

Rowan University

Rowan Digital Works

Theses and Dissertations

5-5-1998

A study analyzing the ability of the Yellow Brick Road Screening Test to predict future academic achievement

Nancy Lazzaro
Rowan University

Follow this and additional works at: <https://rdw.rowan.edu/etd>



Part of the [Educational Psychology Commons](#)

Recommended Citation

Lazzaro, Nancy, "A study analyzing the ability of the Yellow Brick Road Screening Test to predict future academic achievement" (1998). *Theses and Dissertations*. 1961.

<https://rdw.rowan.edu/etd/1961>

This Thesis is brought to you for free and open access by Rowan Digital Works. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Rowan Digital Works. For more information, please contact graduateresearch@rowan.edu.

A STUDY ANALYZING THE ABILITY OF THE YELLOW
BRICK ROAD SCREENING TEST TO PREDICT
FUTURE ACADEMIC ACHIEVEMENT

by
Nancy Lazzaro

A Thesis

Submitted in partial fulfillment of the requirements of the
Master of Arts Degree in the Graduate Division
of Rowan University
May 5, 1998

Approved by

Date Approved 5/5/98

ABSTRACT

Nancy Lazzaro

A STUDY ANALYZING THE ABILITY OF THE YELLOW BRICK ROAD SCREENING TEST TO PREDICT FUTURE ACADEMIC ACHIEVEMENT

Dr. John Klanderman, Project Advisor

Dr. Roberta Dihoff, Project Advisor

Seminar in School Psychology

May 5, 1998

The purpose of this study was to examine the ability of the Yellow Brick Road kindergarten screening instrument to predict future academic achievement using CAT scores as indicators of student performance. There were 31 subjects used in this study. Subjects were selected randomly from the respective elementary school population. The study followed a longitudinal design. Yellow Brick Road scores were obtained, in addition to CAT scores in grades 2, 4, and 6. All scores were recorded anonymously with the permission of the respective school Board of Education.

The data was analyzed using a paired sample T test. Based on the analysis, the null hypothesis was rejected. With significance set at a .05 level, a positive, significant relationship was found between test scores. Based on these findings, the Yellow Brick Road can serve as a useful instrument for determining eligibility for kindergarten entrance, and also as an indicator of future academic achievement.

Mini-Abstract

Nancy Lazzaro

A STUDY ANALYZING THE ABILITY OF THE YELLOW
BRICK ROAD SCREENING TEST TO PREDICT
FUTURE ACADEMIC ACHIEVEMENT

Dr. John Klanderman, Project Advisor

Dr. Roberta Dihoff, Project Advisor

Seminar in School Psychology

May 5, 1998

The purpose of this study was to examine the ability of the Yellow Brick Road screening instrument to predict future academic achievement using CAT scores as indicators of student performance. The results indicated that a significant, positive relationship existed between between test scores on the Yellow Brick Road screening test and scores on the CAT in grades 2, 4, and 6.

Acknowledgments

I would like to thank Dr. Barbara Williams for guiding me through this research study. I can not begin to sufficiently thank my parents for encouraging and supporting me through my education. Finally, thank you to William Scully for being with me and believing in me every step of the way.

Table of Contents

Chapter	Page
Chapter 1: The Problem	1.0
Purpose	1.2
Hypothesis	1.2
Background	1.2
Assumptions	1.6
Limitations	1.6
Definition of terms	1.7
Overview	1.8
Chapter 2: Review of Literature	2.9
Yellow Brick Road	2.9
Table 1	2.11
Table 2	2.14
Predicting Future Academic Success	2.15
Screening	2.17
Summary	2.18
Chapter 3: Design of the Study	3.19
Sample	3.19
Measures	3.19
Design	3.21
Testable Hypotheses	3.22
Analysis	3.22
Summary	3.22
Chapter 4: The Results	4.23
Introduction	4.23
Results	4.23
Table 3	4.24
Table 4	4.26
Table 5	4.27
Table 6	4.28
Summary	4.29

Table of Contents (continued)

Chapter	Page
Chapter 5: Summary and Conclusions	5.30
Summary	5.30
Discussion of Findings	5.30
References	32
Appendix A	35
Appendix B	36

Chapter 1: The Problem

There exists a continuous debate over the use of early screening instruments to determine school entry. Many questions linger about the utility of such measures. Some professionals argue that screening instruments lack the predictive validity that is necessary to determine kindergarten readiness (Judy, 1986). Overall validity and reliability of such instruments are also considered questionable. It is argued that "few of the tests meet 'acceptable standards of reliability and validity' for making predictions about children's future achievement" (Gold, 1988).

Readiness tests are often the focus of this debate. These tests serve as a means for assessing children before the start of their formal education. In an article entitled, "How is My Child Doing", Samuel J. Meisels considered readiness tests to be as accurate as the toss of a coin. He felt that the results of such tests are used as grounds for separating "children from their peers and keep[ing] them in preschool for an extra year (1993).

On the opposite end of the argument, in 1992, the National Educational Goals Panel approved the development of a "comprehensive early-childhood assessment system to help gauge children's readiness for school" (Cohen, 1992). This resolution championed the concept of readiness testing in hopes of reaching the nation's first educational goal by the year 2000 (Cohen). See Appendix A for a complete listing of the

National Educational Goals. The panel agreed that such measures would help to ensure a proper educational setting for all children entering kindergarten (Rothman, 1991).

Purpose:

In 1992, 23 states required school readiness testing (Wallace, 1992). In addition to increased testing being done, academic standards in kindergarten have also increased. These two factors combined, a greater number of children are being deemed unready to enter kindergarten. The purpose of this study was to examine the ability of the Yellow Brick Road kindergarten screening instrument to predict future academic achievement using California Achievement (CAT) scores as indicators of student performance.

Hypothesis:

There will be a significant, positive relationship between scores on the Yellow Brick Road kindergarten screening test and future academic achievement as measured by CAT scores in grades 2, 4, and 6.

