card with the right picture. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had showed the student all immediate recognition items and recorded the response for all items.

After the immediate recognition section, the proctor continued with the spatial memory section. The proctor showed the first spatial memory item (i.e., SR-01), which is a card containing a series of pictures, to the student for five seconds. After the time limit, the proctor put away the picture card and showed the student a set of answer cards.

The proctor asked the student to arrange the cards in the right order to match the picture shown earlier. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had shown the student all spatial memory items and recorded the response for all items.

After the spatial memory section, the proctor continued with the delayed recognition section. The proctor began with the first delayed recognition item (i.e., DR-01), which consists only of the answer cards. The proctor then asked the student to recall the picture given in the first immediate recognition question and asked the student to form an answer from the answer cards. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had shown the student all delayed recognition items and recorded the response for all items.

After the delayed recognition section, the proctor continued with the general knowledge subtest. The proctor read the first general knowledge question (i.e., GK-01) to the student and read the answers. The proctor then asked the student to choose the correct answer. The proctor then recorded the response on the proctor's answer sheet. This process repeated until the proctor had shown the student all general knowledge items and recorded the response for all items. At the end of general knowledge subtest, the main study administration ended. The proctor escorted the student back to the classroom. The proctor returned the answer sheets to the researcher. And the researcher then translated the responses into raw scores of "1" being a correct response and "0" being an incorrect response.

The test administration involved three separate sessions, during each of which two subtests were administered and each lasted approximately two hours. The administration followed the same sequence (i.e., verbal, logical, visual, math, memory, and general knowledge). Some students were allowed to follow a different sequence if they so stated their preference. There were three sections of verbal, five sections of visual, three sections of memory, three sections of math, three sections of logical, and one section of general knowledge subtest. There were 595 items in the whole battery (See Appendix 13 for the number of items for each subtest in the main study). Unless expressing his or her wish otherwise, every student was asked to complete every item. The scores of the students who asked to discontinue the administration were not included in the analysis.

CHAPTER 4

Results

This chapter presents the results of the data analyses of the main dissertation study. Results of the data analyses are presented by subtest in the following order: Visual, Verbal, Memory, Math, Logical, and General Knowledge Subtest. First, the data analyses for each subtest involved the initial Rasch analyses to determine a workable but reduced set of items. Second, analyses of dimensionality and fit were performed to determine whether the original subtests were measuring one dimension and whether items in the subtests showed appropriate fit. Third, exploratory factor analyses were performed to identify if each subtest was in fact measuring more than one dimension and to identify items that reflected the first dimension more strongly, if multiple dimensions were found. Fourth, all items loading on factors other than the most dominant one were removed to form shortened subtests. Fifth, reliability analyses were performed to determine the level of reliability of the subtests and to determine the level of item fit to the subtests. Sixth, item invariance analyses were performed to determine if gender played a role in the responses of subjects with different genders while answering items. If an item showed a significant difference in the item responses of subjects with different genders, the item was removed to form the final subtest. Finally, descriptive statistics and correlation between subtests and between subtest and genders are presented.

Verbal Subtest

The researcher first conducted 11 iterations of Rasch analyses to determine a workable set of items for the Verbal Subtest. Initially, there were 436 children and 140 items. After 11 iterations, the researcher found 70 items with poor fit. Therefore, these items were deleted from the original Verbal Subtest.

Table 19
Misfitting Items in Verbal Subtest

Misfitting Items by Item Label
RV-01, RV-04, RV-06, RV-07, RV-08, RV-09, RV-10, RV-13, RV-15, RV-16, RV-17, RV-18, RV-20, RV-21, RV-22, RV-24, RV-25, RV-26, RV-27, RV-28, RV-30, RV-32, RV-33, RV-34, RV-35, RV-37, RV-43, RV-45, RV-50, RV-60, RV-64, RV-67, WD-02, WD-03, WD-04, WD-06, WD-09, WD-11, WD-12, WD-13, WD-15, WD-16, WD-18, WD-21, WD-23, WD-24, WD-26, WD-29, WD-30, WD-34, RC-01, RC-02, RC-03, RC-04, RC-05, RC-06, RC-07, RC-08, RC-09, RC-10, RC-11, RC-12, RC-13, RC-14, RC-15, RC-16, RC-17, RC-18, RC-19, and RC-20

Note. Items were dropped if their outfit MNSQ was higher than 1.30.

Dimensionality.

Principal Components Analysis of Residuals. The researcher conducted a Rasch principal components analysis of residuals on the eleventh iteration. Table 20 summarizes results from the principal components analysis of residuals.

Table 20
Table of Standardized Residual Variance in Percent—Verbal Subtest

	Empirical	Modeled
Total raw variance in observations	100.0%	100.0%
Raw variance explained by measures	54.3%	54.5%
Raw variance explained by persons	26.3%	26.4%
Raw variance explained by items	28.0%	28.1%
Raw unexplained variance (total)	45.7%	45.5%
Unexplained variance in 1st contrast	2.7%	5.8%

It shows that the Rasch first dimension explained 54.3% of item variance in the data. This dimensionality finding indicates that Verbal Subtest may consist of more than one dimension because the Rasch first dimension can explain 54.3%, which does not meet the recommended 60% (Linacre, 2007). Since evidence was found of multiple dimensions underlying responses to test items, an exploratory factor analysis was conducted to clarify the dimensional structure.

Exploratory Factor Analysis.

Since the initial analysis of dimensionality indicated the presence of multiple factors underlying Verbal Subtest, the researcher ran an SPSS exploratory factor analysis (with *all* items) and found more than one interpretable factor for the Verbal Subtest. Table 21 below lists the items loading on each factor in a factor analysis with two factors specified and items not loading on either factor.

Table 21
List of Items Loading on Factors and Items Failing to Load on the First Two Factors—Verbal Subtest

Items Loading on Factor One
RV-05, RV-07, RV-08, RV-19, RV-39, RV-47, RV-49, RV-51, RV-52, RV-53, RV-54, RV-55, RV-57, RV-58, RV-61, RV-62, RV-65, RV-66, RV-67, RV-68, RV-69, RV-71, RV-72, RV-73, RV-74, WD-02, WD-03, WD-04, WD-05, WD-06, WD-07, WD-08, WD-10, WD-12, WD-13, WD-14, WD-15, WD-16, WD-17, WD-18, WD-19, WD-20, WD-22, WD-23, WD-24, WD-25, WD-26, WD-27, WD-28, WD-29, WD-31, WD-32, WD-33, WD-35, WD-36, WD-38, WD-40, RC-02, RC-04, RC-09 and RC-12.
Items Loading on Factor Two
RV-01, RV-02, RV-03, RV-04, RV-06, RV-09, RV-10, RV-11, RV-12, RV-13, RV-14, RV-15, RV-16, RV-17, RV-20, RV-21, RV-22, RV-23, RV-24, RV-25, RV-26, RV-27, RV-28, RV-29, RV-30, RV-31, RV-32, RV-33, RV-34, RV-35, RV-36, RV-37, RV-38, RV-40, RV-41, RV-42, RV-43, RV-44, RV-45, RV-46, RV-48, RV-50, RV-56, RV-59, RV-60, RV-63, RV-64, RV-70, RV-75, RV-76, RV-77, RV-78, RV-79, RV-80, RC-08, and RC-19.
Items Loading on Neither Factor
RV-18, WD-01, WD-09, WD-11, WD-21, WD-30, WD-34, WD-37, WD-39, RC-01, RC-03, RC-05, RC-06, RC-07, RC-10, RC-11, RC-13, RC-14, RC-15, RC-16, RC-17, RC-18, and RC-20

Based on the exploratory factor analysis findings, the original Verbal Subtest was measuring more than one domain. Because the researcher's original intention was to have one subtest for each domain, the researcher decided to focus on the items loading on

the first factor and delete the items loading on other factors, which may have been possibly measuring something other than verbal skills, or a related facet of verbal skills.

Once the researcher had decided to create a shortened Verbal Subtest based on items loading on the first factor, the researcher ran Rasch analyses to investigate the level of reliability on the shortened Verbal Subtest. The researcher ran 4 iterations and found 19 items with poor fit (i.e., outfit mean square >1.30: RV-05, RV-07, RV-08, RV-62, RV-67, RV-68, WD-02, WD-03, WD-12, WD-13, WD-15, WD-16, WD-18, WD-24, WD-29, RC-02, RC-04, RC-09, and RC-12), which were deleted from the analyses.

Reliability.

For the shortened Verbal Subtest, the reliability of person separation was 0.93 while the reliability of item separation was 0.99. Table 22 provides information for person and item separation and reliability.

Table 22
Person and Item Reliability and Separation—Verbal Subtest

Statistics	Value
Reliability of person separation	.93
Reliability of item separation	.99
Item Separation (Real RMSE Separation)	3.72
Person Separation (Real RMSE Separation)	12.78

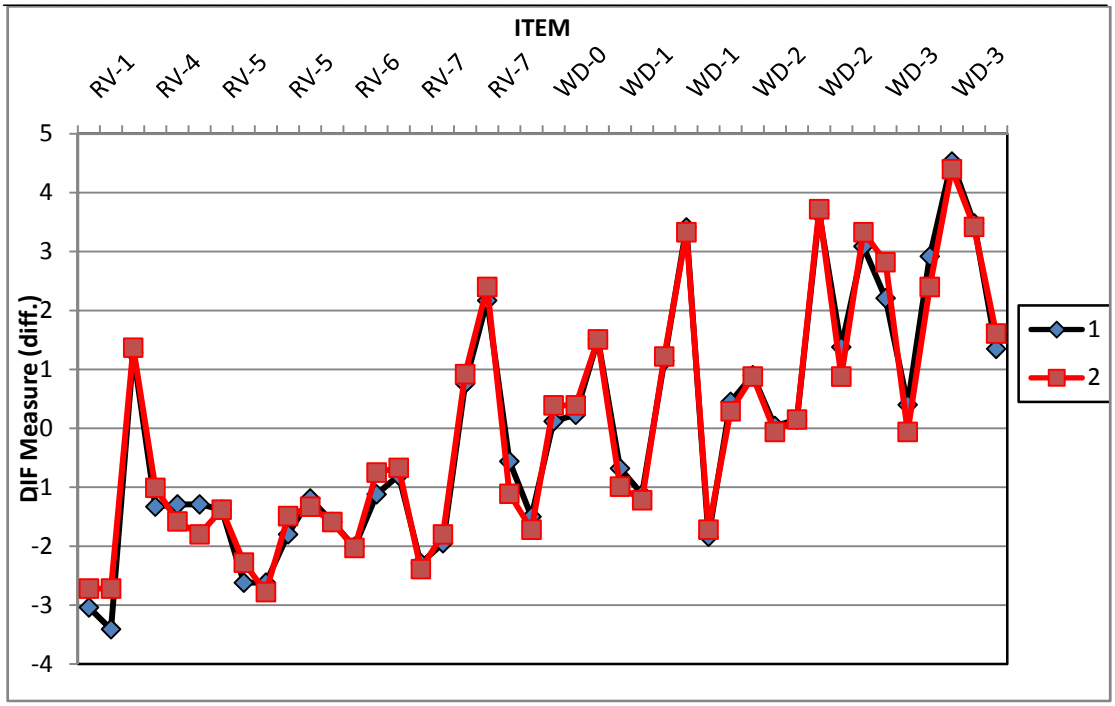
The achieved level of reliability for both person and item separation was high. The items in the shortened Verbal Subtest were able to separate children into almost four distinct ability groups (with a Real RMSE Separation of 3.72) while the children who took the items in the shortened Verbal Subtest were able to separate items into almost thirteen difficulty levels (with a Real RMSE Separation of 12.78). The average outfit

MNSQ was 0.92. The average outfit MNSQ shows that the items in the shortened Verbal Subtest fit the model well.

Item Invariance.

The researcher conducted an analysis of item invariance to determine the effect of gender on children's item responses on the shortened Verbal Subtest. Figure 22 below provides a plot showing item invariance (lack of DIF) of the shortened Verbal Subtest items.

Figure 22
DIF Plot Showing Item Invariance—Verbal Subtest



Based on the figure, it can be inferred that children with different genders did not differ greatly in pattern of responses to items. This is evident in the two lines (blue (1) for boy and red (2) for girl) falling close to each other. However, item WD-04 evidenced

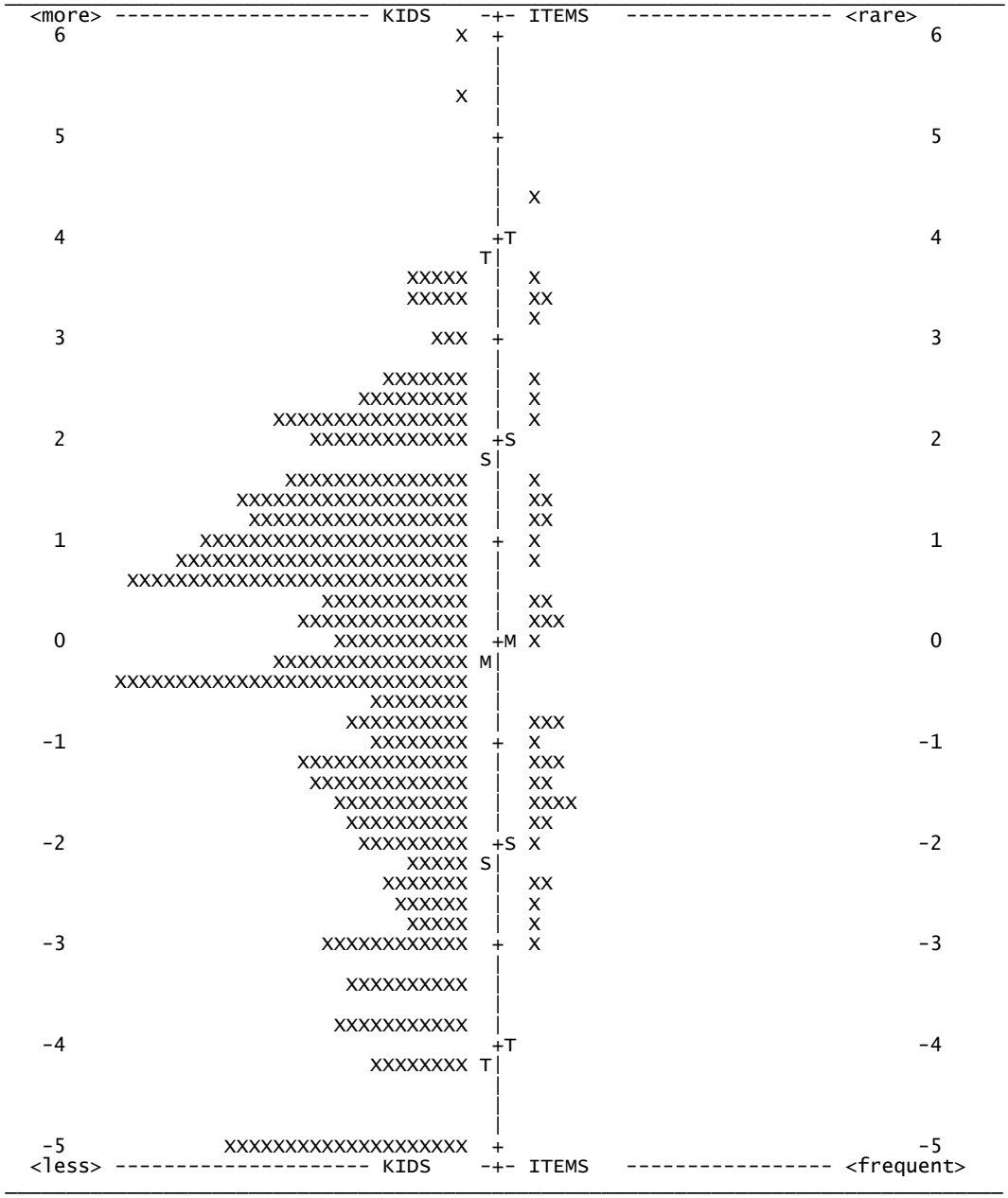
statistically significant DIF ($p<.05$) and hence became a candidate for deletion to arrive at the final scale (See Table 23 below).

Table 23
Item Invariance for Verbal Subtest

Items by Item Label	Sig. of DIF
WD-04	.038

Figure 23 below shows a map of item/child position for the final Verbal Subtest. The map indicates that, on average, children were a little less able than items were difficult, with a child logit average position of -0.22 compared to the arbitrary item mean position of 0.0. In other words, children on average found the items to be slightly difficult.

Figure 23
Item/Child Position Map for Verbal Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Visual Subtest

The researcher first conducted 3 iterations of Rasch analyses to determine a workable set of items for Visual Subtest. Initially, there were 436 children and 131 items. After 3 iterations, the researcher found only 2 items with poor fit (i.e., VI-18 and VR-11). Therefore, these items were deleted from the original Visual Subtest.

Dimensionality.

Principal Components Analysis of Residuals. The researcher conducted a Rasch principal components analysis of residuals on the third iteration. Table 24 summarizes results from the principal components analysis of residuals.

Table 24
Table of Standardized Residual Variance in Percent—Visual Subtest

	Empirical	Modeled
Total raw variance in observations	100.0%	100.0%
Raw variance explained by measures	19.7%	20.8%
Raw variance explained by persons	5.1%	5.4%
Raw variance explained by items	14.5%	15.3%
Raw unexplained variance (total)	80.3%	79.2%
Unexplained variance in 1st contrast	3.8%	4.7%

It shows that the Rasch first dimension explained 19.7% of item variance in the data. This dimensionality finding indicates that Visual Subtest consists of more than one dimension because the Rasch first dimension can explain 19.7%, which does not meet the recommended 60% (Linacre, 2007). Since evidence was found of multiple dimensions underlying responses to test items, an exploratory factor analysis was conducted to clarify the dimensional structure.

Exploratory Factor Analysis.

Since the initial analysis of dimensionality indicated the presence of multiple factor underlying Visual Subtest, the researcher ran an SPSS exploratory factor analysis (with *all* items) and found more than one interpretable factor for the Visual Subtest. Table 25 below lists the items loading on each factor in a factor analysis with two factors specified and items not loading on either factor.

