AN EVALUATION OF A MULTIPLE GATING PROGRAM:
SCREENING FOR DEVELOPMENTAL PROBLEMS
IN A PRESCHOOL POPULATION

DISSertation

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By
Pamela B. Gutter, MA

* * * * *

The Ohio State University
2003

Dissertation Committee:
Professor Antoinette Miranda, Adviser
Professor Wendy Naumann
Professor James L. Moore III

Approved by

Adviser
College of Education
ABSTRACT

This study is a descriptive analysis of archival data collected for the purpose of ascertaining the effectiveness of a multiple gating program implemented by the Child Development Council (CDC) Head Start of Franklin County. Multiple gating incorporates a stepwise assessment procedure which begins with a large-scale screening for purposes of identifying children who have a low base-rate problem, such as a developmental delay in the preschool-age population. Further, screenings are completed, which are more intensive in nature and are conducted on a smaller group of children who had been identified by the preceding large-scale screening. Multiple gating procedures have rarely been used in assessing for developmental delays and the research literature on this technique is severely lacking in the field of education.

This analysis is a replication of the methodology used by Loeber, Dishion, and Patterson (1984) in their study on identifying youth who were at-risk for delinquency. In this current study, Gate I consists of screening the majority of the CDC Head Start children with the Denver II developmental screening instrument. Children enrolled in the CDC Head Start were screened using the Denver II within the first three months of their enrollment. Gate II involves screening those children identified on the Denver II as suspect with the Early Screening Inventory – Revised (ESI-R). This screening instrument identifies a child in one of three categories. Either ok, rescreen (meaning that the child
must be rescreened in an eight-week time period), or refer. Gate III involves an additional evaluation completed by Columbus Public Schools (CPS). This evaluation is completed on those children who performed within the refer range on the ESI-R and includes a parent interview, child observation, and screening using the Brigance Preschool Screen.

Although the CDC Head Start of Franklin County implemented a multiple gating program during the 2002-2003 school year, the results obtained indicate that the current multiple gating program may not be adequate in identifying those children who are at-risk for a developmental delay due to a lack of correlation between the criterion measure and the two developmental screening instruments.
Dedicated to my mother
ACKNOWLEDGMENTS

I wish to thank my advisor, Antoinette Miranda for guiding me through the school psychology program.

I thank Wendy Naumann for her amazing knowledge in statistics, her patience with my limited knowledge, and for her encouragement.

I thank James Moore for the long discussions pertaining to youth.

I thank Melissa Schmidgall for her ability to read all of my work while still managing to work on her own Ph.D. program.

I also wish to thank my partner, Tammy for her support while I worked on completing all of my requirements and for allowing me to have moments of complete insanity.
VITA

August 6, 1972 ……………………Born – Columbus, Ohio

2000 ……………………………….M.A. School Psychology, The Ohio State University.

2001-present ………………………School Psychologist,
            Columbus Public Schools
            Columbus, Ohio, USA

2001-2003 ………………………..Graduate Assistant
            The Ohio State University

PUBLICATIONS

Research Publication


FIELDS OF STUDY

Major Field: Education
            School Psychology
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>Chapter 1</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>5</td>
</tr>
<tr>
<td>Purpose of Study</td>
<td>6</td>
</tr>
<tr>
<td>Limitations of Study</td>
<td>7</td>
</tr>
<tr>
<td>Research Questions</td>
<td>8</td>
</tr>
<tr>
<td>List of Terms</td>
<td>9</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>11</td>
</tr>
<tr>
<td>Unique Issues in Assessing Preschool Children</td>
<td>11</td>
</tr>
<tr>
<td>Purpose of Conducting a Developmental Screening</td>
<td>12</td>
</tr>
<tr>
<td>Problems with Developmental Screeners</td>
<td>13</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>3</td>
<td>Introduction</td>
</tr>
<tr>
<td></td>
<td>Participants</td>
</tr>
<tr>
<td></td>
<td>Research Design</td>
</tr>
<tr>
<td></td>
<td>Instruments</td>
</tr>
<tr>
<td></td>
<td>Procedures</td>
</tr>
<tr>
<td></td>
<td>Data Analysis</td>
</tr>
<tr>
<td></td>
<td>Ethics</td>
</tr>
<tr>
<td>4</td>
<td>Research Questions</td>
</tr>
<tr>
<td></td>
<td>Results of Analysis</td>
</tr>
<tr>
<td>5</td>
<td>Summary</td>
</tr>
</tbody>
</table>

Assessing the Preschool Population .........................................................18

Alternative Methods of Assessing Preschool Children ..............................21

Dynamic Assessment .........................................................................................22

Ecobehavioral Model .........................................................................................23

Multiple Gating .................................................................................................28

Chapter Summary ...............................................................................................30

Chapter 3 ............................................................................................................32

Introduction .........................................................................................................32

Participants ...........................................................................................................33

Research Design ..................................................................................................34

Instruments ...........................................................................................................36

Gate I: Denver II ..................................................................................................36
Gate II: ESI-R .......................................................................................................39
Gate III: CPS ........................................................................................................44
Criterion-Measure: Galileo ...................................................................................50

Procedures ............................................................................................................53

Data Analysis .......................................................................................................53

Ethics ....................................................................................................................58

Chapter 4 ............................................................................................................59

Research Questions .............................................................................................59

Results of Analysis ..............................................................................................61

Chapter 5 ............................................................................................................66

Summary ..............................................................................................................66
Reasons for poor results on the multiple gating program .......................... 68
Galileo ........................................................................................................ 68
Denver II ..................................................................................................... 72
Examiner Bias and Instrumentation Threat .................................................. 73
Effectiveness of the CDC Multiple Gating Program .................................. 73
Program Evaluation and Accountability ...................................................... 76
Recommendations ......................................................................................... 77
Limitations ..................................................................................................... 79
Directions for Future Research ................................................................. 79
Conclusion ................................................................................................... 80
References ..................................................................................................... 82
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Computational Formulas for Measures of Predictive Efficiency ......54</td>
</tr>
<tr>
<td>3.2</td>
<td>Intercorrelates between Independent and Dependent Variables ..........55</td>
</tr>
<tr>
<td>3.3</td>
<td>Prediction table and Summary Statistics for Gate I ..................56</td>
</tr>
<tr>
<td>3.4</td>
<td>Prediction table and Summary Statistics for Gate II ..................57</td>
</tr>
<tr>
<td>3.5</td>
<td>Prediction table and Summary Statistics for Gate III ..................57</td>
</tr>
<tr>
<td>3.6</td>
<td>Predictive Outcome for Multiple-Gating Procedure .....................58</td>
</tr>
<tr>
<td>4.1</td>
<td>Intercorrelates between Independent and Dependent variables ..........62</td>
</tr>
<tr>
<td>4.2</td>
<td>Prediction Table and Summary for Gate I .............................63</td>
</tr>
<tr>
<td>4.3</td>
<td>Prediction Table and Summary for Gate II .............................64</td>
</tr>
<tr>
<td>4.4</td>
<td>Results of children referred to CPS .................................65</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A description of the multiple-gating procedure</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>Sample Criterion Language checklist used by CPS</td>
<td>46</td>
</tr>
</tbody>
</table>
The passage of the 1986 Education of the Handicapped Act, Public Law 99-457 (amended by Public Law 102-119) has had a profound effect on working with pre-school aged children (Preator & McAllister, 1995). Children who exhibited developmental delays in either receptive language, expressive language, self-help, socialization skills, fine motor, gross motor, or cognition were now eligible for services that fall under the Individual with Disabilities Act (IDEA) (Preator & McAllister). This public policy mandates that these children and their families be given early intervention services (Greenwood, Luze, & Carta, 2002). As a result, pre-school children are to be located, identified, and evaluated for purposes of creating an intervention program (Meisels, 1991; Preator & McAllister). According to the U.S. Department of Education, approximately 4.8% of 3 – to 5 year-olds are currently receiving special education services (as cited in Lerner, Lowenthal, & Egan, 2003). Although this number may vary across states, from 1.3 to 9.4 percent, during the 2000-2001 school-year, 13% of all children enrolled in Head Start had a disability (i.e., emotional disturbance, language and speech impairments, learning disabilities, mental retardation, visual handicaps, health impairments, hearing impairments, and orthopedic handicaps) (U.S. Department of
Health and Human Services, 2002). Currently, there has been some controversy over how best to evaluate and identify pre-school children, specifically in terms of what type of assessments are the most appropriate.

Bagnato and Neisworth (1992) argue that preschool intelligence testing does not follow the criteria set forth in the federal mandates of IDEA. The author’s stated that early intelligence tests often do not lead to intervention, they lack predictive validity, and the testing often happens in isolation (i.e., not in the natural environment). According to Lerner et al. (2003) there are two main reasons for evaluating preschool children for possible delays. First, to gather information pertaining to the child which will help in the determination of eligibility and second, an evaluation should be conducted for purposes of planning interventions to assist in educating the child. P.L. 99-457 requires that early intervention services be provided to children with disabilities. Preschool children benefit greatly from early intervention and early detection of possible delays. The goal therefore, is to ascertain whether or not a problem exists, identify the child’s weaknesses and strengths, and finally address how to help with these concerns.

One quick way to assess a large group of children is by using a developmental screening instrument. The purpose of a developmental screening is to answer the question: “Is there a possibility that this child has a problem?” (Lerner et al., 2003, p. 77). The results of these screeners indicate whether or not the child may need a further, more in-depth evaluation. A developmental screener used to assess a large group of children should be inexpensive, brief, and the scoring method should be objective (Gredler, 1997).
Following IDEA guidelines, the Child Development Council (CDC) Head Start of Franklin County actively attempts to locate and evaluate preschool children for possible delays by screening all students and referring those students whose screening results indicate that further evaluation is necessary to the CDC mental health consultants who then conduct an additional evaluation. Mental health consultants are employees of the CDC and their primary responsibility is to screen children for possible developmental delays, conduct behavioral interventions, and to provide in-services to parents and staff members. The child is evaluated and based upon the results either referred to Columbus Public Schools for a further evaluation or he or she is identified as being developmentally appropriate. This three-tiered system utilized by the CDC Head Start is a multiple gating program (see Figure 1 for a pictorial representation), which is often seen in the behavioral and criminology research.

Preschool age children have a low rate of those identified as having a developmental delay. A low base rate problem such as this one tends to result in a high number of false positive predictive errors (children identified as at risk for developmental delays who are actually not at risk). Multiple gating is a technique which has shown to increase the number of valid positives (identifying those who are actually at risk) while lowering the number of false positives. This technique initially uses a screening instrument that is easy to give, inexpensive, and one which tends to have a high number of false positives. The purpose of this is to identify a large number of children. The second gate then uses a screening instrument that is more time consuming and one which does not have as high rate of false positives. This second screening procedure is one that
Figure 1.1: A Description of the Multiple-Gating Program utilized for this current study
is meant to lower the number of children identified as at-risk. Finally, the third gate is one that is more intrusive then the first two gates and one that is also more extensive. This gate may entail an observation of the child, parent interview, and further assessments (i.e., norm-referenced).

The purpose of the gates is to correctly identify those children who may have a developmental delay. Multiple gating is a technique which has historically been used within the behavioral literature and therefore limited research is available within the developmental literature. Yet, this technique is one that can be applied to help identify children who may be at-risk for a developmental delay or academic problems.