Background:

There are various unique components that comprise the field of psychological testing. Specifically, these tests need to be uniform, objective, and quantifiable. Uniformity and objectivity are also referred to as standardization. The presentation and scoring of test items needs to be exact. Scoring criteria also need to be exact to ensure that differences in test results are due to subjects' individual differences rather than flaws in the test itself. Quantifiability in psychological testing means that results are given in numerical scores. Psychological tests attempt to yield quantifiable scores to nonquantifiable concepts, like visual or fine motor skills. Using numerical scores for this purpose serves to provide a more accurate picture of a verbally-described concept.

Because of this, results are easier to communicate. Communication of results is also facilitated by the use of arithmetic manipulation of the numerical scores (Wodrich, 1984).

Test validation is another significant area in psychological testing. Tests should not be used unless appropriate norming information exists on that test. This is especially important when tests are being given to preschoolers or young children where appropriate instruments are often lacking. It is sometimes the case when norming information for tests given to school-aged children is applied to preschoolers. It should be noted that instruments could be lacking this necessary information due to inconsistencies of preschoolers' behavior and development patterns (Nuttall, Romero, and Kalesnik, 1992).

To ensure proper test validation, many factors need to be examined. Validity indicated accuracy of a given test. Results are more convincing coming from a test with high validity (Meisels, 1989). Overall, validity ratings determine whether the test measures what it was designed to measure (Wodrich, 1984). Meisels writes, "Screening tests with unknown or limited validity should not be used". Measures of predictive validity indicate whether or not "current test scores foretell some independent criterion in the future" (Wodrich). Related to validity is the idea of reliability. Reliability is used to demonstrate "how consistently or how often identical results can be obtained with the same screening instrument (Meisels). Measures of reliability show the extent to which discrepancies in test scores are related to chance or to differences in test takers (Meisels).

Informed consent and confidentiality play a key role in the testing of young children. These issues gained significance with the passing of P.L. 94-142, also called the Education for All Handicapped Children Act. This act is often referred to as both a piece of educational and civil rights legislation. The act focuses on educational needs of children, but also acknowledges due process and equal protection for children and their parents (Nuttall, Romero, and Kalesnik, 1992). Parental decision making expanded due to P.L. 94-142. Nuttall et al. write that this law requires parental consent before an

evaluation be made on a child. Informed consent is described as "the legal requirement to obtain permission to perform some action (such as assessing a preschool child) only after fully explaining what is to be done, the purpose of the action, and the potential risks" (Nuttall et al.). Voluntariness, competence, and knowledge are three aspects of informed consent. Voluntariness ensures that parents are intentionally giving their permission. A Parent must also be of legal competence to give permission. Knowledge has received the most attention concerning informed consent. Schools are obligated to inform parents on what is intended for the child, why this action was decided, and how the school plans to accomplish it. In addition, there is an overall ethical responsibility for psychologists to keep client information confidential. However, when minor children are the clients involved, the issue of confidentiality is not as clear-cut. Parents have a legal right to any information on their child except in instances such as child abuse allegations (Nuttall et al.).

In testing preschoolers and young children, there are generally two types of tests used: developmental screening tests and readiness tests. Each serves a particular purpose and measures different abilities. Meisels (1989) states, "The major difference between the two procedures is the difference between *skill acquisition* and the *ability to acquire skills*". Readiness tests lend information "about the extent to which a child has acquired knowledge and skills considered to be important entry criteria for a particular program" (Wilson Hills, 1987). These tests look at skill acquisition and, results effect "class placement and curriculum planning" (Meisels). A readiness test would be given to predict preparedness for a certain educational program. Conversely, developmental screening tests would be given to measure a child's developmental potential. With these instruments, the focus is on one's ability to acquire specific skills (Meisels). Developmental screening tests yield information about performance in areas of normal development and the potential for further skill development (Wilson Hills). It is also

thought that because of the differences between these measures, they should not be used as complementary instruments (Meisels).

The use of screening programs is on the rise. These programs are put into effect to predict those students who may be in need of special services, like special education (Wilson Hills, 1987). Preschool screening programs are designed to be the first step in a process leading to helpful intervention or special education services (Nuttall, Romero, and Kalesnik, 1992). Nuttall et al. (1992) view screening as a "relatively brief and inexpensive assessment of large numbers of children in terms of their vision, hearing, physical health, and development in speech and language, motor skills, social and emotional growth, and cognitive skills". It is necessary to not that screening and assessment are not used synonymously. While screening is a process where all children are being looked at for further educational problems, assessment refers to the next step in the process. It is a deeper evaluation that is required for those children who are eligible for kindergarten to determine eligibility for school entry. These screening programs are designed to establish those children who are ready to begin school and those who are not. With higher standards in the lower grades, these programs are necessary to find those children who may experience difficulties in school (Wilson Hills).

In general, it would seem that screening programs can only be measured in terms of their predictive value. Meisels (1989) gives highlights of various tests that do possess reliability and validity data, but still tends to lose their predictive accuracy over a two year period. Combined with the multitude of standards that screening instruments must meet, such as cost efficiency, brevity, and standardization, it would seem obvious that consistent predictive accuracy is virtually unattainable. Professionals argue that preschool screening programs should not be viewed as comprehensive measures (Meisels). Wilson Hills (1987) lists other sources of valuable information, such as parents and/or informal checklists, that could be useful when deciding on the future of preschool children. Though the predictive value of screening is limited, such programs have served to

somewhat bridge the gap for those children who may not be ready to start their formal schooling or for those children who may be suspected of needing special education services somewhere in their academic career (Nuttall, Romero, and Kalesnik, 1992).