Table 25
List of Items Loading on Factors and Items Failing to Load on the First Two Factors—Visual Subtest

Items Loading on Factor One
VI-01, VI-03, VI-05, VI-07, VI-08, VI-09, VI-11, VI-12, VI-13, VI-14, VI-15, VI-16, VI-17, VI- 18, VI-19, VI-20, VI-21, VI-23, VI-24, VI-25, VI-26, VI-27, VI-28, VI-29, VI-30, VI-31, VI-32, VI-33, VI-35, VI-36, VI-37, VI-38, VI-39, VR-03, VR-04, VR-05, VR-07, VR-08, VR-09, VR-13, VR-14, VR-15, VR-16, VR-17, VR-18, VR-22, VR-25, and VR-26
Items Loading on Factor Two
MF-00, MF-01, MF-03, MF-04, MF-05, MF-06, MF-07, MF-09, MF-10, MF-11, VM-01, VM-03, VM-04, VM-05, VM-06, VM-07, VM-08, VM-10, VM-12, VM-13, VM-14, VM-16, VM-17, VM-18, VM-20, VM-21, VM-22, VM-23, VM-24, VM-26, VM-27, VM-28, VM-29, VM-32, VM-33, VM-34, VM-35, VR-29, and VR-30

Items Loading on Neither Factor
MF-02, MF-08, MF-12, MF-13, MF-14, MF-15, MF-16, MR-00, MR-01, MR-02, MR-03, MR-04, MR-05, MR-06, MR-07, VI-02, VI-04, VI-06, VI-10, VI-22, VI-34, VI-40, VM-02, VM-09, VM-11, VM-15, VM-19, VM-25, VM-30, VM-31, VM-36, VR-01, VR-02, VR-06, VR-10, VR-11, VR-12, VR-19, VR-20, VR-21, VR-23, VR-24, VR-27, and VR-28

Based on the exploratory factor analysis findings, the original Visual Subtest was measuring more than one domain. Because the researcher's original intention was to have one subtest for each domain, the researcher decided to focus on the items loading on the first factor and delete the items loading on other factors.

Once the researcher had decided to create a shortened Visual Subtest based on items loading on the first factor, the researcher ran Rasch analyses to investigate the level of reliability on the shortened Visual Subtest. The researcher ran 5 iterations and found twenty items with poor fit (i.e., outfit mean square >1.30: VI-01, VI-03, VI-05, VI-09, VI-12, VI-15, VI-16, VI-17, VI-18, VI-26, VI-29, VI-30, VI-31, VI-35, VI-38, VR-08, VR-18, VR-22, VR-25, and VR-26), which were deleted from the analyses.

Reliability.

For the shortened Visual Subtest, the reliability of person separation was 0.43 while the reliability of item separation was 0.83. Table 26 provides information on reliability and separation for items and persons.

Table 26
Person and Item Reliability and Separation—Visual Subtest

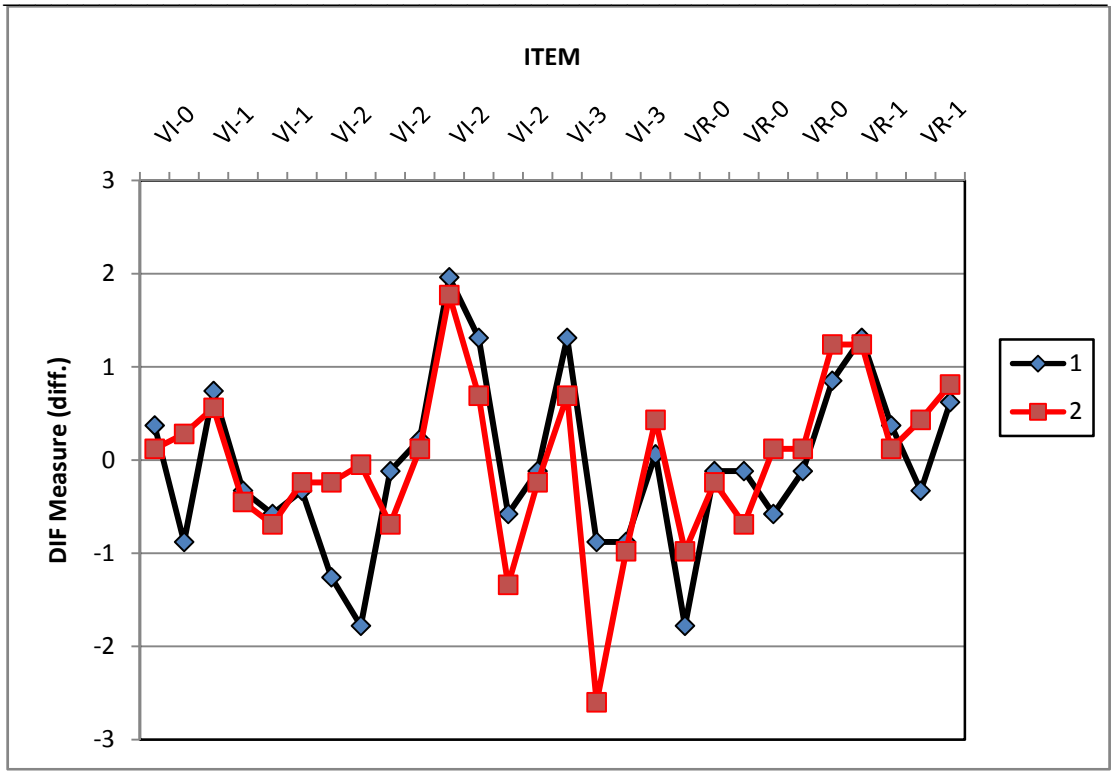
Statistics	Value
Reliability of person separation	.43
Reliability of item separation	.83
Item Separation (Real RMSE Separation)	0.87
Person Separation (Real RMSE Separation)	2.18

The achieved level of reliability for person separation was low and it was high for the item separation. The items in the shortened Visual Subtest were not able to separate children into groups (with a Real RMSE Separation of 0.88) while the children who took the items in the shortened Visual Subtest were able to separate items into more than two difficulty levels (with a Real RMSE Separation of 2.18). The average outfit MNSQ was 0.91. The average outfit MNSQ shows that the items in the shortened Visual Subtest fit the model well.

Item Invariance.

The researcher conducted an analysis of item invariance to determine the effect of gender on children's item responses on the shortened Visual Subtest. Figure 24 below provides a plot showing item invariance of the shortened Visual Subtest items.

Figure 24
DIF Plot Showing Item Invariance—Visual Subtest

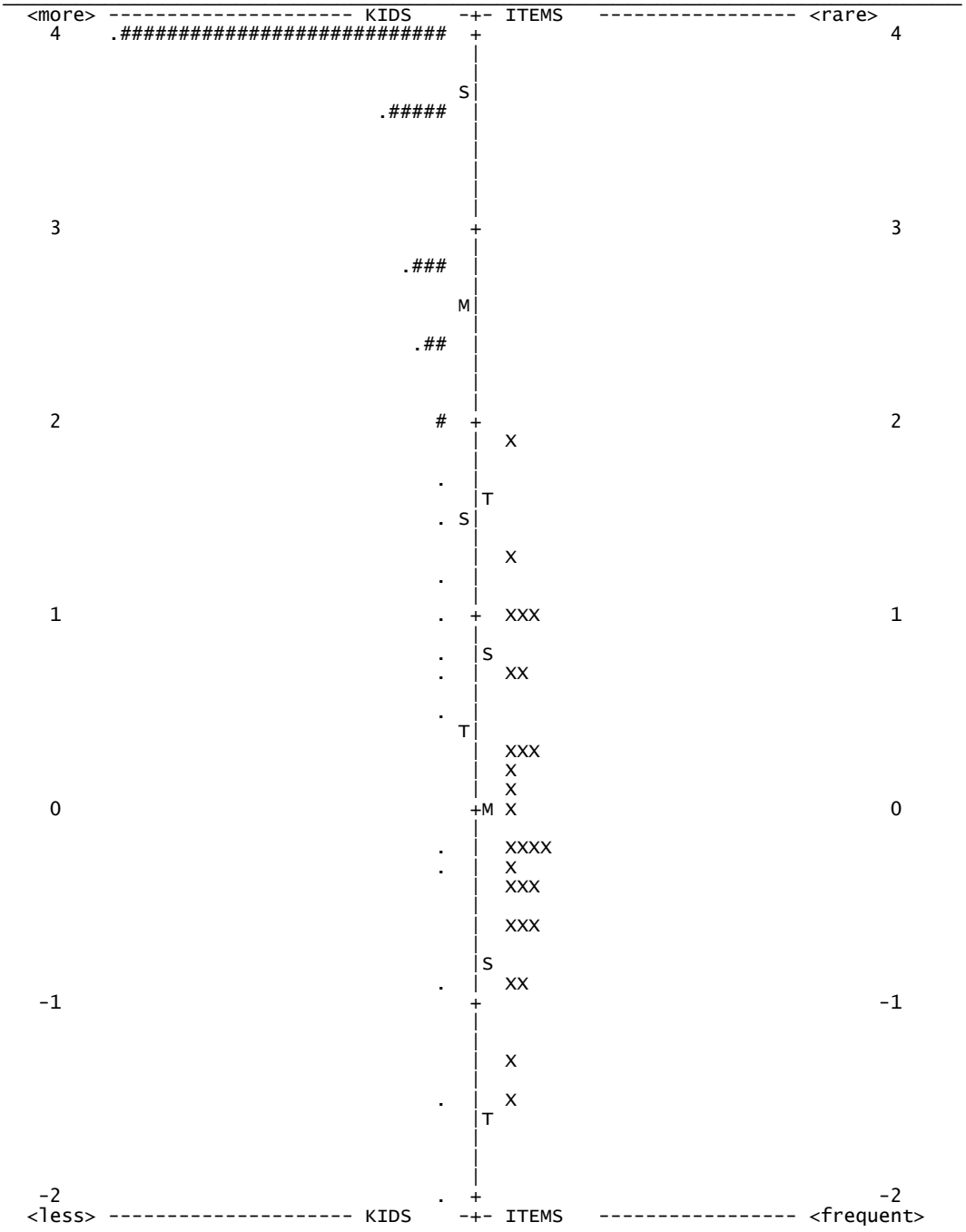


Based on the figure, it can be inferred that children with different genders did not differ greatly in pattern of responses to items overall, though there were clearly items where responses did differ. This is evident in the two lines (blue (1) for boy and red (2) for girl) falling close to each other. However, there was no item which evidenced statistically significant DIF ($p<.05$).

Figure 25 below shows a map of item/child position for the final Visual Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 2.61. In other words, children on average

found the items to be very easy. Thus, this item set was poorly targeted for these children, which led to low separation and reliability.

Figure 25
Item/Child Position Map for Visual Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Memory Subtest

The researcher first conducted 2 iterations of Rasch analyses to determine a workable set of items for Memory Subtest. Initially, there were 436 children and 90 items. After two iterations of analyses, the researcher found 2 items with poor fit (i.e., DR.-01 and DR-08). Therefore, these items were deleted from the original Memory Subtest.

Dimensionality.

Principal Components Analysis of Residuals. The researcher conducted a Rasch principal components analysis of residuals on the second iteration. Table 27 summarizes results from the principal components analysis of residuals.

Table 27
Table of Standardized Residual Variance in Percent—Memory Subtest

	Empirical	Modeled
Total raw variance in observations	100.0%	100.0%
Raw variance explained by measures	16.7%	17.0%
Raw variance explained by persons	8.9%	9.1%
Raw variance explained by items	7.8%	7.9%
Raw unexplained variance (total)	83.3%	83.0%
Unexplained variance in 1st contrast	6.2%	7.4%

It shows that the Rasch first dimension explained 16.7% of item variance in the data. This dimensionality finding indicates that Memory Subtest may consist of more than one dimension because the Rasch first dimension can explain 16.7%, which does not meet the recommended 60% (Linacre, 2007). Since evidence was found of multiple

dimensions underlying responses to test items, an exploratory factor analysis was conducted to clarify the dimensional structure.

Exploratory Factor Analysis.

Since the initial analysis of dimensionality indicated the presence of multiple factor underlying Memory Subtest, the researcher ran an SPSS exploratory factor analysis (with *all* items) and found more than one interpretable factor for the Memory Subtest. Table 28 below lists the items loading on each factor in a factor analysis with two factors specified and items not loading on either factor.

Table 28
List of Items Loading on Factors and Items Failing to Load on the First Two Factors—Memory Subtest

Items Loading on Factor One
DR-01, DR-03, DR-04, DR-05, DR-06, DR-07, DR-08, DR-09, DR-10, DR-11, DR-12, DR-13, DR-14, DR-15, DR-16, DR-17, DR-19, DR-20, DR-21, DR-22, DR-23, DR-25, DR-26, DR-27, DR-29, IR-03, IR-04, IR-05, IR-06, IR-07, IR-08, IR-10, IR-11, IR-14, IR-15, IR-16, IR-17, IR-19, IR-21, IR-22, IR-23, IR-26, IR-27, and IR-29
Items Loading on Factor Two
SR-02, SR-03, SR-04, SR-05, SR-06, SR-07, SR-08, SR-09, SR-10, SR-11, SR-12, SR-13, SR-14, SR-15, SR-16, SR-17, SR-18, SR-19, SR-20, SR-21, SR-22, SR-23, SR-24, SR-25, SR-26, SR-27, SR-28, SR-29, and SR-30

Items Loading on Neither Factor
DR-02, DR-18, DR-24, DR-28, DR-30, IR-01, IR-02, IR-09, IR-12, IR-13, IR-18, IR-20, IR-24, IR-25, IR-28, IR-30, and SR-01

Based on the exploratory factor analysis finding, the original Memory Subtest was measuring more than one domain. Because the researcher's original intention was to have one subtest for each domain, the researcher decided to focus on the items loading on the first factor and delete the items loading on other factors, which may have been possibly measuring something other than memory skills, or a related facet of memory skills.

Once the researcher had decided to create a shortened Memory Subtest based on items loading on the first factor, the researcher ran Rasch analyses to investigate the level of reliability on the shortened Memory Subtest. The researcher ran 1 iteration and found no items with poor fit.

Reliability.

For the shortened Memory Subtest, the reliability of person separation was 0.81 while the reliability of item separation was 0.90. Table 29 provides item and person separation and reliability statistics.

Table 29
Person and Item Reliability and Separation—Memory Subtest

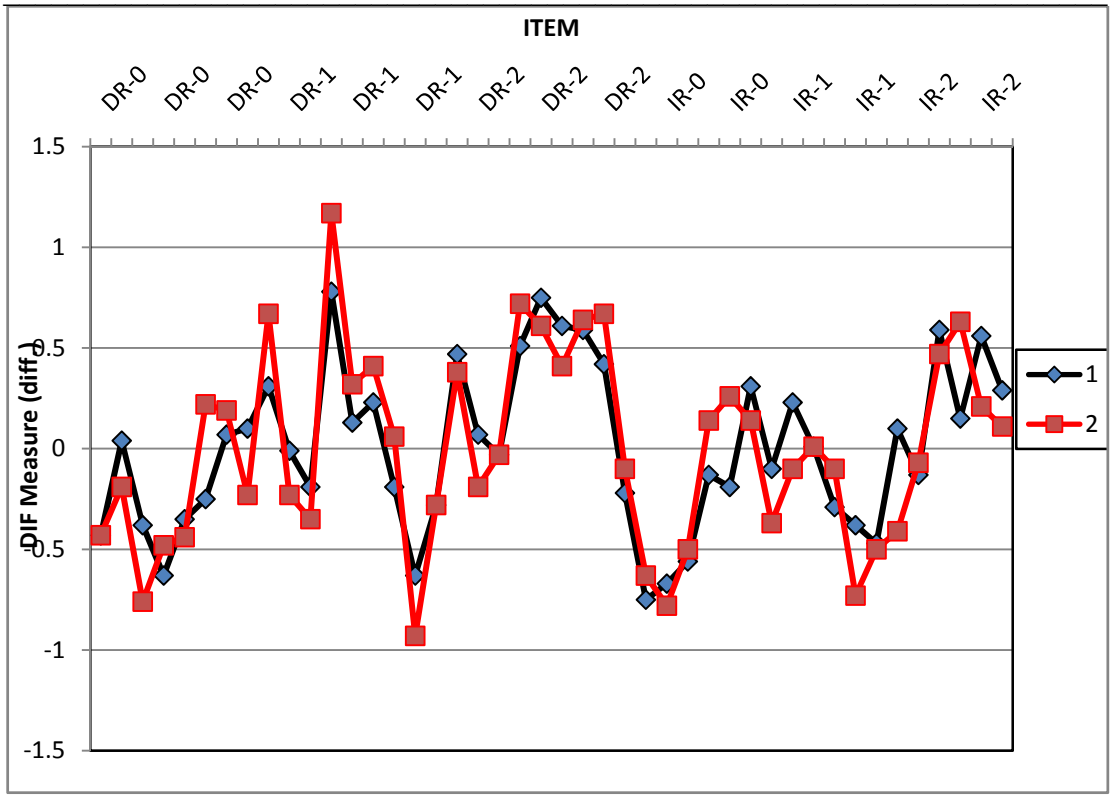
Statistics	Value
Reliability of person separation	.81
Reliability of item separation	.90
Item Separation (Real RMSE Separation)	2.07
Person Separation (Real RMSE Separation)	3.06

The achieved level of reliability for both person and item separation was high. The items in the shortened Memory Subtest were able to separate children into two distinct ability groups (with a Real RMSE Separation of 2.07) while the children who took the items in the shortened Memory Subtest were able to separate items into three difficulty levels (with a Real RMSE Separation of 3.06). The average outfit MNSQ was 0.97. The average outfit MNSQ shows that the items in the new Memory Subtest fit the model well.

Item Invariance.

The researcher conducted an analysis of item invariance to determine the effect of gender on children's item responses on the shortened Memory Subtest. Figure 26 below provides a plot showing item invariance of the shortened Memory Subtest items.