Statement of the Problem

The literature pertaining to the assessment of preschool-age children has emphasized the involvement of multiple methods, settings, and people while conducting an evaluation. The Law states that school districts are federally mandated to locate, evaluate, and serve preschool-age children (IDEA, 1997). The evaluation process is a key component of locating preschool-age children with developmental delays, yet there has been some disagreement on how best to conduct this. There has been a shift from conducting norm-referenced assessments (i.e., standardized tests) to incorporating more ecobehavioral aspects of the child (e.g., home environment, socioeconomic background, and maternal education). Multiple gating is one method that encompasses both norm-referenced and ecobehavioral assessments while simultaneously locating and evaluating children. This technique is one that filters children through a series of gates which allows for identifying as many children as possible in an efficient manner. The CDC Head Start
of Franklin County uses a multiple gating program when assessing for developmental delays for all children currently enrolled.

Purpose of Study

The purpose of this study is to evaluate the effectiveness of a multiple gating program currently implemented by the CDC Head Start of Franklin County in identifying preschool-age children who were at-risk for a developmental delay. The CDC Head Start used a multiple gating program during the 2002-2003 school-year. This program involved screening all children with the Denver II developmental screener, then screening those children identified as suspect on the Denver II with the Early Screening Inventory – revised (ESI-R). Excluded were those children who were referred based upon the Denver II for delays in expressive and/or receptive language. Although the CDC Head Start conducts a multiple gating screening of those students referred for expressive and/or receptive language difficulties it differentiates between these children from those referred based upon potential delays in the areas of socialization skills, self-help, gross motor, fine motor, and/or cognition. Those children at-risk for receptive and/or expressive delays are labeled as speech referrals and are evaluated with different instruments and follow a different gating technique then those children at-risk for delays in socialization skills, self-help, gross motor, fine motor, and/or cognition who are labeled as developmental referrals. The children referred for a speech language evaluation were evaluated by using a different multiple gating program and therefore were not included in this current analysis.
Finally an additional screening instrument, the Brigance Preschool Screen was completed by Columbus Public Schools (CPS) on those who were in the refer range on the ESI-R. In addition, CPS also completed a play-based observation, parent interview, and a language checklist. This multiple gating program was used to identify those children who were at-risk for a possible developmental delay, as defined by the CDC Head Start.

In addition, predictive and concurrent validity will also be assessed by analyzing the student’s results on the Galileo, a criterion-referenced assessment, which the CDC Head Start uses to assess the educational level of all students. The Galileo was given to all of the CDC Head Start students two times during the 2002-2003 school-year.

Limitations of Study

This study utilizes archival data collected during the 2002-2003 school year from the CDC Head Start of Franklin County and from Columbus Public Schools. The data were collected by a variety of staff members, including teachers, educational specialists, and school psychologists; and in a variety of settings. Although this may result in possible measurement error, this data collection method occurs in many agencies and also reflects the methods utilized by Loeber, Dishion, and Patterson (1984), which this current study is modeled after.

This study is an initial analysis assessing a method rarely used in evaluating developmental delays (multiple gating), consisting of preschool children attending a large urban preschool program, and addressing a low-base rate problem. As such the results of
this study can only be generalized to those organizations with similar characteristics to the CDC Head Start of Franklin County.

Research Questions

This study is a replication of the methodology used by Loeber et al. (1984) in their study on identifying youth who were at-risk for delinquency. This current study differs from the Loeber et al. study in several ways. First, this study assesses for the predictive validity of the multiple gating method by using a criterion measure, the Galileo. Second, this study focuses on developmental delays. Finally, this study assesses the effectiveness of a multiple gating model, when used with a preschool population. The goal for this present study is to determine whether or not preschool-age children can be economically and efficiently identified through using a multiple step screening process that successively narrows down a group of children who are high-risk for a developmental delay. This current study will attempt to answer the following questions:

1. Do the number of false positives decrease when the number of gates increase?
2. Does multiple gating improve the rate of valid positives?
3. How many of the valid positive children were initially identified by the first screener?
4. How many of the valid positive children were identified by the combination of the first two screeners?
5. What are the error rates of the false positives and false negatives?
**List of Terms**

1. **concurrent validity** – when criterion data and instrument data are collected near the same time and the results are then compared.

2. **criterion-referenced instrument** – instrument that assesses the child based upon curriculum data (Lerner et al., 2003).

3. **developmental screener** – tests used to assess whether or not a group of children may be at-risk for a developmental delay (Lerner et al., 2003).

4. **dynamic assessment** – an approach or model that involves an interaction between the examinee and the examiner, which focuses on the metacognitive processes of the learner and how the learner responds to intervention, and uses a pretest-intervention-posttest format (Lidz, 1997).

5. **ecobehavioral assessment** – evaluation of the child by considering their cultural and family background, life experiences, and development in cognitive, motor, adaptive, language, and social/emotional areas (Meisels, 1996).

6. **false negative** – a child initially identified as performing within the normal range when in fact they may have a possible developmental delay.

7. **false positive** – a child initially identified as possibly being at-risk for a developmental delay when in fact they are performing within the normal range.

8. **index of sensitivity** – number of children who perform poorly in school and who were identified as at risk on a screening instrument (Rafoth, 1997).

9. **index of specificity** – number of children performing adequately in school and who were identified as not at risk (Rafoth, 1997).
10. **multiple gating technique** – a technique used for screening a large number of children to identify those who may be at-risk for a possible delay. This technique involves multiple measures to ensure that only those who truly are at-risk for a possible developmental delay are identified.

11. **positive predictive validity** – computation derived from four possible outcomes of an administration of a screening instrument: the number of children who were identified as at risk and who later performed poorly, number of children identified as not being at risk who later actually performed poorly, number initially identified as being at risk who later performed adequately, and finally, the number identified as not being at risk who later performed adequately (Gredler, 1997)

12. **predictive validity** – the percentage of children who were identified as at risk and who later actually developed problems (Gredler, 1997)

13. **valid negative** - child is identified as not at risk for a possible delay whose later performance is adequate

14. **valid positive** - child is identified as at risk for a possible delay who later performs poorly
CHAPTER 2

REVIEW OF LITERATURE

Unique Issues in Assessing Preschool Children

Preschool children have historically been assessed by similar methods used for assessing older children yet preschool children are a unique group who often do not perform as well under the same standardized conditions as older children (McConnell, 2000). When assessing preschool children examiners must understand early childhood developmental milestones (Greenwood et al., 2002) and normal child development (Paget & Nagle, 1986). Children’s performance will vary over time and situations due largely to the influence of the school and home environments and due to their rapid cognitive development (Schakel, 1986).

Preschool children have difficulty staying on-task, do not respond as well to novel situations and examiners (Greenwood et al., 2002), and are not as vested in “doing their best” for the examiner. The behaviors of preschool children may have a major impact on their performance and thus their score. Also a child’s performance can vary from day to day.

Another issue in testing young children is that their behavioral patterns are not stable. For example, testing done too early in the year may not reflect problem behaviors that could manifest later in the school year. Children should be able to adjust to the
school environment and be given the opportunity to build a positive relationship with their teacher (Gredler, 1997). There are instances when a child is identified with a delay or deficit yet this delay or deficit may not be present later (Gredler). Child development occurs on a continuum and assessment of children should reflect this (Meisels & Atkins-Burnett, 2000).

**Purpose of Conducting a Developmental Screening**

Developmental assessment is designed for deepening our understanding of a child’s resources and competencies, and of his/her learning and care giving environments which will help in furthering the child’s full developmental potential (Greenspan & Meisels, 1996). Understanding a child’s care-giving environment within a cultural context is necessary for understanding the child’s personality and knowledge (Greenspan & Meisels). Children’s culture and environment must be understood in order to fully understand how education and child rearing is viewed. Each culture is different and people may have different expectations that they place on their child in terms of school readiness skills.

The purpose of a developmental screening is to be a brief measure of a child’s developmental abilities, which are related to their future school success (Meisels, 1987) while also being “culture-free”. Several researchers (Gredler, 1997; Meisels, 1991, Satz & Fletcher, 1988) state that an effective developmental screener has characteristics, which include ease of administration, an adequate standardization sample, low cost, adequate reliability and validity, culture-free, and appropriate content.

Screening in education is based upon the notion that behavioral and learning problems can be predicted accurately by conducting an early childhood assessment and
also that the potential problems of a child can be measured reliably (Gredler, 1997). The accuracy of these predictions may be somewhat problematic. Developmental screenings are often used to assess where a child’s developmental level is in terms of their potential for acquiring a skill, instead of the extent to which a skill has been acquired (Meisels, 1994). These screenings often measure memory of visual sequences, language comprehension, motor coordination, and social-emotional status (Meisels).

Problems with Developmental Screeners

The purpose of screening is for the identification of those who may be eligible for a more in-depth diagnostic evaluation (Carran & Scott, 1992). Screening procedures themselves are not in-depth evaluations and are not very extensive which often results in several errors (Paget & Nagle, 1986). Included in this is when a child is identified as being “at-risk” for a problem yet no significant problem actually exists (false positive) or the failure to identify a child when he/she may actually have a problem (false negative) (Gredler, 1997; Paget & Nagle). According to Gredler (1997) the false positive rate in a number of screening measures is quite high so the users of these measures must use caution when interpreting the results. A high number of false positive results can lead to wastes of diagnostic resources and unnecessary parental anxiety (Glascoe & Byrne, 1993). Yet, Gredler does recommend that when completing an initial screening of a large group of children the instrument should have a high rate of false positives in an attempt to rule out any possible delays in as many children possible.

Gredler (1997) stated that it is essential when assessing the utility of a screening measure to include an index of sensitivity (a/a + c) (a= valid positive, c= false negative) for determining the number of children who performed poorly on a screening
instrument and who had later difficulty (a retrospective view). Also recommended is to include the index of specificity \( \frac{d}{b + d} \) \((b= \text{false positive}, d= \text{valid negative})\), which compares the number of those who performed satisfactorily with those who were considered by the screening measure as not at risk (also considered a retrospective view). Glascoe and Byrne (1993) recommend an index of sensitivity of 80% and an index of specificity of 90%. The positive value \( \frac{a}{a + b} \) \((a= \text{valid positive}, b= \text{false positive})\) is an indication of the number of those children who were identified as being ‘at risk’ and who demonstrated to have actual learning problems (i.e., positive predictive value - percentage of children who were originally identified as being at risk who later did develop difficulties.). \( \frac{d}{c + d} \) \((d= \text{valid negative}, c=\text{false negative})\) is an indicator of those children who performed satisfactorily and were predicted not at risk.

These indices are all important in deciding if a screening instrument is effective yet the positive predictive value is the index considered the most important by school personnel (Gredler, 1997). Schools often act on those who are identified as “at risk”. When the positive predictive value for a screening instrument is depressed (i.e., .20 to .65) and children are placed in a transition room or special needs preschool based on the results, then some will have been placed incorrectly because they do not actually have a behavioral or learning deficit (Gredler).

Glascoe and Byrne (1993) conducted a study assessing how accurate three developmental measures were. The three measures chosen were the Academic Scale of the Developmental Profile – II (DP-II), the Battelle Developmental Inventory Screening Test (BDIST) and the Denver Developmental Screening Test II (Denver II). The author’s measured accuracy by comparing performance on the screening measures to standards for
screenings and then the measures were compared to a series of diagnostic tests. The standards of the measures included the sensitivity (80% is preferable), specificity (90% is preferable) and positive predictive value (70% is preferable) of the three measures. Eighty-nine children were involved in the study and all eighty-nine were given the three developmental screening tests. Of the 89 children involved in the study, 18 were identified as meeting diagnostic criteria for a developmental delay and 71 of the children did not meet the criteria. The findings were that the Academic Scale of the DP-II did not identify the majority of those who had difficulties. The Denver II had the opposite problem. This screening instrument over-identified children as having difficulties. A concern advocated by Glascoe and Byrne is that the Denver-II did not have a requisite validity study completed, which may have been helpful in the authors discarding inefficient items. Glascoe and Byrne did, however, find that the Denver-II detected more subtle problems with children. Finally, the authors reported that the BDIST was a better screening test than the other two yet it also tended to over-refer children. The authors also reported that the BDIST was difficult to administer and was lengthy.