The screening instrument that this study looked at was the Yellow Brick Road. This instrument is made up of four batteries of tests used to judge motor, visual, auditory, and language skills. The four test batteries consist of six subtests with six items per subtest. The items are in order of increasing difficulty. Items on the test were designed to coordinate with the popular movie the *Wizard of Oz*. Test items were arranged to allow children ranging from 4 years, 9 months to age 6 years, nine months to complete the majority of test items successfully. With this instrument, up to twenty-four children can be evaluated at one time. Children move individually to consecutive testing stations. The sequence of testing stations begins with non-verbal motor tests, followed by visual and auditory tests. The children finish with verbal tests. These tests were placed last to allow children the opportunity to become confident with the testing situation (Kallstrom, 1975).

Assumptions:

For the purpose of this study, it should be assumed that all subjects were selected randomly. The subjects' kindergarten screening test scores were taken from the Yellow Brick Road screening program that was/is offered annually in a school district located in Camden County. In addition, CAT scores were also taken from the records of this school district. Confidentiality was also a serious issue concerning this study. Readers should also assume that permission was granted to the author of the study providing that student records remained anonymous. See Appendix B for permission information.

Limitations:

The overwhelming majority of the respective district's population was, and still is, homogeneous. Students' came from a white, middle class background. There was no racial diversity among the student population used in the study.

Definition of Terms:

Developmental Test- An age -related norm-referenced assessment of skills and behaviors that children have acquired (compared to children of the same chronological age).

Sometimes such tests are inaccurately called developmental screening tests (Meisels, 1989).

Objective- This term implies that administration, scoring, and interpretation of scores are independent of the subjective judgment of the particular examiner (Anastasi and Urbina, 1997).

Predictive Validity- Validation of a measure by prediction to an external criterion (Nuttall, Romero, and Kalesnik, 1992). Evidence of criterion-related validity in which scores on the criterion are observed at a later date (Meisels, 1989).

Reliability- The consistency of a measuring instrument (Frankfurt-Nachmias and Nachmias, 1992). The degree to which test scores are consistent, dependable, or repeatable; that is the degree to which test scores can be attributed to actual differences in test takers' performance rather than to errors of measurement (Meisels, 1989).

Screening- A process for determining whether further reevaluation is needed. Its purpose is to identify children who are not within normal ranges of development and need further evaluation and who may be candidates for early intervention programs (Nuttall, Romero, and Kalesnik, 1992).

Screening Test- A test used to identify children who may be in need of special services, as a first step in identifying children in need of further diagnosis; focuses on the child's ability to acquire skills (Meisels, 1989).

Standardization- This term implies uniformity of procedure in administering and scoring the test (Anastasi and Urbina, 1997).

Standardized Test- An instrument composed of empirically selected items that has definite instructions for use, adequately determined norms, and data on reliability and validity (Meisels, 1989).

Testing- The administration, scoring, and interpretation of scores of a standardized test (Meisels, 1989).

Overview:

In the following chapter, a review of existing literature on kindergarten screening programs will be presented. Key studies in this area were examined and discussed in relation to the study at hand. In chapter 3, an outline of the research design used in this study will be discussed. The sample will be described, as will be the testable hypotheses and variables that were used. The results of this study will be introduced in chapter 4. The information given in chapter 1 will serve to lay the foundation for the literature review to be presented in the following chapter.

Chapter 2: Review of Literature

This chapter contains a review of the existing literature that was relevant to this study. I begin by discussing some key elements of the Yellow Brick Road. Information on the test manual and scoring the test are presented. Remarks and feelings from two reviews of the test from the Eighth Mental Measurements Yearbook (Buros, 1978) follow the specific Yellow Brick Road information. Finally, results from a study done on the Yellow Brick Road are presented. The next section of chapter 2 highlights previous studies that have looked into the predictive validity of screening instruments. Studies done on the Metropolitan Readiness Test (MRT), the Kindergarten Screening Battery (KSB), and the Aptitude Test for School Beginners are discussed regarding their ability to predict academic achievement. Also, a longitudinal study that was based on predicting reading success is covered. Other various studies showing implications for the uses of screening are also presented.

Yellow Brick Road:

The Yellow Brick Road is a screening instrument that was developed in 1975 by Christine Kallstrom. Tests in Print, Volume 3 Mitchell, 1983) described it as "a gross screening tool for special learning needs". The test is comprised of four batteries that yield twenty-nine scores overall. According to the Eighth Mental Measurements Yearbook (Buros, 1978), no reliability data, norming information, or supporting information regarding cutoff scores exists for the instrument. In the yearbook, the Yellow Brick Road was reviewed by Robert P. Anderson of Texas Tech University and James C.

Reed of Tufts University. Based on the movie the *Wizard of Oz*, the instrument tests motor, visual, auditory, and language skills. The testing area is divided into four stations. The motor battery is represented by the Scarecrow, visual battery by the Lion, auditory battery by the Tin Man, and the language battery by the Munchkins. The children move to the respective stations along a "yellow brick road". At each station, specific skills are tested. The skills tested are outlined in Table 1.

Both reviewers noted a unique feature of this instrument to be the ability of both professionals and nonprofessionals being capable of administering this test; the instrument was developed to ensure this. Anderson added, "The manual is easy for a lay person to read and understand. It states that persons administering the tests should be given at least one practice session" (Buros, 1978). He also remarked that the amount of emotional involvement of parents intending to administer the tests should be considered during training to "offset [possible] scoring biases" (Buros). James C. Reed felt that scoring done by nonprofessionals required an even greater amount of observation and training than for professionals using the test (Buros).

All items are scored on a pass-fail basis. A perfect score on individual batteries would be 36, yielding a perfect score of 144 for the whole instrument. Cutoff scores were supplied to suggest levels of "adequate functioning". A score of 18 or below in any battery for children ages three to four indicate the need for a professional evaluation. Kindergarten or first-grade success is expected with a score of 130 or above for five and six year olds. Though these scores were listed, it appeared that they were set arbitrarily by standards created by the author; no statistical information existed as backing for them.

