A predictive validity study of the DABERON: a screening test used for identifying kindergarten children who may be at-risk for academic failure.

Marianne F. LaRoche
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation
https://scholarworks.umass.edu/dissertations_1/4456

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
A PREDICTIVE VALIDITY STUDY OF THE DABERON:
A SCREENING TEST USED FOR IDENTIFYING KINDERGARTEN
CHILDREN WHO MAY BE AT-RISK FOR ACADEMIC FAILURE

A Dissertation Presented
by
MARIANNE F. LAROCHE

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 1989

School of Education
Copyright by Marianne F. LaRoche 1969

All Rights Reserved
A PREDICTIVE VALIDITY STUDY OF THE DABERON:
A SCREENING TEST USED FOR IDENTIFYING KINDERGARTEN
CHILDREN WHO MAY BE AT-RISK FOR ACADEMIC FAILURE

A Dissertation Presented
by
MARIANNE F. LAROCHE

Approved as to style and content by:

Ena Vazquez Nuttall
Chairperson of Committee

Maria Brassard, Member

Gail McGee, Member

Marilyn Haring-Hidore, Dean
School of Education
ACKNOWLEDGEMENTS

Appreciation is extended to my dissertation director, Dr. Ena Nuttall, and to the members of my committee, Dr. Marla Brassard and Dr. Gail McGee, for their suggestions, help, and encouragement in this project.

I also wish to thank Dr. Jill Adels and Joanne Kalesnik, dear friends and mentors, whose support, advice, and input have been invaluable to me on so many occasions. Special thanks also to my family and friends, whose patience and understanding have been limitless.

This work is dedicated to my mother, Margaret M. LaRoche, whose unwavering encouragement, love, and pride propelled me forward in this endeavor.
ABSTRACT

A PREDICTIVE VALIDITY STUDY OF THE DABERON:
A SCREENING TEST USED FOR IDENTIFYING KINDERGARTEN
CHILDREN WHO MAY BE AT-RISK FOR ACADEMIC FAILURE

MAY 1989

MARIANNE F. LAROCHE, B.A., UNIVERSITY OF MASSACHUSETTS
M.Ed., UNIVERSITY OF MASSACHUSETTS
Ed.D., UNIVERSITY OF MASSACHUSETTS
Directed by: Dr. Ena Vazquez Nuttall

This study addresses the problem of whether the DABERON, an easily administered and objectively scored screening test, has a useful degree of validity in identifying children at risk for educationally handicapping conditions. The research population consisted of 165 first, second, and third grade students, each of whom had been administered the DABERON at the beginning of their kindergarten year as part of a state-mandated kindergarten screening program. The sensitivity and specificity of the DABERON as a screening instrument was assessed, as well as its overall percentage of correct predictions, its error rate (percentage of misclassifications), and its positive and negative predictive values. Additionally, subjects' DABERON scores were correlated with two criterion measures thought to be predictive of academic success in the early elementary grades, the Metropolitan Readiness Test (MRT), and the Kindergarten Progress Report (KPR), a teacher rating. The impact of subject variables such as gender, age, SES, cultural
background, and prior educational experience on DABERON performance was also evaluated.

The results yielded modest sensitivity and high specificity rates for the DABERON (.31 and .98, respectively), with 81% of its total predictions found to be correct. In relationship to the estimated prevalence of educationally handicapping conditions in the population, the predictive value of a positive DABERON result was found to be 84%, and the predictive value of a negative DABERON result was 81%. Significant moderate correlations were obtained with both the MRT ($r=.68$, $p<.0001$) and the KPR ($r=.44$, $p<.0001$). Based on these results, it was concluded that the DABERON appears to be a useful and valid screening measure for identifying children with potentially handicapping conditions, at least for the population studied.
**TABLE OF CONTENTS**

ACKNOWLEDGEMENTS .......................................................... iv

ABSTRACT .............................................................................. v

LIST OF TABLES ...................................................................... ix

Chapter

I. INTRODUCTION .................................................................. 1

  Statement of the Problem ....................................................... 2
  Purpose and Significance ....................................................... 3
  Outline of the Remaining Chapters ....................................... 5

II. REVIEW OF THE LITERATURE ............................................. 7

  Historical Background .......................................................... 7
  Why Screen? ........................................................................ 11
  The Screening Process ......................................................... 13
  Selecting Appropriate Screening Instruments ....................... 15
  Developmental Screening vs. Readiness Tests ....................... 17
    Readiness Tests .................................................................. 18
    Developmental Screening Tests .......................................... 19
  Evaluation of Screening Instruments ...................................... 20
    Examples of Predictive Validity Studies .............................. 22
  Summary and Implications ..................................................... 24

III. METHODOLOGY ................................................................. 26

  Research Questions ................................................................ 26
  Assumptions ........................................................................ 31
  Scope and Limitations ........................................................ 31
  Population ........................................................................... 33
  Instruments .......................................................................... 34
    DABERON .......................................................................... 34
    Metropolitan Readiness Test .............................................. 35
    Kindergarten Progress Report .......................................... 36
  Data Collection .................................................................... 37
  Data Analysis ....................................................................... 38
  Significance .......................................................................... 40
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV. RESULTS</td>
<td>42</td>
</tr>
<tr>
<td>Demographic Data</td>
<td>42</td>
</tr>
<tr>
<td>Results</td>
<td>43</td>
</tr>
<tr>
<td>V. DISCUSSION</td>
<td>57</td>
</tr>
<tr>
<td>Summary of Results</td>
<td>57</td>
</tr>
<tr>
<td>Suggestions for Further Research</td>
<td>69</td>
</tr>
<tr>
<td>Conclusions</td>
<td>70</td>
</tr>
<tr>
<td>APPENDIX: METHOD FOR CALCULATING POSITIVE AND NEGATIVE PREDICTIVE VALUES USING BAYES' THEOREM</td>
<td>71</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>72</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1. Relationship Between Pass/Fail Performance on the DABERON and the Presence or Absence of a Diagnosed Handicapping Condition .............................................. 43
2. Ranges, Means, Standard Deviations, and Medians for the DABERON, MRT, and KPR Total Scores .......................................................... 46
3. Correlation Coefficients of DABERON Total Scores and Subsection Scores with MRT and KPR Total Scores ........................................... 46
4. Correlations Between DABERON Total Scores and KPR Total Scores by Individual Teacher ................................................................. 47
5. Results of a T-Test for Significant Differences Between Boys' and Girls' Mean Total Scores on the DABERON ................................... 49
6. Relationship Between Gender and Pass/Fail Performance on the DABERON ......................................................................................... 50
7. Results of a T-Test for Significant Differences Between the Mean Ages of Subjects Who Passed the DABERON vs. Those Who Failed .......... 51
8. Results of a T-Test for Significant Differences Between Mean DABERON Total Scores of Low Income vs. Not Low Income Subjects ........................................................................ 52
9. Relationship Between Pass/Fail Performance on the DABERON and Family Income Status of the Subjects ........................................ 52
10. Results of a T-Test for Significant Differences Between Mean DABERON Total Scores of Hispanic vs. Non-Hispanic Subjects ................................................................. 54
11. Relationship Between Pass/Fail Performance on the DABERON and Subject Cultural Background (Hispanic vs. Non-Hispanic) ................................................. 54
12. Results of a T-Test for Significant Differences Between Mean DABERON Total Scores of Subjects with Prior Educational Experience vs. Those With No Prior Educational Experience ...... 56
13. Relationship Between Prior Educational Experience and Pass/Fail Performance on the DABERON ................................................................. 56
Predicting the academic achievement of young children has been a concern of psychologists and educators for many years, since the development of the first Binet scales. Every September in the United States more than three million children begin formal schooling with their first day of kindergarten (Shepard & Smith, 1986), differing tremendously in their inherent abilities, early experiences, developmental maturity, and readiness to learn. Kindergarten screening, and other early assessment programs, can be viewed as the first step in an educational process that focuses on fostering success in school. It begins with the early identification of those children who, because of problems of development and/or experience, may be least able to meet the typical expectations of school. For these children, school can become an unhappy, failure-ridden experience, but through early screening programs many of them can presumably be identified at a young age, and be given help to prevent failure (Rafoth, 1984; Zeitlin, 1976).

With the passage of Public Law 94-142 in 1975, Congress officially recognized the responsibility of the federal, state, and local governments to adequately meet the educational needs of all children, and required states to institute "child find" procedures to identify children needing special education services. A major consequence of this law has been a demand on the part of both state government officials and educators for earlier and earlier recognition or detection of handicapping conditions prior to entry into grade one (Barnes, 1982).
This has led to the implementation of screening programs in school systems across the country, and the early identification of children prone to school difficulties has thus become an educational problem of primary significance.

In Massachusetts, Chapter 766 (passed in 1972) requires that "each school committee make a continuous and systematic effort to identify children in need of special education" (303.0), and specifically calls for kindergarten entry screening, as well as screening for children ages 3 and 4 whose parents have requested such screening (304.3) (Massachusetts Department of Education, 1986). Therefore, every school system in Massachusetts must devise an efficient and effective early screening program to identify those children at risk for school failure who may need special education services. The selection of appropriate instruments and procedures for this task is left to the discretion of the local school committees.

Statement of the Problem

Numerous screening tests for school readiness have been published in the last 20 years and subsequently adopted by schools for use in their mandated screening programs, however their effectiveness for properly identifying high risk children is often not well established (Barnes, 1982). Although the manuals for most of these screening instruments report reliability and normative data, typically these tests have not been examined with regard to their predictive validity (Piersel & Kinsey, 1984; Rafoth, 1984; Tsushima, Onorato, Okumura & Sue, 1983).
In addition, instruments are sometimes selected which are not suitable for the intended purpose. Accepted standards for screening instruments include the following: norm-referenced, developed on a representative and systematically-determined sample; valid and reliable; sensitive, i.e., correctly identifying children possibly at-risk; and specific, i.e., correctly excluding others from further assessment (Cadman, Chambers, Walter, Feldman, Smith & Ferguson, 1984; Meisels, Wiske & Tivnan, 1984). According to Meisels (1984), if the goal is to predict quickly and accurately whether a child might have difficulty succeeding in school, or could profit from a specialized educational placement or services, then the screening test selected should be one which has developmental content (i.e., which samples developmentally age-appropriate abilities, rather than specific accomplishments that indicate academic readiness, such as being able to recite the alphabet), normative standardization, and established predictive validity. Other obviously desirable characteristics of a screening test would include face validity, to help ensure cooperation and enthusiasm on the part of screening teams, attractiveness to children, and ease and speed of administration.

**Purpose and Significance**

The main purpose of this study is to evaluate the predictive validity of the DABERON (Danzer, Gerber & Lyons, 1982), subtitled "A screening test for school readiness," which has been used as the kindergarten screening instrument in the city of Northampton, Massachusetts for the past five years. The primary question addressed is whether the
DABERON accurately predicts those children at high risk for academic failure who may need special education services.