According to Meisels (1987) the most common abuse of a developmental screening comes from using measures that lack established validity and reliability. Barnett, Macmann, and Carey (1992) advanced concerns regarding screening pre-school age children. First, the author’s reported a concern with developmental screenings lack of identifying specific target behaviors, which results in a limited treatment utility. Another concern regarding developmental screening is that there is a lot of variation between different screening instruments and a child may perform differently on each screening. A
study done by Hall and Barnett (1991) found that when looking at classification of risk, for every one case where there was agreement, there were four cases of disagreement.

Even those tests that are expected to measure the same skills of children can have different results. According to Bracken (1988) there are ten reasons why tests that are similar can produce dissimilar results. First, he states that tests need to be able to distinguish between those who are not delayed from those who are handicapped, delayed, or disabled. A test must have enough easy items to evaluate low-level success. If a test has a psychometrically weak floor, it may cause significant differences between two tests.

A second reason tests may not produce similar results pertains to the ceiling of the tests. A test must have an adequate number of items that are difficult so to differentiate between those children who are average or high-average. Another issue to be cognizant of is the item gradients. An item gradient describes how sharply test items align within a test, and it reflects the test’s content validity. The steeper the item gradient, the bigger the differences are in the standard score. This may explain why two tests have such drastic results. Failing or passing one or a few items may have a large effect on the outcome of the test or measurement (Barnett et al., 1992). Bracken (1988) recommends consulting the norm tables for the different ages for any gaps in the standard scores.

The fourth reason given by Bracken (1988) pertains to how the different norm tables are presented. There are some instruments that have unique characteristics that must be kept in mind when using them. For example, some tests, due to their norm tables, result in drastic differences for the same child even if the child was tested within two successive days. Bracken gives the example of the McCarthy scales where a child could lose eleven intelligence quotient points if tested on two successive days.
A fifth reason given by Bracken (1988) relates to using age or grade equivalents for placement decisions. Bracken states that age and grade equivalents should never be used for making placement or diagnostic decisions because these equivalents do not have the same psychometric properties as standard scores. There is variability within the age equivalents on one test and thus differences between age equivalents on two separate tests are even greater.

The sixth reason is that tests have differences in how reliable they are. Those tests that are not very reliable result in more measurement error and as a result they have larger confidence intervals around the child’s “actual score”. Sometimes the low reliability may explain some of the differences in the scores between the two instruments (Bracken, 1988).

Seventh, although tests may report that they measure the same global area, they may not adequately overlap in the specific skills that are being assessed. The eighth difference is that the content across the tests may be different. Ninth, the instruments may have different publication dates and this may lead to significant differences in performance. Those instruments that were normed some time ago tend to result in higher scores compared to those that were normed more recently. The time between publication dates is directly related to the size of the differences in the scores. Finally, related to the norming of the test has to do with the population that the test was normed on. The older the test the less representative they tend to be. Tests must be chosen that include those from the population that one is preparing to administer it to (Bracken, 1988). The ten reasons given by Bracken reflect all tests, not just those used with the preschool population.
Assessing the Preschool Population

Traditional standards of assessment have historically functioned in terms of a deficit model by sorting children according to different categories (Barnett et al., 1992; Meisels & Atkins-Burnett, 2000; Neisworth & Bagnato, 1992; Schakel, 1986). The areas measured include gross/fine motor development, cognitive development, social/emotional development, and communication development (McLean, Smith, McCormick, Schakel, & McEvoy, as cited in Barnett et al., 1992). There has been a shift in how assessments are completed with the pre-school population. It has changed from highly specified, rigid procedures, which were administered in a formal environment (Meisels & Atkins-Burnett, 2000) to more recent approaches that use everyday experiences that provide children with more of an opportunity to show what they are capable of doing and what they know (Greenspan & Meisels, 1996) and are also conducted in a more natural environment. P.L. 99-457 requires the identification of children, and this has often been done by assessing children using traditional assessment techniques (i.e., norm-referenced tests). Meisels and Atkins-Burnett (2000) advocate a collaborative, ongoing assessment process of systematic analysis and observation.

Accurately defining those who are at risk depends on how one defines success and the different expectations for children entering school (Gredler, 1997). The teaching environment, attendance of the student, number of school changes, and background of the child must all be analyzed when determining who is “at risk”. Screeners are used in attempting to determine if a child is at risk for a future learning or behavior problem. According to Gredler, this is speculative at best because poor
performance has not yet been established and it is not certain whether or not the child will have a behavioral or learning problem.

Early intelligence testing with the pre-school population has resulted in numerous other concerns including the construct of early intelligence, prediction of later performance, professional acceptability, standardization, and appropriateness in decision making (Neisworth & Bagnato, 1992). There is a lack of agreement on how to define early intelligence. Thus, comparing children to a normative sample has been somewhat problematic (Neisworth & Bagnato). Children vary tremendously in their development profiles (Greenspan & Meisels, 1996). Therefore, identifying children based on norm-referenced tests may be problematic due to this variation between children.

Another concern regarding early intelligence testing involves the predictive validity of these instruments (Gredler, 1992; Neisworth & Bagnato, 1988; Satz & Fletcher, 1988). The assumption is that the child’s performance on the intelligence test will remain relatively stable and predict how they will perform when they are older. Yet, there are other variables that are not considered. (i.e., the environment of the child). There have been predictive studies conducted, which have shown that there is little predictive relationship which exist for preschool children who have handicaps, mild problems, or who are at risk for difficulties (Neisworth & Bagnato; Satz & Fletcher). For example, Gredler (1997) found that the Denver II had a positive predictive value of .23. Gredler also analyzed 12 screening instruments with the results indicating that 45 out of 100 children identified as at risk for reading problems were actually capable of reading adequately.
Goodman (1990) reported that the best predictor was the child’s socioeconomic status (SES), as long as the child’s SES did not change. Maternal education level has also been shown to be a predictor in school achievement (Rafoth, 1997; Schakel, 1986). Thus, early intelligence tests should not be the sole determinant in placing a child in a special education class. Another concern with using norm-referenced measures (e.g., intelligence tests) with young children pertained to the lack of children with special needs being included in the norming sample. According to Neisworth and Bagnato (1992) it is rare to include children with special needs in the norming groups of any of the major preschool or infant intelligence tests that are often used by psychologists. There has also been some concern regarding how professionals feel about the assessment of preschool children and the utility of standardized assessments.

A national survey was conducted by Bagnato and Neisworth (1994) of 185 preschool psychologists regarding early intelligence testing of preschoolers who had developmental deficits and the social and treatment validity of these tests. The authors reported three conclusions to their study. First, the early intelligence tests had a high rate of “untestable” children (42%), which resulted in these test having little value in terms of intervention planning. A separate survey conducted by Bagnato, Neisworth, and Butler (1991) resulted in similar results with school psychologist reporting a large number of children deemed as “untestable” (40%). Second, the psychologists who were surveyed reported that there were functional limitations with the tests and specifically the instruments did not accommodate the response limitations of most young children (i.e., sensory, motor, behavioral, or language deficits). Finally, the results indicated that early
intelligence tests did not meet the primary purpose of early childhood assessment, which is to ascertain the need for early intervention services.

*Alternative Methods of Assessing Preschool Children*

The assessment of preschool age children recommendations vary depending upon the literature one consults (Schakel, 1986). The school psychology literature has historically focused on psychometric techniques while the early childhood literature focus has been theories proposed by the likes of Piaget. Finally, the special education literature has placed an emphasis on criterion-referenced tools and assessment for program planning (Schakel). Over the past two decades, the school psychology literature has begun to merge these three viewpoints and advocate a more ecologically based assessment of preschool children.

Although there have been some changes, norm-referenced tests continue to be the primary method when conducting early childhood assessment (Bagnato, Neisworth, & Munson as cited in Gettinger, 2001). Gettinger advocates for professionals to conduct performance-monitoring of children based upon how the staff view child development and learning. Included in this is the understanding that children develop differently and that these differences must be kept in mind when evaluating a child and when creating interventions.

Numerous other researchers have advocated the use of a variety of assessment practices when working with preschool children (Bagnato & Neisworth, 1992; Greenwood et al., 2002; McConnell, 2000; Meisels, 1996; Paget & Nagle, 1986; Schakel, 1986). Assessment should go beyond just testing and include other activities, such as observation, parental interviews, authentic assessment, play-based assessment, and must
be linked to intervention. According to McConnell, the purpose of early education and intervention is to improve the competencies, adjustment, or skills for children and their families. McConnell advocates an ecobehavioral model and states that this type of assessment helps to identify the child’s developmental liabilities and resources, which results in an expanded notion of treatment including both informal supports and formal services in naturalistic environments and in structured settings. An analysis of the child and their environment is key when matching the assessment results to an intervention (Paget & Nagle).

*Dynamic Assessment*

Dynamic assessment is an approach which follows the theory that the best way to understand how a child learns, is through engaging them in the learning process. The purpose is to observe the child apply his or her cognitive functions within the natural learning environment. The inclusion of intervention is unique to dynamic assessment. The process entails observing a child within her natural learning environment, implement interventions targeting any areas that may need remediation, and finally, assessing the child again within the natural environment. Although dynamic assessment allows for children to be observed within the natural environment, there appears to be some psychometric concerns for several reasons. One is the use of gain scores, which tends to be unreliable. Yet, traditional reliability and validity measures are not appropriate for analyzing dynamic assessment. For example, high test-retest reliability indicates low validity for dynamic assessment because the purpose of dynamic assessment is to bring about change not to be stable (Lidz, 1997). Dynamic assessment has shown evidence for improving predictions on standardized achievement tests for students who are lower-
performing specifically, between following intervention (posttests) and achievement test scores (Lidz).

*Ecobehavioral Model*

An ecobehavioral model requires many different perspectives when problem solving. The preschool environment and the family must be understood in terms of their impact on the child and also in their potential for facilitating social and personal development (Barnett et al. 1992). The natural environment also is important in the ecobehavioral model because it allows for providing numerous chances for children to learn alternative or new skills to replace their maladaptive skills (Bailey & McWilliam, 1990). Naturalistic observations can lead to naturalistic interventions, which may occur in the classroom or at home. The result is an assessment plan that can be directed toward the identification of a) a variety of treatment options that are based on functional analysis and research; b) naturally occurring teacher or parent intervention plans that tend to have high rates of success; or c) interventions that can be tailored to fit individual parenting or teaching styles (Barnett et al.).

The goal of assessment is to a) define the problem situation; b) determine the strategies and goals for changing the identified behaviors; and c) evaluate the possible roles of family members, peers, and caregivers, and the environment for supporting change (Barnett et al., 1992). When selecting the targeted behavior and subsequent intervention, the process involves a plan that is created, implemented, continued, evaluated, and changed when necessary.

Barnett et al. advocate an ecobehavioral model, which focuses on the roles and contributions of peers and caregivers who have the most opportunity to be involved
with the child who is experiencing the problems. “Ecobehavioral assessment improves our ability to identify conditions associated with desired or undesired developmental outcomes, identifying potentially effective treatment conditions or components and extending our definition of ‘treatment settings’ to include a wider variety of naturalistic settings” (McConnell, 2000, p. 46).

Barnett et al. (1992) recommend three assessment techniques when conducting an ecobehavioral assessment. First, ecobehavioral interviews, which the authors state have two purposes: a) consider problem circumstances and behaviors and b) analyze the problem circumstances and behaviors in more depth. These interviews allow for a clear description of the problem behaviors across a variety of settings. Teachers and parents are included to gather information, which may include different situations and settings, such as waking up in the morning, and transitioning from one activity to another. The interviews allow for the identification of important time periods, circumstances, persons, and settings that may be related to the child’s behavior and their caregiver’s concerns.