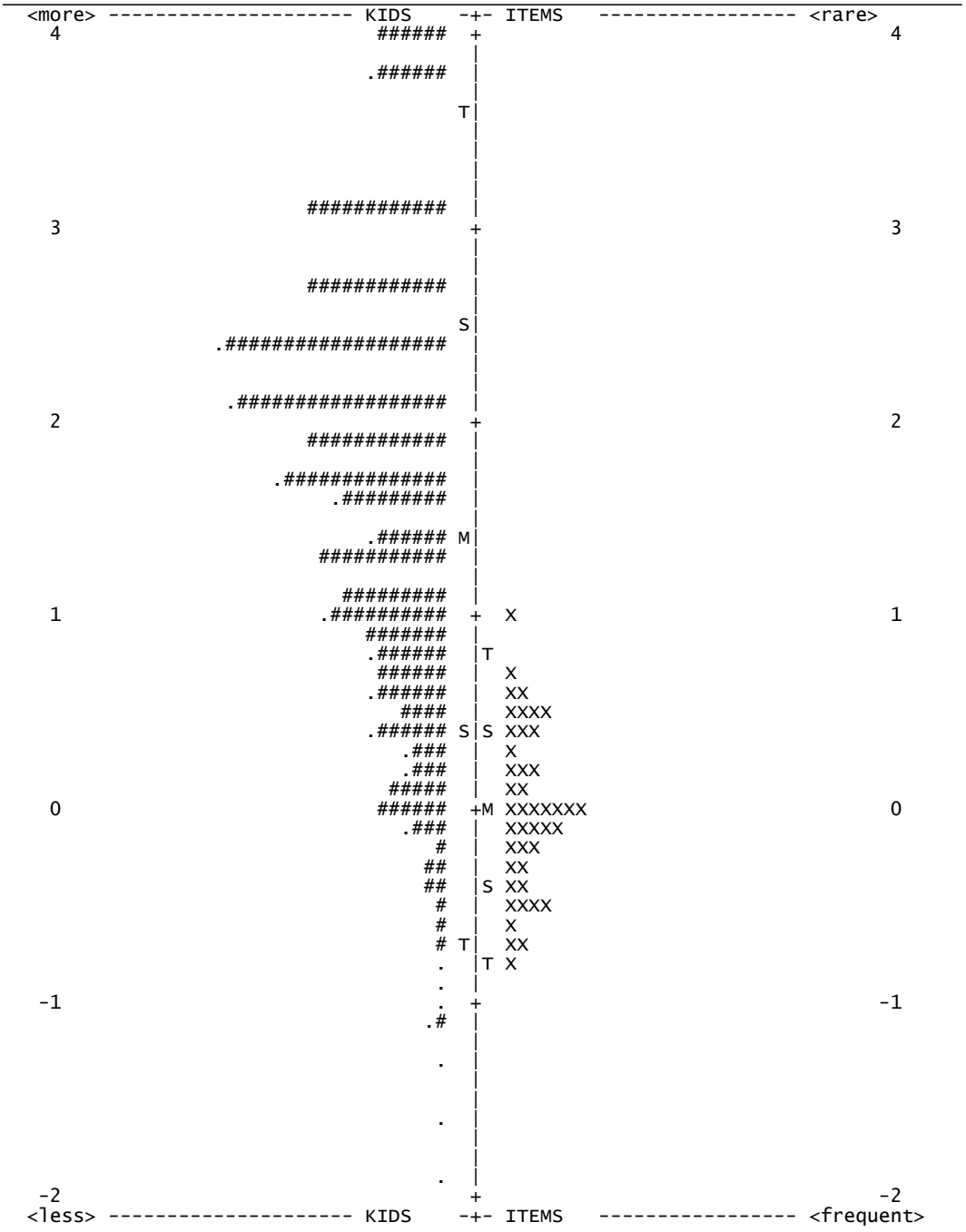
Figure 26
DIF Plot Showing Item Invariance—Memory Subtest



Based on the figure, it can be inferred that children with different genders did not differ greatly in pattern of responses to items. This is evident in the two lines (blue (1) for boy and red (2) for girl) falling close to each other. In addition, there was no item which evidenced statistically significant DIF ($p<.05$).

Figure 27 below shows a map of item/child position for the final Memory Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit mean position of 1.43. In other words, children on average found the items to be very easy.

Figure 27
Item/Child Position Map for Memory Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Math Subtest

The researcher first conducted 9 iterations of Rasch analyses to determine a workable set of items for Math Subtest. Initially, there were 436 children and 97 items. After nine iterations of analyses, the researcher found 49 items with poor fit (See Table 30 Below). Therefore, these items were deleted from the original Math Subtest.

Table 30
Misfitting Items in Math Subtest

Misfitting Items by Item Label
MC-01, MC-02, MC-11, MC-41, MC-42, MC-43, MC-44, MC-45, MC-46, MC-47, MC-48, MC-50, MCV-01, MCV-02, MCV-03, MCV-04, MCV-06, MCV-07, MCV-08, MCV-09, MCV-10, MCV-11, MCV-12, MCV-13, MCV-14, MCV-15, WP-03, WP-05, WP-06, WP-07, WP-08, WP-09, WP-11, WP-12, WP-13, WP-14, WP-15, WP-16, WP-17, WP-22, WP-23, WP-24, WP-26, WP-27, WP-28, WP-29, WP-30, WP-31, and WP-32

Note. Items were dropped if their outfit MNSQ was higher than 1.30.

Dimensionality.

Principal Components Analysis of Residuals. The researcher conducted a Rasch principal components analysis of residuals on the ninth iteration. Table 31 summarizes results from the principal components analysis of residuals.

Table 31
Table of Standardized Residual Variance in Percent—Math Subtest

	Empirical	Modeled
Total raw variance in observations	100.0%	100.0%
Raw variance explained by measures	48.2%	48.4%
Raw variance explained by persons	22.4%	22.6%
Raw variance explained by items	25.7%	25.9%
Raw unexplained variance (total)	51.8%	51.6%
Unexplained variance in 1st contrast	4.0%	7.7%

It shows that the Rasch first dimension explained 48.2% of item variance in the data. This dimensionality finding indicates that Math Subtest may consist of more than one dimension because the Rasch first dimension can explain 48.2%, which does not meet the recommended 60% (Linacre, 2007). Since evidence was found of multiple dimensions underlying responses to test items, an exploratory factor analysis was conducted to clarify the dimensional structure.

Exploratory Factor Analysis.

Since the initial analysis of dimensionality indicated the presence of multiple factor underlying Math Subtest, the researcher ran an SPSS exploratory factor analysis (with *all* items) and found more than one interpretable factor for the Math Subtest. Table 32 below list the items loading on each factor in a factor analysis with two factors specified and items not loading on either factor.

Table 32
List of Items Loading on Factors and Items Failing to Load on the First Two Factors—Math Subtest

Items Loading on Factor One
MC-03, MC-04, MC-05, MC-06, MC-07, MC-08, MC-13, MC-14, MC-15, MC-16, MC-17, MC-18, MC-22, MC-23, MC-26, MC-27, MC-28, MC-30, MC-31, MCV-01, MCV-02, MCV-03, MCV-04, MCV-07, MCV-08, MCV-09, MCV-10, MCV-11, MCV-12, MCV-13, MCV-14, MCV-15, WP-01, WP-02, WP-03, WP-04, WP-06, WP-07, WP-08, WP-09, WP-10, WP-13, WP-18, WP-19, WP-20, WP-21, WP-22, WP-23, WP-24, and WP-28.
Items Loading on Factor Two
MC-01, MC-09, MC-10, MC-11, MC-12, MC-19, MC-20, MC-21, MC-24, MC-25, MC-29, MC-32, MC-33, MC-34, MC-35, MC-36, MC-37, MC-38, MC-39, MC-40, MC-41, MC-42, MC-43, MC-44, MC-45, MC-46, MC-47, MC-48, MC-49, MC-50, MCV-05, WP-05, WP-11, WP-12, WP-14, WP-17, WP-25, WP-27, WP-29, WP-30, WP-31, and WP-32.
Items Loading on Neither Factor
MC-02, MCV-06, WP-15, WP-16, and WP-26

Based on the exploratory factor analysis finding, the original Math Subtest was measuring more than one domain. Because the researcher's original intention was to have one subtest for each domain, the researcher decided to focus on the items loading on the first factor and delete the items loading on other factors, which may have been possibly measuring something other than math skills, or a related facet of math skills.

Once the researcher had decided to create a shortened Math Subtest based on items loading on the first factor, the researcher ran Rasch analyses to investigate the level of reliability on the shortened Math Subtest. The researcher ran 8 iterations and found 34 items with poor fit (i.e., MC-03, MC-04, MC-05, MC-06, MC-07, MC-08, MC-16, MC-17, MC-18, MCV-01, MCV-02, MCV-03, MCV-04, MCV-07, MCV-08, MCV-09, MCV-10, MCV-11, MCV-12, MCV-13, MCV-14, MCV-15, WP-03, WP-04, WP-06, WP-07, WP-08, WP-09, WP-10, WP-13, WP-22, WP-23, WP-24, and WP-28), which were deleted from the analyses.

Reliability.

For the shortened Math Subtest, the reliability of person separation was 0.84 while the reliability of item separation was 0.99 (See Table 33 below).

Table 33
Person and Item Reliability and Separation—Math Subtest

Statistics	Value
Reliability of person separation	.84
Reliability of item separation	.99
Item Separation (Real RMSE Separation)	2.26
Person Separation (Real RMSE Separation)	11.67

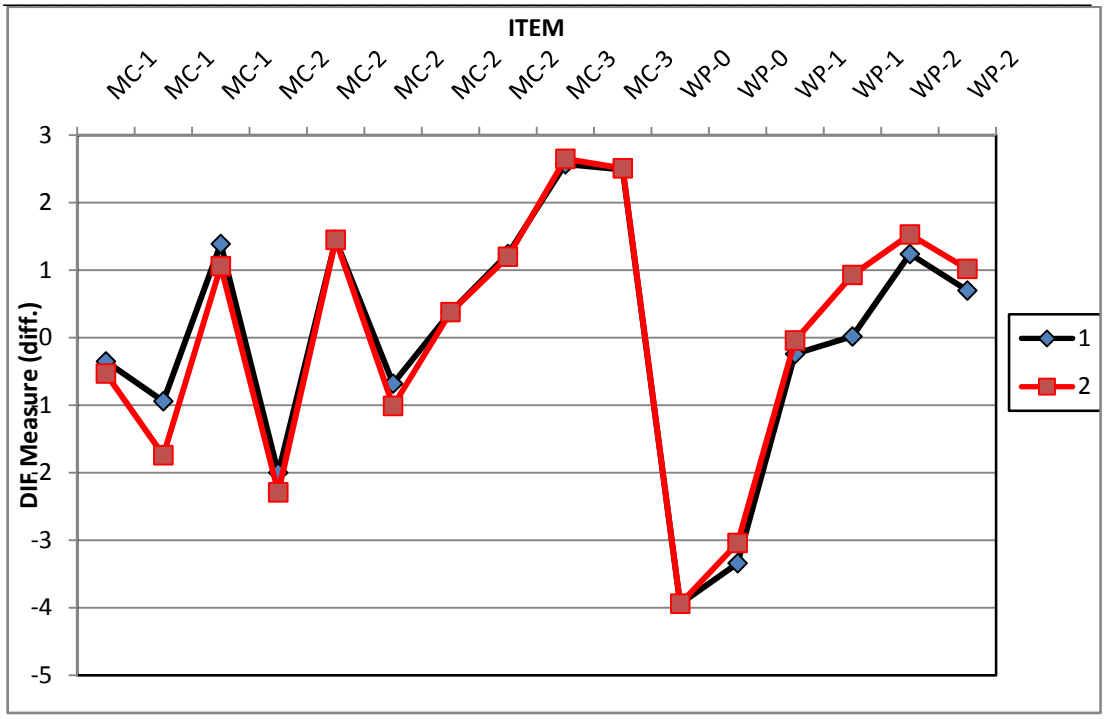
The achieved level of reliability for both person and item separation was high. The items in the shortened Math Subtest were able to separate children into more than two distinct ability groups (with a Real RMSE Separation of 2.26) while the children who took the items in the shortened Math Subtest were able to separate items into more than eleven difficulty levels (with a Real RMSE Separation of 11.67). The average outfit

MNSQ was 0.95. The average outfit MNSQ shows that the items in the shortened Math Subtest fit the model well.

Item Invariance.

The researcher conducted an analysis of item invariance to determine the effect of gender on children's item responses on the shortened Math Subtest. Figure 28 below provides a plot showing item invariance of the shortened Math Subtest items.

Figure 28
DIF Plot Showing Item Invariance—Math Subtest



Based on the figure, it can be inferred that children with different genders did not differ greatly in their pattern of responses to items. This is evident in the two lines (blue (1) for boy and red (2) for girl) falling close to each other. However, two items (i.e., MC-

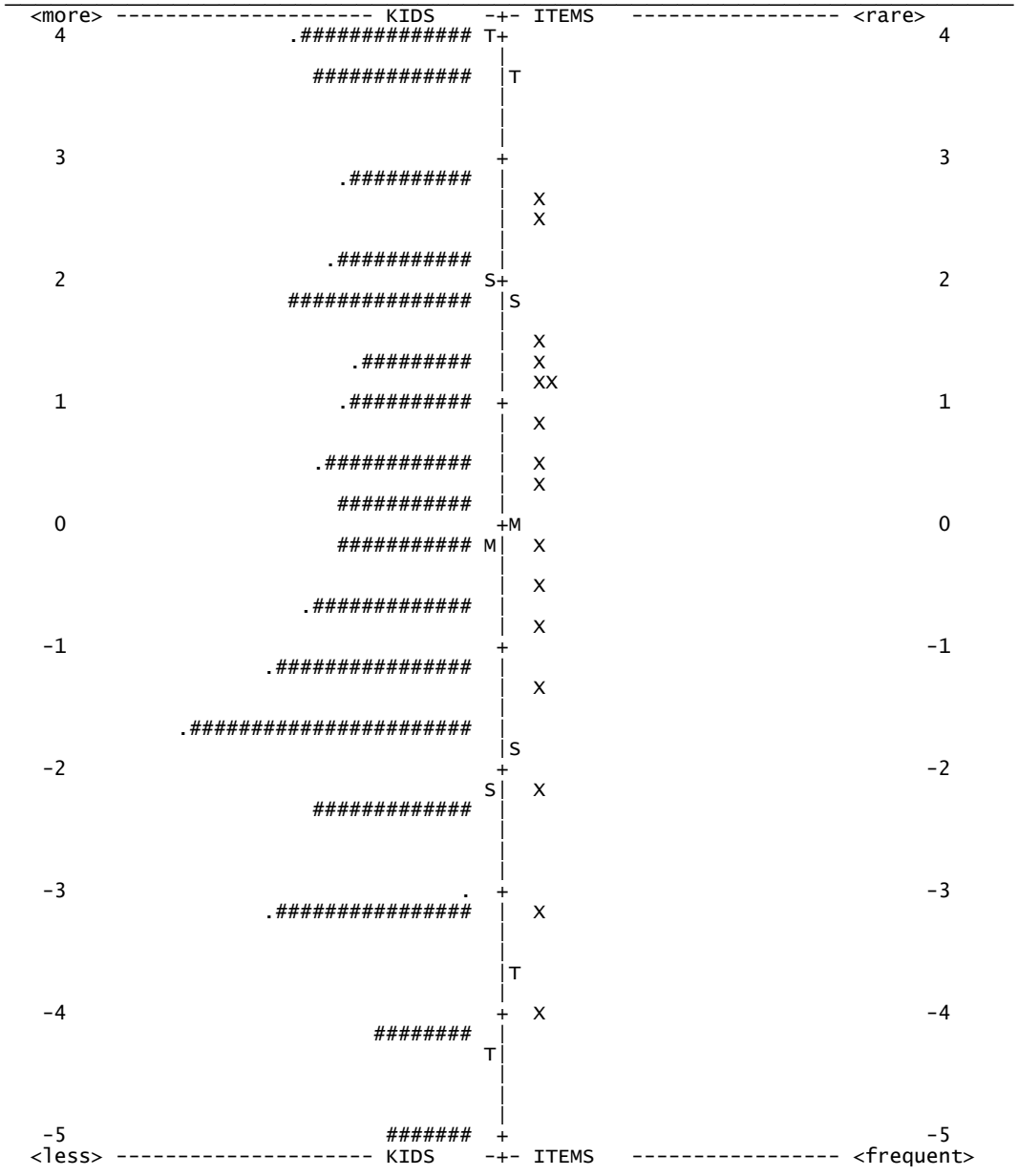
14 and WP-19) evidenced statistically significant DIF ($p<.05$) and hence became candidates for deletion to arrive at the final scale (See Table 34 below).

Table 34
Item Invariance for Math Subtest

Items by Item Label	Sig. of DIF
MC-14	.004
WP-19	.001

Figure 29 below shows a map of item/child position for the final Math Subtest. The map indicates that, on average, children were a little less able than items were difficult with a child logit mean position of -0.12. In other words, children on average found the items to be a little difficult.

Figure 29
Item/Child Position Map for Math Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Logical Subtest

The researcher first conducted 1 iteration of Rasch analyses to determine a workable set of items for Logical Subtest. Initially, there were 436 children and 62 items. After 1 iteration of analyses, the researcher found no item with poor fit. Therefore, all items were maintained in the original Logical Subtest.

Dimensionality.

Principal Components Analysis of Residuals. The researcher conducted a Rasch principal components analysis of residuals on the one iteration. Table 35 summarizes results from the principal components analysis of residuals.

Table 35
Table of Standardized Residual Variance in Percent—Logical Subtest

	Empirical	Modeled
Total raw variance in observations	100.0%	100.0%
Raw variance explained by measures	27.0%	26.7%
Raw variance explained by persons	11.4%	11.2%
Raw variance explained by items	15.6%	15.5%
Raw unexplained variance (total)	73.0%	73.3%
Unexplained variance in 1st contrast	3.8%	5.2%

It shows that the Rasch first dimension explained 27% of item variance in the data. This dimensionality finding indicates that Logical Subtest may consist of more than one dimension because the Rasch first dimension can explain 27%, which does not meet the recommended 60% (Linacre, 2007). Since evidence was found of multiple dimensions underlying responses to test items, an exploratory factor analysis was conducted to clarify the dimensional structure.

Exploratory Factor Analysis.