Robert P. Anderson discussed the norming population in his review of the Yellow Brick Road. This instrument was normed on children ranging from four years, nine months to six years, nine months in a "suburban community in a large metroplex" (Buros, 1978). In the norming data, no information existed regarding "sex, social status,

Table 1

Battery Breakdown of Skills Tested
By The Yellow Brick Road (Buros, 1978)

Motor Battery	7 Scores: imitation, movement, body parts, spatial relationships, right-left, draw-a-person, total; 1 form (11 pages)
Visual Battery	7 Scores: tracking, fusion, visual discrimination, visual memory, visual motor, figure-ground, total; 1 form (15 pages)
Auditory Battery	7 Scores: auditory discrimination, how many, copy cat, sequence, automatic associations, guess what I am, total; 1 form (11 pages)
Language Battery	7 Scores: auditory discrimination, how many, copy cat, sequence, automatic associations, guess what I am, total; 1 form (11 pages)

or ethnic background" for the norming sample. Anderson assumed from information given in the manual that the sample was taken from a middle class background. He did issue caution when generalizing any of this information to children from African-American, "working-class", or "non-English speaking backgrounds" although Kallstrom's manual does not. In addition, Kallstrom presented no reliability data for the respective subtests. Statistics showing predictive validity were also absent. Though correlation's for concurrent validity were reported in the manual to range from .43 to .80, the studies citing these numbers were not gathered from an exhaustive literature review. In his review, Reed also remarked on the lack of "technical justifications for the test" (Buros). He stated,

There is no evidence, such as factor analysis or anything else, cited to show that the separate batteries do, indeed, measure different areas of functioning or these particular areas are related to or necessary for success (Buros, 1978).

An additional area of concern listed by Anderson was found in the last section of the test manual. Developmental activities were given as possible strategies to "train children in those areas where they have alleged deficits" (Buros, 1978).

Consequently, the training suggestions were backed by no information. He argued,

The development of a teaching prescription based on subtest scores with no reported reliability and only clinical judgments of face validity is a highly questionable procedure (Buros, 1978).

Overall, both reviewers offered strong caution for using the Yellow Brick Road. Anderson found no studies lending support for the use of this instrument (Buros, 1978). Anderson remained unsure about the "functional" use of the information that the instrument provided. He felt that a short observation period with a child could reveal

similar information on a child's functioning. The only directly positive statement cited by Anderson was the benefit of involving parents with groups of children. He felt this allowed a great opportunity for additional "support and involvement in early childhood screening" (Buros). James C. Reed felt that the test should be recommended to no one. In his thoughts, existing data did not meet the standards for accurate test development. His interesting analogy to the *Wizard of Oz* read,

Dorothy, Tin Man, Lion, and Scarecrow walked the Yellow Brick Road to Emerald City. They found disappointment because the Wizard was only a weakling incapable of providing a solution to their imaginary problems. Imaginary because Oz itself was only a dream. As an assessment instrument, the *Yellow Brick Road* may be no better than a dream (Buros, 1978).

For the purpose of this project, one study was found done on the Yellow Brick Road. Conducted at the Edmund Guidance Center in Edmund, Oklahoma, France, Couch, Cauthen, Carpenter, Jones, Jordan, Morgan, Lottinville, and Neph (1979) looked at the predictive factors within the Yellow Brick Road. The instrument was chosen for the "assessment of the first-grade readiness in the kindergarten population of a school district within the greater Oklahoma City area" (France et al., 1979). The authors found total scores for each subject and then divided the overall scores into four battery scores. Norms were needed for the sample population. Standard deviations and means were established for the 504 subjects (See Table 2 for results). High correlations revealed strong relationships between the battery items. A principal-components analysis was used to find this information. As revealed by the analysis and intercorrelation matrix, the subscales "were tapping a common factor, which accounted for 49% of the variability" (France et al.). Of the subscale scores listed, the auditory battery score showed the highest correlation ($r=.80$). These results showed that the auditory battery was the "best

Table 2

Mean and Standard Deviation Scores for
the France et al. study on the
Yellow Brick Road

Scores	M	SD
Total	115.90	13.03
Motor Battery	28.08	3.82
Visual Battery	30.16	4.50
Auditory Battery	24.63	6.36
Language Battery	32.84	3.32

overall predictor" (France et al.). It was also noted that with the high intercorrelation's found on the subscales, results could suggest that performance on any of the individual batteries could be "contingent on the abilities that are measured on other subscales" (France et al.). The authors felt that, for example, auditory battery scores could be dependent upon verbal and motor functioning and hearing ability. France et al. felt that the results could support the use of the auditory battery as a possible short form of the Yellow Brick Road. They also felt that more extensive research needs to be conducted before concrete statements on the instrument can be made (France et al.).

Predicting Future Academic Success:

Many studies have concentrated on the predictive validity of screening instruments. With an increase in school readiness testing over the past fifteen to twenty years, it seems necessary to provide information showing the benefits of such tests. Predictive validity scores are of particular interest to many involved in testing. As Oosthuizen, Van Resenberg, and Jordaan (1997) wrote, "It is also important to have sufficient information on the predictive validity of tests from a practical point of view since counseling psychologists and educational psychologists make recommendations and predictions, based on test scores, that can influence children's lives". In a study done by Oosthuizen et al. (1997), the authors compared predictive validity scores on the Metropolitan Readiness Test (MRT) and the Aptitude Test for School Beginners. 120 white, African-speaking boys and girls were administered both tests. The tests were given during the first quarter of the 1992 school year. The sample populations were selected randomly. Correlation scores for both tests were found to be significant at a .01 level. However, correlation's did compliment the MRT. The authors reported, "For the [MRT] out of 98 correlation's 82 moderate and 6 larger values were also significant..." (Oosthuizen et al.). They also concluded that predictive validity ratings for each test did reach satisfactory levels. Overall, the study did favor the MRT in areas concerning cutoff

scores and variance. For the MRT, 50% of the variance was explained by the indicated predictors, whereas 29% was explained for the Aptitude Test for School Beginners. The authors did state that research could be further expanded upon by looking at various language groups and by using larger populations. They also suggested using other tests for future comparisons of predictive validity (Oosthuizen et al.).