Second, the authors state that observations are imperative for understanding the cultural and ecological impact, the analysis of specific behaviors, skills, interactions, the order of the interactions, and for analyzing the effectiveness of the interventions (Barnett et al., 1992). Preschool children may be observed in a variety of activities and settings. For example, during play activities, while interacting with their peers or adults, and during an academic or learning task.

Finally, Barnett et al. (1992) recommend using curriculum-based measurement when assessing preschool children. This type of measurement involves
analyzing the child’s functional developmental skills while occurring within the instructional environment.

Meisels and Atkins-Burnett (2000) advocate using either a functional-based assessment or play-based assessment when assessing preschool children. Assessments should begin with, first, establishing working, reliable alliances with those who are significant in the child’s life. The goal of an assessment is to gain accurate and useful information regarding a child in order to promote the child’s development (Meisels & Atkins-Burnett). The assessment process should be collaborative with professionals and parents working together and should also involve developmental monitoring. Developmental monitoring involves watching a child’s development closely to ascertain if an assessment is even needed (Dworkin, 1989; Meisels & Atkins-Burnett).

Meisels and Atkins-Burnett (2000) advocate five components for early childhood assessment. First, the target of the assessment should view the child within the familial context. The authors state that children should not be assessed in isolation with no regard to their familial relationships due to numerous studies that link how parental interactions with their children can have an effect on their development (Barnard, Morisset, & Spieker, 1993; Beckwith, 1990). The second component pertains to the context of the assessment. The context of the assessment brings about issues pertaining to social or treatment validity. If social validity is not accounted for, then the probability of obtaining reliable results decreases. Meisels and Atkins-Burnett (2000) state that assessments which incorporate context provide different information than those assessments that attempt to be culture- or context-free. “When we engage in ecobehavioral analysis and contextual assessment,
and when we are concerned about ethnic validity, it follows that our assessments will be marked by multiple sources and multiple methods” (Meisels & Atkins-Burnett, p. 241).

The third component advocated by Meisels and Atkins-Burnett (2000) involves those methods included in the assessment. The goal is to use more ecologically valid assessment methods, including parental and teacher interviews, and naturalistic observation, as opposed to the more traditional methods of assessment of diagnostic decisions based on norm-referenced assessments.

Fourth, the assessment personnel are important and Meisels and Atkins-Burnett (2000) advocate a collaborative model, which includes professionals and parents. Parents need to be actively involved in the assessment process (Meisels, 1991). By including many different viewpoints, such as parents, school psychologist, and teachers the functional abilities of the child can then be evaluated (Fleischer, Belgredan, Bagnato, & Ogonosky, 1990; Paget & Nagle, 1986).

Finally, Meisels and Atkins-Burnett (2000) recommend integration of the assessment with an intervention. There are three assumptions behind this integration. First, is that the assessment process is a dynamic endeavor, which receives information from a variety of sources that are collected at different times to reflect the array of experiences that children have. Second, putting the intervention into practice allows for more information to be acquired so to refine the assessment while simultaneously enhancing the intervention. The third assumption emphasizes the purpose of the assessment process and how it is related to implementing an intervention. This fusion of assessment and intervention results in a performance based assessment (Meisels & Atkins-Burnett).
Bagnato and Neisworth (1992) have advocated a Convergent Assessment Model for preschoolers, which consists of an individually created integration of developmentally based, curricular, eco-, norm-, and judgment based scales that examine different functional domains that are not hypothesized traits, but are actual skills. Examples of alternative functional scales provided by the authors include, parenting stress, functional developmental competencies, and comprehensive development. Alternative approaches to assessing preschool children are advocated by Bagnato and Neisworth primarily due to the link between alternative assessments and the creation of interventions. The authors further advocate for alternative assessments due to demonstrated treatment and social validity (Bagnato & Neisworth, 1994).

Paget and Nagle (1986) focus on Bronfenbrenner’s model of ecological systems and how individuals interact across and within different systems. Children are influenced by their environment and this impacts their development (Greenspan & Meisels, 1996). Therefore, Paget and Nagle also recommend an ecological approach when assessing preschool children. This method is used to help increase the social validity of the interventions implemented since the important environmental influence should be ascertained during the assessment process.

An ecological based method of assessment is a viewpoint which allows for the assessment of preschool children within the natural environment and includes their caregivers who observe their child in numerous settings and can be a valuable asset in the development of interventions. Alternative forms of assessment allow for the child to be viewed holistically and for their culture and socioeconomic background to be considered.

A study conducted by Brooks-Gunn, Klebanov, and Duncan (1996) looked at
483 Black and White premature children from birth to five-years of age to examine any differences in intelligence scores. All of the children were assessed by the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). The results indicated that the Black children performed 1SD below the White children. The authors then adjusted for social and economic differences between the two groups and this resulted in all but completely eliminating the differences between their IQ scores. The results of this study emphasize the importance of looking at children within an ecological framework because of the impact that the home environment, family background, and economic status played in how the children performed on the intelligence test and thus can also have an impact upon how they perform in the classroom and on the creation of suitable interventions.

*Multiple Gating*

A technique which combines traditional norm-referenced and ecobehavioral assessment is a technique called multiple gating. The technique was originally referred to as sequential assessment and had been developed for using psychological tests to make personnel decisions (Cronbach & Gleser, 1965). This technique involved multi-stage plans for the purposes of making decisions after each stage. These multiple stages are called gates and involve using multiple means for completing an assessment. The purpose was to assess as many people as possible as quickly as possible while not spending too much money. Once the first test was completed, the group was then divided into three areas: those who were accepted, those who were rejected, and those who needed to be given a second test. This process was used in making quick personnel decisions.

This technique was later referred to as a sequential gating model and was utilized for dealing with multiple assessment instruments for assessment and
classification in child psychopathology (Reid, Baldwin, Patterson, & Dishion, 1988).
Here the information gathered during Gate I is used for the purposes of making important
and reliable differentiations. The information may be used to differentiate between those
children who have no significant problems, from those who have profound problems,
from those who indicate that there may be a problem yet further assessments are
necessary. Gate II is then used to assess those children who were identified as needing
further assessments. This Gate is more extensive than Gate I. The process is similar to
Gate I in that it separates the children according to those who have no significant
problems from those who need a more extensive assessment. This gating process is a
method which allows for the utilization of multiple assessment measures when assessing
children and is efficient and cost-effective (Reid et al.). According to Reid et al. “Such a
system holds the possibility for allowing clinicians to use multiple assessments in a
sequential manner utilizing present decision rules for interpreting the data, and shows
promise in increasing classification accuracy of childhood disorders” (pp. 189-190).
Historically, though, multiple gating is a technique which has been used primarily in the
social-emotional literature and not within the academic or developmental literature.

In following best practices, it is imperative that those who are identified as
“at-risk” on a screening measure have an appropriate follow-up evaluation conducted.
This is even more essential when the initial screening involved only one test, which was
not conducted in conjunction with a more ecologically based evaluation (Pyle, 2002).
Multiple gating is a technique which allows for the screening of a large number of
children for the purposes of identifying those who may be at-risk for a developmental
delay. This technique initially uses a screening instrument that is easy to give,
inexpensive, and one which tends to have a high number of false positives. The purpose of this is to identify a large number of children. The second gate then uses a screening instrument that is more time consuming and one which does not have as high rate of false positives. This second screening procedure is one that is meant to lower the number of children identified as at-risk. Finally, the third gate is one that is more intrusive than the first two gates and one that is also more extensive. This gate may entail an observation of the child, parent interview, and further assessments. The purpose of these gates is to correctly identify those children who may have a developmental delay. Multiple gating is a technique which has historically been used within the behavioral literature. Yet, this technique is one that can be applied to help identify children with developmental delays or academic problems.

Chapter Summary

The purpose of this review of literature was to present information on developmental screening of the preschool population. Historically, preschool age children were assessed by using the same means as assessing older children. Preschool age children often present different issues than older children and also respond differently to the testing environment. Ecobehavioral assessments have been advocated for when assessing preschool age children. This type of assessment allows for the inclusion of the child’s family and also the observation of the child in multiple settings.

Preschool age children have a low rate of those identified as having a developmental delay. A low base rate problem such as this one tends to result in a high number of false positive predictive errors (children identified as at risk for developmental delays who are actually not at risk). Multiple gating is a technique which has shown to
increase the number of valid positives (identifying those who are actually at risk) while lowering the number of false positives. This technique infuses elements of traditional assessment and also ecobehavioral assessment. This technique is more common in the behavioral field than in the academic yet it is one that can be applied to help identify children with developmental delays or academic problems.
CHAPTER 3

METHODOLOGY

Introduction

The purpose of this study is to evaluate the effectiveness of a multiple gating program in identifying preschool-age children enrolled at the CDC Head Start of Franklin County who were at-risk for a developmental delay. The CDC Head Start differentiates between those children referred based upon potential receptive and/or expressive language delays from those referred based upon potential delays in the areas of socialization skills, self-help, gross motor, fine motor, and/or cognition. Those children at-risk for receptive and/or expressive delays are referred to as speech referrals while those children at-risk for delays in socialization skills, self-help, gross motor, fine motor, and/or cognition are referred to as developmental referrals. Although receptive and expressive language delays fall under the federal definition of a developmental delay, those children referred for a speech/language assessment were assessed using different measurements than those included in this current evaluation and are categorized by the CDC as being at-risk for a speech/language delay and therefore were excluded from this current analysis.
The CDC Head Start used a multiple gating program during the 2002-2003 school-year. This program involved initially screening all children with the Denver II developmental screener, then screening those children identified as suspect in the areas of self-help, socialization skills, gross motor, fine motor, or cognition on the Denver II with the Early Screening Inventory – revised (ESI-R), and finally an additional screening instrument, the Brigance Preschool Screen was completed by Columbus public schools on those who were in the refer range on the ESI-R. In addition, Columbus Public Schools also completed a play-based observation, parent interview, and a language checklist. This multiple gating program was used to identify those children who were at-risk for a possible developmental delay. Predictive and concurrent validity will also be assessed by using the Galileo developmental assessment scales, which is a criterion referenced assessment. Every student who attended the CDC Head Start during the 2002-2003 school year was screened with the Galileo at least two times throughout the year.

Participants

The CDC Head Start of Franklin County is an agency which served approximately 2400 pre-school aged children during the 2002-2003 school year. The racial make-up of children consisted of Black, non-Hispanic (n= 1479), White, non-Hispanic (n= 498), Hispanic (n=116), Asian or Pacific Islander (n= 125), multi-racial (n=118), and American Indian (n=9). Approximately 85% of the children attending the CDC Head Start met the federal definition for living in poverty.

This study will consist of an archival data analysis of the results of all children who were evaluated with the Denver II developmental screening instrument, ESI-R screening instrument, and had an evaluation completed by Columbus Public Schools
During the 2002-2003 school year while attending the CDC Head Start. Since the purpose of this study is to ascertain the effectiveness of this multiple gating program implemented by the CDC Head Start, all of these screening instruments will be included. In addition, the CDC Head Start evaluated all of the children attending the CDC Head Start during the 2002-2003 school year with the Galileo, a criterion referenced assessment. These assessments occurred at the beginning and end of the school year. The Galileo results will be used for evaluating the concurrent and predictive validity of the multiple gating method. Due to the CDC Head Start’s screening students for the purpose of correctly identifying those who meet criteria for a low base-rate problem, it is expected that the selection ratio of the participants will be reduced during Gate II to Gate III since the overall accuracy is expected to increase.