The early identification of children with learning handicaps is not only desirable for the purposes of early intervention, but is also mandated by law. Since the advent of P.L. 94-142, most states have implemented periodic, systematic screening of young children entering school in order to identify those with potential handicapping conditions. Thus there is clearly a need for reliable and valid means of identifying children early in their school career who have or may develop learning difficulties, and many screening instruments have been designed for this purpose. What is of concern, however, is the frequent lack of adequate validity studies to assist in defining the strengths and limitations of these assessment devices.

The DABERON (developed in 1972, revised 1982) is a nationally standardized, norm-referenced screening test for four-, five-, and six-year-old children, which was designed to be a simple means of predicting readiness for school. The revised manual (1982) states that a high percentage of accurate responses on the DABERON indicates school readiness, while inaccurate responses may indicate potential problem areas, and the need for further diagnostic evaluation. However a search of the relevant literature did not turn up any empirical evidence to clearly establish the predictive validity of this screening instrument. Since the DABERON is an easily administered test which has face validity and is appealing to most children, empirical evidence of its predictive validity would be a useful contribution to the field of early screening and assessment. More specifically, since it is
Currently the instrument being used for all kindergarten screening in
the demographically diverse city of Northampton, Massachusetts (where
the author is one of three school psychologists responsible for adminis-
tering and conducting the kindergarten screening process), evidence
for or against the usefulness and appropriateness of the DABERON for
this purpose would be extremely helpful in determining whether to con-
tinue using this test for screening.

Outline of the Remaining Chapters

This first chapter has briefly reviewed the purposes of early
childhood educational screening, the legal mandates related to the
establishment of "child find" procedures, and the resultant national
movement toward the implementation of screening programs for the early
identification of children with special education needs. The paucity
of sufficiently validated screening instruments for this purpose was
cited as the primary reason for undertaking this study. Also discussed
was the educational significance and expected benefits to be derived
from this research.

Chapter II reviews the significant literature published on screen-
ing for early identification of learning handicaps and prediction of
early academic achievement. Issues pertaining to the screening pro-
cess, as well as to the selection of appropriate screening instruments,
are discussed. Several studies which have examined the predictive
validity of various screening measures are also reviewed.

Chapter III reviews the research questions informing this study,
and discusses its inherent assumptions and limitations. A detailed
description of the measurements, population sample, and statistical methods employed is also included in this chapter.

Chapter IV details the evaluative data, reports the results of each statistical analysis, and presents the answers to the questions posed by this investigator.

Chapter V summarizes the data obtained from this study, and discusses its implications and its relationship to previously published research. Suggestions for further research are included.
CHAPTER II

REVIEW OF THE LITERATURE

This chapter first presents a review of the historical background relevant to the increasing use of early childhood screening instruments and procedures in the United States. It then goes on to examine the goals and processes of early developmental screening, the selection of appropriate screening instruments, and the distinction between developmental screening measures and readiness tests. Finally, it discusses the empirical evaluation of screening instruments, citing several examples of predictive validity studies which are similar to the present study.

Historical Background

When popular education became widespread approximately 125 years ago, those children who were difficult to educate simply did not attend school (Zeitlin, 1976). However in the early 1900s, the French psychologist Alfred Binet urged that children who failed to respond to normal schooling be examined before dismissal and, if deemed educable, be assigned to special classes (T.H. Wolf, 1973, cited in Anastasi, 1988). An ultimate outcome of his position was the development of the first Binet intelligence scales, which began the movement toward intelligence testing in schools. The first objective tests for screening large populations were developed during the World War I period, to screen the suitability of men who were being inducted into the armed forces. When the upsurge of child guidance clinics came about in the
late 1930s, interest became directed toward developing screening measures for preschool children (Barnes, 1982). One of the best known pioneers in this area was Arnold Gesell, who created and published the Gesell Developmental Schedules (Gesell, 1940; Gesell & Armatruda, 1941).

During the past three decades, early childhood education and assessment has attracted much national interest. Increased attention has been focused on the long-held notion that the preschool years are a period of particular importance in developing the foundations essential for later school learning and success (Bloom, 1964; Bruner, 1960). Research in early cognitive development, the growing number of preschool education programs, and widespread concern about the effects of cultural differences on a child's ability to profit from school instruction have all contributed to the increased interest in this area (Anastasi, 1988). A push for more detailed screening measures and more comprehensive screening programs has come from the demands of professionals in the fields of health and social welfare, and government legislation has responded to these demands and provided much of the impetus for the initiation of widespread screening programs for young children (Barnes, 1982).

The first piece of legislation having an impact on early childhood screening practices in this country was passed in 1967, as an amendment to the Social Security Act. It mandated health screening for all children from birth to 21 years of age, whose families received Social Welfare benefits (Hepner & Kaufman, 1985).
In 1968, a two year pilot program known as the Wellesley project was undertaken, the goals of which were to identify young children who showed signs of developmental delay or deviation and to provide special services if needed, both before and upon entry to school. The professed hope was that the early identification and training of high risk children would contribute to the prevention of school failures in later years (Zeitlin, 1976).

In 1970, the Joint Committee on Mental Health of Children published a report advocating legal, medical, and educational reforms in an attempt to diminish inequality in justice, care, and opportunity for children. The report proposed a model for delivering services to children and families, which included the screening and assessment of young children in school for early signs of dysfunction (Zeitlin, 1976).

In 1972, the President's Committee on Mental Retardation called for the development of screening and assessment programs as a preliminary stage to a general program of prevention and remediation (Meier, 1973).

In 1973 Congress passed the Rehabilitation Act, requiring school districts to find all qualified handicapped children not receiving a public education and notify parents or guardians of their right to attend school (Hepner & Kaufman, 1985). The same year, the American Academy of Pediatrics issued a statement supporting the establishment of early identification programs in each community, to follow children considered to be potential high risks because of past history or present unusual behavior which might interfere with normal learning. And in 1975, the State of California authorized massive expenditures and
expansion of resources to the very young, on the theory that it is better to diagnose and correct learning problems during a child's first few years of school than to provide costly and less effective remedial programs later (Zeitlin, 1976).

The piece of legislation which has had the most profound effect on the screening movement is Public Law 94-142, the Education For All Handicapped Children Act, which was passed in November of 1975. This piece of legislation was designed to ensure that all handicapped children would be educated at a level commensurate with their individual needs, in the least restrictive environment available. A major consequence of this Act has been a demand on the part of both state government officials and educators for earlier and earlier recognition or detection of handicapping conditions prior to entry into grade one (Barnes, 1982).

The introduction of P.L. 94-142 has thus further focused attention on the identification of high-risk children, i.e., those likely to encounter difficulty once they enter the mainstream of formal education. As of September 1978, P.L. 94-142 applies to all handicapped children, ages 3 to 21 inclusive. Because it is difficult to differentiate a "developmental lag" that falls within the range of expectancy from "risk" at the younger ages (Reeve & Holt, 1987; Zeitlin, 1976), the aim of this law at the preschool level is to identify children with potential learning problems, to arrive at diagnostic decisions, and to plan appropriate programs of instruction (Boehm & Sandberg, 1982).

As a result of P.L. 94-142, public schools across the nation have become responsible for providing appropriate educational opportunities
for preschool handicapped children ages three to five. Although states have interpreted the law in different ways, the mandate has resulted in nationwide efforts aimed at the identification of potential "risk" in the 3- to 5-year-old preschool child, the ultimate aim being prevention or remediation of problems associated with early school failure (Boehm & Sandberg, 1982).

Why Screen?

The primary goal of early childhood screening, simply stated, is to predict quickly whether a child might have difficulty succeeding in school, for whatever reason. Children thus identified are then referred for further diagnostic evaluation, and determination of need for special educational intervention or placement. Zeitlin (1976) cites statistics estimating that 25% of all children entering school will show some signs of developmental deviation. Of those children, 15% can be expected to exhibit some form of mild learning disability calling for intervention, with another 3% to 5% having more severe problems. Other researchers have reported similar statistics, but note that the actual percentages may vary tremendously across school districts (Thurlow, O'Sullivan & Ysseldyke, 1986).

Another reason to screen young children for potential educational handicaps has to do with recognition of the importance of early intervention, i.e., that there are specific stages in a child's development when intervention is more likely to be effective (Hepner & Kaufman, 1985; Powers, 1974; Zeitlin, 1976). It further stands to reason that identifying children at risk for failure, and providing them with
appropriate help early in their educational careers will be ultimately beneficial in terms of long-term services that will not be required, increased educational productivity, and the enhanced self-concept of children who otherwise might have experienced academic failure before services could be provided (Judy, 1986).

Koppitz (1971), in a five year study of children with learning disabilities, concluded that extremely immature and otherwise vulnerable children should be identified at the time of school entry and be given help before they develop serious learning and emotional problems. She recommended that "all children should be screened prior to their enrollment in kindergarten."

Judy (1986) posited that early screening is essential for educational accountability, because it provides educators with information critical to successful learning for the population of children screened. Screening measures provide information about children's areas of strengths as well as weaknesses, and can be used to plan curriculum, or to devise remediation strategies which capitalize on strengths to develop the weaker modalities. Meisels (1986) concurs that using appropriate tests will yield information that, when linked to individualized program planning, can significantly improve a child's early school experience.

Uses for screening may include any or all of the following: (1) to identify children who may have special learning needs; (2) to refer identified children for further assessment; (3) to identify children who should be monitored over a period of time to determine if they need special help or are just immature in their development;
(4) to give an overview of the developmental range of the screening population; and (5) to facilitate reexamination of existing programs. Initial screening should never be used for placement or exclusion of children (Meisels, 1986; Zeitlin, 1976).

The Screening Process

The screening process has been defined as "a process of early detection for all those preschool children, who, for a variety of reasons (social, emotional, intellectual, biological, physical, linguistic, environmental, or any combination of such), will be unable to attain optimum growth and/or normal development" (Barnes, 1982, p. 7). It is used to identify those children in the general population who may be at-risk for a specific disability, or who may otherwise need special services or programs in order to develop to their maximum potential. The screening process has become a critical component of "child find," as mandated by P.L. 94-142 (Boehm & Sandberg, 1982).

In considering the screening process, the major problems appear to be definition and instrumentation. What is the goal of screening? When should screening take place? What does the high risk learner look like, and how do we assess those characteristics? What should screening measure or observe? How should children's abilities be evaluated? How do we determine a cutoff point at which a potential problem exists?

Many schools utilize a standard battery of screening instruments that is administered to all children entering kindergarten. It does not involve in-depth testing, but rather the use of inexpensive and easily administered group or individual instruments. The determination
of a cutoff point is frequently resolved by using a percentage as a
guideline, e.g., the lowest 10 or 15 percent is referred for further
evaluation. Many screening programs also establish a "watch and wait,"
or "rescreen" category for those scoring in the questionable range on
the continuum (Meisels, 1986; Zeitlin, 1976).

Most school screening programs concentrate either on those about
to enter school or those already in kindergarten. An advantage of
screening before entry is that staff can then determine the educational
needs of the group of children who will be entering kindergarten that
year, and plan curriculum accordingly. On the other hand, the advan-
tages of screening early in the kindergarten year are that the children
are easily accessible, are more comfortable in the school environment,
and are usually somewhat familiar with the adults who do the screening.