Weller, Schnittjer, and Tuten (1992) also focused on the MRT in a study they conducted. This study looked at scores on the MRT to predict future academic performance in grades three through ten. The authors found data on 415 students entering first grade in 1976.¹ The students also spent their twelve years of schooling in one district. The dependent instruments used as indicators of academic performance were various standardized achievement tests given in grades three, six, nine, and ten. The predictor variables in the study were scores from the MRT, individual information such as, socioeconomic status, race, gender, IQ, remediation, and age at school entrance. Correlation's indicated that IQ scores generated the most support for predicting academic performance. The MRT did show statistically significant correlation's as a predicting variable. Weller et al. (1992) wrote, "Consequently, these results suggest the MRT may have significant potential for contributing to readiness decisions in the early grades".

The Kindergarten Screening Battery (KSB) was the focus of a study trying to predict existing learning difficulties in bright kindergartners. In a study done on 215 New York City entering kindergartners, Kelly and Peverly found that the KSB could serve as an "effective predictor of the 1st and 2nd grade criteria" (1992). In addition, they found the battery to be beneficial as a counterpart to intelligence tests when screening gifted students for existing "educational needs" (Kelly and Peverly). The *Catch-'Em Early* preschool screening instrument was used to explore predictive validity in testing as well. Nicely-Leach, Redebaugh, Morrill, and Shreeve (1987) used the instrument as a means of predicting reading success in grades two, five, and eight. They found that the ability to predict such success increased after second grade and, also that the useful predictors also

changed at this time. The predictors of reading success varied at each grade level. Nicely-Leach et al. (1987) used these results to champion the need for longitudinal research to pinpoint changes during a child's academic career (Nicely-Leach et al.)

Screening:

Several articles have been written on the issues surrounding screening. Genevieve M. Fedoruk (1989), in her article "Kindergarten Screening for 1st-Grade Learning Problems: The Conceptual Inadequacy of a Child-Deficit Model", discussed the implications of generalizing screening results onto the whole child. She stated that,

It may be, however, that the conceptual foundation upon which the practice of screening tests is incompatible with educational phenomena and that this may explain, at least partially, the predictive inaccuracies that characterize kindergarten screening (Fedoruk).

She compared educational screening to screening for medical reasons. With both, the screening is used to predict possible interventions. In medicine, the etiology of physical problems can often be determined. Fedoruk argued that the etiology of learning and school problems is not so clear. School performance "encompasses a complex, multifaceted process" (Fedoruk). She felt that screening provides answers to global questions rather than yielding the "predictive accuracy" needed to guide a young person's development. She argued against following a "child deficit model" whereby "learning problems are conceptualized as the exclusive property of the student" (Fedoruk). In this model, outside situational variables that affect the child are ignored. Fedoruk felt that with the knowledge that exists on teacher and classroom variability, screening to predict academic achievement should look beyond properties solely within the child (Fedoruk).

Relating to Fedoruk's work, a study by Gallerani, O'Regan, and Reinharz (1982) examined the variables in certain preschool screening instruments that predict readiness for first grade. Their findings indicated that many variables were at play when

predicting readiness for first grade. In addition to cognitive and verbal abilities, parental ratings and certain demographic information provided valuable insight when screening young children (Gallerani et al., 1982).

Summary:

This chapter served to provide information on screening instruments used on kindergarten-aged children. Created in 1975, the Yellow Brick Road is a kindergarten screening instrument still in use today. For the most part, data on this test do not lend favorable justifications for its use. While the manual is easily understood and nonprofessionals can become active in the screening process, statistical support is lacking. Predictive validity was also discussed in this chapter to show the importance that it has in screening young children. With future interventions and placements being made on the basis of screening results, predictive validity for such measures should be strong. In addition to validity issues, situational variables also play a role in the educational performance of children. The studies in chapter 2 indicated the need for determining the predictive ability of the Yellow Brick Road. Determining the predictive validity of this test could impact the testing population who still use this test, specifically the suburban school population used for this study.

Chapter 3: Design of the Study

Sample:

The sample population was comprised of thirty-one students. The population was selected from an elementary school within a suburban school district in Camden County. This district operates two elementary schools. All of the student data was taken from only one of the elementary schools, called school B for the purpose of this study. This school houses kindergarten through sixth grade. Based on information from the 1998 school district report card, this elementary school serves 532 students and is 100% English proficient. The class sizes average twenty-two students per class. The student mobility rate is 13%. Of the thirty-one subjects selected, fifteen were male and sixteen were female. Special education students were included in the subject population (District Report Card, 1998).

Measures:

Four independent variables were used to examine academic achievement in the sample population. The kindergarten screening instrument, the Yellow Brick Road, is administered to children from the ages four years, nine months to six years, nine months. The test is given to shed light on to the child's functioning in the areas of motor, visual, auditory, and language skills. Scores are also used to identify those children who may be in need of further evaluation. The test was designed to follow the theme of the movie,

Wizard of Oz. The author coordinated the test in this manner to encourage participation by the children. The Yellow Brick Road was constructed to allow as many as twenty-four children to be evaluated at one time. The four test batteries each contain six subtests with six items in each. The testing stations are to be arranged in order of increasing difficulty. Children proceed along a yellow brick road to each location. Each battery is represented by a *Wizard of Oz* character. The Yellow Brick Road was constructed to allow professionals and nonprofessionals to administer the test. It was noted in the testing manual that nonprofessionals should receive some prior training before executing the test. A perfect score on this test is 144. Each item counts for one point. A score of 130 indicates sufficient performance in the test batteries for children at the kindergarten or first grade levels (five to six years of age). Three and four year old children are functioning adequately if scores are twenty-four or above in an individual battery (Kallstrom, 1975). In terms of statistical evidence for the Yellow Brick Road, no reliability data, norming information, or data supporting or justifying the establishment of cutoff scores was presented in the test manual (Buros, 1978). The suburban school district being looked at for this study administers the Yellow Brick Road to all children in the district reaching kindergarten eligibility. The test is given in the spring preceding kindergarten entrance. For the selected population used in this study, the Yellow Brick Road was administered in May of 1990. Scores are continued to be used for determining kindergarten readiness.