During the 2002-2003 school-year, 8 of the CDC Head Start sites chose to implement the LAP-D screener instead of the Denver II. There are no differences between those children attending these sites from those sites that use the Denver II. The CDC Head Start is discontinuing the use of the LAP-D after the 2002-2003 school year. As a result, this study will exclude all of those children who were not evaluated by the Denver II developmental screening instrument (n= 449).

Research Design

This study is a descriptive analysis of archival data collected for the purpose of ascertaining the effectiveness of a multiple gating program currently implemented by the Child Development Council (CDC) Head Start of Franklin County. Multiple gating incorporates a stepwise assessment procedure which begins with a large-scale screening for purposes of identifying children who have a low base-rate problem such as a
developmental delay in the preschool-age population. Further screenings are completed, which are more intensive in nature and are conducted on a smaller group of children who had been identified by the preceding large-scale screening. Multiple gating procedures have rarely been used in assessing for developmental delays and the research literature on this technique is severely lacking in the field of education.

This study is a replication of the methodology used by Loeber et al. (1984) in their study on identifying youth who were at-risk for delinquency. In this current study, Gate I consists of screening the CDC Head Start Children with the Denver II developmental screener. These children were screened using the Denver II within the first three months of their enrollment. Gate II involves screening those children identified on the Denver II as suspect with the Early Screening Inventory – Revised (ESI-R). Excluded are those children who are referred based upon potential delays in expressive or receptive language. Those children are screened with the Bracken Basic concepts, not the ESI-R. The ESI-R screening instrument identifies a child in one of three categories; either ok, rescreen (meaning that the child must be rescreened in an eight-week time period); or refer. The third gate involves an additional evaluation completed by Columbus Public Schools. This evaluation is completed on those children who performed within the refer range on the ESI-R and includes a parent interview, child observation, and screening using the Brigance Preschool Screen.

The criteria used to interpret the results of this study are similar to the ones used by Loeber et al. (1984).
1. Each consecutive level of screening should result in an overall increase in discriminative efficiency, and thus simultaneously decreasing the false positive error rate while increasing the valid positive rate in the risk group.

2. The multiple-gating process should result in a risk group of preschool-age children with the proportion of valid positive larger than 50% and, while at the same time, the proportion of the false positive error rate should be less than 50%.

3. A majority of those who are actually developmentally delayed should be located within the risk groups (this results in a false negative rate that is less than 50%).

4. When comparing a single-gate prediction to the multiple-gating process, the results should not be an overall loss of predictive validity.

The present study demonstrates the utility of the three gates in screening preschool-age children for a developmental delay.

*Instruments*

*Gate I: Denver II Developmental Screening Instrument.* The Denver II is a brief, easy to administer developmental screening instrument which is administered by either the CDC Head Start classroom teacher or the CDC Head Start Educational coordinator. According to a study completed by Glascoe & Byrne (1993) the Denver II tends to have a high rate of false positives, which Gredler (1997) advocates for initial screening instruments.

The purpose of the Denver II is to identify those preschool-age children who are atypical in one of four areas: gross-motor (walking, sitting, jumping, and their overall
use of large muscle movements), person-social (how well the child gets along with others and how they care for their personal needs), language (using, understanding, and hearing language), and fine motor/adaptive (manipulation of small objects, eye-hand coordination, and problem-solving) compared to their same-age peers. The child’s behavior during the assessment process is also assessed for the purposes of determining how well they use their abilities (Frankenburg et al., 1992).

According to the Denver II Training Manual (Frankenburg et al., 1992) a quota sample was used with the Denver-II to control for ethnicity, residence, and maternal education within age groups. The data were collected by 17 screeners during 1988. All of the children included in the study were from urban, rural, and semi-rural regions in Colorado. Only one child per family and those who were born full term without any obvious defects – were in the sample of 2,096 preschool-age children.

The norms were computed by analyzing the data through logistic regression for determining the ages at which the children passed each item. Each item had been administered approximately 440 times and some items up to 1,309 times (u = 783, SD = 190.7). The analysis was completed to determine when the children passed a test item (specifically when 25%, 50%, 75%, and 90% of the children passed an item). A child can perform in either the Normal, Suspect, or Untestable range. The Normal range indicates that the child currently does not exhibit any delays. The Suspect range indicates that the child may need a more intensive screening to determine if they are at-risk for a developmental delay. The Untestable range means that a child refused to complete a testing item. If a child tests within the Suspect or Untestable range, the child is to be rescreened in 1 to 2 weeks to ensure that the child’s performance was not the result of
illness, fear, or fatigue. If the child again tests within the Untestable or Suspect range, it is recommended for the child to be referred for a more intensive screening to determine if they are at-risk for a developmental delay (Frankenburg et al., 1992).

The reliability of the Denver II was completed by recruiting thirty-eight children from the 10 age groups and evaluating the children on two separate occasions which were separated by seven to ten days. Two different types of reliability were assessed and they were test-retest stability and concurrent examiner-observer reliability. “The mean examiner-observer reliability for the items included in the Denver II is .99 with a range of .95-1.00 and a standard deviation of .016. The mean 7 to 10 day test-retest reliability for the same items is .90 with a range of .50-1.00 and a standard deviation of .12.” (Frankenburg et al., 1992, p. 4).

There are 125 test items and they are represented on the test form by a bar which spans across the ages. The ages include birth to six years of age. Until 24 months, each space on the scale represents one month, after 24 months, each space is representative of 3 months. The bar also depicts the percentages of the number of children in the standardization sample who passed that particular item (i.e. 25%, 50%, 75%, and 90%) (Frankenburg et al., 1992).

Validity studies of the Denver II were not conducted. Instead, the author’s chose to base the validity of the Denver II on the content validity studies that were completed on the original Denver Developmental Screening Test (DDST). According to Glascoe & Byrne (1993) this lack of validity studies may contribute to the Denver II producing more incorrect than correct identifications yet, the author’s stated that the Denver II is capable of identifying those children who have more subtle problems.
Gate II: Early Screening Inventory- Revised (ESI-R). The ESI-R is a developmental screening instrument designed to be used with children ages 3 to 6. The ESI-R and Denver II differ in that the goal of the ESI-R is to identify those preschool-age children who may be in need of special education services. The ESI-R is to be used only for identifying the possibility of a handicapping or learning condition which may affect a child’s potential for being successful in school (Meisels, Marsden, Wiske, & Henderson, 1997).

The purpose of the ESI-R is to provide a brief overview of the child’s development within three areas: language and cognition (language comprehension and verbal expression; the child’s ability to count and reason; and how well they remember auditory sequences), visual motor/adaptive (assess fine motor skills, hand-eye coordination, and short-term memory skills), and gross motor (evaluates large muscle coordination). Children are classified in one of three categories: “Refer”, which means that the child should be assessed further as soon as possible; “Rescreen”, which indicates that another ESI-R assessment should be administered within 8 to 10 weeks; or “Ok”, which indicates that the child does not need a further assessment (Meisels et al.). During this second gate, the child is evaluated by a CDC mental health consultant.

This current edition of the ESI was standardized on approximately 6,000 pre-school age children and there are two different age groups represented, the ESI-P for those children who are ages 3 to 4 1/5 and the ESI-K for those children who are 4 1/5 to 6. The author’s stated that they created two different ESI tests because the two age groups differ (Meisels et al., 1997). The ESI-P and ESI-K standardization, validity, and reliability were completed separately and reported separately too.
The data was collected during 1986-1990 and again during 1992-1994 to increase the sample size. The total data for both the ESI-P and ESI-K was collected in 10 states using 60 different classrooms. The total sample size used consisted of 5,034 preschool-age children. The sample contained approximately equal females to males and the ethnic/racial representation was approximately 16% African American, 70% White/non-Hispanic, and 14% classified as other. The education of the mother was reported as approximately 80% completed high school while less than 20% did not complete high school. The ESI-K data reported that around 52% of the children attended public school, 32% attended a Head Start program, and the rest of the children attended some other type of preschool program (Meisels et al., 1997).

A predictive validity sample was completed on the ESI-K by drawing a sample of 251 of the children and administering the McCarthy Scales of Children’s Abilities (MSCA). The sample used children from 12 different sites. Follow-up testing was then completed seven to nine months following the initial screening (M = 7.79 months). The non-validity sample was comparable to the validity sample in terms of fathers’ occupation, maternal education, and race. In relation to the MSCA outcome the results indicated a sensitivity of .93 and a specificity of .80. This indicates that 93% of those children who were identified as at-risk were identified correctly. Also, 4 out of 5 children were correctly identified as not being at-risk and as a result, were not referred for more testing (Meisels et al., 1997).

An analysis was then completed on the ESI-K and included ascertaining the mean scores for each item by age so to establish norms. An item analysis was run for the
purpose of determining if discrimination occurs between the Ok/Rescreen and Refer
groups and to demonstrate the reliability on the combined samples (Meisels et al, 1997).

The results of the item analysis showed a gradual increase in the success on
items with age. Older children were more likely to perform better on the items when
compared to younger children.

An item analysis was then completed to ascertain if the items on the ESI-K
discriminated between those children who passed the ESI-K screening compared to those
children who needed further evaluation. The item analysis included the scores of all
5,034 children in the sample. Of the entire sample, 4.2 % or 212 children were referred
for a more extensive evaluation and 95.8% or 4,822 children passed the screening. An
independent sample T-test was completed and indicated that a significant difference
occurred on each of the items on the ESI-K between those children who were in the Ok
group from those identified as in the Refer group (p <.001) (Meisels et al., 1997).

Inter-rater reliability was determined by performing a Cronbach’s alpha.
According to Meisels et al (1997) when assessing for inter-rater reliability a Cronbach’s
alpha is equal to a Guttman’s split-half coefficient. The data used were from 586 tester-
observer pairs. The correlations for the observers and testers were above .97. The alphas
were also obtained for tester observer pairs and the results were .98. The test-retest
reliability was obtained by having a child assessed by two different examiners seven to
ten days apart. One-hundred and seventy-four test-retest pairs were used with twenty-two
examiners. Counter-balancing was used with the examiners for order effects. The whole
group resulted in a Cronbach reliability coefficient of .87. This indicates that the ESI-K is
consistent and stable. Finally, the standard error of measurement was obtained and based
on 586 total cases, amounted to .067. This standard error of .067 indicates little test-based error (Meisels et al., 1997). The index of sensitivity was .93 while the index of specificity was .80. An index of sensitivity of 80% and index of specificity of 90% is recommended (Glascoe & Byrne, 1993).

The Early Screening Inventory – Revised Preschool (ESI-P) was validated and standardized separately from the ESI-K. The ESI-P had data collected from 16 sites during 1993 to 1996. Ten of the sites were Head Start programs. 977 children were included in the sample and it was divided into increments of six-months and included ages 3.0 to 4.5. The sample consisted of approximately 48% female and 52% male. The ethnic/racial make-up of the sample consisted of 21% African-American, 26% identified as “other” (i.e., American Indian, Asian, etc), and 53% White/non-Hispanic.

Approximately 29% of the fathers and 3% of the mothers did not finish high school (Meisels et al., 1997).

A predictive validity sample was completed on the ESI-P by drawing a sample of 130 of the children and administering the MSCA. All of those children who were within the Refer range were chosen to complete follow-up testing and a random equivalent-sized sample of those children identified within the Ok or Rescreen range were also selected for the MSCA. Follow-up testing was then completed four to six months following the initial screening (M = 5.62 months). The non-validity sample was comparable to the validity sample in terms of father’s occupation, maternal education, and race. In relation to the MSCA outcome the results indicated a sensitivity of .92 and the specificity of .80. This indicates that 92% of those children who were identified as at-
risk were identified correctly. Also, 4 out of 5 children were correctly identified as not being at-risk and as a result, were not referred for more testing (Meisels et al., 1997).