Since the initial analysis of dimensionality indicated the presence of multiple factors underlying Logical Subtest, the researcher ran an SPSS exploratory factor analysis (with *all* items) and found more than one interpretable factor for the Logical Subtest. Table 36 below list the items loading on each factor in a factor analysis with two factors specified and items not loading on either factor.

Table 36
List of Items Loading on Factors and Items Failing to Load on the First Two Factors—Logical Subtest

Items Loading on Factor One
CF-02, CF-03, CF-04, CF-05, CF-06, CF-07, CF-12, PF-01, PF-02, PF-03, PF-04, PF-05, PF-06, PF-07, PF-08, PF-10, PF-11, PF-12, PF-13, PF-14, PF-15, PF-16, PF-17, SO-02, SO-04, SO-05, SO-06, SO-07, SO-08, and SO-09.
Items Loading on Factor Two
CF-01, CF-09, CF-10, CF-13, CF-14, CF-15, CF-16, CF-17, CF-18, CF-19, CF-20, CF-22, CF-23, CF-24, CF-25, PF-09, PF-18, PF-19, PF-20, PF-21, PF-22, SO-14, and SO-15.
Items Loading on Neither Factor
CF-08, CF-11, CF-21, SO-01, SO-03, SO-10, SO-11, SO-12, and SO-13

Based on the exploratory factor analysis finding, the original Logical Subtest was measuring more than one domain. Because the researcher's original intention was to have one subtest for each domain, the researcher decided to focus on the items loading on

the first factor and delete the rest of items loading on other factors, which may have been possibly measuring something other than logical skills, or a related facet of logical skills.

Once the researcher had decided to create a shortened Logical Subtest based on items loading on the first factor, the researcher ran Rasch analyses to investigate the level of reliability on the shortened Logical Subtest. The researcher ran 3 iterations and found 2 items with poor fit (i.e., SO-06 and SO-07), which were deleted from the analyses.

Reliability.

For the shortened Logical Subtest, the reliability of person separation was 0.79 while the reliability of item separation was 0.98 (See Table 37 below).

Table 37
Person and Item Reliability and Separation—Logical Subtest

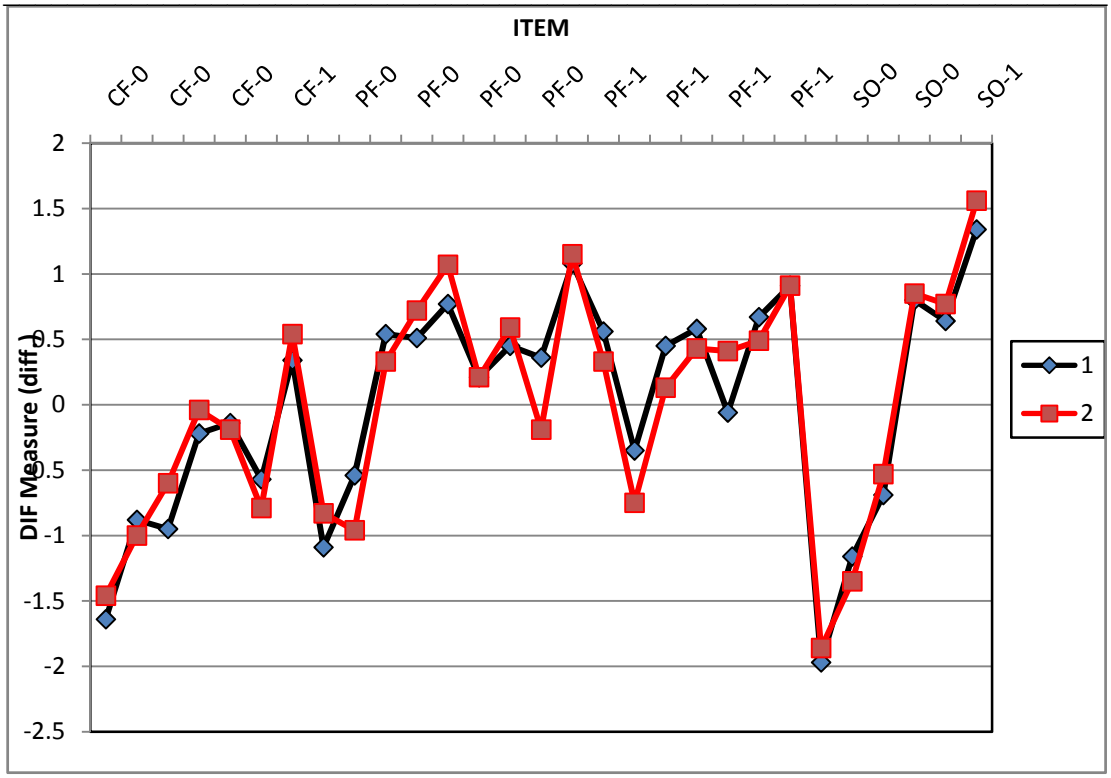
Statistics	Value
Reliability of person separation	.79
Reliability of item separation	.98
Item Separation (Real RMSE Separation)	1.95
Person Separation (Real RMSE Separation)	6.57

The achieved level of reliability for both person and item separation was high. The items in the shorten Logical Subtest were able to separate children into almost two distinct ability groups (with a Real RMSE Separation of 1.95) while the children who took the items in the shortened Logical Subtest were able to separate items into more than six difficulty levels (with a Real RMSE Separation of 6.57). The average outfit MNSQ was 0.99. The average outfit MNSQ shows that the items in the shortened Logical Subtest fit the model well.

Item Invariance.

The researcher conducted an analysis of item invariance to determine the effect of gender on children's item responses on the shortened Logical Subtest. Figure 30 below provides a plot showing item invariance (lack of DIF) of the shortened Logical Subtest items.

Figure 30
DIF Plot Showing Item Invariance-Logical Subtest



Based on the figure, it can be inferred that children with different genders did not differ greatly in pattern of response to items. This is evident in the two lines (blue (1) for boy and red (2) for girl) falling close to each other. However, two items (i.e., PF-08 and

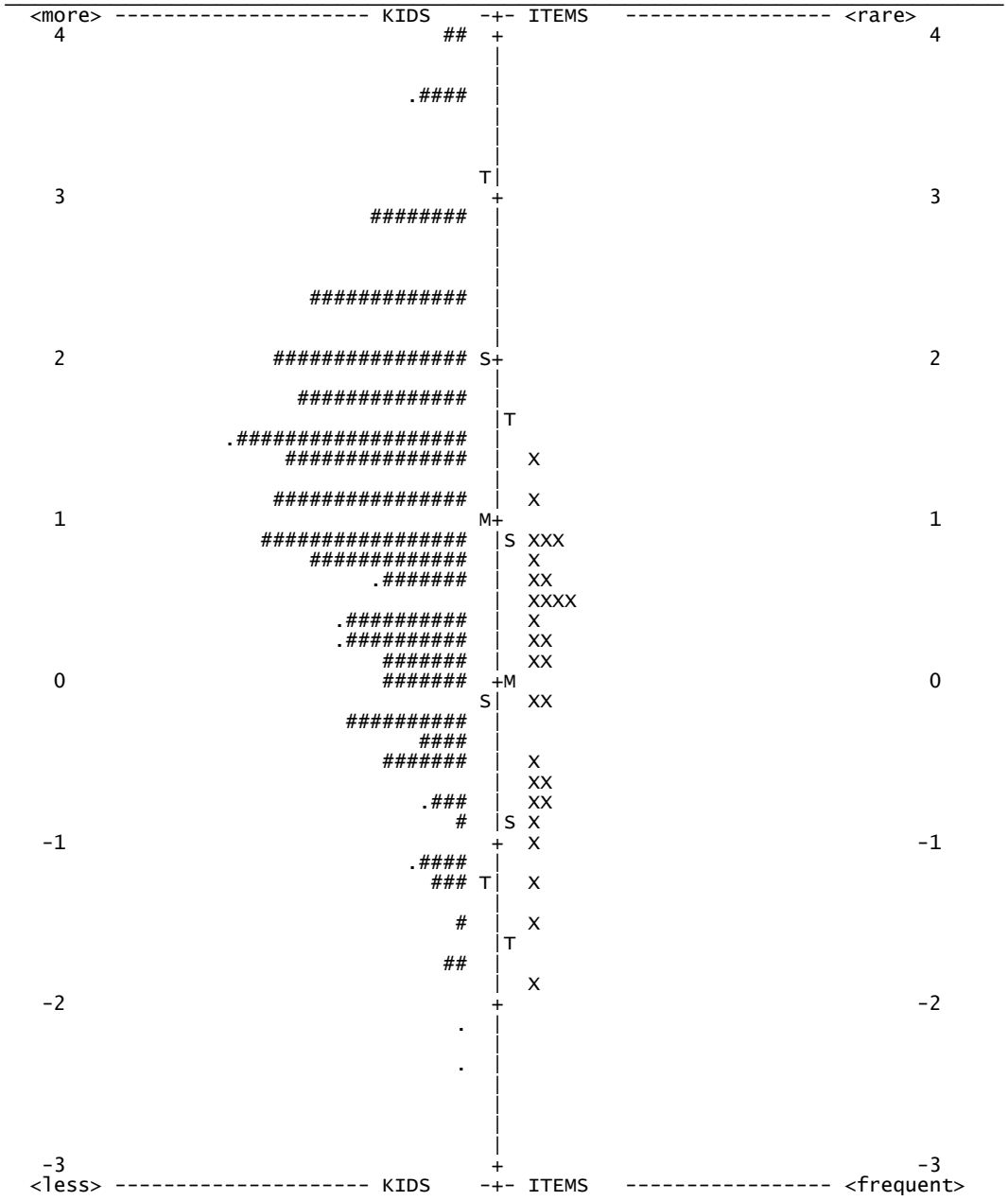
PF-15) evidenced statistically significant DIF ($p<.05$) and hence became a candidate for deletion to arrive at the final scale (See Table 38 below).

Table 38
Item Invariance for Logical Subtest

Items by Item Label	Sig. of DIF
PF-08	.017
PF-15	.038

Figure 31 below shows a map of item/child position for the final Logical Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit average position of 0.95. In other words, children on average found the items to be very easy.

Figure 31
Item/Child Position Map for Logical Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

General Knowledge Subtest

The researcher first conducted 4 iterations of Rasch analyses to determine a workable set of items for General Knowledge Subtest. Initially, there were 436 children and 75 items. After four iterations of analyses, the researcher found 3 items with poor fit (i.e., GK-06, GK-53, and GK-72). Therefore, these items were deleted from the original General Knowledge Subtest.

Dimensionality.

Principal Components Analysis of Residuals. The researcher conducted a Rasch principal components analysis of residuals on the fourth iteration. Table 39 summarizes results from the principal components analysis of residuals.

Table 39
Table of Standardized Residual Variance in Percent—General Knowledge Subtest

	Empirical	Modeled
Total raw variance in observations	100.0%	100.0%
Raw variance explained by measures	20.2%	20.7%
Raw variance explained by persons	11.7%	12.0%
Raw variance explained by items	8.5%	8.7%
Raw unexplained variance (total)	79.8%	79.3%
Unexplained variance in 1st contrast	3.5%	4.3%

It shows that the Rasch first dimension explained 20.2% of item variance in the data. This dimensionality finding indicates that General Knowledge Subtest may consist of more than one dimension because the Rasch first dimension can explain 20.2%, which does not meet the recommended 60% (Linacre, 2007). Since evidence was found of

multiple dimensions underlying responses to test items, an exploratory factor analysis was conducted to clarify the dimensional structure.

Exploratory Factor Analysis.

Since the initial analysis of dimensionality indicated the presence of multiple factors underlying General Knowledge Subtest, the researcher ran an SPSS exploratory factor analysis (with *all* items) and found more than one interpretable factor for the General Knowledge Subtest. Table 40 below lists the items loading on each factor in a factor analysis with two factors specified and items not loading on either factor.

Table 40
List of Items Loading on Factors and Items Failing to Load on the First Two Factors—General Knowledge Subtest

Items Loading on Factor One
GK-01, GK-02, GK-03, GK-04, GK-05, GK-06, GK-07, GK-08, GK-09, GK-12, GK-13, GK-15, GK-16, GK-17, GK-18, GK-20, GK-21, GK-22, GK-27, GK-28, GK-30, GK-31, GK-32, GK-36, GK-41, GK-48, GK-50, GK-51, GK-52, GK-54, GK-61, GK-62, GK-63, GK-64, GK-65, GK-66, GK-68, GK-69, GK-70, and GK-71.
Items Loading on Factor Two
GK-10, GK-11, GK-14, GK-19, GK-24, GK-25, GK-26, GK-29, GK-33, GK-34, GK-35, GK-37, GK-38, GK-39, GK-40, GK-42, GK-43, GK-44, GK-46, GK-47, GK-49, GK-55, GK-56, GK-57, GK-58, GK-59, GK-60, GK-73, and GK-75.
Items Loading on Neither Factor
GK-23, GK-45, GK-53, GK-67, GK-72, and GK-74

Based on the exploratory factor analysis finding, the original General Knowledge Subtest was measuring more than one domain. Because the researcher's original intention was to have one subtest for each domain, the researcher decided to focus on the items loading on the first factor and delete the rest of items loading on other factors, which may have been possibly measuring something other than general knowledge, or a related facet of general knowledge.

Once the researcher had decided to create a shortened General Knowledge Subtest based on items loading on the first factor, the researcher ran Rasch analyses to investigate the level of reliability on the shortened General Knowledge Subtest. The researcher ran 2 iterations and found 2 items with poor fit (i.e., GK01 and GK-54), which were deleted from the analyses.

Reliability.

For the shortened General Knowledge Subtest, the reliability of person separation was 0.73 while the reliability of item separation was 0.95 (See Table 41 Below).

Table 41
Person and Item Reliability and Separation—General Knowledge Subtest

Statistics	Value
Reliability of person separation	.73
Reliability of item separation	.95
Item Separation (Real RMSE Separation)	1.66
Person Separation (Real RMSE Separation)	4.34

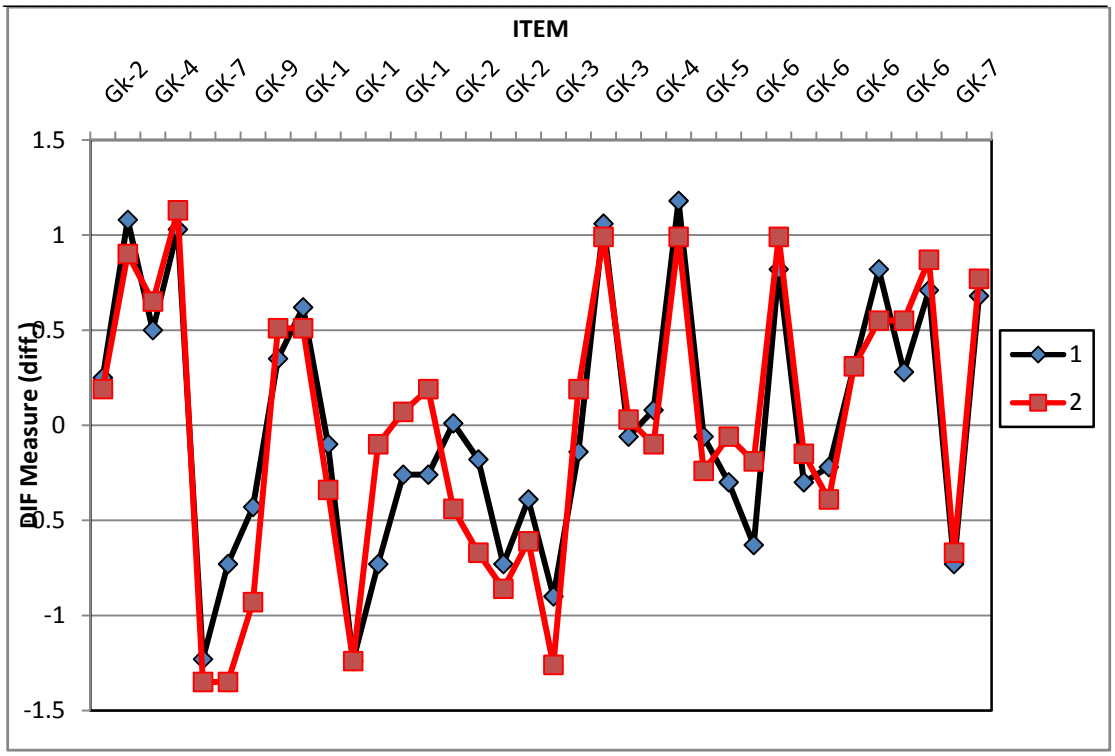
The achieved level of reliability for both person and item separation was high. The items in the shortened General Knowledge Subtest were able to separate children

into almost two distinct ability groups (with a Real RMSE Separation of 1.66) while the children who took the items in the shortened General Knowledge Subtest were able to separate items into more than four difficulty levels (with a Real RMSE Separation of 4.34). The average outfit MNSQ was 0.97. The average outfit MNSQ shows that the items in the shortened General Knowledge Subtest fit the model well.

Item Invariance.

The researcher conducted an analysis of item invariance to determine the effect of gender on children's item responses on the shortened General Knowledge Subtest. Figure 32 below provides a plot showing item invariance (lack of DIF) of the shortened General Knowledge Subtest items.