The second independent variable included in this study are California Achievement Test (CAT) scores. The CAT can be used from kindergarten through twelfth grade. Forms E and F have replaced C and D which were created in 1978. The purpose of the CAT is to "measure achievement in the basic skills commonly found in state and district curricula"(Conoley and Kramer, 1989). The CAT uses a multiple-choice format at all testing levels. Several changes were made when forms E and F were put in use. High school testing was included, science and social studies tests were made optional for grades two through twelve, tests were lengthened, and objectives were revised to bring

procedures "more in line with current practice and emphasis" (Conoley and Kramer). Test items have been reviewed to eliminate any existing ethnic and/or sexual bias. Forms E and F were normed in the fall of 1984 and spring of 1985. The norming population consisted of 300,000 students from kindergarten through twelfth grade. The spring and fall samples were compared to reveal that they did represent the school population of interest. As noted in The Tenth Mental Measurements Yearbook (Conoley and Kramer), reliability and validity information was presented in the test manual. Intercorrelations on subtests range between .5 to .8. In addition, correlation's between total scores on Forms E and F range from .6 at level 13 of the CAT to a .8 at level 16. Using a Kuder Richardson 20, internal consistency reliability's were shown to be in the .8's and .9's for testing levels above the first grade. Standard errors also lended evidence of reliable measurements. The CAT, Forms E and F, were reviewed by Peter W. Airasian and James L. Wardrop in The Tenth Mental Measurements Yearbook (Conoley and Kramer). Both reviewers gave the CAT favorable reviews. Airasian wrote, "The CAT, Forms E & F, is technically state of the art, well constructed, and well documented". Though Waldrop mentioned some concerns about the technical manual, the Objective Performance Indices, and the Classroom Management Guide, he considered the CAT to be "one of the best standardized achievement batteries available" (Conoley and Kramer).

Design:

This study was considered a predictive study. Yellow Brick Road scores were compared to CAT scores from grades two, four, and six in order to examine the overall predictive ability of the Yellow Brick Road. A longitudinal design was followed. Students who were given the Yellow Brick Road in May of 1990, prior to their kindergarten entrance, who were also given the CAT through grade six were randomly selected for the sample population. The students must have continued their education in

the district from kindergarten through sixth grade. Student CAT scores from second, fourth, and sixth grades were compared to their Yellow Brick Road scores.

Testable Hypotheses:

Null Hypothesis: No relationship exists between scores on the Yellow Brick Road kindergarten screening test and future academic achievement as measured by CAT scores in grades two, four, and six.

Alternate Hypothesis: A positive, significant relationship exists between scores on the Yellow Brick Road kindergarten screening test and future academic achievement as measured by CAT scores in grades two, four, and six.

Analysis:

A paired sample T test was used to test the hypotheses of the study.

Summary:

In this study, the predictive ability of the kindergarten screening test, the Yellow Brick Road was examined. Scores on the CAT in grades two, four, and six served as measures of academic performance. Thirty-one students who completed through sixth grade in one of the district's elementary schools, school B, were selected for the sample population. By comparing student scores on the kindergarten entrance test to scores on the CAT at various grade levels, the predictive ability of the Yellow Brick Road was identified.

Chapter 4: The Results

Introduction:

The purpose of this study was to examine the ability of the Yellow Brick Road kindergarten screening instrument to predict future academic achievement using California Achievement Test (CAT) scores as measures of student performance. The testable hypothesis used in this study stated that a significant, positive relationship existed between scores on the Yellow Brick Road screening test and future academic achievement as measured by CAT scores in grades two, four, and six.

Thirty-one subjects were selected randomly from the student population who were given the Yellow Brick Road in May of 1990. The test was administered to a total of sixty-nine students in the respective elementary school. Mean scores and standard deviations for the sixty-nine students were calculated prior to the study (See Table 3). A perfect score for individual batteries would be thirty-six, yielding a perfect score of 144 for the whole instrument. Kindergarten or first-grade success is expected with a score of 120 or above. Mean scores for this population were below such designated cutoff scores on all test batteries. Scores on the CAT, Form E, in grades two, four, and six were used as indicators of student performance.

Results:

Table 3

Yellow Brick Road Results for 1990
Mean Scores Reported for Subtests and Totals
School B Only (n=69)

Test Battery	M	SD
Motor	26.75	4.08
Visual	28.43	4.04
Auditory	26.11	5.25
Language	30.83	3.42
Total Battery	112.26	13.61

For the purpose of this study, total battery scores on the CAT were recorded from grades two, four, and six. The scores were reported as national percentiles. In general, percentiles are utilized to show a subject's position within a standardization sample. Percentile scores can be advantageous when used as selected measures in research studies. Overall, they tend to be easily understood and are applicable to various populations and tests (Anastasi and Urbina, 1997). Total battery scores from the Yellow Brick Road were also recorded for use in this study.

Scores were analyzed using a paired sample T test. This type of statistic is generally used when data is comprised of paired observations existing for the subject population (Voelker and Orton, 1993). Yellow Brick Road scores were paired with CAT scores from second, fourth, and sixth grades. Correlation's for paired variables ranged from .135 to .874. This analysis revealed positive correlation's between kindergarten screening scores and CAT scores at each grade level included in this study.