An analysis was then completed on the ESI-P and included establishing the mean scores for each item by age so to establish norms; running an item analysis for the purpose of determining if discrimination occurs between the Ok/Rescreen and Refer groups; and demonstrating the reliability on the combined samples. The results of the item analysis showed a gradual increase in the success on items with age. Older children were more likely to perform better on the items when compared to younger children (Meisels et al., 1997).

An item analysis was then completed to ascertain if the items on the ESI-P discriminated between those children who passed the ESI-P screening compared to those children who needed further evaluation. The item analysis included the scores of all 977 children in the sample. Of entire sample, 20.7% or 202 children were referred for a more extensive evaluation and 79.3% or 775 children passed the screening. An independent sample T-test was completed and indicated that a significant difference occurred on each of the items on the ESI-P between those children who were in the Ok group from those identified as in the Refer group (p < .05). Those children in the Ok group performed better on each item compared to the Refer group (Meisels et al., 1997).

A Cronbach’s alpha was used to determine the inter-rater reliability. The data used were from 35 tester-observer pairs. The correlations for the observers and testers were above .99. The test-retest reliability was obtained by having a child assessed by two different examiners seven to ten days apart. Five test-retest pairs were used with two examiners. A correlation of .98 was obtained between the test and retest scores. The
whole group resulted in a Cronbach reliability coefficient of .98. This indicates that the ESI-P is consistent and stable. Finally, the standard error of measurement was obtained and based on 35 total cases, amounted to .20. This standard error of .20 indicates little test-based error (Meisels et al., 1997).

The ESI-P and ESI-K both were analyzed to ascertain the false positive and false negative rate. According to Meisels et al (1997) both measurements tend to over refer students (false positives) while simultaneously exhibiting a low false negative rate. “This means that the ESI-P, similar to the ESI-K, not only refers a high proportion of children who are actually at-risk, but also excludes most of the children who are not at-risk from further evaluation” (Meisels et al., 1997, p. 164).

Gate III: Columbus Public Schools Multifactored Evaluation. Columbus Public Schools conducts an additional evaluation on those children who are referred by the CDC Head Start. Their evaluation consists of a parent interview, play-based observation of the child, administration of the Brigance Preschool Screen, and the completion of a criterion based language checklist. The child is evaluated in terms of their adaptive behavior (how well they separate from their parent, can they take off their coat, etc.), cognition (do they know their colors, can they count, how elaborate do they play), communication (see Figure 2: Columbus Public Schools Criterion Language Checklist), motor skills (large muscle coordination), and social-emotional behavior.

The parent interview is a non-standardized form completed to assess the parent’s view of their child’s strengths and weaknesses. This interview is then written up and summarized as part of the evaluation of the child. The process provides an ecobehavioral view of the child and includes how their parent views their development
and their performance in the home compared to how they perform in the classroom. An ecobehavioral evaluation allows for a more holistic view of the child by using multiple means for gathering information while being completed in the child’s natural environment.

The child is then assessed by the examiner while he/she plays in a classroom. Some areas the examiner assesses are how well does the child separate from the parent or guardian, how the child plays (i.e., do they play appropriately), his/her ability to problem-solve in the classroom, and how well he/she moves around the classroom. Observations of a child in his/her classroom are best practices. Assessment of a child in the classroom is also conducted when completing a multifactored evaluation.

The child is also given the Brigance Preschool Screen. This screening measure is used for the purposes of identifying those preschool age children who may need to be referred for services that address developmental difficulties.

The Brigance Preschool Screen is a criterion-reference measure which has been recently standardized (Glascoe, 1995). The Brigance Screens originated from the Brigance Diagnostic Inventory of Early Development (IED, 1979). Items were selected from the IED by diagnosticians, teachers, and curriculum supervisors throughout the United States and those items that were nominated by approximately 90 percent of these professionals were kept to create the Brigance Screens. These items were then field tested around the United States for purposes of refining directions and clarifying item content.
Birth to 1 year:  

_____ Turns to look at others
_____ Correctly responds to requests for objects

1 year to 2 years:

_____ Uses between 3 and 20 words
_____ Asks for “more”

2 years to 3 years:

_____ Uses no in combination with other words
_____ Gives labels for a variety of objects

3 years to 4 years:

_____ Makes effort to count objects
_____ Describes a sequence of two events

4 years to 5 years:

_____ Tells stories
_____ Follows a series of 3 commands

5 years to 6 years:

_____ Uses complete, grammatical sentences in all verb tenses
_____ Uses language for a variety of purposes

Figure 2.1: Sample Criterion Language Checklist Used by Columbus Public Schools
The Brigance Screens is made up of the Brigance Preschool Screen, Brigance Early Preschool Screen, and the Brigance K & 1 Screen. The Brigance Preschool Screen involves a Three-Year and Four-Year form. The Brigance Early Preschool Screen includes a Two-Year and Two-and-a-Half-Year form. Finally, the Brigance K & 1 Screen has a Kindergarten and First-Grade form, plus an additional form for the End-of First-Grade.

An independent study was commissioned by Curriculum Associates (Glascoe, 1995) on the Brigance Screens for the purposes of: a) defining how children perform across socioeconomic levels, geographical boundaries, and other variables; b) identifying the best cutoff score for the identification of those children who are at-risk for school failure and also for those who may be in need of being assessed for special education services; c) determining the predictive validity; and d) determining how reliable the Brigance Screens are in terms of test-retest and internal consistency and inter-examiner reliability.

The standardization occurred in four different sites located throughout the United States. The sample size included 408 children. Children were recruited for purposes of representing the United States based upon their ethnicity/race, socioeconomic status, and parental occupation.

The validation study completed on the Brigance Screens used Guttman scalability coefficients for each form. These coefficients were used to indicate that each of the forms and items are unidimensional, homogenous, and hierarchical measures of readiness and academic skills. The results of the Guttman coefficients were as follows: 2-
year form = .96; 2.5-year form = .98; 3-year form = .97; 4-year form = .99; K – form = .99; 1<sup>st</sup> grade form = .90; and end of 1<sup>st</sup> grade = .52 (Glascoe, 1995).

The Standard Error of Measurement was also completed to specify how well the sample mean approximates the population mean. The Standard error of measurement results were: 2-year form = 1.88; 2.5-year form = 2.70; 3-year form = 1.43; 4-year form = 1.29; K- form = 1.26; 1<sup>st</sup> grade form = 1.06; and End of 1<sup>st</sup> = 2.52 (Glascoe, 1995).

The test-retest of the Brigance Screens primarily comes from the validation study done on the Inventory of Early Development (IED) which is the larger instrument that the Brigance Screen items come from. The IED was administered on two separate occasions to 1,156 children, who had been stratified according to their ethnicity (11% were Hispanic and 14% were African American), gender, and geographic region. The test-retest agreement during this validation study was 86%. Coefficients averaged between 87% and 98% in all but two areas of the test. Research was completed on the K & 1 Screen with test-retest agreement at 82%, even though the interval was 2-3 months after the initial screening (Glascoe, 1995).

An evaluation of the inter-rater reliability was conducted by Mantzicopoulos and Jarvinen (as cited in Glascoe, 1995). These authors looked at a sample of 134 kindergarten and first-grade children who were from low socioeconomic status backgrounds living in Indiana. Thirty-seven of these children were given the Brigance screens by two examiners and they were tested 3 months apart. A test-retest of .97 was obtained.

The content validity of the Brigance Screens has been addressed significantly in the research literature (Brennan; Helfeldt; Schearer as cited in Glascoe, 1995) the
results indicating that the Brigance Screens measures what it purports to measure. The Brigance Screens’ means do increase as the age of the child increase thus there is age-discriminating power. The power to discriminate between children based upon their age is another way of establishing content validity.

The Brigance Screens were assessed to determine their predictive validity. According to Glascoe (1995) those children who were identified on the Brigance Screens continue to have difficulty in school 5 months and even up to 6 years after the administration of the screening. Mantzicopoulos (1999) conducted an outcome/prediction analyses to explore how accurate the Brigance was in predicting the special education status of children at the end of their preschool years, the results indicated that the Brigance had a fairly high false negative rate while also having a reasonable true positive rate.

An index of sensitivity and specificity were also conducted for purposes of determining the cutoff scores for identifying those children who may be at-risk for academic difficulty or a disability. Separate cutoffs were used for older and younger children. The Index of Specificity for the Brigance Screen forms are as follows: Two-year old = 100%; Two-and-a-half Year old = 92%; Three-year old = 76%; Four-year old = 74%; Kindergarten = 75%; First-grade = 73%; End of First = 77%. The overall Index of Specificity was .86. The Index of Sensitivity for the Brigance Screen forms are as follows: Two-year old = 100%; Two-and-a-half Year old =75%; Three-year old = 74%; Four-year old = 73%; Kindergarten = 77%; First-grade = 72%; End of First = 78%. The overall Index of Sensitivity was .75 (Glascoe, 1995). Although both the index of sensitivity and specificity were slightly lower than the ideal (80 and 90% respectively)
the Brigance screening is only one component of the Columbus Public Schools screening of pre-school children.

The addition of more ecologically based forms of assessment, including a play-based observation, parent interview, and a criterion language assessment is done to ensure that those children who are truly at-risk for a developmental delay (valid positive) are identified and referred for a complete multifactored evaluation while those who are not truly at-risk (valid negative) are not referred for a multifactored evaluation.

Columbus Public Schools completes the evaluation and then determines if the child needs further assessment, which entails a full multifactored evaluation or if the child does not need any further assessment. Gate III is used in determining if a full multifactored evaluation is needed for those children who exhibit behaviors that indicate that they may have a developmental disability.

*Criterion-Measure- The Galileo Scales:* The Galileo was used as the criterion-measure because it is an academic measure utilized by all of the CDC Head Start classrooms and therefore is a naturally occurring assessment of each child’s academic performance and occurs at least twice a year. The Galileo is considered a dynamic assessment because the Galileo is used as a pre- post-test with the results of each student’s performance being used to create a curriculum individually tailored to address each child’s academic needs. The CDC Head Start has chosen the Galileo for several reasons, one of which is the CDC’s belief of a Galileo-Denver connection (personal communication, June 9, 2003).

The purpose of the Galileo Scales is to link assessment to instruction by evaluating each child using multiple means. The child’s development is assessed by the
classroom teacher through direct observation during instructional activities and by conducting a one-on-one evaluation. This information is then used to provide corresponding learning opportunities to support development. Several areas are evaluated by the Galileo: math (counting skills, shapes, patterns, time-concepts, and graphing); language and literacy (understanding and listening, speaking and communicating, phonological awareness, book knowledge and appreciation, and alphabet knowledge); social development (pro-social behavior, social responsibility, and social understanding); approaches to learning (initiative and curiosity, learning about events and objects, engagement and persistence, goal setting and planning, reasoning and problem-solving); gross and fine motor development (cognitive and perceptual-motor development, physical development and motor development, balance and gross motor control, gross motor coordination, fine motor control and dexterity, and hand-eye coordination); self-help (eating and nutrition, exercise practice, hygiene practice, dressing and toileting, health and safety practices) (Bergan, J.R. et al. 2003). An average score is 500 with a standard deviation of 50. For the purposes of this study, only the Early Math, Language/literature, Fine/Gross motor, and Social Development measures were included due to these four measures being the only measures that are required by the state of Ohio and therefore completed by all of the CDC Head Start sites.