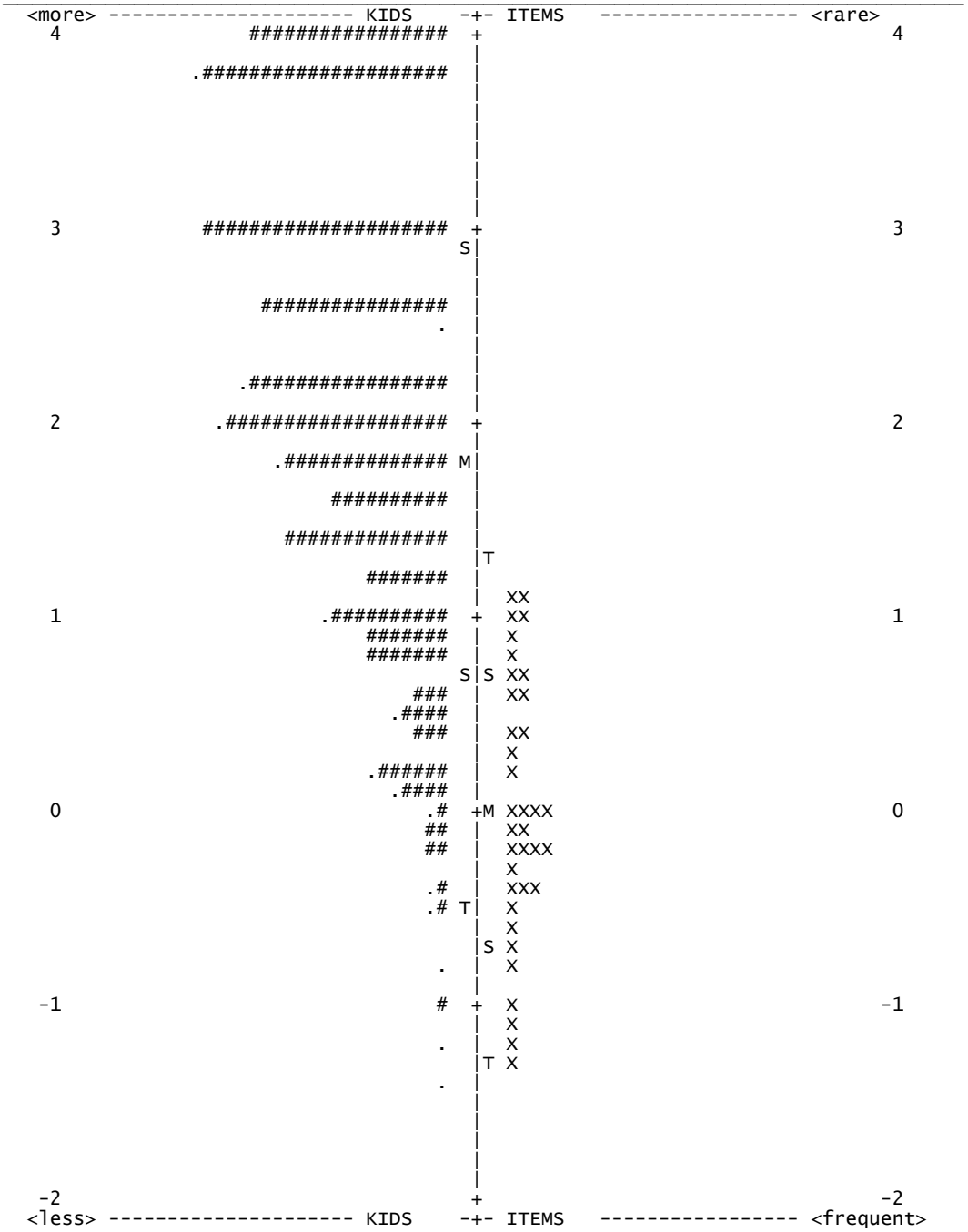
Figure 32
DIF Plot Showing Item Invariance—General Knowledge Subtest



Based on the figure, it can be inferred that children with different genders did not differ significantly ($p<.05$) in behavior while answering items. This is evident in the two lines (blue (1) for boy and red (2) for girl) falling close to each other. In addition, there was no item which evidenced statistically significant DIF ($p<.05$).

Figure 33 below shows a map of item/child position for the final General Knowledge Subtest. The map indicates that, on average, children were much more able than items were difficult, with a child logit average position of 1.81. In other words, children on average found the items to be very easy.

Figure 33
Item/Child Position Map for General Knowledge Subtest



Note. "x" denotes a child. "#" denotes an item. "M" denotes means. "S" denotes standard deviation. "Q" denotes two standard deviations.

Descriptive Statistics

Appendix 14 provides a summary of descriptive statistics for each subtest by gender, of correlation statistics between subtests, and of t-test statistics for each subtest by gender. The results suggest that, except for visual subtest, responses to all subtests were relatively normally distributed. The correlations between subtests suggest that each subtest correlated with others but at a low to moderate level and that each subtest is measuring a different trait. The t-test found no difference in response level between boys and girls.

CHAPTER 5

Discussion

This chapter provides a summary of the findings for each subtest. The researcher established that some TAR subtests had adequate levels of reliability and validity. Findings are then discussed in light of the literature and limitations of the TAR. The researcher also points out the limitations of this study and recommends further research topics for verification and improvement of TAR.

Summary of Findings

The purpose of this dissertation was to develop a new scale, called TAR, which measures different components of academic readiness for first grade instruction in Thailand. The researcher used exploratory factor analysis and the Rasch model to identify appropriate items and assess the level of reliability; two groups of experts were recruited to assess item content to ensure content validity.

Reliability

Verbal Subtest

The level of reliability for person separation was 0.93. The reliability of person separation is the ability of a set of items to reliably discriminate among people based on their trait level. Therefore, this means that TAR contains a set of verbal items with a high

level of ability to reliably discriminate among children's verbal ability level. The person separation for verbal items was 3.72. Person separation is an indicator of the spread in person measures. This index indicates the number of distinct levels into which the sample of persons can be classified. This degree of separation indicates that the difference in scores is due to the differences in the magnitudes of the person's underlying attribute (Bode & Wright, 1999). Since a variable exists when the measure assesses different amounts of the trait, the researcher established that the verbal subtest is a variable as it can measure different amounts of verbal skills. In this case, the verbal subtest can differentiate the amount of children's verbal skills into almost four groups.

The verbal items did not evidence significant differential item functioning (DIF), indicating that boys or girls did not understand items differently (or were influenced by other factors) while responding to verbal items because of their gender. In other words, differing gender does not create an unfair advantage for a specific gender group to perform better than the other. DIF quantifies the difference in the probability of an item response pattern for two groups in questions (Stark & Chernyshenko, 2001). DIF may occur due to factors underlying the gender difference. For example, boys may be more able to answer a logical item while girls may be more able to answer a visual item or vice versa. Nonetheless, this was not the case for the verbal items.

The child's logit mean position for verbal subtest was -0.22, as compared to the arbitrary item logit mean position of 0.00. The mean person position indicates how well the measure is targeted on the sample. Since the child's logit mean position was -0.22,

the difference from the item mean of 0.0 was minimal and indicates that verbal subtest was well-targeted. In other words, the verbal items had an appropriate level of difficulty.

Visual Subtest

The level of reliability for person separation was 0.43. This means that TAR contains a set of visual items with a low level of ability to reliably discriminate among children's visual ability level. The person separation for visual items was 0.87. In this case, the visual subtest cannot differentiate the amount of children's visual skills into groups. The visual items did not evidence significant DIF, indicating that boys or girls do not experience any different level of item difficulty while responding to visual items because of their gender. The child's logit mean position was 2.61. Such a difference is large and it indicates that the visual items did not have an appropriate level of difficulty for the sample. More specifically, the visual items were too easy for these children and they were, as a result, poorly targeted for the children's ability level.

Memory Subtest

The level of reliability for person separation was 0.81. This means that TAR contains a set of memory items with a moderate level of difficulty to reliably discriminate among children's memory ability level. The person separation for memory items was 2.07. In this case, the memory subtest differentiated the level of children's memory skills into two groups. The memory items did not evidence significant DIF, indicating that boys or girls did not experience any different level of item difficulty while responding to memory items because of their gender. The child's logit mean position was 1.43. Such a

difference indicates that the memory items were quite easy for these children overall. More specifically, the memory items are adequately targeted for these children's memory ability levels but more difficult items may be useful should the scale be revised.

Math Subtest

The level of reliability for person separation was 0.84. This means that TAR contains a set of math items with a moderately high level of ability to reliably discriminate among children's math ability level. The person separation for math items was 2.26. In this case, the math subtest differentiated the level of children's math skills into groups. The math items did not evidence significant DIF, indicating that boys or girls did not experience any different level of item difficulty while responding to math items because of their gender. The child's logit mean position was -0.12. Such a difference from the item mean of 0.0 is minimal and it indicates that the math items overall had appropriate levels of difficulty for this sample of children. More specifically, the math items were well targeted for these children's math ability level.

Logical Subtest

The level of reliability for person separation was 0.79. This means that TAR contains a set of logical items with a moderate level of ability to reliably discriminate among children's logical ability level. The person separation for logical items was 1.95. In this case, the logical subtest differentiated the level of children's logical skills into almost two groups. The logical items did not evidence significant DIF. The child's logit mean position was 0.95. Such a difference indicates that the logical items overall were

somewhat easy for the sample. More specifically, the logical items were not perfectly targeting children's logical abilities.

General Knowledge Subtest

The level of reliability for person separation was 0.73. This means that TAR contains a set of general knowledge items with a moderate level of ability to reliably discriminate among children's general knowledge levels. The person separation for general knowledge items was 1.66. The general knowledge items did not evidence significant DIF. The child's logit mean position was 1.81. Such a difference indicates that the general knowledge items overall were quite easy for the sample. More specifically, the general knowledge items are not perfectly targeting children's general knowledge level.

TAR as a Scale

TAR contains six subtests (i.e., verbal, visual, memory, math, logical, and general knowledge). With the exception of the visual subtest, TAR has sufficient reliability levels to be worthwhile of future research projects to improve and provide further validation information for TAR as a measure of academic readiness for first grade instruction in Thailand. Although not all subtests had a high level of reliability, the achieved levels of reliability were improvements over the earlier version of TAR. Table 42 provides a summary of statistics for each subtest in the second pilot study and in the main study

Table 42
Subtests' Level of Reliability in Second Pilot Study and Main Study

Verbal Subtest	Second Pilot Study	Main Study	Change
Person Reliability	0.94	0.93	Similar
Person Separation	3.81	3.72	Similar
Child Logit Position	1.22	-0.22	Improved

Visual Subtest	Second Pilot Study	Main Study	Change
Person Reliability	0.72	0.43	Worsened
Person Separation	1.59	0.87	Worsened
Child Logit Position	2.85	2.61	Slightly Improved

Memory Subtest	Second Pilot Study	Main Study	Change
Person Reliability	0.51	0.81	Improved
Person Separation	1.02	2.07	Improved
Child Logit Position	2.09	1.43	Improved

Math Subtest	Second Pilot Study	Main Study	Change
Person Reliability	0.76	0.84	Improved
Person Separation	1.77	2.26	Improved
Child Logit Position	-2.60	-0.12	Improved

Logical Subtest	Second Pilot Study	Main Study	Change
Person Reliability	0.86	0.79	Worsened
Person Separation	2.50	1.95	Worsened
Child Logit Position	1.23	0.95	Slightly Improved

General Knowledge Subtest	Second Pilot Study	Main Study	Change
Person Reliability	0.00	0.73	Improved
Person Separation	0.00	1.66	Improved
Child Logit Position	1.79	1.81	Similar

Validity

Content validity refers to "the degree to which the scores yielded by a test adequately represents the content, or conceptual domain, that these scores purport to measure" (Gall, et al., 1996, P250). One way to establish the content validity for TAR is through a literature review. The researcher conducted an extensive literature review and found both literature outlining different domains for school readiness and extant tests (both for the U.S. and for Thailand) that contain domains similarly found in TAR (Charoensuk, 1973; Cizek, 2001; Doherty, 1997; Fairbank, 1998; Fitzmaurice & Witt, 1989; Ineay, 2004; Kagan et al., 1995; Katz, 1991; Kunesh & Farley, 1993; Linn, 1989; Malunpong,1982; Morrogiello, 1997; Nurss, 1987; ONPEC, 1991; ONPEC, 1998; OPEC,1990; Paget; 1989; Panich,1988; Pinjinda, Jongpayuha, & Charoensuk,1973; Pluksawan,1975; Seefeldt & Barbour, 1994; Sintuwej,1984; Stinnett, 1989). Therefore, the content coverage of the tests reviewed in Chapter 3 was similar to the content coverage of the TAR.

In addition to the literature review, the researcher recruited two groups of experts to provide feedback regarding components of academic readiness and to provide sample items representing each component. The results of expert review suggested that the components included in TAR and their items were representative of the academic readiness domains.

Discussion of Results of Scale Development of the TAR

The scale development of the TAR followed the steps recommended in Chapter 3. By ensuring the TAR was developed based on reviews of the literature regarding

academic readiness for first grade both in the U.S. and in Thailand and of similar readiness tests both developed for the U.S. and for Thailand, the researcher was able to ensure that TAR has necessary and similar characteristics of academic readiness domains and of academic readiness tests reviewed. This is evident in the achieved level of reliability and the established content validity discussed above.

Nonetheless, the achieved level of reliability was sufficient only for proof of TAR as a credible and worthwhile research project for an academic readiness assessment tool. It provides a strong beginning for researchers, who want to further develop the TAR and has adequate psychometric properties for research use. Further development may include recommendations found in the next section. With the achieved level of reliability, some TAR subtests are not ready for use as a tool for school administrators to make decisions about intervention program for individual students. TAR needs to achieve a higher level of reliability in order to provide users of the scale more confidence when making intervention decisions.

While TAR is recommended for limited usage, it has proven to be a good start for a researcher to develop an academic readiness assessment tool for Thailand. TAR's achieved level of reliability, for the most part, has improved since the first pilot study. The review of the literature and of the readiness tests has confirmed that TAR contains similar set of domains found in the literature and the readiness tests. TAR's components include verbal, visual, memory, math, logical, and general knowledge. For example, review of the literature in the U.S. found that Doherty (1997) proposed two components of school readiness (i.e., language skills and general knowledge skills). The two

components fall in line with two of the six academic readiness components found in TAR. Morrogiello (1997) proposed general knowledge as one of the school readiness components. Morrogiello's recommendation falls in line with the general knowledge component of TAR. Nurse (1987) proposed language as a school readiness component. Nurse's recommendation falls in line with the verbal component of TAR.

Review of the literature in Thailand found that Malumpong (1982) proposed four components of school readiness (i.e., verbal, visual, math, and logical). The four components are in line with four academic readiness components found in TAR. Panich (1988) proposed four school readiness components (i.e., verbal, visual, math, and logical). The four components are in line with four academic readiness components found in TAR. Pinjinda, Jongpayuha, and Charoensuk (1973) proposed two school readiness components (i.e., memory and logical). The two components fall in line with two academic readiness components found in TAR.

Review of the tests in the U.S. found that BTBCR measures two similar readiness components (i.e., verbal and logical skills) as those of TAR (Fitzmaurice and Witty, 1986). CBRT-R measures a similar readiness component (i.e., visual skills) as that of TAR (McCarthy, 1985; Proger, 1985). BABERON-2 measures five similar readiness components (i.e., verbal, visual, math, logical, and general knowledge skills) as those of TAR. DIAL-3 measures two similar readiness components (i.e., verbal and logical) as those of TAR (Cizek, 1998; Fairbank, 1998). Review of two tests in Thailand found that LRTCEPS1 and IRTPC measure a similar readiness component (i.e., verbal skills) as that

of TAR. IRTPC, especially, measures a similar set of readiness components (i.e., verbal, visual, math, logical, and general knowledge).

Most tests developed for the U.S. population contain 3-5 subtests with 76-142 items and take from 10 minutes to 1 hour to administer. The two tests developed for the Thai population contain 1 and 9 subtests with 55 and 220 items and take 45 minutes to 1.5 hours to administer. TAR contains six subtests with 595 items and takes 12 hours. Even though TAR contains number of subtests within the similar range as that found in other tests, TAR has more than double the number of items and six times the number of hours needed for administration. As a result, an improvement in TAR in terms of reduction in number of items is necessary in order to reduce time for administration.

Many of the reviewed tests used one or more of the readiness components found in TAR. It should also be noted that DIAL-3 was developed using the Rasch model, which is the same as was done for TAR. Two tests (i.e., CBRT-R, DABERON-2, and IRTPC) achieved a high level of reliability. Three tests (i.e., BDI, DIAL-3, and LRTCEPS1) achieved a minimum of moderate level of reliability. BTBCR achieved a low level of reliability. BDI achieved an excellent rating for validity from test reviewers while BTBCR, DABERON-2, DIAL-3, LRTCEPS1, and IRTPC claimed content validity. CBRT-R and IRTPC claimed concurrent and predictive validity. Therefore, with TAR containing similar readiness components as those tests in the U.S., TAR has a similar level of content validity as found in those tests.

Limitations and Further Research Topics

Small Sample Size

A relatively small sample size was used for in the TAR administration, though the sample was much larger than that selected for development of the IRTPC. While classical test theory requires a smaller sample size for standardization, IRT and the Rasch model dictates a large enough sample size to confidently draw conclusions about reliability. While the researcher was able to recruit 436 participants, such a number is relatively small for a scale development study. In further research, the researcher recommends that a larger sample size be used in order to gain more precision regarding the reliability of the measure and its targeting.

Sampling Method

The researcher used convenience sampling during recruitment of participants. While the number of participants obtained was adequate, the geographical diversification was minimal. In other words, the sample was obtained from graduating kindergarten students at two private schools in Thailand. Although the demographic characteristics of these participants are not at all different from students in other private schools, the characteristics may be different than for public school students or students from other localities (e.g., upcountry provinces). For further study, the researcher recommends that samples be drawn from a wider variety of locations including private and public schools nationwide.

Item Pool Generation

While the researcher was able to create more than double the number of items in main study as compared to that of the second pilot study, the deletion process during the analyses resulted in TAR having gaps in the child/item position map. Relating to the level of reliability obtained for TAR, new items would provide increased levels of reliability as long as the new items prove to represent the domains they are purported to measure and as long as they are thought to fill the gaps. While the former task is rather easy to accomplish by asking content experts to provide additional items thought to measure the intended domains, the latter task is more challenging. As the researcher was revising items following the second pilot study, the researcher recruited a group of experts to provide sample items, which were thought to both measure the domain and to fill the gap. The researcher was disappointed when new items were found to be redundant with the existing ones, to have poor fit, or not to fill the gap. As there have been three versions of TAR (i.e., first pilot, second pilot, and the main study), the researcher is certain that there are more to be developed in order to improve the quality of the scale. For further research, the researcher recommends that more items and in particular more difficult items be generated and administered. The analyses will provide insights into whether the new items contribute to the better quality of a revised version of TAR.