With a .05 level of probability established, accepting or rejecting the null hypothesis would be determined at this cutoff level. Based on data obtained through the paired sample T test conducted for this study, the null hypothesis was able to be rejected. With degrees of freedom set at 30 for paired variables and an alpha level of .05, t scores exceeded critical values found on a standard t distribution [$t_{(30df)}=2.042, p > .05$] yielding a significant relationship between variables (Spatz, 1997). For the analysis, Yellow Brick Road scores were documented as variable 1, second grade CAT scores as variable 2, fourth grade CAT scores as variable 3, and sixth grade CAT scores as variable 4.

Table 4 outlines mean scores and standard deviations for paired samples in the study. Similarly, Table 5 lists means and standard deviations calculated for paired differences. Mean scores for the selected population showed a higher average than was achieved for the original 1990 testing population. The sample population mean of

Table 4

Mean Scores and Standard Deviations Found for Paired Samples

n=31

Pair		M	SD
Pair 1	Variable 1	117.1935	9.2138
	Variable 2	67.8065	21.5413
Pair 2	Variable 1	117.1935	9.2138
	Variable 3	72.4839	18.8359
Pair 3	Variable 1	117.1935	9.2138
	Variable 4	73.7742	19.4931

Note. Variable 1= Yellow Brick Road Scores, Variable 2= Grade 2 CAT scores, Variable 3= Grade 4 CAT scores, Variable 4= Grade 6 CAT scores.

Table 5

Mean Scores and Standard Deviations Found for Paired Differences

n=31

Pair		M	SD
Pair 1	Variable 1- Variable 2	49.3871	22.2549
Pair 2	Variable 1- Variable 3	44.7097	18.4376
Pair 3	Variable 1- Variable 4	43.4194	17.9848

Note. Variable 1= Yellow Brick Road scores, Variable 2= Grade 2 CAT scores, Variable 3= Grade 4 CAT scores, Variable 4= Grade 6 CAT scores.

Table 6

Paired Samples t Test Significance Ratings

n=31

Pair		t	df	Sig. (2-tailed)	r
Pair 1	Variable 1- Variable 2	12.356	30	.000	.135
Pair 2	Variable 1- Variable 3	13.501	30	.000	.287
Pair 3	Variable 1- Variable 4	13.442	30	.000	.394

Note. t=t value, df= degrees of freedom, r= correlation coefficient, sig.= significance.

Variable 1= Yellow Brick Road scores, Variable 2= Grade 2 CAT scores, Variable 3= Grade 4 CAT scores, Variable 4= Grade 6 CAT scores.

117.1935 on the Yellow Brick Road falls close to the score of 120 which the testing manual designated as the cutoff score for kindergarten or first-grade success. Table 6 highlights significance ratings found for analyzed data. Specified t values for paired variables exceeded critical values, thus allowing the rejection of the null hypothesis. Correlation's were also listed on this table demonstrating the positive relationship between paired variables.

Summary:

This study looked at the predictive ability of a specific kindergarten screening instrument. The testable hypothesis used stated that a significant, positive relationship would exist between scores on the Yellow Brick Road screening test and future academic achievement as measured by California Achievement Test scores in grades two, four, and six. A paired sample T test was selected to analyze scores. Set at a .05 level of probability, results across paired samples yielded significant findings. Based on these findings, the null hypothesis was rejected.

Chapter 5: Summary and Conclusions

Summary:

The purpose of this study was to examine the ability of the Yellow Brick Road kindergarten screening instrument to predict future academic achievement using CAT scores as indicators of student performance. There were 31 subjects used in this study. Subjects were selected randomly from the respective elementary school population. The testable hypothesis used stated that a significant, positive relationship existed between scores on the Yellow Brick Road screening test and future academic achievement as measured by California Achievement Test scores in grades two, four, and six. The study followed a longitudinal design whereby student Yellow Brick Road scores were compared to their later CAT scores in second, fourth, and sixth grades. All scores were recorded anonymously with the permission of the respective school Board of Education.

The data was analyzed using a paired sample T test. Based on the analysis, the null hypothesis was rejected. With significance set at a .05 level, a positive, significant correlation was found between scores.

Discussion of Findings:

In a time when academic excellence continues to receive attention in the public eye, this study can be used to further research that is being done on kindergarten screening programs. With the first National Education goal focusing on children beginning school ready to learn, greater evidence shedding light on specific screening

instruments seems crucial to the future of education. The review of literature in this study discussed various screening tests used for young children. Few research studies have been conducted on the Yellow Brick Road. In spite of some of its major limitations, such as a lack of norming, standardization, validity, and reliability information, this study may be a reason to re-investigate the use of this instrument. Perhaps a more recent, better-constructed edition of the Yellow Brick Road could be implemented if more research studies indicated the benefits of its use in the student population.

Future research could be done in a variety of educational areas to expand the findings demonstrated in this study. For instance, the longitudinal design of this study could be carried on further in the educational careers of the subjects selected for this study. Yellow Brick Road scores for this population could be compared to CAT scores in elementary and high school, classroom performance in high school, and other standardized tests taken at the high school level, such as the High School Proficiency Test. Additional variables could also be added to studies specifically focusing on elementary populations; examples of such variables could be teacher-made observations and/or classroom evaluations. Research on the Yellow Brick Road could be compared between populations to lend additional reliability and/or validity data. Finally, studies could also be conducted to the special education population and the utility of using the Yellow Brick Road in determining those students in need of further assessment services; follow-up studies could then track the placements and progress of those students. While possible future studies could benefit kindergarten screening programs overall, the author of this study does recommend using screening results as complementary measures when deciding on the future of young children. Professionals should consider other aspects of a child's life as well. Parent questionnaires or direct observations could be possible measures used in conjunction with kindergarten screening test scores when making critical decisions about the educational careers of our youngest students.