The personnel involved in the screening process may vary widely in
different locales. Common approaches include training volunteers to
conduct the screening under supervision, a team or individual approach
using school specialists and teachers, or hiring outside specialists
specifically for screening. Barnes (1982) recommends the training of
paraprofessional and nonprofessional volunteers in the use of objective
measures as the most efficient in terms of cost, time, and personnel.

In a state by state survey of screening requirements and practices
(Gracey, Azzara & Reinherz, 1984), it was noted that 33 states mandate
some form of screening in the preschool or early school years, with
nearly all screening programs conducted by the school systems. Of
these states, 22 require screening in more than one domain (physical,
language, cognitive, behavior), and four more recommend this practice
in their guidelines. Only 11 states require comprehensive screening programs covering all four of the major domains surveyed. The format of the screening programs (e.g., whether screening is conducted more than once, one or multiple raters used, information collected from more than one source, etc.) and the type of information collected (e.g., SES, parents' educational level, developmental history, etc.) varies widely. Based on a longitudinal study conducted by one of the survey authors (Reinherz, Gordon, Morris & Anastas, 1983) it was recommended that screening programs should ideally consist of ongoing monitoring and referral, with multiple raters and multiple assessments in content areas found to be predictive of later school difficulties.

Selecting Appropriate Screening Instruments

Many writers in the field have offered suggestions for selecting an appropriate screening instrument, which are usually based on developmental theory, research, and/or experience. Some of the more widely recommended characteristics of screening tests are outlined below.

When the purpose of screening is to identify children in need of more intensive assessment, the most essential properties of the selected instrument must be: (1) high sensitivity (i.e., yielding a positive result in a high proportion of children who actually have subsequent school problems); (2) high specificity (i.e., yielding a negative result in a high proportion of those who do not have subsequent school problems; and, most importantly, (3) high predictive value, resulting in a high proportion of children with and without subsequent learning problems being correctly identified as such.
Therefore one characteristic of a good screening instrument is a demonstrated low rate of both false positives and false negatives, i.e., that it identifies children who are at-risk with predictive accuracy (Barnes, 1982; Bradley & Caldwell, 1974; Cadman, Chambers, Walter, Feldman, Smith & Ferguson, 1984; Mercer, 1979).

In a related vein (and the source of some of the most common criticisms levied against many screening instruments), such tests should be technically adequate. They should adequately measure what they are supposed to measure (validity), produce stable measures (reliability), and be based on an appropriate sample (norming) (Chew & Morris, 1984; Collegan, 1976; Kaplan, 1985; Klein, 1977; Meisels et al., 1984; Piersel & Kinsey, 1984; Thurlow et al., 1986). The normative sample must include children like those to be tested, from various geographic areas, racial, and ethnic groups, and types of schools, and should include a wider age range than the one for which the test is constructed (Graue & Shepard, 1988; Meisels, 1986; Powers, 1974; Zeitlin, 1976).

Specific recommendations for screening instruments generally include many or most of the following characteristics. They should be inexpensive, brief, simple to administer, and easy to interpret (Barnes, 1982; Chew & Morris, 1984; Hills, 1987; Kaplan, 1985; Meisels et al., 1984; Obrzut et al., 1981; Zeitlin, 1976). Tests should be individually administered, and scoring should be objective, based on observable behavior rather than subjective judgment (Chew & Morris, 1984; Powers, 1974; Zeitlin, 1976). Screening instruments should have a high utility value (i.e., be cost-efficient and yield useful
results), and be noninvasive or not highly objectionable to the child (Barnes, 1982; Cadman et al., 1984; Chew & Morris, 1987; Powers, 1974).

With regard to content, screening tests should be multidimensional (i.e., covering several areas of development) (Graue & Shepard, 1988; Meisels, 1986; Meisels et al., 1984; Zeitlin, 1976), and contain a broad range of item difficulty so that there are adequate base and ceiling levels (Chew & Morris, 1987; Powers, 1974). In addition, they should be culturally non-biased, as well as noncategorical (i.e., identifies high risk children regardless of the reason for the potential learning problem) (Chew & Morris, 1984; Meisels, 1985; Zeitlin, 1976).

**Developmental Screening vs. Readiness Tests**

As stated by Meisels et al. (1984), the distinction between school readiness tests and developmental screening instruments "rests on the difference between a child's ability to acquire skills, in contrast to a child's current level of skill achievement and performance" (p. 26). These two types of tests should not be considered interchangeable, as poor performance on a readiness test may reflect only limited prior experience rather than an impairment that affects the child's ability to acquire knowledge. Thus it is crucial that any screening test selected be appropriate for its intended purpose, but unfortunately there tends to be a great deal of confusion over how to distinguish these two types of tests. Terms such as "screening," "readiness," and "developmental" are used in descriptions of both developmental screening and school readiness tests, making their purpose difficult to
ascertain (Bradley & Caldwell, 1974; Graue & Shepard, 1988). Characteristics of these two types of tests are outlined below.

Readiness Tests

Historically, readiness assessment has dealt with academic-process testing of abilities which are thought to underlie most academic skills. According to Boehm & Sandberg (1982), readiness tests assess a child’s academic skills as indicators of preparedness for entry into kindergarten or first grade. The tests generally measure specific behaviors and concepts that are the building blocks for basic academic skills, as well as for general instruction in the primary grades. They have much in common with intelligence tests in these younger grades, but usually place more emphasis on the abilities found to be important in learning to read. Some attention is also given to the prerequisites of numerical thinking and to the sensorimotor control required in learning to write. Among the specific functions frequently covered are visual and auditory discrimination, motor control, aural comprehension, vocabulary, quantitative concepts, and general information (Anastasi, 1988).

Readiness tests are sometimes norm-referenced (Reeve & Holt, 1987), although criterion-referencing is more typical, since that makes it possible to specifically identify those concepts and skills lacking and needing to be acquired for immediate school success. The content of a selected readiness test should therefore be consistent with the values and curriculum of the school the child is entering if it is to
be of any use for informing classroom instructional decisions (Boehm & Sandberg, 1982; Graue & Shepard, 1988).

Salvia and Ysseldyke (1985) noted that one of two orientations is commonly adopted in devising these instruments. A skills orientation takes the position that readiness involves basic skills development, i.e., mastery of directly school-related skills (e.g., ability to hold a pencil, to count, to name letters, etc.). A process orientation, on the other hand, leads to the construction of readiness tests which attempt to assess the underlying processes (e.g., intelligence) believed necessary for the acquisition of academic skills and knowledge.

In simple terms, readiness tests look at skills a child has acquired. They are more a content- than a process-oriented test, and are best used for curriculum planning and class placement decisions, although many schools mistakenly use them as screening tests (Warren, 1988). Examples of several well-known readiness tests include the Metropolitan Readiness Tests (Nurss & McGauvran, 1976), the Gesell School Readiness Test (Ilg & Ames, 1972), the Brigance K and 1 Screening Test (Brigance, 1982), and the Cognitive Skills Assessment Battery (Boehm & Slater, 1977).

**Developmental Screening Tests**

Also referred to as multifunction screening tests (Boehm & Sandberg, 1982), these are intended to identify children with potential learning difficulties who might benefit from remedial programs or special placements. These tests are designed to assess a child's
ability or potential to acquire skills, not their current skill achievement or level of general knowledge. They attempt to discover what processes a child has, not the facts he or she knows. A good screening test should determine if a child can reason, retain information, comprehend language, express himself or herself using language, and be aware of objects spatially (Warren, 1988). It should be norm-referenced, sample a child's performance in multiple developmental areas, and have good reliability and predictive validity (Meisels, 1986).

Many developmental screening tests have been developed specifically to aid in P.L. 94-142 mandated identification of high-risk children. Examples of this type of instrument include the Early Screening Inventory (Meisels & Wiske, 1983), the McCarthy Screening Test (McCarthy, 1978), and the Minneapolis Preschool Screening Inventory (Lichtenstein, 1980).

Evaluation of Screening Instruments

The effectiveness of screening measures may be evaluated in two main ways. One is in terms of how well they have been constructed and standardized, and their levels of consistency over time. The second way is in terms of how accurately their scores predict to certain outcome measures, which will be the focus of the present discussion of research in this area.

Because of the nature of the task it is designed to do, the results obtained on a screening test should correlate highly with such outcome measures as professional diagnosis and/or the results of
selected, valid criterion measures. One of the most common criticisms levied against various screening measures concerns their low rates of predictive validity, or the lack of any predictive validity data at all (Gallenani, O'Regan & Reinherz, 1982; Lindeman, Goodstein, Sacks & Young, 1984; Shepard & Smith, 1986; Telegdy, 1977; Tsushima et al., 1983). One of the keys to a screening measure's effectiveness is the number of false positives and false negative predictions it generates, as well as the actual number of true positives and negatives. The lower the number of false positives and negatives, the higher the screening test's predictive accuracy for identification of at-risk children (Satz, Taylor, Friel & Fletcher, 1979).

Predictive validity estimates are frequently based on the correlation between an individual's test score and some subsequent criterion measure. These are generally standard measures which have been validated and accepted by professionals in the field as accurate indicators of the abilities or behaviors under investigation. In all predictive validity studies, the outcome measure follows only after some weeks or months have elapsed, or some treatment procedure has been completed. Test score data and outcome measure data that are collected at approximately the same time can be used to estimate concurrent validity, but not predictive validity. The studies which are described below were selected because of their similarity in purpose and procedure to the present study.
Examples of Predictive Validity Studies

Lindeman and his associates (1984) conducted a study which compared results obtained on the Yellow Brick Road Test (YBRT) to those from the Metropolitan Readiness Test (MRT), and found a correlation of .67. The YBRT predicted those who perform well with 100% accuracy, but misdiagnosed 83% of those who were predicted to do poorly. The authors noted that it was rare for a screening test to report its predictive validity, i.e., the frequency rate of accurate and inaccurate predictions, and recommended that a prediction-performance comparison study be done on all screening instruments before they are considered valid and efficient.

In 1983, Tsushima et al. gave the Screening Test of Academic Readiness (STAR) to 59 incoming kindergarten boys at a private school for children with high I.Q.s. The Metropolitan Readiness Test (MRT) and teachers' rankings were used to assess the predictive validity of the STAR. Except for a low but significant correlation (.24) between the STAR and the MRT Numbers subtest, no substantial relationship was found between the STAR and the MRT. The correlation between the STAR and teachers' rankings was statistically significant but low (.23). The investigators concluded that the STAR was not effective for screening children with high I.Q.s, and that local validation should be done before the test is used in other populations.

Meisels et al. (1984) conducted a longitudinal investigation of the predictive validity of the Early Screening Inventory (ESI) in which results obtained from kindergarten screening with the ESI were compared with a variety of measures of school success for 465 students from
kindergarten through grade four. The outcome measures used included a cumulative score derived from report card grades, an indication of whether the child was referred for or received special education services, and whether the child was promoted or retained at the end of the school year. Correlation and multiple regression analyses were performed, yielding high correlations between the ESI and success in kindergarten (.70, p<.001), and moderate correlations with first and second grade performance (.50 and .52, respectively, p<.001). Sensitivity and specificity of the ESI were found to be moderate to high in kindergarten and first grade, becoming less stable thereafter. The authors concluded that the ESI is a valid and reliable developmental screening instrument.