The Galileo Scales is a criterion referenced assessment, which follows a problem-solving model for purposes of assisting in the creation of educational goals for each child. The Galileo is scored on the computer and consists of a number of software applications. Included are assessment, goal setting, quality control, training, planning, implementation, monitoring, evaluation, and communication. As a result of the computer-
based scoring program, the Galileo is consistently updating the standardization sample.

Standardization occurred in 1994 with a sample size of 2638 children who were recruited for purposes of representing the United States based upon their ethnicity/race, and socioeconomic status. Approximately 50% of the students were female and 50% were male. 34% came from the south, 24% from the Midwest, 28% from the northeast, and 22% from the west. 36% were Caucasian, 31% were African American, 30% were Hispanic, 1% Asian, less than 1% was Native American, and less than 1% was considered other. Minorities were over-sampled purposefully to better represent the population served by the Galileo.

The validation study completed on the Galileo scale computed Coefficient Alpha for each form. These coefficients were used to indicate that each of the forms and items are unidimensional, homogenous, and hierarchical measures of readiness and academic skills. The results of the coefficients were as follows Early Math Scale = .93; Language and Literacy = .94; Nature and Science Scale = .94; Social Development Scale = .96 (Bergan, J. R.et al. 2003).

An evaluation of the inter-observer reliability was conducted on the Math scales and Language and Literacy scales for purposes of controlling for measurement error. Observations occurred on three randomly selected students from each of the 318 classrooms participating in a Head Start program in the State of Ohio. The two observers had very similar results. Observer one had a Math Scale mean of 10.93 while Observer two had a mean of 10.86. Observer one had a Language and Literacy scale mean of 14.72 while Observer two had a mean score of 14.57.
The content validity of the Galileo Scale has been addressed by the authors’ of the Galileo Scales and the results indicate that the Galileo Scales measures what it purports to measure. The Galileo Scales means do increase as the age of the child increase thus there is age-discriminating power. The power to discriminate between children based upon their age is another way of establishing content validity.

**Procedures**

The multiple-gating process being evaluated in this study is pictorially represented in Figure 1. All of the children who have been evaluated by the Denver II and subsequently the ESI-R results are kept by the mental health manager and mental health consultant supervisor at the CDC Head Start. Columbus Public Schools keeps the information on those children who had a multifactored evaluation conducted. The number of children who continue through the three gates and their results will be evaluated for the effectiveness of this program. No identifying information of any of the children will be needed in completing this study.

**Data Analysis**

SPSS software will be used for purposes of analyzing the data. The formulas used for computing measures of predictive efficiency are summarized in Table 1. These formulas are the same used by Loeber et al. (1984) in their study. The base rate depicts the prevalence of the actual developmental status (at-risk for a developmentally delayed vs. not at-risk for a developmentally delayed) in the sample prior to the screening. The selection ratio is the number of preschool-age children predicted to be developmentally delayed based on a screening measure.
(Based on a two-by-two prediction table; $N = A + B + C + D$)

<table>
<thead>
<tr>
<th>Actual academic rating</th>
<th>official developmental status</th>
</tr>
</thead>
<tbody>
<tr>
<td>At-Risk</td>
<td>Not At-Risk</td>
</tr>
</tbody>
</table>

**Predictive Behavior**

<table>
<thead>
<tr>
<th>At-Risk Group</th>
<th>Valid Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$B$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not At-Risk Group</th>
<th>False Negative</th>
<th>Valid Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>$D$</td>
<td></td>
</tr>
</tbody>
</table>

**Base Rate (BR)** = $A + C/N$

**Selection Rate (SR)** = $A + B/N$

**Observed Percent Correct (OPC)** = $A + D/N$

**Chance Percent Correct (CPC)** = $BR \times SR + (1 - BR) \times (1 - SR)$

**Maximum Possible Percent Correct (MP)** = $A + D/N$ when either $B$ or $C$ are set to 0 (If SR exceeds BR, C is set to 0; otherwise B is set to 0).

**Improvement over chance (IOC)** = $OPC - CPC$

**Valid Pos. Rate** = $A/A + B$

**False Pos. Rate** = $B/A + B$

**False Neg. Rate** = $C/C + D$

Table 3.1: Computational Formulas for Measures of Predictive Efficiency

To demonstrate the effectiveness with which the gating method differentiates between those who are at-risk for a developmental delay from those not at-risk for a developmental delay, several interrelated indices are utilized (see Table 3.1 for the computational formulas). The Observed Percent Correct (OPC) compiles the percentage of those correctly identified as at-risk for a developmental delay and those not at-risk (valid negatives plus valid positives). According to Loeber et al. (1984), “the RIOC index takes into account the range of possible correct identifications due to disparities between the selection ratio and base rate” (p. 21).
Two types of error that are important for this study are false positives and false negatives. A false positive is an error which pertains to a child being identified as developmentally delayed when they are not. A false negative refers to a developmentally delayed child who is actually identified as not being developmentally delayed (Gredler, 1997).

The independent variables in this study are the three different Gates while the Dependent variable is the official developmental status. The degree of correlation between the independent and dependent measures will be determined in this study by computing a point-biserial correlation for the official developmental status criteria and a product-moment correlation will be computed for the rest of the variables (see Table 3.2: Intercorrelations between the independent variables and dependent variables for preschool-age children). A point-biserial correlation was chosen due to the official developmental status being a dichotomous variable.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Denver Results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>ESI-R Results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>CPS – Screening Results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Galileo results</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Intercorrelates between Independent and Dependent Variables

Each Gate will be evaluated by using prediction tables and summary statistics. Both chi-square and phi coefficient will be conducted to assesses whether or not the screening instruments conducted during each gate differentiates between those at-risk for
a developmental delay and those not at-risk for a developmental delay (see Table 3.3: Prediction Table and Summary Statistics for Gate I Screening with the Denver II; see Table 3.4: Prediction Table and Summary Statistics for Gate II Screening with the ESI-R; and see Table 3.5: Prediction Table and Summary Statistics for Gate III Columbus Public Schools Multifactored evaluation). Finally, the prediction outcome for the entire procedure will be conducted and summarized for the entire sample (see Table 3.6: Prediction Outcome for Multiple-Gating Method Summarized for the entire sample).

<table>
<thead>
<tr>
<th>Actual Academic Rating</th>
<th>At-Risk</th>
<th>Not At-Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Status</td>
<td>Valid Pos.</td>
<td>False Pos.</td>
</tr>
<tr>
<td>At-Risk Group</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Not At-Risk Group</td>
<td>False Neg.</td>
<td>Valid Neg.</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number in Risk Group</th>
<th>Valid Positive Rate</th>
<th>False Pos. Rate</th>
<th>False Neg. Rate</th>
<th>Total Observed Correct</th>
<th>Chi-Square</th>
<th>Phi-coefficient</th>
</tr>
</thead>
</table>

p

Table 3.3: Prediction Table and Summary Statistics for Gate I: Screening with the Denver (N=1888)
### Table 3.4: Prediction Table and Summary Statistics for Gate II: Screening with the ESI-R (N=29)

<table>
<thead>
<tr>
<th>Actual Academic Rating</th>
<th>At-Risk</th>
<th>Not At-Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At-Risk Group</td>
<td>Valid Pos.</td>
<td>False Pos.</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Not At-Risk Group</td>
<td>False Neg.</td>
<td>Valid Neg.</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Number in Risk Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid Positive Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Pos. Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Neg. Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Observed Correct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-Square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phi-coefficient</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

p

Table 3.5: Predictive table and Summary Statistics for Gate III: CPS screening (N=14)
Table 3.6: Predictive Outcome for Multiple-Gating Procedure, Summarized for the Full Sample (N=)

**Ethics**

This current study is an archival data analysis which is being conducted to ascertain the effectiveness of a naturally occurring assessment program. No identifying information will be published for the purposes of this study. Although individual children’s performance on the developmental screening measures will be needed to determine if this process is effective, none will have any personal information disclosed and only whether or not they needed further assessment will be evaluated.
CHAPTER 4

RESULTS

The purpose of this study was to evaluate the effectiveness of a multiple gating program currently used by the CDC Head Start of Franklin County in identifying pre-school students who were at-risk for a developmental delay. The data analysis and procedures used here were a replication of the procedures and analysis used by Loeber et al. (1984). A developmental delay is defined by the CDC as including deficits in self-help, socialization skills, gross motor, fine motor, or cognition. Initially, 1888 children were assessed with the Denver II developmental screening instrument. Of these students, 29 were referred to the CDC mental health consultants for a further evaluation (see Table 4.1: Prediction table and summary statistics for Gate I: Screening with the Denver II). Fourteen were found to need a further more intensive evaluation while fifteen were identified by the mental health consultants as scoring within the Ok range on the ESI-R and therefore no further evaluation was conducted (see Table 4.2: Prediction table and summary statistics for Gate II: Screening with the ESI-R).

Research Questions

An attempt was made to answer the following questions:
1. Do the number of false positives decrease while the number of gates increase?
2. Does multiple gating improve on the rate of valid positives?
3. How many of the valid positive children were initially identified by the first screener?

4. How many of the valid positive children were identified by the combination of the first two screeners?

5. What are the error rates of the false positives and false negatives?

**Question 1: Do the number of false positives decrease while the number of gates increase?**

“The end result of the multiple-gating procedure should be a risk group of children with a valid positive rate greater than 50% and, conversely, a false positive error rate of less than 50%” (Loeber et al., 1984, p. 12). Following the methods of Loeber et al. the false positive rate for Gate I was identified as 17.8% while the false positive rate (see Table 4.1) for Gate II was 81.8% (see Table 4.2), which indicated a high false positive rate during Gate II.

**Question 2: Does multiple gating improve on the rate of valid positives?**

The valid positive rate for Gate I was 82.1% yet it was 18.1% during Gate II. Of all of the children who were initially screened with the Denver II developmental screening instrument, 23 were identified as a valid positive during Gate I. This number decreased drastically by Gate II with only 2 children considered being a valid positive.

**Question 3: How many of the valid positive children were initially identified by the first screener?**

Only Gates I and Gates II were completed. The valid positive rate decreased between Gate I and Gate II. The overall valid positive rate was unable to be obtained due to Gate III not being completed.
Question 4: How many of the valid positive children were identified by the combination of the first two screeners?

The overall valid positive rate was unable to be obtained due to Gate III not being completed. The valid positive rate for Gate I was 82.1% yet 18.1% for Gate II.

Question 5: What are the error rates of the false positives and false negatives?

The error rates of the false positives and false negatives were unable to be completed due to Gate III not being conducted.

Results of Analysis

The Galileo results of all children who were referred for an evaluation based upon their Denver II developmental screening was analyzed. This analysis was conducted to ascertain whether the multiple gating program was effective in correctly identifying those children who were actually at-risk for a developmental delay. One child’s information was not included because there was no Galileo information available on this student. The Galileo was chosen because it was a naturally occurring criterion-referenced assessment conducted within the classroom. Many states use standard deviations (SD) as a way to quantify a delay and 1.5 SD or 2.0 SD below the mean are accepted as the norm (Lerner, Lowenthal, & Egan, 2003). The cut-off score used for this study was 1.5 standard deviations below the mean.

The Denver II and Galileo results were taken from a convenient sample of 685 children from 23 different sites. This sample represents 36% of the total sample of 1888 children. Because the 29 children who originally fell in the refer range on the Denver II were all from these 23 sites, the percentage of students who passed through Gate I is overrepresented in this sample. In order to address this over representation, the data from
these results were extrapolated to represent the entire sample of 1888. The extrapolation process went as follows. First, the 685 children were divided up into their specified groups: 541 valid positives, 24 false positives, 144 false negatives, and 5 valid negatives. Next, the 29 children were subtracted from 1888 because they were over represented in the sample. The number of false negatives, 144, were then divided by the total number of the convenient sample, 685 to obtain the percentage of false negatives. Then, this percentage, 21% was multiplied by 1859 (1888-29) to obtain the new false negative rate of 390. Finally, the total number of children, 1859 was subtracted from the new false negative rate of 390 to obtain 1469, or the false positive rate.