References

- Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River: Prentice Hall.
- Buros, O. K. (Ed.). (1978). The eighth mental measurements yearbook. Highland Park: The Gryphon Press.
- Conoley, J. C., & Kramer, J. J. (Eds.). (1987). The tenth mental measurements yearbook. Lincoln: The University of Nebraska Press.
- Cohen, D. L. (1992, April, 8). Goals panel adopts plan to develop early-childhood assessment system. Education Week.
- Fedoruk, G. M. (1989). Kindergarten screening for 1st-grade learning problems: the conceptual inadequacy of a child-deficit model. Childhood Education, 66 (1), 40-42.
- France, G. A., Couch, J., Jones, D., Jordan, P., Morgan, C., Lotinville, E., & Neph, L. (1979). Yellow brick road revisited. Journal of Consulting & Clinical Psychology, 47, (4), 760-761.
- Gallerani, D., O'Regan, M., Reinharz, H., (1982). Prekindergarten screening: how well does it predict readiness for first grade?. Psychology in the Schools, 19, (2), 175-182.
- Gold, D. L., (1988, March 30). Early testing said to have long term negative effects. Education Week, 6.
- Judy, J. (1986). Early screening is essential for educational accountability : response to

- Shepard & Smith. Education Leadership, 87-88.
- Kallstrom, C. Ph.D., Yellow brick road testing manual. Austin: Learning Concepts.
- Kelly, M. S., & Peverly, S. T. (1992). Identifying bright kindergartners at risk for learning difficulties: predictive validity of a kindergarten screening tool. School Psychology, 30, 245-258.
- Meisels, S.J. (1993, August). How is my child doing?. Education Week, 12, (40), 40-41.
- Meisels, S. J. (1989). Developmental screening in early childhood: a guide (3rd ed.). Washington, DC: National Association for the Education of Young Children.
- Mitchell, J. V., Jr. (Ed.). (1983). Tests in print 3. Lincoln: The University of Nebraska Press.
- Nicely-Leach, J., Redebaugh, M. R., Morrill, C. M., & Shreeve, W. C. (1987). Predicting reading success: a nine-year longitudinal study. Early Childhood Development & Care, 28, 47-57.
- Nuttall, E. V., Romero, I., & Kalesnik, J. (Eds.). (1992). Screening preschoolers: psychological & educational dimensions. Needham Heights: Allyn & Bacon.
- Oosthuizen, S., Van Resburg, E., & Jordaan, C. (1997). Comparison of the predictive validity of the aptitude test for school beginners and of the metropolitan readiness test. Perceptual & Motor Skills, 84, 59-65.
- Rothman, R. (1991, June 5). Bush panel members endorse readiness assessment. Education Week, 10
- The National Education Goals Panel. (1997). The national education goals report summary. Washington, DC: Author.
- Spatz, C. (1997). Basic statistics tales of distributions (6th ed.). Pacific Grove: Brooks/Cole Publishing Company.
- Voelker, D. H., M.A., & Orton, P. Z., Ed.M. (1993). Statistics. Lincoln: Cliffs Notes

Incorporated.

Wallace, J. (1992). Ready or not- can a test tell? State Education Leader, Spring.

Weller, L. D., Schnittjer, C. J., & Tuten, B. A. (1992). Predicting achievement in grades three through ten using the metropolitan readiness test. Association for Childhood Education International.

Wilson Hills, T. (1987). Screening for school entry. ERIC Digest.

Wodrich, D. L. (1984). Children's psychological testing- a guide for nonpsychologists.
Baltimore: Paul H. Brooks Publishing Co..

Appendix A

The National Education Goals

1. By the year 2000, all children in America will start school ready to learn.
2. By the year 2000, the high school graduation rate will increase to at least 90 percent.
3. By the year 2000, all students will leave grades 4, 8, and 12 having demonstrated competency over challenging subject matter including English, mathematics, science, foreign languages, civics and government, economics, arts, history, and geography, and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our Nation's modern economy.
4. By the year 2000, the Nation's teaching force will have access to programs for the continued improvement of their professional skills and the opportunity to acquire the knowledge and skills needed to instruct and prepare all American students for the next century.
5. By the year 2000, United States students will be first in the world in mathematics and science achievement.
6. By the year 2000, every adult American will be literate and will possess the knowledge and skills necessary to compete in a global economy and exercise the rights and responsibilities of citizenship.
7. By the year 2000, every school in the United States will be free of drugs, violence, and the unauthorized presence of firearms and alcohol and will offer a disciplined environment conducive to learning.

8. By the year 2000, every school will promote partnerships that will increase parental involvement and participation in promoting the social, emotional, and academic growth of children.

Appendix B

November 5, 1997

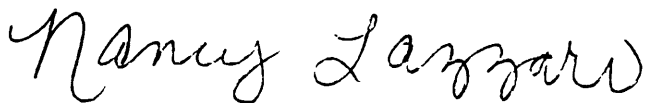
Dear

I am writing this as both a 1992 graduate of and a full-time graduate matriculated into the Master of Arts in School Psychology program at Rowan University. In addition, I will also be continuing my education in Rowan University's School Psychology certification program next year. As part of my requirements for the Seminar in School Psychology, I am required to conduct a research study.

I have met with Dr. Barbara Williams in order to discuss topics which would be feasible for me to do my thesis and would also assist the district in evaluating a program. Together, we have outlined a research project involving comparing incoming kindergarten students' Yellow Brick Road scores with future achievement.

I am requesting permission to conduct the study using data from cumulative files. This will be done under Dr. Williams' supervision. I am hopeful that you will grant me permission to complete my study here.

Sincerely,

A handwritten signature in cursive script that reads "Nancy Lazzaro".

Nancy Lazzaro

17. RESEARCH PROJECT - N. LAZZARO

To approve Ms. Nancy Lazzaro's research project for Rowan University under the supervision of Dr. Williams. indicated this would be "blind" research which would not require researcher to view student folders. This response was to a question from concerning privacy of student records.

VOTE FOR ITEMS 13 - 17

Item #13, 17 - Approved by unanimous voice vote.

Item #14, 15, 16 - Approved by unanimous voice vote. 8-0-1.

abstained.