The predictive validity of the Gesell School Readiness Tests (GSRT) was examined by Graue and Shepard (1988) by correlating measured developmental age from the GSRT with performance in first grade. A sample of 45 students referred by their teachers for developmental testing and a random sample of 106 students were tested with the GSRT, and a small positive correlation was found between GSRT developmental ages and first grade report card grades (.23). Additional outcome measures (Metropolitan Readiness Test, Comprehensive Test of Basic Skills) were available for a subgroup of the total sample. Correlations between the GSRT and the MRT and CTBS were .40 and .40 (p<.05), thus indicating that the GSRT has only modest predictive validity for standardized tests. The authors concluded that the low predictive validity of the GSRT does not support its use for school readiness assessments leading to placement decisions.
The ability of the Vane Kindergarten Test (VKT) to predict academic achievement was assessed with 289 pre-kindergarten and kindergarten children (Powers, 1974). One year after the administration of the VKT, the children were given either the MRT or the Stanford Achievement Test (SAT). MRT-VKT coefficients were all significant (p<.01), and ranged from .40 to .53. Several of the SAT-VKT coefficients were significant, the largest being between the VKT and SAT vocabulary subtests (.47). Although most of the correlations were significant, the author posited that the VKT did not have sufficiently high predictive validity for use in assessment and program planning for individual children.

Swanson, Payne, and Jackson (1981) used a sample of 136 entering first grade students to assess the predictive validity of the group-administered MRT and the individually-administered Meeting Street School Screening Test (MSSST) against end-of-first grade Metropolitan Achievement Test (MAT) scores. Validity coefficients obtained ranged from .73 to .84, with the MRT tending to yield slightly higher correlations than the MSSST on the average. Both tests were judged to be valid predictors of first grade achievement.

**Summary and Implications**

This chapter has examined some of the reasons for the growth of early childhood screening practices in the United States, and discussed guidelines for selecting and evaluating screening instruments. It was observed that the research literature consistently points out the need for more and better evaluative studies of screening tests, particularly
of their predictive validity, before they are presented as valid and reliable for use in a screening program. Studies which have been carried out on various screening measures indicate that the predictive validity of these instruments can vary widely, and must not be assumed without empirical data.

The DABERON was selected as the screening instrument to be used in the Chapter 766 mandated kindergarten screening program in Northampton, Massachusetts, yet there appears to be little empirical data to either support or negate its validity as a predictor of academic success in the primary grades. Nor is there any data available with regard to its predictive accuracy for identifying children with special academic needs. Therefore, it seems necessary and appropriate to undertake to obtain such data, in order to justify the continued use of this instrument for screening the local population of kindergarten children. The results of this study may be expected to generate broader interest as well, since a major test distributor (Western Psychological Services) has recently begun marketing the DABERON nationwide.
CHAPTER III

METHODOLOGY

The main purpose of this study was to evaluate the predictive validity of the DABERON (Danzer, Gerber & Lyons, 1982), an instrument which has been labeled by its authors as "a screening test for school readiness." The primary question addressed was whether the DABERON is able to accurately predict those children entering kindergarten who are likely to be at high risk for academic failure, and who may need special education services. Both DABERON total scores and subcomponent scores were examined for their predictive value. Of secondary interest was the question of whether there was any significant relationship between subjects' performance on the DABERON and individual characteristics such as their gender, age, cultural background (Hispanic vs. non-Hispanic), socioeconomic status, or prior educational experience.

Research Questions

The following nine research questions were examined in this study:

Question 1: Does performance on the DABERON (pass or fail) accurately predict the presence or absence of an educationally handicapping condition requiring Special Education services?

Subjects' Special Education status was defined as whether or not they were placed on an individualized education plan (IEP) at any time during their school tenure to date, which was assumed to signify the presence of a diagnosed educational handicap. It was hypothesized that
kindergarten children who passed the DABERON would be significantly less likely to have an IEP in the early elementary years, and those who failed the DABERON would be significantly more likely to be placed on an IEP at some time during their early school years. Thus, subjects who passed the DABERON and were never placed on an IEP were considered to be "true negatives," while those who passed but did end up being placed on an IEP at some point were considered to be "false negatives." Subjects who failed the DABERON and were subsequently placed on an IEP at some time during their first few years of school constituted the "true positives" group, while those who failed but were not given an IEP were the "false positives" group.

Question 2: Does a statistically significant relationship exist between the DABERON total scores and the Metropolitan Readiness Test (MRT) (Nurss & McGauvran, 1976) composite scores?

The subjects in this study were administered the DABERON at the beginning of their kindergarten year, and the MRT at the end of that year. The MRT, a norm-referenced group test, is one of the most widely used measures of early academic skills. Its predictive validity with regard to success in first grade has been fairly well established (Bolig, 1973; Lessler & Bridges, 1973; Swanson, Payne & Jackson, 1981; Telegdy, 1977), and it has frequently been used as a standard against which alternate screening instruments have been judged (Chew & Morris, 1984; Hayes, Mason & Covert, 1975; Obrzut, Bolocofsky, Heath & Jones, 1981; Scourfield, 1982; Seda & Michael, 1971; Tsushima et al., 1983). It was hypothesized that children who performed well on the DABERON would also tend to perform well on the MRT, and vice versa.
Question 3: Does a statistically significant relationship exist between the DABERON total scores and the Kindergarten Progress Report (KPR) total scores?

Kindergarten teachers filled out a progress report checklist for each child in their class at the end of the school year, rating their performance on a variety of academic and non-academic behaviors. It was hypothesized that children who scored higher on the DABERON at the beginning of the year would also be those who would ultimately tend to receive a greater number of points from their teachers on the KPR.

Question 4: Is there a statistically significant relationship between any of the DABERON subsection scores and the MRT composite scores and/or KPR total scores?

The DABERON test is composed of several discrete sections (described in greater detail under Instruments), which for the purposes of this study are termed "subsections." This research question was designed to evaluate whether performance on any of these individual subsections was significantly related to performance on either of the two criterion measures (MRT and KPR), irrespective of subjects' total DABERON performance.

Question 5: What differences are there in DABERON performances related to subject gender, and how significant are these differences?

Girls tend to mature faster, and are often more verbal than boys. Since the DABERON is a heavily language-based test, girls may tend to score higher than boys on the average, regardless of whether or not an educationally handicapping condition exists. This question was
designed to assess whether this subject variable had a significant impact on the overall results obtained from this study.

Question 6: What differences are there in DABERON performances related to subject age at the time of testing, and how significant are these differences?

The DABERON has norms for children from 4 to 6 years of age. The younger a child is when tested, the fewer correct responses he or she must give in order to pass. However the kindergarten child who is relatively older when tested (e.g., 5.6 to 6.0 years) must respond correctly to most of the items in each subsection--failing one or two of the subsections might be enough to cause the child to fail the entire test. Since the DABERON does contain subsections which tap into acquired knowledge (e.g., colors, number concepts) as opposed to underlying developmental processes, it was hypothesized that some of the older kindergarten children may have been more likely to fail the DABERON because of inadequate exposure to such knowledge, regardless of whether or not an actual handicapping condition exists. This question was therefore designed to assess whether this subject variable had a significant impact on the overall results obtained from this study.

Question 7: What differences are there in DABERON performances related to subject socioeconomic status (SES), and how significant are these differences?

Children from low SES families often tend to be deprived of a broad range of environmental and cultural experiences, and are usually considered at high risk for educational problems in the early school
years. Children with low SES status may therefore do less well on the DABERON than middle or high SES children, although this may not imply the existence of an actual handicapping condition warranting further diagnostic evaluation. This variable was tested by examining the relationship between SES (low income vs. not low income) and performance on the DABERON.

Question 8: What differences are there in DABERON performances related to subject cultural background (Hispanic vs. non-Hispanic), and how significant are these differences?

A significant proportion of the research sample were children of Hispanic origin (16%). Although there is a Spanish translation of the DABERON (which was administered to any Hispanic child who failed the initial screening in English), it is possible that there are other variables related to cultural background (besides language) which might affect subjects' performance on the DABERON. If so, poor performance on the DABERON by an Hispanic child should not necessarily be considered indicative of a need for further diagnostic evaluation.

Question 9: What differences are there in DABERON performances related to subjects' prior educational experiences, and how significant are these differences?

Young children today are frequently exposed to some type of preschool, daycare, or organized playgroup experiences prior to entering kindergarten. Those who have had such exposure may be likely to perform better on the DABERON than those who have not had such prior experiences, regardless of whether or not an actual difference in
ability or school readiness exists. This variable was tested by examining the relationship between prior educational experience and performance on the DABERON.

Assumptions

1. It is possible for a screening test to accurately predict a high proportion of children who will ultimately be diagnosed as having an educationally handicapping condition.

2. Performance on the MRT is indicative of subjects' ability to achieve satisfactorily in their early school years; therefore satisfactory performance on the MRT is contraindicative of an educationally handicapping condition.

3. KPR ratings are indicative of subjects' ability to achieve satisfactorily in first grade; therefore the higher the KPR total, the less likely it is that the subject will be found to have an educationally handicapping condition.

4. Subjects who are on an IEP have an educationally handicapping condition which warrants Special Education services.

Scope and Limitations

The scope of this study was to determine whether kindergarten children's performance on the DABERON screening test predicts, at an acceptable level, whether or not they may have an educationally handicapping condition requiring Special Education intervention. The following limitations exist:
1. The results of this study can only be applied to similar populations, that is, kindergarten children attending public school in small northeastern cities. National applicability cannot be claimed.

2. The population chosen represents a very heterogeneous sample, which makes it difficult to account for all of the possible variables which might influence test performance. There is some research which indicates that the predictive validity of a screening instrument may vary depending upon the specific population being tested (Tsushima et al., 1983).

3. A certain proportion of the target population of first, second, and third graders (28%) could not be included in the study because they had not attended public kindergarten in Northampton.

4. Some of the subgroupings obtained may not have been large enough to test some of the research questions adequately.

5. Social/emotional factors which might hamper academic performance and lead to a Special Education referral were not assessed through the screening process, since the DABERON does not specifically address social/emotional functioning. This may not be as serious a limitation as it appears, however, since emotional disturbance serious enough to constitute a "handicapping condition" within the meaning of the law is likely also to affect performance on a test which samples a fairly broad range of behaviors.

6. The teachers who were rating the children at the end of their kindergarten year had participated in the initial screening, and were therefore aware of subjects' performance on the DABERON.
7. It is not known how consistently the DABERON was administered across the research sample, since screening teams may have been composed of different people in different years. However scoring on the DABERON is objective, and subjective interpretation is quite limited, thus minimizing the consistency factor compared to a number of other popular screening tests.