Predictive Validity

This study did not perform predictive validity analyses. Therefore, there is a clear limit in TAR's ability to predict the level of success for first grade instruction in Thailand

for children who take TAR and earn scores signifying readiness for first grade instruction. In order to gain a wider acceptance of a valid and accurate measure (or predictor) of academic readiness for first grade instruction in Thailand, predictive validity needs to be established. This can be done by correlating the results of TAR scores for each child with his or her scores in first grade. If there is a sufficient correlation between TAR scores and the scores in first grade, it can be inferred that TAR provides a valid prediction of success (due to the child's academic readiness) in first grade in Thai schools.

Standardization of TAR

This study did not include a norming process, which should be performed by region in Thailand, gender, and age. Once the TAR is normed, the scores received from TAR administration can be used to infer a child's ability based on a wider population, based on both genders, and based on a more focused age group entering first grade level. In addition, in order to ensure that there is not a difference in children's scores due to difference in administration procedures, a standard administration manual needs to be developed. More specifically, a manual that includes but is not limited to the purpose, number of subtests involved, types of items, question and item formats, briefing and debriefing procedure, answer sheets and scoring instructions. Once a manual is created, scores obtained from TAR administration will have a higher level of credibility as well.

TAR Administration

As previously mentioned, TAR contains too many items and takes too long to administer. It is discouraging for users of readiness test to consider using TAR. The preparation and the time requirements of the participants and of the research sites increase as the time to administer TAR increases. Such factors become drawbacks for TAR. It is therefore necessary to further develop TAR by removing additional redundant items to possibly create a shorter version of TAR.

Bibliography

Aonjumras, P. (1985). *Development of teaching package through systemic analysis to increase level of readiness and ability in writing Thai alphabets in first graders.* Graduate Thesis, Department of Elementary Education. Faculty of Education. Srinakarinwirot University, Bangkok, Thailand.

Axford, S. N. (1992). Review of the DABERON-2: Screening for School Readiness. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurement yearbook* (pp. 256-259). Lincoln, NE: The Buros Institute of Mental Measurements.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *The American Journal of Occupational Therapy, 36,* 789-799.

Bode, R. K., & Wright, B. D. (1999). Rasch measurement in higher education. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research.* (Vol.14, pp. 287-316). Edison, NJ: Agathon Press.

Boonruang, S. (1991). *Intellectual Readiness Test for Pre-school Children.* Graduate Thesis, Faculty of Education, Chingmai University, Chaingmai, Thailand.

Boonsawat, P., Issarangkul Na Ayudhaya, I., Laungsuwan, P., & Tosupan, S. (1980). *Learning Development in Kindergarten Children.* Bangkok, Thailand: Erawan Publishing.

Bradley, G. (1984, April). Ways to help your child succeed in kindergarten. *PTA Today, 9,* 11-12.

Bupawes, T. (1984). *Studies of relationship between readiness and grades.* Graduate Thesis, Faculty of Education, Konkein Univeristy, Konkein, Thailand.

Charoensuk, S. (1986). *Educational Psychology: Supplementary Training.* Bangkok, Thailand: Charoenwit Publishing.

Chonchop, S. (1982). *Thai language teaching in elementary level.* Department of Elementary Education, Faculty of Education, Chiangmai University, Chiangmai, Thailand.

Chuthai, P. (1982). *Psychology in teaching.* Bangkok, Thailand: Faculty of Education, Kasetsart University.

Cizek, G. J. (2001). Review of the Developmental Indicators for the Assessment of Learning (3rd Ed.). In B. S. Plake & J. C. Impara (Eds.), *The fourteenth mental measurement yearbook* (pp. 394-398). Lincoln, NE: the Buros Institute of Mental Measurements.

Clymer, T. & Barrett, T. C. (1966). *The Clymer-Barrett Readiness Test Revised Edition.* Santa Barbara, CA: Chapman Brook and Kent.

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

Conoley, J. C., & Kramer, J. J. (Eds.). (1989). *The tenth mental measurement yearbook.* Lincoln: NE: the Buros Institute of Mental Measurements.

Cronbach, L. J. (1970). *Essentials of Psychological Testing.* (3rd ed.). New York: Harper & Row.

Danzer, V. A., Gerber, M. F., Lyons, T. M., & Voress, J. K. (1967). *DABERON-2 Screening for School Readiness.* Austin, TX: PRO ED Inc.

De Cos, P. L. (1997, December). *Readiness for Kindergarten: What Does It Mean?* Sacramento: CA: California Research Bureau.

Department of Curriculum and Instruction Development. (1997). *Pre-Elementary Curriculum.* Bangkok, Thailand: Ministry of Education.

DeVellis, R. F. (1991). *Scale Development.* Newbury Park, CA: SAGE.

Doherty, G. (1997, May). *Zero to Six: The Basis for School Readiness (Technical Paper R-97-8E).* Quebec, Canada: HRDC Publication Centre.

Downing, J. & Thackrey, D. (1971). *Reading Readiness.* New York: London, University of London Press.

Dunn, L. M., and Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised (Manual).* Minnesota: American Guidance Service.

Elliot, C. D. (1983). *British Ability Scales (Manual 1: Introductory Handbook).* Windsor, Berks: NFER-Nelson Publishing Company Ltd.

Fairbank, D. W. (2001). Review of The Developmental Indicators for the Assessment of Learning (3rd Ed.). In B.S. Plake & J. C. Impara (Eds.), *The fourteenth measurement yearbook* (pp. 394-398). Lincoln, NE: the Buros Institute of Mental Measurements.

First Grade Exams (1997). Bangkok, Thailand: Books Athen.

Fitzmaurice, C., & Witt, J. C. (1989). Review of Boehm Test of Basic Concepts-Revised. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurement yearbook* (pp. 98-102). Lincoln, NE: the Buros Institute of Mental Measurements.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational Research: An Introduction, Sixth Edition.* White Plains, NY: Longman Publishers.

Gardner, H. (1993). *Frames of Mind: The Theories of Multiple Intelligences.* NY: Basic Books.

Good, C. V. (1973). *Dictionary of Education.* New York: McGraw-Hill.

Gracey, G. R., Carey, V. A., & Reinherz, H. (1984, Summer). Screening Revisited: A Survey of U.S. Requirements. *The Journal of Special Education, 18,* 101-107.

Green, K .E. (1996). Applications of the Rasch model to evaluation of survey data quality. *New Directions for Evaluation, 70,* 81-92.

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Education Measurement: Issues and Practice, 12,* 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, R. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

Helair, F. (1996). *Daroon Suksa, Pre-elementary.* Bangkok, Thailand: Thai Watanapanit, Ltd.

Hughes, S. (1992). Review of the DABERON-2: Screening for School Readiness. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurement yearbook* (pp. 256-259). Lincoln, NE: The Buros Institute of Mental Measurements.

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

Ineay, K. (2004). *Development of Language Readiness Test for Continuing Study in Prothom Suksa 1*. Graduate Thesis, Faculty of Education, Chiangmai University, Chaingmai, Thailand.

Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington DC: American Council on Education.

Kagan, S. L. (1990, December). Readiness 2000: Rethinking Rhetoric and Responsibility. *Phi Delta Kappan, 72,* 272-279.

Kagan, S. L. (1992, November). Readiness past, present and future: Shaping the agenda. *Young Children, 48,* 48-53.

Kagan, S. L., Moore, E., & Bredekamp, S. (Eds). (1995, June). *Reconsidering Children's Early Development and Learning: Toward Common Views and Vocabulary.* Washington, DC: National Education Goals Panel.

Kangpenkae, N., & Tonnue, W. (1986). *Guidelines for Assessment and Evaluation of Academic Readiness.* Bangkok, Thailand: Kurusapaladprao Publishing.

Kaosim, J. (1994). *Studies of influence of types of creative and corner activities on creativity and problem-solving skills in pre-elementary students.* Konkein Kindergarten, Konkein, Thailand.

Katz, L. G. (1991). Readiness: Children and Schools. *Eric Digest.* Urbana, IL: Eric Clearinghouse on Elementary and Early Childhood Education.

Kisawatkon, R. (2000). *Relationship between child rearing and readiness in first year kindergarteners.* Graduate Thesis, Department of Educational Psychology, Faculty of Education, Mahasarakam University, Mahasarakam, Thailand.

Kramer, J. J., & Conoley, J. C. (Eds.). (1992). *The eleventh mental measurement yearbook.* Lincoln: NE: the Buros Institute of Mental Measurements.

Kunesh, L. G., & Farley, J. (1993). Collaboration: The Prerequisite for School Readiness and Success. *Eric Digest.* Urbana, IL: Eric Clearinghouse on Elementary and Early Childhood Education.

Lamberty, G., & Crnic, K. (1994, April). School Readiness Conference: Recommendations. *Early Education and Development, 5,* 165-176.

LeCompte, M. D. (1980, Fall). "The Civilizing of Children: How Young Children Learn to Become Students." *Journal of Thought, 15,* 105-27.

Lewit, E. M., & Baker, L. S. (1995, Summer/Fall). School Readiness. *The Future of Children, 5,* 128-139.

Linn, R. L. (1989). Review of Boehm Test of Basic Concepts-Revised. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurement yearbook* (pp. 98-102). Lincoln, NE: the Buros Institute of Mental Measurements.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Welsley Publishing Company.

Mardell, C. C. & Goldenberg, D. S. (1983). *Developmental Indicators for the Assessment of Learning (3rd Ed.).* Pines, MN: American Guidance Service Inc.

Maikeaw, B. (2000). *Factors considered by parents while selecting a Bangkok public kindergarten.* Graduate Schools, Kasetsart University, Bangkok, Thailand.

Malumpong, W. (1982). *Assessment of Kindergarten Learning.* Bangkok, Thailand: Kurusapaladprao Publishing.

Mati Silapin, B. (1987). *School Readiness Workbook: Motor, Visual, Memory, and Logical.* Bangkok, Thailand: Aksorn Bandit, Ltd.

McCarthy, J. (1985). Review of Clymer-Barrett Readiness Test Revised Edition. In J. V. Mitchell (Ed.). *The ninth mental measurement yearbook* (pp. 347-350). Lincoln, NE: the Buros Institute of Mental Measurements.

Mitchell, J. V. (Ed.). (1985). *The ninth mental measurement yearbook.* Lincoln: NE: the Buros Institute of Mental Measurements.

Moopung, J. (1982). *Kindergarten Education.* Bangkok, Thailand: Tipauksorn Publishing.

Morrongiello, B. A. (1997, September). *Tapping School Readiness in the NLSCY: Measurement Issues and Solutions (Technical Papers T-98-1E).* Quebec: Canada: Human Resources Development Canada.

Newborg, J., Stock, J. R., Wnek, L., Guidubaldi, J., & Swinicki, J. (1984). *Battelle Developmental Inventory.* Chicago, IL: The Riverside Publishing Company.

Nilarun, P. (1987). *Thai language learning readiness in kindergarteners of language problem localities through language experiences.* Graduate Thesis, Department of Kindergarten Studies, Faculty of Education, Srinakarinwirot University, Bangkok, Thailand.

Nilwichian, H. (1989, October). *"Early Age Education: Who is it for?" Sanpatanaluksoot, 91*(3), 35-39.

Nurss, J. R. (1987). Readiness for Kindergarten. *Eric Digest.* Urbana, IL: Eric Clearinghouse on Elementary and Early Childhood Education.

Nurss, J. R., & Hodges, W. L. (1982). Early Childhood Education. In H. E. Mitzel (Ed.), *Encyclopedia of Educational Research.* (5th ed., Vol. 2, pp. 489-507). New York: The Free Press.

Office of the National Education Commission (1994). *Educational Administration Manual for Kindergarten Education.* Bangkok, Thailand: Kurusapa Ladprao Publishing.

Office of the National Education Commission. (1999). *Learning in Kindergarteners.* Bangkok, Thailand: Seven Printing Group Co., Ltd.

Office of the National Primary Education Commission. (1982). *Research findings in Management of Pre-school Center in Thailand.* Bangkok, Thailand: Chaorenpol Publishing.

Office of the National Primary Education Commission. (1983). *Management of Pre-school Center.* Bangkok, Thailand: Thaiwattanapanit Publishing.

Office of the National Primary Education Commission. (1991). *Assessment and evaluation manual for pre-elementary students.* Bangkok, Thailand: Kurusapaladprao Publishing.

Office of the National Primary Education Commission. (1998). *Educational Administration Manual for Kindergarten Education.* Bangkok, Thailand: Kurusapa Ladprao Publishing.

Office of Private Education Commission (1990). *Developmental Assessment Manual for Kindergarten 1-3.* Bangkok, Thailand: Srimuang Publishing.

Nontapuk., P. (2009). *Logical Exercises for Primary One Entrance Exam.* Bangkok, Thailand: Rung Ruang Sarn.

Paget, K. D. (1989). Review of Battelle Development Inventory. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurement yearbook* (pp. 66-72). Lincoln, NE: the Buros Institute of Mental Measurements.

Panich, S. (1988). *Preparing early age children.* Rachaburi Community School, Ratchaburi, Thailand.

Panpum, W. (2004). *Development of assessment instrument for social-emotional readiness in kindergarteners through plays.* Graduate Thesis, Department of Measurement and Evaluation, Faculty of Education, Konkein University, Konkein, Thailand.

Pengsawat, W. (2001). *Research in Kindergarten Education.* Bangkok, Thailand: Suweriyasas Publishing.

Pinjinda, S., Jongpayuha, N., & Charoensuk, S. (1973). *Manual of Developmental Studies.* Bangkok, Thailand: Pikanes Publishing.

Pinyo Anantapong, S. (1993). *Tests of Spatial Skills, Form 3.* Bangkok, Thailand: Thai Watanapanit, Ltd.

Pinyo Anantapong, S. (1996). *Tests of Thai Language.* Bangkok, Thailand: Thai Watanapanit, Ltd.

Pluksawan, B. (1975). *Teaching guidelines in first grade and early grade teacher manual.* Bangkok, Thailand: Thaiwatanapanit.

Pangwiruitrak., C. (2007). *Primary One Entrance Exam Exercises.* Bangkok, Thailand: Duang Kamol Samai.

Prawalpruk, W. (1975). Assessment in Kindergarten. *Journal of Education, 4,* 3-18.

Proger, B. B. (1985). Review of Clymer-Barrett Readiness Tests Revised Edition. In J. V. Mitchell, Jr. (Ed.), *The ninth measurement yearbook* (pp. 347-350). Lincoln, NE: the Buros Institute of Mental Measurements.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Ratana, P. (1992). *Kindergarten Education Management.* Bangkok, Thailand: Kurusapaladprao Publishing.

Roid, G. H., & Miller, L. J. (1997). Leiter International Performance Scale-Revised: Examiner's Manual. In G. H. Roid and L. J. Miller (Eds.), *Leiter International Performance Scale-Revised.* Wood Dale, IL: Stoelting CO.

Sang Asanee, A., Boon Urapeepinyo, S., Wong Wijit Sin, S., Ruji Rek, N., & Apichartimanon, P. (1990). *The Master Group's Thai Workbook, Series 5, Grade A1. Bangkok,* Thailand: Institute of Curriculum and Instruction Development.

Sangmali, B. (1986). *Day Care Center and Kindergarten Management.* Nontaburi, Thailand: Sukothaitamatirat Publishing.

SECA Public Policy Institute Report (1993, Fall). Children Are Born Learning: Schools Must Make Ready to Celebrate and Nurture What Children Instinctively Can and Will Do. *Dimensions, 22,* 5-8.

Seefeldt, C., & Barbour, N. (1994). *Early Childhood Education: An Introduction.* Netherlands: Swets and Zeitlinger B.V.

Setsukko, T (1981). *School Readiness Training.* Bangkok, Thailand: Thaiwatanapanit.

Sintuwej, S. (1986). *Assessment of language skills in elementary level* (2nd ed.). Bangkok, Thailand: Office of Educational Testing, Division of Academic Development.

Snyder, S., & Sheehan, R. (1992,Winter). Research Methods. The Rasch Measurement Model: An Introduction. *Journal of Early Intervention, 16,* 87-95.

Stark, S., & Chernyshenko, S. (2001, April 28). Methods for detecting differential item/test. Slides presented at the 2001 Society for Industrial and Organizational Psychology Symposium on a Practical Guide to IRT. PowerPoint presentation retrieved on April 24th, 2009. from http://work.psych.uiuc.edu/irt/IRT%20Tutorial_Intro.ppt

Stinnett, J. O. (1989). Review of Battelle Development Inventory. In J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurement yearbook* (pp. 66-72). Lincoln, NE: the Buros Institute of Mental Measurements.

Suwanatat, J. (1996). *What a good kindergarten should teach kids.* Bangkok, Thailand: Plan Printing Co., Ltd.

Tangjitsomkit, W. (1996). *Education in Thailand.* Bangkok, Thailand: Odion Store Press.

Tantrirat, S. (1988). Teaching Thai language in elementary level. Department of Elementary Education. College of Education. Konkein University.

Thorndike, R. M. (1997). *Measurement and Evaluation in Psychology and Education, Sixth Edition.* NJ: Prentice Hall, Inc.

Tongdee, S., & Kanjanakij, S. (1994). *Evolution of Kindergarten Education. Teaching materials for principle and thoughts on kindergarten education.* Nontaburi, Thailand: Sukothaitamatirat University.

Tongsawat, R. (1994). *Academic Administration in Kindergarten Education.* Nontaburi, Thailand: Sukothaitamatirat University.

Trium Sop Por 1 (2009). *Primary One Entrance Exam Exercises.* Bangkok, Thailand: Poom Bundit.

Wai Prip Trium Sop (2009). *Techniques for Entrance Exam.* Bangkok, Thailand. Pailin Dek Keng.

Witantam, W. (1990). *Expectation of Kindergarten Curriculum from Parents in Public and Private Kindergarten in Bangkok.* Graduate Thesis, Faculty of Education, Chulalongkorn Univeristy, Bangkok, Thailand.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design.* Chicago, IL: Mesa Press.

Yimyong, S. (2001, April). Organization of Kindergarten Experiences. *Journal of Kindergarten Education, 5,* 46.