8. Parents' educational level was not known for the subjects in this study. This has been cited as an important criterion in the determination of SES (Kaufman & Kaufman, 1983). Some youngsters who qualify for free or reduced school lunches (and would thus be classified as low SES in this study) have parents who are well-educated university students currently living on a fixed low income while they attend school. These children may not experience the type of environmental and cultural deprivation typically associated with low SES students.

**Population**

The population for this study was comprised of children attending first, second, or third grade at a large public elementary school in September 1988. The school, located in the city of Northampton, Massachusetts (population 30,000), had a total enrollment of 536, of which 78% were Caucasian, 15% were Hispanic, and 7% included all other minorities (American Indian, Black, and Asian). The final sample included all of the school's first, second, and third graders in September 1988 for whom the relevant data were available in their cumulative record folders. Specifically excluded were children who
did not attend public kindergarten in Northampton (and thus were not administered the DABERON). Out of a total population of 228 first, second, and third grade students, 165 met the above criteria for inclusion in this study.

**Instruments**

**DABERON**

The DABERON was developed as an individually-administered, norm-referenced screening battery to predict preschool children's readiness for school, and it is appropriate for children who range in age from 4 to 6 years old. It has a total of 122 items, surveying knowledge of body parts (12 items), color and number concepts (33 items), the functional use of prepositions and plurals (12 items), ability to follow directions (7 items), general knowledge (30 items), visual perception and visual motor integration (10 items), gross motor development (12 items), and the ability to categorize (6 items). The content of the DABERON was reported to be compiled from a study of child development research in specific areas. Authors reviewed included Gesell, Beery, Bangs, Piaget, Terman, Merrill, and other recognized experts. Administration time is approximately 20-30 minutes.

The DABERON was standardized on a sample of 1,358 preschool and school-aged children from 15 different states across the United States, drawn from a population defined as "normal children from the mainstream of preschool and school aged education" (Danzer et al., 1982, p. 30). The sample was representative in that it was balanced to conform closely to U.S. Census distributions (1976) for family income, gender,
and ethnicity (including American Indian, Caucasian, Black, Asian, and Hispanic). Data obtained from the national standardization study was used to determine the content validity and the criterion related validity of the DABERON.

The authors of the DABERON Test instrument also conducted a predictive validity study by correlating it with an independent behavioral checklist. The ratings of 15 experts were averaged to compose a checklist of school readiness skills. According to the authors, the items were drawn from the literature and refined by experts, and in no way reflected the DABERON test or its specific test items. All 1979-1980 kindergarten children who were part of the national standardization sample were contacted for evaluation with this newly-developed behavioral checklist for school readiness. Their first grade teachers completed the checklists, which were then correlated with their 1979-80 DABERON test scores. The predictive validity correlation between the kindergarten DABERON test scores and the followup behavioral checklists was reported to be $r(216) = .84; p<.001$ (Danzer et al., 1982).

**Metropolitan Readiness Test**

The MRT is a widely used, group administered readiness battery that is available in two levels: Level I for use from the end of pre-kindergarten to the middle of kindergarten; Level II for use at the end of kindergarten and the beginning of first grade. Level II was the form which had been administered to the children in this study. This test battery is orally administered, requiring the child to make simple marks in a test booklet. A practice booklet is administered in
advance, to acquaint children with the materials and procedures they will encounter in the test. Administration time is 80-90 minutes, usually divided into two or more testing sessions.

Level II of the MRT is composed of eight subtests: beginning consonants, sound-letter correspondence, visual matching, finding patterns, school language, listening, quantitative concepts, and quantitative operations. Scores on these subtests are grouped into component scores labeled Auditory, Visual, Language, and Quantitative. A composite score, identified as a Pre-Reading Skills Composite, is calculated from the first three component scores. National norms, derived from a representative national sample of about 30,000 kindergarten and first grade children, are available for finding percentile ranks and standard scores. Several studies with earlier editions of the MRT yielded correlations in the .70s with group intelligence tests, and consistently high validities have been reported when the MRT is checked against end-of-year achievement tests (Anastasi, 1988). The MRT is generally considered to be a valid and reliable predictor of success in the early primary grades.

Kindergarten Progress Report

A Kindergarten Progress Report (KPR) is issued for every kindergarten child in the middle and at the end of the school year. It consists of the teacher's ratings of the child's abilities and progress in the following seven areas: Personal, Social Development, Work Habits, Reading Readiness, Language Development, Mathematics Readiness, and Physical Development. A check system is used, in which two checks
indicates "satisfactory performance," one check indicates "having difficulty," and 0 checks indicates "no judgment made." Only the totals obtained on the final (i.e., end of the year) KPRs of the research subjects were utilized in this study.

Data Collection

All of the data was obtained from the cumulative record folders of the children who comprised the research sample. The data collected for each subject included the following variables: gender; age at time of testing with the DABERON; socioeconomic status (low income vs. not low income, as determined by eligibility for free or reduced school lunches); cultural background (Hispanic vs. non-Hispanic); prior educational experience (preschool or organized group day care); total DABERON score; pass/fail status on the DABERON; DABERON subsection scores; MRT subsection scores and Composite score; quantified total of Kindergarten Progress Report (total of 0, 1, and 2 checks) as a whole, and for each individual area rated; referral status (referred/not referred for diagnostic evaluation); Special Education status (IEP or no IEP); and identity of kindergarten teacher.

As specified by Northampton kindergarten screening criteria, subjects' pass/fail status had been determined by comparing the total score obtained on the DABERON to the established norms for his or her age (to the nearest month). Subjects whose total DABERON score was more than one standard deviation below the norm for their age were classified as screening failures, and were referred for more intensive diagnostic evaluation. An exception to this rule was applied to
Hispanic subjects, in that any Hispanic child who failed the initial DABERON screening was subsequently rescreened with the Spanish version of the DABERON before a decision was made whether to refer for further evaluation. In cases where the cumulative record folder of an Hispanic child contained both the English and the Spanish DABERON protocols, the higher score was utilized as data for this study.

**Data Analysis**

Both descriptive and inferential statistics were used for analysis of the data. Frequency distribution tables were first calculated for all of the data. 2x2 contingency tables (crosstabulations) were then constructed for all variables having the characteristics of being discrete (categorical) and numeric. Appropriate inferential statistics were employed to address each research question, depending upon whether the data was categorical or continuous. Statistics that were used included t-tests, Pearson product moment correlations, and chi-square, each of which were accompanied by a level of significance. Additionally, the phi statistic was computed for several of the categorical variables. In the event of missing data, sub-analyses were conducted to make use of all the available data.

Due to the categorical and nominal nature of some of the research data, 2x2 contingency tables were constructed and the chi-square statistic was used to determine if the variables examined were likely to be independent or related. Since there was only one degree of freedom, Yates correction for continuity was used in the calculation of chi-square, as recommended by Minium (1978, p. 437). When the probability
level of chi-square indicated a significant relationship between two
variables, the phi statistic was calculated in order to assess the
strength of that relationship. These statistics were applied to the
following research questions:

Question 1: Does performance on the DABERON (pass/fail) accurately
predict the presence or absence of educationally handicapping condi-
tions requiring Special Education services?

Question 5: What differences are there in DABERON performances
related to subject gender, and how significant are these differences?

Question 7: What differences are there in DABERON performances
related to subject SES (low income vs. not low income), and how sig-
nificant are these differences?

Question 8: What differences are there in DABERON performances
related to subject cultural background (Hispanic vs. non-Hispanic), and
how significant are these differences?

Question 9: What differences are there in DABERON performances
related to subject prior educational experiences (preschool or organ-
ized group daycare), and how significant are these differences?

Data collected for the following research questions were continu-
ous in nature and were assumed to be linearly related, therefore
Pearson product-moment correlations were used to analyze them. These
research questions included:

Question 2: Does a statistically significant relationship exist
between the DABERON total scores and the MRT composite scores?

Question 3: Does a statistically significant relationship exist
between the DABERON total scores and the KPR total scores?
Question 4: Does a statistically significant relationship exist between any of the DABERON subsection scores and the MRT composite scores and/or the KPR total scores?

Question 6: What differences are there in DABERON performances related to subject age at the time of testing, and how significant are these differences?

In addition to the statistics reported above, t-tests were performed for the variables in Questions 5, 6, 7, 8, and 9, in order to assess whether age, gender, SES, cultural background, and/or prior educational experience affected total DABERON scores. In order to ensure the most conservative estimates, separate rather than pooled variance estimates were used for t-tests in which the standard error estimates of the two sample means were significantly different (i.e., greater than a 1:2 ratio).

In conducting the data analysis, the SPSS Program was utilized through the computer services at the University of Massachusetts. Consultation services were also obtained for the statistical analyses.

Significance

The study is of potential benefit for three reasons:

1. It determines if there is a significant relationship between children's performance on the DABERON and eventual diagnosis of an educationally handicapping condition requiring Special Education services.

2. The presence or absence of such a relationship has implications for the use of the DABERON as a screening instrument in mandated
kindergarten screening programs (such as the one in Northampton) which are intended to screen for educationally handicapping conditions.

3. If a significant relationship is found to exist, it provides empirical evidence for the predictive utility of the DABERON as a screening instrument.
CHAPTER IV
RESULTS

This chapter presents the results obtained from the statistical analyses of the data, and addresses the research questions in a sequential and systematic fashion. Both descriptive and inferential statistics of the study are reviewed.

Demographic Data

The demographics of the research sample (N=165) were as follows. The sample included 78 males and 87 females (47% and 53% respectively), whose age at the time of DABERON testing ranged from 57 to 76 months (X=63 months, SD=3.8 months). Of the total sample, 81% were Caucasian (n=133), 16% were Hispanic (n=26), and 3% (n=6) represented other minority groups (Black, Asian, American Indian). A total of 63 subjects (38%) were considered low SES (determined by eligibility for free or reduced school lunches), while all others (n=102, 62%) were categorized as "not low SES." Of the Hispanic youngsters, 85% were classified as low SES, while only 29% of non-Hispanic youngsters were so classified. Fifty-three percent of the sample (n=88) were documented as having had some type of formal educational experience prior to kindergarten, and 16% (n=26) did not have any prior educational experience. Prior educational history was unknown for the remaining 31% of the sample (n=51). Of those for whom prior educational history was known, 84% of the Hispanic youngsters had attended preschool or organized
group day care prior to kindergarten entry, as had 76% of the non-Hispanic children.

Results

Question 1: Does performance on the DABERON (pass/fail) accurately predict the presence or absence of educationally handicapping conditions requiring Special Education services?

In order to answer this question, the variable of Pass/Fail on the DABERON was crosstabulated with the variable IEP/no IEP, which was assumed to signify the presence of a diagnosed handicapping condition. The chi-square statistic (with Yates correction) was then applied in order to determine whether those two variables are independent or related, and the phi statistic was utilized to assess the strength of the relationship. The results are presented in Table 1.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>DABERON</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N=165)</td>
<td>Pass</td>
<td>Fail</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n=150 (91%)</td>
<td>n=15 (9%)</td>
</tr>
<tr>
<td>IEP</td>
<td></td>
<td>n=29</td>
<td>n=13</td>
</tr>
<tr>
<td>n=42 (25%)</td>
<td></td>
<td>18%</td>
<td>8%</td>
</tr>
<tr>
<td>false negatives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no IEP</td>
<td></td>
<td>n=121</td>
<td>n=2</td>
</tr>
<tr>
<td>n=123 (75%)</td>
<td></td>
<td>73%</td>
<td>1%</td>
</tr>
<tr>
<td>true negatives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>false positives</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chi-square = 29.13, N=165 (1), p<.0001
Phi = .42
Sensitivity = .31
Specificity = .98
Hit Rate = .81

In this study, the overreferral rate for the DABERON (false positives/false positives and true positives) was .13 or 13%. The underreferral rate (false negatives/false negatives and true negatives) was 19%. Thus, 13% of those who failed the DABERON and were consequently referred for further evaluation were not diagnosed as handicapped, while 19% of the subjects who passed the DABERON screening were later diagnosed as having an educationally handicapping condition.

Question 2: Does a statistically significant relationship exist between the DABERON total scores and the Metropolitan Readiness Test (MRT) composite scores?

Question 3: Does a statistically significant relationship exist between the DABERON total scores and the Kindergarten Progress Report (KPR) total scores?

These two research questions were addressed by correlating the DABERON total scores with the MRT and KPR total scores, using Pearson product moment correlations. Descriptive data regarding these scores are presented in Table 2, and the results of the statistical analysis can be viewed in Table 3. A highly significant positive correlation of .68 (p<.0001) was obtained between the DABERON and MRT total scores, indicating that there was a moderately strong predictive relationship between subjects' performance on the DABERON screening test and their subsequent MRT performance. A less strong but statistically significant positive correlation of .44 (p<.0001) was found between the
Table 2

Ranges, Means, Standard Deviations, and Medians for the DABERON, MRT, and KPR Total Scores

<table>
<thead>
<tr>
<th></th>
<th>DABERON N=165</th>
<th>MRT N=160</th>
<th>KPR N=140</th>
</tr>
</thead>
<tbody>
<tr>
<td>(possible range)</td>
<td>(0-122)</td>
<td>(0-97)</td>
<td>(0-122)</td>
</tr>
<tr>
<td>obtained range</td>
<td>47-121</td>
<td>15-72</td>
<td>41-122</td>
</tr>
<tr>
<td>mean</td>
<td>108.6</td>
<td>51.1</td>
<td>102.6</td>
</tr>
<tr>
<td>standard deviation</td>
<td>12.2</td>
<td>14.7</td>
<td>17.9</td>
</tr>
<tr>
<td>median</td>
<td>112.0</td>
<td>54.0</td>
<td>114.0</td>
</tr>
</tbody>
</table>

Table 3

Correlation Coefficients of DABERON Total Scores and Subsection Scores with MRT and KPR Total Scores

<table>
<thead>
<tr>
<th>Total Scores</th>
<th>MRT N=160</th>
<th>KPR N=140</th>
</tr>
</thead>
<tbody>
<tr>
<td>DABERON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Score</td>
<td>.68***</td>
<td>.44***</td>
</tr>
<tr>
<td>(Subsections)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body Parts</td>
<td>.47***</td>
<td>.33***</td>
</tr>
<tr>
<td>Colors</td>
<td>.46***</td>
<td>.40***</td>
</tr>
<tr>
<td>Number Concepts</td>
<td>.65***</td>
<td>.49***</td>
</tr>
<tr>
<td>Prepositions</td>
<td>.36***</td>
<td>.17</td>
</tr>
<tr>
<td>Follow Directions</td>
<td>.44***</td>
<td>.19</td>
</tr>
<tr>
<td>Plurals</td>
<td>.32***</td>
<td>.27**</td>
</tr>
<tr>
<td>General Knowledge</td>
<td>.60***</td>
<td>.37***</td>
</tr>
<tr>
<td>Visual Motor</td>
<td>.57***</td>
<td>.38***</td>
</tr>
<tr>
<td>Gross Motor</td>
<td>.09</td>
<td>.12</td>
</tr>
<tr>
<td>Categorization</td>
<td>.39***</td>
<td>.02</td>
</tr>
</tbody>
</table>

Two-tailed significance
***p<.0001
**p<.001
*p<.01
DABERON and KPR total scores, suggesting that subjects who scored higher on the DABERON were more likely to subsequently receive higher teacher ratings than lower scoring subjects.

The correlations between DABERON total scores and KPR total scores by individual teachers are displayed in Table 4. Statistically significant, moderate correlations were found for all teachers with the exception of one, for whom insufficient data was available. These results suggest that the pooled KPR/DABERON total score correlations reported in Table 3 were not significantly skewed by any individual differences in teachers' rating styles.

Table 4

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Number of Students (N=140)</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher 1</td>
<td>51</td>
<td>.45**</td>
</tr>
<tr>
<td>Teacher 2</td>
<td>37</td>
<td>.57***</td>
</tr>
<tr>
<td>Teacher 3</td>
<td>27</td>
<td>.61**</td>
</tr>
<tr>
<td>Teacher 4</td>
<td>8</td>
<td>.43</td>
</tr>
<tr>
<td>Teacher 5</td>
<td>21</td>
<td>.54*</td>
</tr>
</tbody>
</table>

Two-tailed significance:

***p<.0001
**p<.001
*p<.01
Question 4: Is there a statistically significant relationship between any of the DABERON subsection scores and the MRT composite scores and/or KPR total scores?

This question was addressed by correlating each of the DABERON subsection scores individually with both the MRT Composite scores and the KPR total scores, using Pearson product moment correlations. The reader is again referred to Table 3 for the results of this analysis. The strongest correlation for both of the criterion measures was with the DABERON Number Concepts subsection (r=.65 for the MRT and .49 for the KPR, both of which were significant at the .0001 level). These results indicate that, at least for the subjects in this study, performance on the Number Concepts subsection of the DABERON had about the same predictive value with regard to their performance on the two criterion measures as did subjects' total DABERON test scores.

Additionally, several other moderately strong correlations were found between individual DABERON subsections and the MRT Composite scores, most notably the General Knowledge and Visual Motor subsections (r=.60 and .57 respectively, p<.0001). Other subsections showing moderate correlations with the MRT included Body Parts (.47, p<.0001), Colors (.46, p<.0001), and Following Directions (.44, p<.0001). The Prepositions, Plurals, and Categorization subsections yielded significant but low correlations (.36, .32, and .39 respectively, p<.0001) with the MRT. Statistically significant but weak correlations were also obtained between five of the DABERON subsection scores and the KPR total scores: Body Parts (.33), Colors (.40), Plurals (.27), General Knowledge (.37), and Visual Motor (.38).
Question 5: What differences are there in DABERON performance related to subject gender, and how significant are these differences?

In order to find out whether girls tended to score significantly higher on the DABERON than boys, or vice versa, a t-test was utilized to determine if there was a difference between the mean scores obtained by each group (Table 5). In addition, crosstabulations, chi-square, and the phi statistic were performed to see if one group was more likely than the other to pass or fail the DABERON screening (Table 6). The results of these analyses indicate that there were no significant differences related to gender in either total scores obtained on the DABERON or the probability of passing or failing the DABERON screening.

Table 5

Results of a T-Test for Significant Differences Between Boys' and Girls' Mean Total Scores on the DABERON

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Cases (N=165)</th>
<th>Mean</th>
<th>SD</th>
<th>T value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DABERON Total Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1 (males)</td>
<td>78</td>
<td>108.9</td>
<td>12.4</td>
<td>.39</td>
<td>.697</td>
</tr>
<tr>
<td>Group 2 (females)</td>
<td>87</td>
<td>108.2</td>
<td>12.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6
Relationship Between Gender and Pass/Fail Performance on the DABERON

<table>
<thead>
<tr>
<th></th>
<th>Pass n=150</th>
<th>Fail n=15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1 (males)</td>
<td>n=72</td>
<td>n=6</td>
</tr>
<tr>
<td></td>
<td>92%</td>
<td>8%</td>
</tr>
<tr>
<td>Group 2 (females)</td>
<td>n=78</td>
<td>n=9</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Chi-square = 0.10, N=165 (1), p=.749

Question 6: What differences are there in DABERON performances related to subject age at the time of testing, and how significant are these differences?

Since age was recorded as a continuous rather than a categorical variable, the answer to this question was sought through statistical means other than those used for the other variables related to individual characteristics of the subjects (i.e., gender, cultural background, SES, and prior educational experience). The results of a t-test performed to determine if there was any significant age difference between those subjects who passed the DABERON screening and those who failed are presented in Table 7. It shows that there was no significant difference in pass/fail performance on the DABERON related to subject age. In addition, a Pearson product moment correlation was performed between the total DABERON scores and the subjects' age at the
Table 7

Results of a T-Test for Significant Differences Between the Mean Ages of Subjects Who Passed the DABERON vs. Those Who Failed

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of cases (N=165)</th>
<th>Mean</th>
<th>SD</th>
<th>T value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in Months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1 (Pass)</td>
<td>150</td>
<td>63.6</td>
<td>3.8</td>
<td>1.74</td>
<td>.098</td>
</tr>
<tr>
<td>Group 2 (Fail)</td>
<td>15</td>
<td>62.0</td>
<td>3.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

time of testing. This analysis yielded a negligible positive correlation (r=.19), which was significant at the .05 level.

Question 7: What differences are there in DABERON performances related to subject SES (low income vs. not low income), and how significant are these differences?

To address this question, a t-test was performed to assess whether there were significant differences in DABERON total scores related to SES, and crosstabulation, chi-square, and the phi statistic were used to determine if the likelihood of a subject passing or failing the DABERON screening was significantly related to their family income status. The results are presented in Tables 8 and 9.

The results of both statistical analyses indicate that there was a significant relationship between subjects' family income status and their performance on the DABERON screening test. The results of the t-test (Table 8) inform us that children from low income families tended to score significantly lower overall on the DABERON than
Table 8
Results of a T-Test for Significant Differences Between Mean DABERON Total Scores of Low Income vs. Not Low Income Subjects

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Cases (N=165)</th>
<th>Mean</th>
<th>SD</th>
<th>T  Value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DABERON Total Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1 (not low income)</td>
<td>102</td>
<td>112.6</td>
<td>8.8</td>
<td>6.02</td>
<td>.0001</td>
</tr>
<tr>
<td>Group 2 (low income)</td>
<td>63</td>
<td>102.0</td>
<td>14.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9
Relationship Between Pass/Fail Performance on the DABERON and Family Income Status of the Subjects

<table>
<thead>
<tr>
<th>DABERON (N=165)</th>
<th>Pass n=150</th>
<th>Fail n=15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Low Income</td>
<td>n=99</td>
<td>n=3</td>
</tr>
<tr>
<td>(n=102)</td>
<td>97%</td>
<td>3%</td>
</tr>
<tr>
<td>Low Income</td>
<td>n=51</td>
<td>n=12</td>
</tr>
<tr>
<td>(n=63)</td>
<td>81%</td>
<td>19%</td>
</tr>
</tbody>
</table>

Chi-square = 10.35, N=165 (1), p=.0013  
Phi = .25
children from moderate or higher income homes. There was also a statistically significant relationship between subjects' family income status and whether they passed or failed the DABERON screening (Table 9), with lower income children being somewhat more likely to fail. However the strength of the relationship between family income and DABERON pass/fail (indicated by the Phi statistic) was not compelling, as children from both groups were far more likely to pass than fail the DABERON. Thus, although low income subjects' scores may have been below the average of the other subjects overall, in most cases enough points were achieved to pass the screening.