Appendices

Appendix 1

First Pilot Panel of Content Experts

Names	Education	Experience	Years	Current Position
Mrs. Kwunyeon Kasintorn	B.A. in Education	Teaching, Curriculum Planning, Educational Administration	31	School Director
Ms. Jitinan Nitatsatian	B.A. in Education-Kindergarten	Teaching	15	Kindergarten Teacher
Ms. Umporn Saisee	B.A. in Nutrition	Teaching	16	Kindergarten Teacher

Appendix 2

Suggestions from Expert Panels for First Pilot Items

Suggestions
Suggestions from Content Experts 1. Items are too easy. Need more difficult items in Math and Verbal 2. Some sections need more items 3. The level of difficulty of some other items in motor, visual, and logical-mathematical domains is moderate to high.

Appendix 3

List of Sample Items Provided by the Content Experts

Verbal

RV-01	กวาดบ้านถูบ้าน
RV-02	สระว่ายน้ำ
RV-51	อวัยวะ
RV-52	สวรรค์
WD-39	มหาศาล
WD-40	ปฐมพยาบาล

มานีอยู่กับพ่อ ตอนเช้ามานีออกไปช่วยพ่อทำนา ในนามีกาลงมากินข้าว มานีชอบตกปลาเอามาเป็นอาหาร

RC-01. มานีอยู่กับใคร

ก.ตา ข.พ่อ ค.อา

RC-02. มานีออกไปทำนาเวลาใด

ก.ตอนเย็น ข.ตอนเช้า ค.ตอนเที่ยง

Math

MCV-01	จงเขียน “ห้าลบสอง” เป็นเลขอาราบิค
MCV-15	จงเขียน “หกสิบสี่ลบสิบเอ็ดๆเจ็ดสิบห้า” เป็นเลขอาราบิค
MC-08	798 + 235 = ___
MC-09	608 + 456 = ___
MC-28	___ - 67 = 37
WP-20	เราใช้แบงค์ยี่สิบซื้อน้ำส้มสองแก้ว ๆ ละ สามบาท เราต้องได้เงินทอนเท่าใด?
WP-22	มีแอบเปิ้ลสิบเก้าผล เราต้องการแอบเปิ้ลทั้งหมดยี่สิบห้าผล เราต้องการแอบเปิ้ลเพิ่มอีกกี่ผล?

Appendix 4

Second Pilot Panel of Content Experts

Names	Education	Experience	Years	Current Position
Mrs. Kwunyorn Kasintorn	B.A. in Education	Teaching, Curriculum Planning, Educational Administration	33	School Director
Ms. Jitinan Nitatsatian	B.A. in Education-Kindergarten	Teaching	17	Kindergarten Teacher
Ms. Umporn Saisee	B.A. in Nutrition	Teaching	18	Kindergarten Teacher
Ms. Nutchanok Seilim	B.A. in Education	Teaching	9	First Grade Teacher
Ms. Nongluk Klinhuan	B.A. in Education	Teaching	6	First Grade Teacher

Appendix 5

List of Questions for Content Experts for the Second Pilot Study

Questions
1. Please list the major content areas that are taught in kindergarten/first grade.
2. Please categorize those content areas in terms of first grade subject matters
3. Please review the second pilot items.
4. For each item, please indicate which subject matter category the item belongs. (For the items that do not belong to any category, please leave them alone).
5. Is there a category to which no item belongs?
6. If the answer the number 5 is yes, please provide a sample of items that will best represent the subject matter category?

Appendix 6

Second Pilot Panel of Experts

Names	Expertise
Dr. Kathy Green	Psychometric Theories
Dr. Marty Tombari	Measurement and Evaluation
Dr. Gloria Miller	Child Development
Dr. Lucretia Peebles	Curriculum and Instruction
Ms. Alisa Kasintorn	Child Development

Appendix 7

Suggestions from Other Experts

Suggestions
1. Certain instructions are not clear (e.g., motor, visual5)
2. Poor test materials. (e.g., circles in orange and in red looked very much the same. This could have confused the examinees. The size of small and medium circles is also too close. This also could have confused the examinees. Finally, the quality of the recorded sound used for music domains is poor. Certain sounds were unintelligible.)
3. Too much redundancy of items
4. Too long a test for young children.
5. Lack of clear justification of domain inclusion. (e.g., unclear what construct (or content) the domains were trying to measure.)
6. The test needs some guidelines for discontinuing the administration in the case of examinees not answering a series of questions correctly.

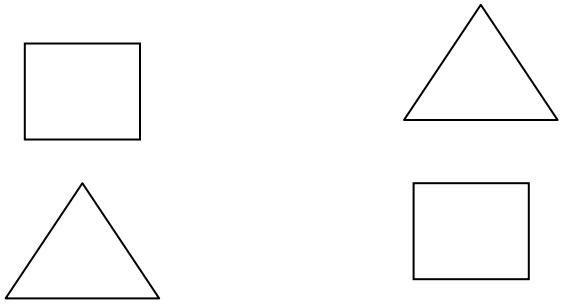
Appendix 8

List of Sample First Pilot Items

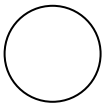
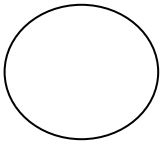
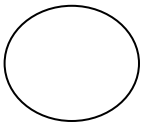
Subtests

Visual Discrimination

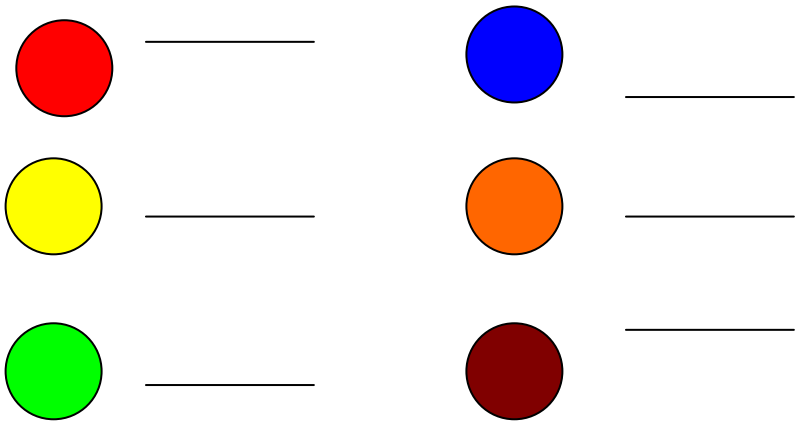
The examiner will provide the student with two columns that contain different shapes: circle, oval, square, diamond, and triangle. Then, the examiner will ask the student to connect the same shape across the column with a straight line.



The examiner will ask the student to identify the biggest and the smallest circles from the 5 circles provided.

1.  2.  3.  4.  5. 

The examiner will ask the student to identify the color of 6 circles provided below



Verbal Skills

The examiner will read to the student a series of words. The student will be asked to spell the word after each word is pronounced.

1. Car
2. Van
3. Jar
4. Walk

The examiner will provide the student with a story (50 words), which is provided below. The student will be asked to read the story out loud.

"Linda is a student. She walks to school everyday. She enjoys playing with her friends at lot. She loves to help her mom prepare dinner after school. She likes to read with her mom. Her mom loves to read to her before bedtime. Linda goes to bed at 8 o'clock."

Logical-Mathematical

The examiner will ask the student to complete the sentences.

1. A cat was running because_____
2. I love my mom because_____

The examiner will ask student to provide answer to following questions as an indication of the student's practical reasoning.

1. What do you do when you are hungry_____
2. What do you say when somebody say something no nice to you_____
   _____

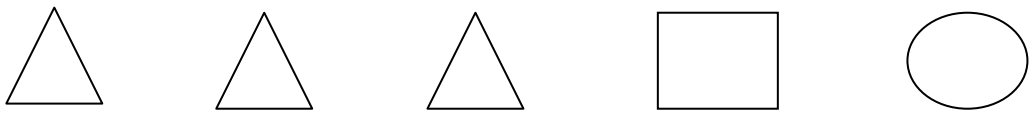
The examiner will ask the student to complete the following addition and subtraction tasks.

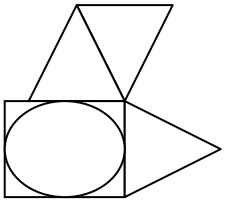
12	4	10	27	52	89
3+	5+	6+	21+	32+	15-
____	____	____	____	____	____

The examiner will have the student read the following word problems. The student will be required to perform computation to arrive at the correct answers.

1. There is one orange on a table. Your mom put another orange on the table. How many orange are there on the table?
   Spatial Skills

The examiner will provide the student with varieties of cardboard-cut shapes. The students will be required to arrange the shapes into the picture depicted for each tasks.





Appendix 9

Number of Items for Each Subtest in First Pilot Study

Subtests	Number of Items
Visual	
Subtest I	3
Subtest II	15
Subtest III	3
Subtest IV	6
Subtest V	3
Subtest VI	12
Subtest VII	5
Verbal	
Subtest I	38
Subtest II	25
Subtest III	75
Subtest IV	5
Subtest V	1
Subtest VI	40
Subtest VII	55
Math	
Subtest I	7
Subtest II	5
Subtest III	22
Subtest IV	4
Spatial	
Subtest I	4

Appendix 10

Example of an Input File for Math Subtest for First Pilot Study

```
&INST
TITLE='Art Math'
;;Deleted: None
NI=38
ITEM1=5
NAME1=1
CODES=012
TABLES=1110011001101000000l000
PERSON=KID
ITEM=TASK
IDELQU=Y
PFILE=ARTVIS.PF
IFILE=ARTVIS.IF
&END
MTH_1_1
MTH_1_2
MTH_1_3
MTH_1_4
MTH_1_5
MTH_1_6
MTH_1_7
MTH_2_1
MTH_2_2
MTH_2_3
MTH_2_4
MTH_2_5
MTH_3_1
MTH_3_2
MTH_3_3
MTH_3_4
MTH_3_5
MTH_3_6
MTH_3_7
MTH_3_8
MTH_3_9
MTH_3_10
MTH_3_11
MTH_3_12
MTH_3_13
MTH_3_14
MTH_3_15
MTH_3_16
MTH_3_17
MTH_3_18
MTH_3_19
MTH_3_20
MTH_3_21
MTH_3_22
MTH_4_1
MTH_4_2
MTH_4_3
MTH_4_4
END NAMES
A1  22222222222222222220222020222220022200
A2  22222222022222222222002200220200222202
A3  22222222222022222222222020222222222202
A4  20222222022222222222220200222222222220
A5  22222222222222222222222222222222222220
A6  22222222222222222222222222222222022222
A7  22222022222222222222220202222222222220
A8  22402222222022221222222210222222222224
A9  22422222124422022202222020222222222221
A10 21212222420522222222221222222222121222
A11 21222222122222202022422202222112222211
A12 22222225212222222022222222220222200
A13 22222221222220222222202002222222222200
A14 41402222412122222222222221222222122222
```

A15 22222222222222222222220202222222222221
A16 21222222422422222222222222222221222221
A17 12222222222222220222222222222222222222
A18 21222222222122222210202020222212222222
A19 22222222212222222002222000221221222241
A20 12222222522200220222222222252222211
A21 22242222422222222220222222220222222201
A22 22222222222222222222222220222222221212
A23 12222222222222222222221222222222222211
A24 22202222222222222222222222222222222222
A25 22222222222222122222022222222222122221
A26 22222222222222212022122222222222222222
A27 55102222502522222222222210002220122115
A28 10222222211522222222222022222222222211
A29 20202222222222222222222222221222222110
A30 20202222242022022222002200212222222212
A31 55252122522022222222222222222222222202
B1 22222222442420022002222222220222222404
B2 12222222242222220202220202220222222240
B3 22222222220222222222222220222222222202
B4 22222222222222222222222222220222221220
B5 22222222520522222222002222222220222202
B6 22402222224022221222022102242222222211
B7 22212222222222222220022200222222200000
B8 01222222552122222220022220222222222220
B9 21240222411222020202222222222220221220
B10 52220222222422222022222220222222222202
B11 22220142221122222222200222222222222220
B12 22222220220222222222222222222220222222
B13 22222222221222222022222202222202222221
B14 22222222222122022222020202222222222202
B15 22222222222222222222222222222222222221
B16 41210222122222222222222222220222222211
B17 22202222222222202222222222222222222200
B18 22222222202122220222222222222222222221
B19 21122221422222222222222222222222222220
B20 22222222222222222022212122220022222201
B21 22222222112222222222222222221222202212
B22 22222222110222222222222222222022222200
B23 22222222222020022022222200220220221102
B24 20202222122222200202202222222020022221
B25 21152222522222220202022220222222222200
B26 22222222221122212222200022222221220200
B27 12202222222222222222222200222222220200
B28 22022222221222222222220202222222220200
B29 22200222022122222222222222222222022002
B30 21122222122222222222202022222222220222

Appendix 11

List of Sample Second Pilot Items

Subtests
Verbal
Reading Vocabulary
อธิการบดี
อุทยาน
อันธพาล
Writing Dictation
กระทรวง
การเกษตร
กาญจนา
Reading Comprehension

วันหนึ่งกระรอกไปเที่ยวที่สวนสาธารณะ และเห็นนก
จึงตัดสินใจแอบในพุ่มไม้ใกล้ต้นไม้นั้น หลังจากที่นกบินไปแล้ว
กระรอกจึงออกมาจากพุ่มไม้และเริ่มปีนขึ้นไปบนต้นไม้ หลังจากที่อยู่บนต้นไม่
กระรอกเห็นผลไม้มากมายห้อยจากกิ่งก้านของต้นไม่ กระรอกจึงตัดสินใจเด็ดผลไม้กิน
หลังจากกินเสร็จ กระรอกจึงเริ่มเก็บผลไม่เพื่อนำกลับบ้านให้ครอบครัวของมัน ระหว่างทาง
กระรอกเห็นกระรอกอีกตัวหนึ่งที่กำลังหิวโหย กระรอกจึงแบ่งผลไม้ให้กระรอกตัวนั้นไป
เมื่อกระรอกมาถึงบ้าน มีเพียงพี่กระรอกสองตัวอยู่บ้าน พ่อ แม่ และน้องสาว ออกไปหาอาหาร
และจะกลับมาในตอนค่ำ
กระรอกจึงตัดสินใจนำผลไม้ไปเก็บและเอาบางส่วยออกมาแบ่งพี่ทั้งสอง
หลังจากนั้นกระรอกรู้สึกเหนื่อย จึงหลับไป ขณะที่หลับอยู่นั้น
กระรอกฝันว่านกกำลังจะมาทำร้าย มันจึงตกใจตื่นขึ้น.

1. เรื่องเป็นเรื่องเกี่ยวกับอะไร

    ก. นก

    ข. หนู

    ค. กระรอก

    ง. สุนัข

2. กระรอกอยู่ตรงไหนขณะที่มันเห็นนก?

    ก. บ้าน

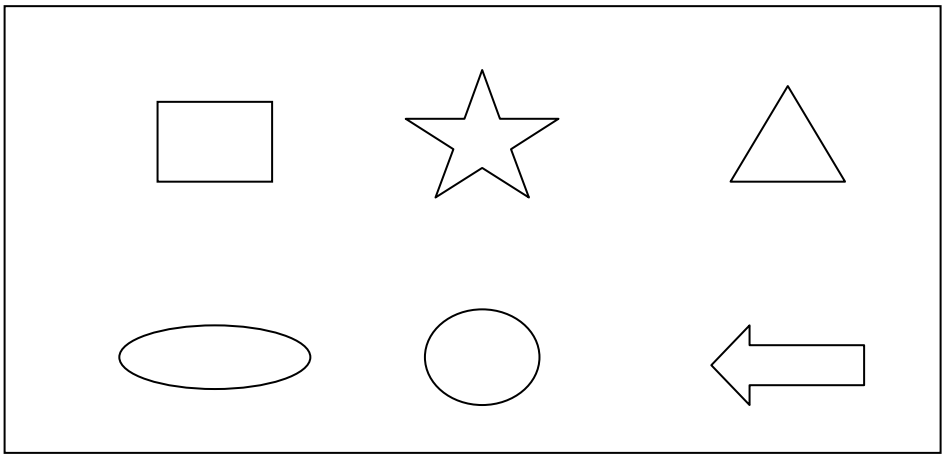
    ข. ถนน

    ค. บนต้นไม้

    ง. ในสวน

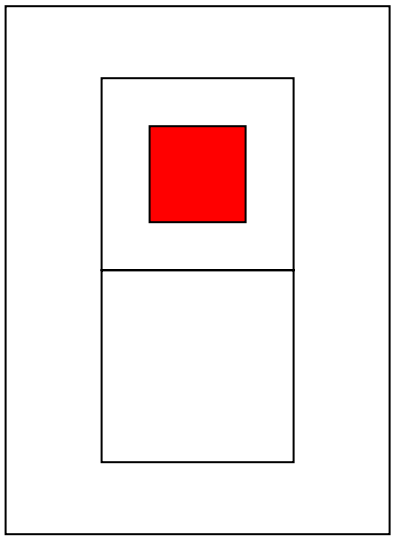
Visual

Visual Identification

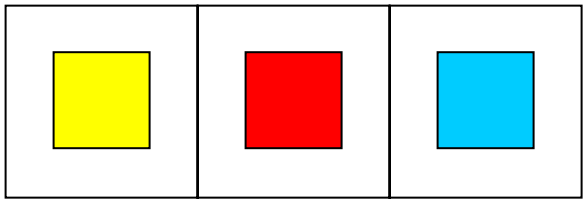


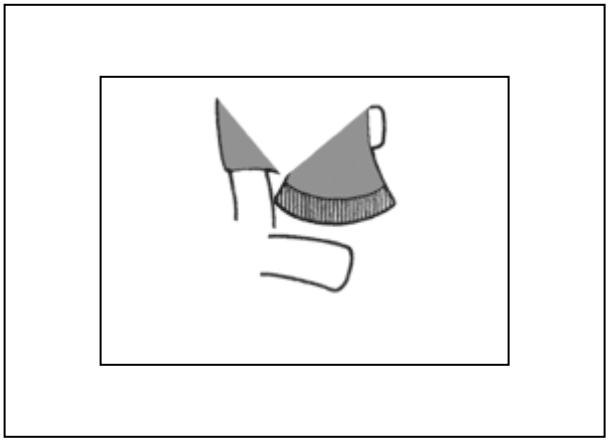
Please point to:
1. An arrow
2. A triangle

Visual Matching

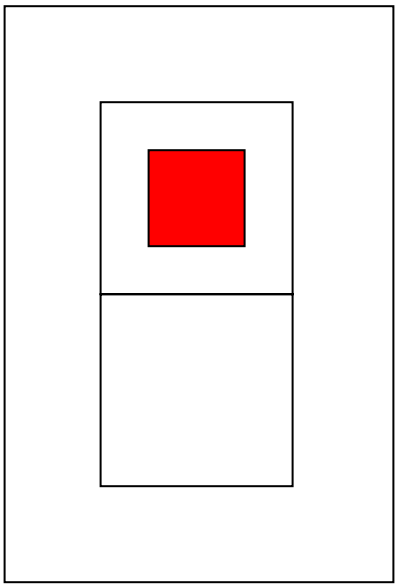


1. **2.** **3.**

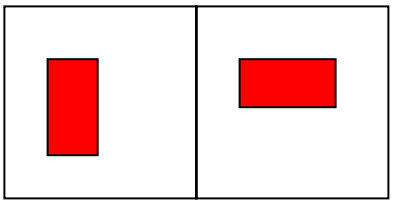




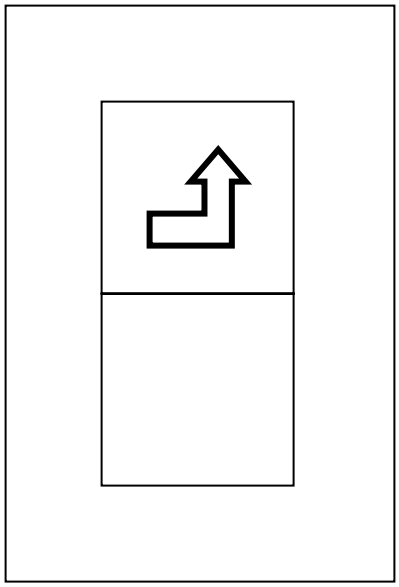
Mental Folding



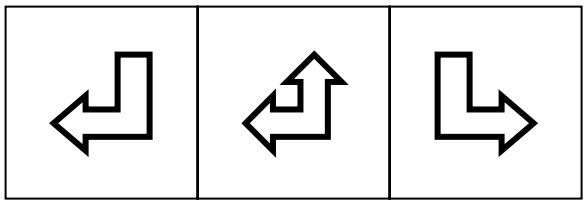
1. **2.**



Mental Rotation



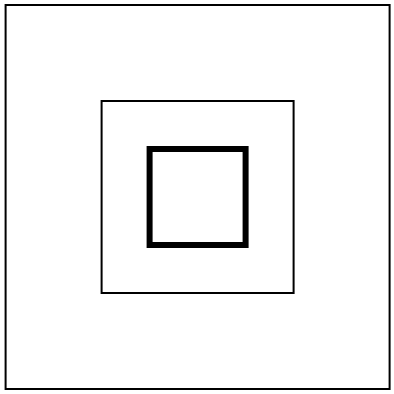
1. **2.** **3.**



Memory

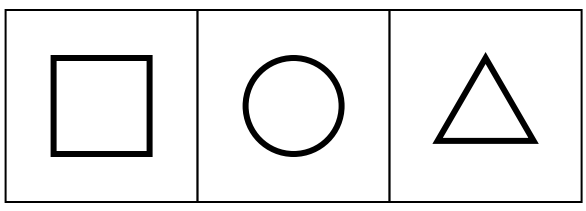
Immediate Recognition

Page 1



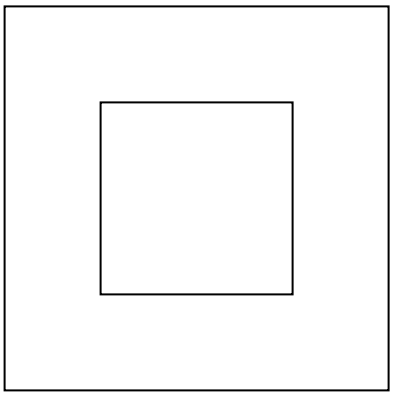
Page 2

1. **2.** **3.**



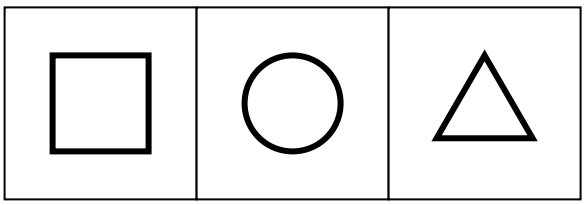
Delayed Recognition

Page 1



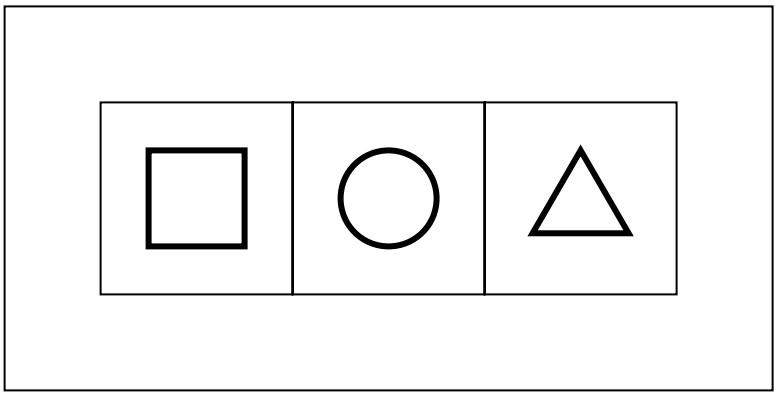
Page 2

1. **2.** **3.**

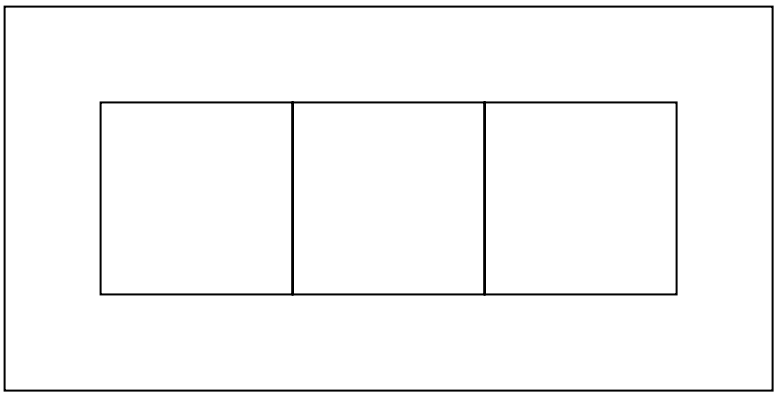


Spatial Relation

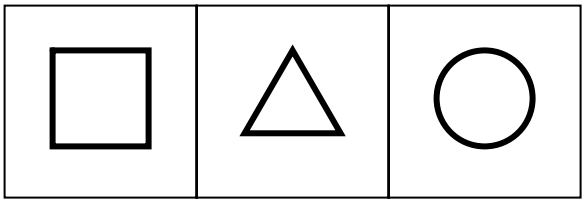
Page 1



Page 2



1. **2.** **3.**



Math

Math Concept and Vocabulary

1. จงเขียน 1+1
2. จงเขียน 2+3

Math Computation

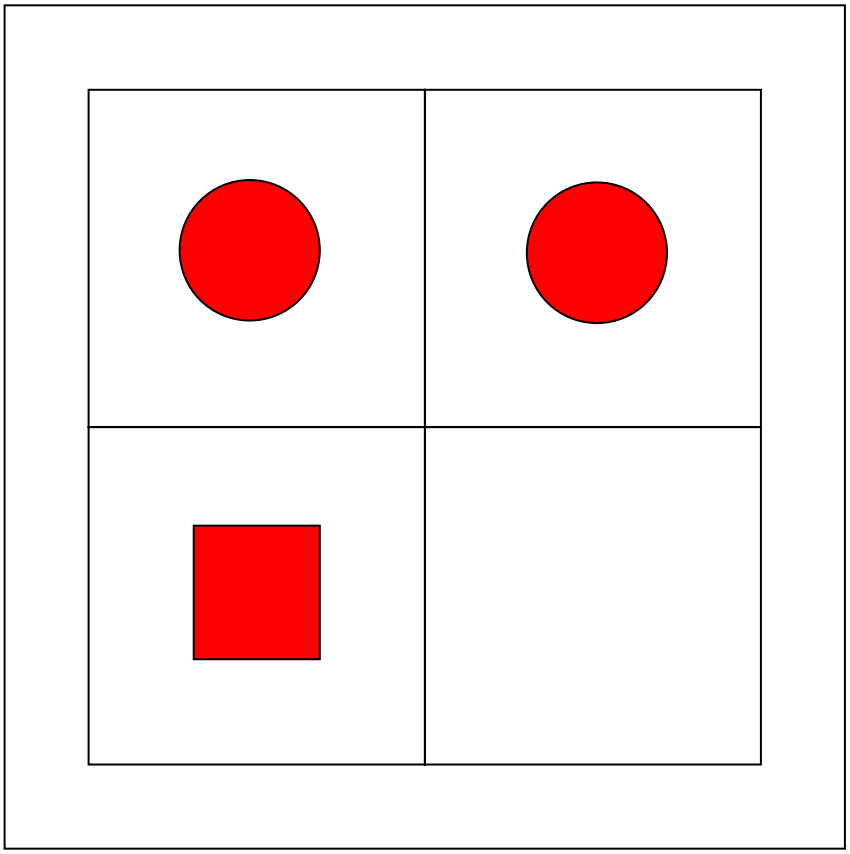
1. 52 - 15 = ___
2. 608 + 456 = ___

Word Computation

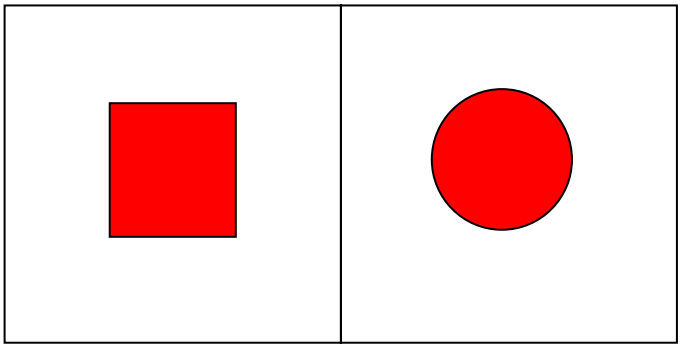
1. ฉันมีลูกแล้ว 12 ลูก แม่ฉันให้มาอีก 24 ลูก พี่ฉันยืมไป 18 ลูก ฉันมีลูกแก้วเหลือกี่ลูก?
2. ปีเตอร์เริ่มหยอดกระปุกอาทิตย์ละ 1 บาท ทุกสุดสัปดาห์แม่ของปีเตอร์จะหยอดเพิ่มให้อีก 2 บาท หลังจากครบ 2 เดือน ปีเตอร์มีเงินในกระปุกเท่าใด?

Logical

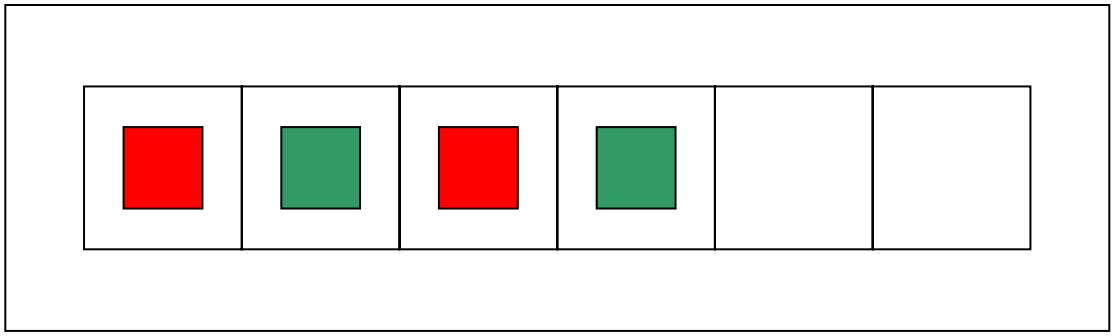
Concept Formation



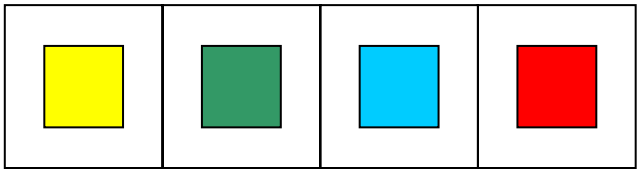
1. 2.



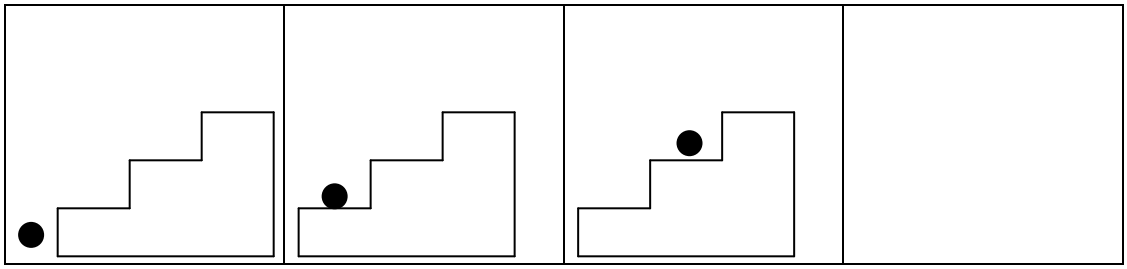
Pattern Finding

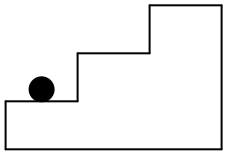


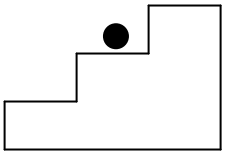
1. 2. 3. 4.

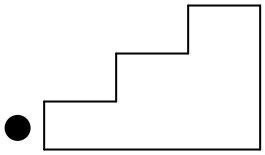


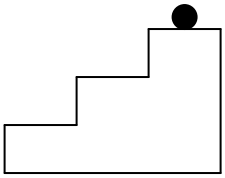
Sequential Order



1. 

2. 

3. 

4. 

General Knowledge

1. นก_____; สุนัข_____

a) ว่ายน้ำ, บิน
b) บิน, ว่ายน้ำ
c) วิ่ง, เดิน
d) บิน, เดิน

Appendix 12

Number of Items for Each Subtest for Second Pilot Study

Subtests	Number of Items
Verbal	
Reading Vocabulary	60
Writing Dictation	29
Reading Comprehension	10
Visual	
Visual Identification	44
Visual Recognition	9
Visual Matching	32
Mental Folding	8
Mental Rotation	7
Memory	
Immediate Recognition	10
Delayed Recognition	11
Spatial Relations	10
Math	
Math Concept and Vocabulary	15
Math Computation	30
Word Problem	20
Logical	
Concept Formation	20
Pattern Finding	17
Sequential Order	19
General Knowledge	10

Appendix 13

Number of Items for Each Subtest for Main Study

Subtests	Number of Items
Verbal	
Reading Vocabulary	80
Writing Dictation	40
Reading Comprehension	20
Visual	
Visual Identification	40
Visual Recognition	30
Visual Matching	36
Mental Folding	17
Mental Rotation	8
Memory	
Immediate Recognition	30
Delayed Recognition	30
Spatial Relations	30
Math	
Math Concept and Vocabulary	15
Math Computation	50
Word Problem	32
Logical	
Concept Formation	25
Pattern Finding	22
Sequential Order	15
General Knowledge	75

Appendix 14

Descriptive Statistics for TAR subtests by genders

Subtests/Statistics	Boys	Girls
Verbal		
Mean	-.274	-.400
Median	.100	-.240
Std. Deviation	2.097	2.250
Minimum	-6.220	-6.220
Maximum	3.660	6.800
Range	9.880	13.020
Skewness	-.7240	-3.100
Kurtosis	.111	.222
Visual		
Mean	4.084	4.008
Median	4.830	4.830
Std. Deviation	1.191	1.302
Minimum	-1.990	-1.460
Maximum	4.830	4.830
Range	6.820	6.290
Skewness	-1.722	-1.642
Kurtosis	3.363	2.260
Memory		
Mean	1.529	1.541
Median	1.560	1.560
Std. Deviation	1.327	1.078
Minimum	-1.910	-1.140
Maximum	5.060	5.060
Range	6.970	6.200
Skewness	.523	.233
Kurtosis	.605	.417

Math		
Mean	.147	-.251
Median	.240	-.770
Std. Deviation	2.748	2.516
Minimum	-5.760	-5.760
Maximum	4.960	4.960
Range	10.720	10.720
Skewness	-.206	.231
Kurtosis	-.530	-.343
Logical		
Mean	1.008	.977
Median	1.150	1.150
Std. Deviation	1.209	1.197
Minimum	-2.300	-2.300
Maximum	4.730	4.730
Range	7.030	7.030
Skewness	.100	.327
Kurtosis	.262	.826
General Knowledge		
Mean	2.125	2.083
Median	2.050	1.820
Std. Deviation	1.431	1.372
Minimum	-1.130	-1.450
Maximum	5.040	5.040
Range	6.170	6.490
Skewness	.321	.458
Kurtosis	-.372	.000

Correlations between Subtests

Subtests	Logical	GK	Verbal	Visual	Memory	Math
Logical	1	0.338	0.339	0.24	0.347	0.364
GK	0.338	1	0.535	0.18	0.04	0.495
Verbal	0.339	0.535	1	0.199	-0.17	0.676
Visual	0.24	0.18	0.199	1	0.106	0.199
Memory	0.347	0.04	-0.017	0.106	1	-0.101
Math	0.364	0.495	0.676	0.199	-0.101	1

Mean Difference (t-tests) by Subtest by Gender

Subtest	t	p
Logical	-1.185	.237
GK	0.314	.754
Verbal	0.606	.545
Visual	0.636	.525
Memory	-0.096	.924
Math	1.560	.119