Question 8: What differences are there in DABERON performances related to subject cultural background (Hispanic vs. non-Hispanic), and how significant are these differences?

A t-test was employed to determine if there were significant differences in DABERON total scores between Hispanic and non-Hispanic students. Crosstabulation, chi-square, and phi were used to assess the relationship between cultural background and pass/fail on the DABERON. The results of these statistical analyses are contained in Tables 10 and 11.

As these tables indicate, Hispanic children tended on the average to score significantly lower than non-Hispanic children on the DABERON (Table 10). In addition, they were also more likely to fail the DABERON screening (Table 11), despite the fact that Hispanic children who failed the initial screening in English were retested in Spanish, and the higher score was entered as data for this study. Although the strength of the relationship between these two variables was again not
Table 10
Results of a T-Test for Significant Differences Between Mean DABERON Total Scores of Hispanic vs. Non-Hispanic Subjects

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of cases (N=165)</th>
<th>Mean</th>
<th>SD</th>
<th>T value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DABERON Total Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1 (non-Hispanic)</td>
<td>139</td>
<td>111.0</td>
<td>9.6</td>
<td>4.83</td>
<td>.0001</td>
</tr>
<tr>
<td>Group 2 (Hispanic)</td>
<td>26</td>
<td>95.4</td>
<td>16.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11
Relationship Between Pass/Fail Performance on the DABERON and Subject Cultural Background (Hispanic vs. Non-Hispanic)

<table>
<thead>
<tr>
<th>DABERON</th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N=165)</td>
<td>n=150</td>
<td>n=15</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>n=132</td>
<td>n=7</td>
</tr>
<tr>
<td>95%</td>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>n=18</td>
<td>n=8</td>
</tr>
<tr>
<td>69%</td>
<td></td>
<td>31%</td>
</tr>
</tbody>
</table>

Chi-square = 14.57, N=165 (1), p=.0001
Phi = .30
overwhelming (Phi=.30), the results do suggest that Hispanic children did less well on the DABERON than non-Hispanic children, perhaps for reasons other than language dominance. It should also be pointed out that these results are probably not independent of the results obtained for family income status (and vice versa), since 85% of the Hispanic children in this study were classified as low income, as opposed to 29% of the non-Hispanic children.

Question 9: What differences are there in DABERON performances related to subject prior educational experience (preschool or organized group daycare), and how significant are these differences?

This question was addressed by performing a t-test to determine if there was any difference in DABERON total scores between children who had prior educational experience and those who did not. Crosstabulation and chi-square were utilized to assess the relationship between prior educational experience and DABERON pass/fail status. Refer to Tables 12 and 13 for the results of these analyses.

The results in Table 12 indicate that there is a statistically significant difference (at the .05 level) between the mean total DABERON scores of children who had prior educational experience and those who did not. On average, children with prior educational experience tended to score slightly higher overall on the DABERON. However the results presented in Table 13 inform us that there was no significant relationship between prior educational experience and whether a child passed or failed the DABERON screening, since children in both groups were likely to earn enough points for a passing score.
Table 12

Results of a T-Test for Significant Differences Between Mean DABERON Total Scores of Subjects with Prior Educational Experience vs. Those With No Prior Educational Experience

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of cases (N=114)</th>
<th>Mean</th>
<th>SD</th>
<th>T value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DABERON Total Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1 (prior education)</td>
<td>88</td>
<td>110.3</td>
<td>10.6</td>
<td>2.17</td>
<td>.038</td>
</tr>
<tr>
<td>Group 2 (no prior education)</td>
<td>26</td>
<td>103.2</td>
<td>15.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13

Relationship Between Prior Educational Experience and Pass/Fail Performance on the DABERON

<table>
<thead>
<tr>
<th>DABERON</th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N=114)</td>
<td>n=104</td>
<td>n=10</td>
</tr>
<tr>
<td>Prior Education</td>
<td>n=82</td>
<td>n=6</td>
</tr>
<tr>
<td></td>
<td>93%</td>
<td>7%</td>
</tr>
<tr>
<td>No Prior Education</td>
<td>n=22</td>
<td>n=4</td>
</tr>
<tr>
<td></td>
<td>85%</td>
<td>15%</td>
</tr>
</tbody>
</table>

Chi-square = .926, N=114 (1), p=.336

These research results and their implications will be discussed further in Chapter V.
CHAPTER V
DISCUSSION

The general purpose of this study was to determine whether the DABERON is an adequate screening measure to use in legally-mandated kindergarten screening programs. The purpose of such screenings is to identify children who are likely to have handicapping conditions warranting Special Education intervention, and refer them for further diagnostic evaluation. The overall results of this study have given strong support to the basic hypothesis that the DABERON has a useful degree of validity for identifying kindergarten children who are subsequently diagnosed with some type of educational handicap warranting the provision of Special Education services. These results and their implications will be summarized and discussed in the next section.

Summary of Results

Research Question 1: Does performance on the DABERON (pass/fail) accurately predict the presence or absence of educationally handicapping conditions requiring Special Education services?

Contingency table analyses of the data indicated that pass/fail performance on the DABERON was significantly related to whether a child was subsequently diagnosed as educationally handicapped (p<.0001). The strength of this relationship was modest (phi=.42), reflecting the undeniable fact that there are many other variables (besides the presence of a handicap) which can impact on DABERON performance. The specificity of the DABERON for excluding non-handicapped children from
further evaluation was excellent (.98), although the DABERON's sensitivity for identifying children with handicapping conditions appeared to be rather low (.31). However the overall percentage of correct predictions yielded by the DABERON (hit rate) was 81%. The error rate of the DABERON in this study was thus only 19%, although error rates of 50% and higher are commonly reported in the literature (Shepard & Smith, 1986; Warren, 1988).

With regard to screening errors, the DABERON tended to underrefer (i.e., yielded false negatives) to a greater extent than it overreferred (yielded false positives), which is acceptable and appropriate for this type of instrument and the purposes for which it is utilized. Screening tests are only designed to pick out children who differ significantly from other children, so they would tend to catch only the more severe handicaps within the general population. It has been shown that when tests are expected to predict mild risks, decision errors (i.e., misclassification of children) increase greatly (Fletcher & Satz, 1982).

Part of the reason for the higher rate of false negatives and somewhat low sensitivity of the DABERON found in this study is therefore likely to be because subjects were included who had been referred for evaluation during first or second grade (i.e., not as a result of screening failure in kindergarten). Most of these who eventually qualified for special education services probably had milder conditions than a screening test could be expected to pick up. This is corroborated by the observation that none of these later-referred subjects were given a 502.3 or 502.4 IEP prototype (which denote more intensive
Special Education services). Additionally, certain handicapping conditions (e.g., Developmental Reading Disorder) can only be diagnosed in older children, and therefore would not be likely to be identified by an early screening measure. Other conditions warranting the implementation of an IEP may be related to social, emotional, or behavioral factors, which can arise at any time and are often of a temporary nature.

A benefit of having a higher rate of false negatives is that it helps to avoid the pitfall of overidentifying children as handicapped. This is considered important because of the possible detrimental social and emotional effects for the child of being labeled as suspected of having a problem (Cadman et al., 1984; Heller, Holtzman & Messick, 1982), although some research reports minimal, if any, labeling effect resulting from early intervention (Reinherz et al., 1983). Mistakenly overreferring non-handicapped children for further evaluation is also very costly in terms of time and personnel involved.

The usefulness of the DABERON as a screening measure may perhaps best be summed up by determining its positive and negative predictive values. The idea behind predictive utility, as it is often explained in the literature, is to evaluate the actual accuracy rates (i.e., true positives and true negatives) of a measure in relation to the actual or estimated prevalence of the disorder or disability which it is attempting to predict. When the prevalence is known or can be accurately estimated, a measure's predictive utility can be computed directly by using the principle of inverse probability, or Bayes Theorem (Barnes, 1982; Cadman et al., 1984; Satz, Fennell & Reilly,
Whether the positive and negative predictive values obtained for any given instrument are considered adequate or not remains a subjective judgment, since it depends upon the purposes and goals of the specific screening program.

For this study, the determinants of the predictive value or usefulness of the DABERON are its sensitivity (.31) and specificity (.98) in relationship to the actual prevalence of educational handicaps in the population. This was estimated from the prevalence rate of subjects having IEPs in the research sample, which was 25%. The calculations yield a positive predictive value of .84 for the DABERON, and a negative predictive value of .81 (see Appendix). This tells us that, for the population studied (i.e., elementary schoolchildren in Northampton, MA), the probability that those who fail the DABERON will actually be found to have a handicapping condition is 84%, while the probability that those who pass the DABERON will actually be free of a handicapping condition is 81%. This would seem to be adequate predictive utility for a screening instrument, especially in light of the fact that most of those misclassified by the DABERON appear to be children with milder, later-developing, or more temporary forms of educationally handicapping conditions.

Meisels et al. (1984), in their predictive validity study of the Early Screening Instrument (ESI), found the sensitivity and specificity rates of the ESI to be .88 and .82 respectively for children in kindergarten and first grade, declining gradually thereafter. These rates were considered to be robust indicators of the ESI's validity. The positive and negative predictive values of the ESI were not given. On
the other hand, Cadman et al. (1984) reported a specificity rate of .99 but a sensitivity rate of only .10 in their large scale study of the Denver Developmental Screening Test. The negative predictive value of the DDST ranged from 79% to 93%, but the predictive value of a positive test (i.e., screening failure) was only 30% to 62%. They therefore concluded that the DDST was not sufficiently sensitive to use as a screening tool for the general population which was studied.

The results obtained for the DABERON in this study fall somewhere between those described in the preceding paragraph. All three tests had high specificity, but varied widely in their sensitivity to handicapping conditions. The DABERON’s sensitivity, although higher than that obtained for the DDST, did not come close to the sensitivity of the ESI, as reported by Meisels and his associates. However even with this relatively low level of sensitivity, given the estimated prevalence of educationally handicapping conditions diagnosed in elementary school children in Northampton, the DABERON, unlike the DDST, was found to have high positive and negative predictive values. It can therefore be considered a predictively useful instrument for this population, although its utility cannot be assumed for other populations of school children.

Research Question 2: Does a statistically significant relationship exist between the DABERON total scores and the Metropolitan Readiness Test (MRT) Composite scores?

The MRT was chosen as a criterion measure because it has been shown to be a reliable and valid predictor of success in the early primary grades, and because its use as a standard against which other
Instruments are evaluated has been well documented in the literature. In this study, subjects' total score on the DABERON correlated moderately well with their subsequent MRT scores (.68, p<.0001). Since the MRT is predictive of success in first grade (and thus higher scores are presumably contraindicative of an educational handicap), these results imply that higher scores on the DABERON may be considered contraindicative of handicapping conditions as well.

Other studies which have used the MRT as a standard for assessing the predictive usefulness of other early childhood screening tests have found significant correlations of .78 between the DIAL and the MRT (Obrzut et al., 1981); .53 between the Vane Kindergarten Test and the MRT (Powers, 1974); and .76 between a screening checklist and the MRT. Screening tests have also been measured against other achievement tests, such as the Stanford Achievement Test and the California Achievement Test. In general, correlations exceeding .50 seem to be accepted as evidence for the predictive validity of the screening instrument being investigated.

Research Question 3: Does a statistically significant relationship exist between the DABERON total scores and the Kindergarten Progress Report (KPR) total scores?

The authors of the DABERON reported it to have high predictive validity, based on a study they performed which correlated subjects' DABERON scores with teacher ratings of subjects (on a behavioral checklist) one year later (Danzer et al., 1982). The content of the checklist was not specified, beyond saying that it rated "school readiness skills." The correlation between kindergarten DABERON test
scores and the followup teacher ratings was reported to be .84 (p<.001). In this study, the correlation between subjects' DABERON scores and the KPR, which is a teacher rating checklist of behaviors considered important for success in first grade, was much weaker (.44, p<.0001), and does not on its own provide much support for the predictive validity of the DABERON. This apparently poor replication of results might be explained by the fact that the content of the teacher rating checklists used in the Danzer study was not specified, and therefore the two rating checklists may not be truly comparable. Another reason for the weak correlation may relate to the constraints of a ceiling effect—a substantial number of children (25%) scored at or within 5 points of the ceiling level on the KPR. This decreased range would tend to attenuate the potential correlations.

Research Question 4: Is there a statistically significant relationship between any of the DABERON subsections scores and the MRT Composite scores and KPR total scores?

Correlations between each of the 10 DABERON subsections and the MRT and KPR total scores indicated that performance on the Number Concepts subsection of the DABERON was the best single predictor of performance on the criterion measure (r=.65 for the MRT and .49 for the KPR, p<.0001). In fact, subjects' performance on this subsection alone had about the same predictive value with regard to the two criterion measures as their total DABERON scores. This finding seems to indicate that a major portion of the DABERON's predictive validity derives more from its academic rather than its developmental content. High scores on the Number Concepts subsection seem to rest primarily
on acquired knowledge, although the developmental underpinnings needed for the acquisition of such concepts must already be in place.

Interestingly, Telegdy (1984) reported on two studies (Silberberg, Iversen & Silberberg, 1968; Silberberg, Silberberg & Iversen, 1972), which may have some bearing on this finding. These two studies reported that the ability to recognize letters and numbers in kindergarten was highly predictive of end of first grade achievement. Using the Gates Reading Readiness Test scores from kindergarten to predict end of first grade reading level, they found that the Letters and Numbers subtests were as valid predictors as the complete test. However they also subsequently found that providing children with specific training in letter and number recognition only increased those skills, but had no beneficial effect on reading levels per se. These results seem to imply that "readiness skills" are not independent of developmental processes, but rather the presence of developmental processes might be assumed from the child's ability to acquire certain readiness skills. The Number Concepts subsection of the DABERON, while it may appear to relate directly to acquired knowledge such as familiarity with numbers, also requires the developmental processes of visual perception, auditory discrimination, abstract reasoning, long term memory, and verbal conceptual ability. These abilities are typically associated with good performance on most standardized achievement tests (Telegdy, 1977), which probably accounts for the strong showing of this subsection in relation to the criterion measures. The common dictum that developmental screening tests must have strictly developmental
content may therefore be less necessary or valid than is generally thought.

Research Question 5: What differences are there in DABERON performance related to subject gender, and how significant are these differences?

Subjects' gender was not found to be significantly related to either their total DABERON scores or their pass/fail status. Since the DABERON is a heavily language-based test, it was hypothesized that girls, who tend to mature faster and be more verbal on the average (Ilg & Ames, 1965), might perform better on the whole than boys. Had this been the case, it may have substantially increased the overreferral rate for boys, thus decreasing the predictive usefulness of the DABERON.

Research Question 6: What differences are there in DABERON performances related to subject age at the time of testing, and how significant are these differences?

The data indicate that subjects' performance on the DABERON (total score and pass/fail status) was not significantly influenced by their age at the time of testing. The fact that age did not impact on pass/fail status is not too surprising, since subjects' age is taken into account in the determination of cutoff scores for passing and failing. However it was feared that since several subsections of the DABERON tap into acquired knowledge, the relatively older kindergarten child could be erroneously identified as at risk, when in fact he or she had simply not been exposed to the relevant information. This fear proved to be unfounded for the subjects in this study.
What may seem more surprising is that there was not a stronger relationship found between age and total score on the DABERON \((r=0.19, \ p<.05)\), since it is expected that children will be able to respond correctly to more DABERON items as they get older. It is likely that the explanation for this relates to the statistical reality that the value of a correlation coefficient is affected by the degree of variation characterizing the two variables. The value of \(r\) is smaller in those situations in which the range of either variable is restricted (Mínium, 1978). While the DABERON standardization sample was composed of children ranging in age from 3 to 7, the subjects in this research sample were from a much more restricted age range (the majority ranged in age from 4-11 to 5-7, a span of only 8 months), which may explain the very weak correlation obtained. It is also possible that, for this population of students, factors other than age were more influential with respect to their performance on the DABERON.

Research Question 7: What differences are there in DABERON performance related to subject SES (low income vs. not low income), and how significant are these differences?

Children from low income families tended to score significantly lower overall on the DABERON than children from moderate or higher income homes. Lower income children were also somewhat more likely to fail the DABERON screening. These results are not surprising, since they correspond to what is known about I.Q. and low SES, i.e., that a disproportionate number of low I.Q. children are from low SES backgrounds (Shepard & Smith, 1986). This of course is entwined with other variables often associated with low SES, such as a lack of environmental
stimulation, inadequate pre- and postnatal care, poor nutrition, and other risk factors associated with the development of learning problems. It has also been pointed out that many tests, of which the DABERON may be one, are biased against low income and minority students (Foglia, 1988).

Research Question 8: What differences are there in DABERON performances related to subject cultural background (Hispanic vs. non-Hispanic), and how significant are these differences?

Hispanic children tended to score significantly lower than non-Hispanic children on the DABERON, and they were also more likely to fail the DABERON screening. It would seem that these results cannot be attributed solely to language dominance, since a Spanish version of the DABERON was administered to any Hispanic child who failed the screening in English. However it might be argued that language development is likely to be slower in bilingual children, and they may be less proficient in both languages during their early years. Since the DABERON is so heavily language-based, this could account for the relatively poorer performance of Hispanic children. Cultural differences may also play a role, although the DABERON standardization sample did include a representative number of Hispanic and other minority children.

There also appears to be some interaction between cultural background and low SES. The rates of identified handicapping conditions were similar for the Hispanic and non-Hispanic populations (26% of the non-Hispanic subjects had an IEP, and 23% of the Hispanic subjects). Eighty-five percent of all Hispanic children in this study were classified as low income, as compared to 29% of non-Hispanic children.
However 100% of the Hispanic children with IEPs were from the low income group, while 53% of the non-Hispanic special needs students were classified as low SES. Thus, for both groups, low income youngsters were far more likely to be identified as educationally handicapped. But when the data is broken down further it is discovered that although nearly half (46%) of all low SES, non-Hispanics have IEPs, only about one-quarter (27%) of low SES Hispanic youngsters are similarly identified. Clearly the variable of low family income accounts for more of the variability in DABERON performance with non-Hispanic youngsters than it does with Hispanic youngsters. The specific reasons why Hispanic youngsters seem to have more difficulty on the DABERON remain unclear, although language may be the primary factor.

Research Question 9: What differences are there in DABERON performances related to subjects' prior educational experience (preschool or organized group daycare), and how significant are these differences?

Children who had prior educational experience tended to achieve significantly higher DABERON scores, on the average, than those with no prior educational experience. However prior educational experience made no difference in whether a child passed or failed the DABERON screening, since children from both groups generally were able to score high enough to pass. This suggests that the DABERON has sufficient developmental content (i.e., material which assesses cognitive processes, rather than acquired knowledge which is specifically taught) to make it adequate for screening purposes. On the other hand, it may in fact demonstrate that developmental screening instruments need not
(and perhaps should not) be restricted solely to developmental content.

**Suggestions for Further Research**

Several other research ideas can be developed from the groundwork laid by the present study. Replication studies should be done to confirm the present findings with regard to the predictive validity of the DABERON. Ideally these studies should be on a larger scale, so that the number of screening failures will be large enough to lend more power to the statistical analyses.

Another research possibility would be to perform similar studies on other popular screening tests, in order to contribute to the body of knowledge pertaining to the predictive validity of such instruments. The demand for tests which have been shown to have predictive validity as screening measures is not likely to subside any time in the near future.

The relationship between SES and minority status with regard to DABERON (or other screening test) performance is another area needing further study. The present investigation did not have a large enough sample of minority subjects who were not Hispanic to assess the impact minority status might have on DABERON performance independent of a specifically Hispanic cultural background. This study was also not able to effectively differentiate the effects of cultural background on DABERON performance independent of low income status. Studies which are designed to specifically address these issues would be most desirable.
Other interesting areas of future research might include attempts to develop profiles of children who are most likely to be misclassified by a particular screening battery or test, which might lead to ways of improving the sensitivity and specificity of screening measures. Another angle might be to try to assess and delineate variables which contribute to the persistence or remission/remediation of learning problems.

Conclusions

The DABERON performed well in every theoretical direction, and does seem to identify children with handicaps as opposed to (or as well as) school readiness. It had a very low error (misclassification) rate (19%) for the subjects in this study, and yielded excellent positive and negative predictive values. The DABERON's correlations with the MRT suggest that it is a moderately valid predictor of academic success in the early primary grades. On the whole, the data from this study suggest that the DABERON can be a useful and valid screening measure for identifying children at risk for educational handicaps that are likely to impede their learning.
Bayes' Theorem of inverse probability was used to calculate the positive and negative predictive values of the DABERON as follows:

Positive Predictive Value =

\[
\frac{(\text{Sensitivity}) \times (\text{Prevalence})}{(\text{Sensitivity})(\text{Prevalence}) + (1-\text{Specificity})(1-\text{Prevalence})}
\]

\[
= \frac{(0.31)(0.25)}{(0.31)(0.25) + (1-0.98)(1-0.25)} = 0.84
\]

Negative Predictive Value =

\[
\frac{(\text{Specificity}) \times (1-\text{Prevalence})}{(1-\text{Sensitivity})(\text{Prevalence}) + \text{Specificity} \times (1-\text{Prevalence})}
\]

\[
= \frac{(0.98)(0.75)}{(0.69)(0.25) + (0.98)(0.75)} = 0.81
\]
REFERENCES