The degree of correlation between the independent and dependent measures were to be computed by a point-biserial correlation for the official developmental status criteria and a product-moment correlation was to be computed for the rest of the variables. A point-biseral correlation was chosen due to the official developmental status being a dichotomous variable. Certain correlation analysis could not be computed because of the instruments use of different categorical criteria. A phi-coefficient was computed for the Denver II and Galileo and for the Galileo and ESI-R only (see Table 4.1: Intercorrelations between the independent variables and dependent variables).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Denver Results</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2. ESI-R Results</td>
<td>-</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3. CPS-Screening Results</td>
<td>-</td>
<td>-</td>
<td>1.0</td>
<td>-</td>
</tr>
<tr>
<td>4. Galileo Results</td>
<td>-.018</td>
<td>.007</td>
<td>-</td>
<td>1.0</td>
</tr>
</tbody>
</table>

*p< .05

Table 4.1: Intercorrelates between Independent and Dependent Variables
At Gate 1, 28 (.015%) of the 1888 children were identified on the basis of the Denver II developmental screening as at risk for a developmental delay. Table 4.2 shows the results of the first gate prediction. Of the 29 at-risk children, 24 were identified as at-risk with an additional 5 (17.2%) not at-risk (false positives). The other 390 of the 414 were misclassified as not at-risk, resulting in a 20.9% false negative rate. The first gate had an overall correct classification rate of 79%. Yet, the first stage of screening did not significantly differentiate between those children who were at-risk for a developmental delay from those children who were not, \( \chi^2 (1) = .241, p > .05; \phi (1) = -.018, p > .05 \).

<table>
<thead>
<tr>
<th>Predictive Status</th>
<th>At-Risk Group</th>
<th>Not At-Risk Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid Pos.</td>
<td>False Pos.</td>
</tr>
<tr>
<td>At-Risk Group</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>Not At-Risk Group</td>
<td>False Neg.</td>
<td>Valid Neg.</td>
</tr>
<tr>
<td></td>
<td>390</td>
<td>1469</td>
</tr>
</tbody>
</table>

Number in Risk Group 28
Valid Positive Rate = \( A/A + B \) 82.1%
False Pos. Rate = \( B/A + B \) 17.8%
False Neg. Rate = \( C/C + D \) 20.9%
Total Observed Correct = \( A + D/N \) 79%
Chi-Square .241
Phi-coefficient -.018

*\( p < .05 \)

Table 4.2: Prediction Table and Summary Statistics for Gate I: Screening with the Denver II (\( N = 1888 \))
The 28 children classified as at-risk during Gate I were then screened on the ESI-R by the CDC mental health consultant. The results were a reduction in the number of children identified as at-risk from 28 to 11 (selection ratio = 39%). The prediction table for Gate 2 is presented in Table 4.3. There was a decrease in the valid positive rate to 18.1% of the at-risk group, while an increase occurred in the false positive rate to 81.8%. A decrease occurred in the false negative rate to 17.6%. Both the phi-coefficient, $(1) = .007, p > .05$, and the $\chi^2 = (1) .001, p > .05$ were not statistically significant and indicated that Gate 2 did not increase the accuracy in classifying children who were at-risk for a developmental delay.

<table>
<thead>
<tr>
<th>Actual Academic Rating</th>
<th>At-Risk</th>
<th>Not At-Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At-Risk Group</td>
<td>Valid Pos. 2</td>
<td>False Pos. 9</td>
</tr>
<tr>
<td>Not At-Risk Group</td>
<td>False Neg. 3</td>
<td>Valid Neg. 14</td>
</tr>
<tr>
<td>Number in Risk Group</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Valid Positive Rate</td>
<td>18.1%</td>
<td></td>
</tr>
<tr>
<td>False Pos. Rate</td>
<td>81.8%</td>
<td></td>
</tr>
<tr>
<td>False Neg. Rate</td>
<td>17.6%</td>
<td></td>
</tr>
<tr>
<td>Total Observed Correct</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td>Chi-Square</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Phi-coefficient</td>
<td>.007</td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$

Table 4.3: Prediction Table and Summary Statistics for Gate II: Screening with the ESI-R (N= 28)
A Prediction Table and Summary Statistics for Gate III: CPS Screening (N=14) was unable to be conducted due to the limited number of students who actually completed Gate 3. Therefore, a table is presented which describes the final results for those who were referred to CPS based upon their ESI-R results (see Table 4.5: Results of children referred to CPS).

<table>
<thead>
<tr>
<th>Student Number</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Referred for MFE</td>
</tr>
<tr>
<td>2</td>
<td>Not Referred for MFE</td>
</tr>
<tr>
<td>3</td>
<td>No Record sent from CDC</td>
</tr>
<tr>
<td>4</td>
<td>No Record sent from CDC</td>
</tr>
<tr>
<td>5</td>
<td>CDC cancelled and never rescheduled</td>
</tr>
<tr>
<td>6</td>
<td>Child from out of district not referred</td>
</tr>
<tr>
<td>7</td>
<td>Child from out of district not referred</td>
</tr>
<tr>
<td>8</td>
<td>CDC cancelled and never rescheduled</td>
</tr>
<tr>
<td>9</td>
<td>Not referred for MFE</td>
</tr>
<tr>
<td>10</td>
<td>No record sent from CDC</td>
</tr>
<tr>
<td>11</td>
<td>Referred for MFE</td>
</tr>
<tr>
<td>12</td>
<td>Qualified for Special Needs Pre-school</td>
</tr>
<tr>
<td>13</td>
<td>Not referred for MFE</td>
</tr>
<tr>
<td>14</td>
<td>CDC cancelled and never rescheduled</td>
</tr>
</tbody>
</table>

Table 4.4: Results of those children referred to CPS based upon ESI-R tests scores.
CHAPTER 5

DISCUSSION

Summary

The purpose of this study was to evaluate the effectiveness of a multiple gating program in identifying preschool-age children enrolled in the CDC Head Start of Franklin County who were at-risk for a developmental delay. The study used archival data collected during the 2002-2003 school year from the CDC Head Start and from Columbus Public Schools (CPS). Upon analysis of the data, the study was unable to draw any conclusions due to a lack of correlation between the criterion measure and the two developmental screening instruments. One goal of this study was to improve the identification of preschool children who are at-risk for a developmental delay by identifying possible predictors. Yet, this lack of correlation indicates that any conclusions would be speculative at best.

Multiple gating is an assessment method that has been shown to be helpful within the behavioral and criminology literature yet less is known within the area of academics. The CDC Head Start has used a multiple gating technique in an attempt to locate and identify as many children as possible who were at-risk for a developmental delay by initially screening a large group of children. All students included in this study were screened with the Denver II developmental screening instrument, an instrument that is known for having a high false positive rate. Having a high false positive rate is
beneficial when attempting to locate a low-base rate problem, such as a developmental delay in pre-school children. The students who scored in the suspect range on the Denver II were referred on for a more intensive evaluation, which was conducted by the CDC mental health consultants. This second gate is more intensive in nature and involves a one-on-one assessment with the child in a private room. This second gate identifies the child as either needing a further evaluation or as being developmentally appropriate and therefore ending the evaluation process. The third gate involves a screening completed by CPS on those children who were referred by the CDC based upon their ESI-R scores. Although an analysis was completed, the results are inconclusive due to the large amount of measurement error found during the Galileo and Denver II screenings.

Approximately 5-7% of the preschool population is actually at risk for a developmental delay, the importance of finding these children is reiterated within the early intervention literature (Meisels & Atkins-Burnett, 2000). During the 2000-2001 school year, 13% of all children enrolled in the national head start program had a disability (i.e., emotional disturbance, language and speech impairments, learning disabilities, mental retardation, visual handicaps, health impairments, hearing impairments, and orthopedic handicaps) (U.S. Department of Health and Human Services, 2002). Of the 1888 children screened with the Denver II developmental screening instrument by the CDC Head Start of Franklin County, a total of 112 were referred for either a developmental assessment or a speech language assessment. Of these 112, 29 were referred for a developmental assessment as defined by the CDC Head Start. The 112 initially referred for a further evaluation do fall within the 5-7% yet, it does not correspond with the national head start average of 13% and without an appropriate
criterion measure, it is difficult to know whether or not those children who are being identified as developmentally appropriate are truly not at risk for a delay.

Reasons for poor results during the multiple gating program

The intercorrelates between the independent and dependent variables were ascertained by computing a phi coefficient (see Table 4.1: Intercorrelates between Independent and Dependent Variables). The Denver II and Galileo had a correlation of -.18. This was not significant and indicates that the Denver II results and the Galileo scores did not correlate at all. The ESI-R and Galileo also did not correlate. The correlation between these two was .007, which was not significant. Those students who were identified as developmentally appropriate were not distinguishable on their Galileo results when compared to those students who were identified as needing further evaluation. This means that no conclusions can be drawn regarding the effectiveness due to the lack of an appropriate criterion measurement. Although the CDC Head Start has implemented a multiple gating program the effectiveness is unable to be evaluated due to the lack of an appropriate criterion measure, as a result of the large amount of measurement error.

Galileo

The Galileo was chosen as a criterion measure because it is a naturally occurring assessment that is implemented by the classroom teacher at-least twice a year. The Galileo is a type of dynamic assessment. It is used for purposes of identifying an area that a student needs remediation and in the creation of corresponding academic interventions. The Galileo was chosen as the criterion measurement in terms of correctly identifying the number of false negatives, false positives, valid positives, and valid
negatives. Since the Galileo is a naturally occurring assessment, the results may not be affected in the same way that norm-referenced assessments are when screening preschool children. For example, a dynamic assessment involves observing the child within the natural environment and therefore does not require meeting with the child one-on-one to administer an assessment. This may alleviate some stress that the child may have from being removed from their classmates. Also, the examiner may witness the child performing a task within the natural environment that they may not do when working with an adult examiner.

An analysis of n= 713 children’s Galileo scores was conducted and the results indicated that the Galileo is a poor criterion reference measurement for evaluating the effectiveness of the CDC multiple gating method. There was no statistically significant difference on the Galileo between those who were referred for a further evaluation from those who were not referred. Although some of the children at Head Start were identified as at-risk for a developmental delay, while others were identified as developmentally appropriate, an instrument used to measure how well the children were performing within the classroom was unable to differentiate between these students. There are several reasons which may explain this.

First, although the Galileo is a standardized measurement, its implementation is not. This allows for a lot of variability in how the evaluation occurs and thus lends itself to possible examiner bias. Each child is observed within the classroom with the intent of the measurement to identify whether or not the child is capable of performing the criterion established by the Galileo instrument. This means that although the Galileo is closely aligned with the Head Start domains, how, when, and how long each child is
observed is left up to some interpretation. There are no pre-set criteria for what time of
day each child should be observed, how long he/she should be observed, and what
activity the child should be observed during. As a result, one examiner may report that a
child is capable of performing a task because they saw her do it once or because they
think that the child is capable of performing the task. The examiner does not come out
and ask a child a specific question or have her complete a specific task. The assessment
completely takes place in the natural environment. Again, a child may be able to perform
a certain task yet if it goes unnoticed then the examiner may report that the child is
unable to complete the problem and as a result the child’s score may not reflect his actual
ability level.

Second, while gathering data for this current analysis it was noticed that there
was a difference between sites in terms of whether or not each child was assessed on all
four of the state mandated components of the Galileo (i.e., Early Math, fine/gross motor,
language/literacy, social development). For example, one site only completed the
fine/gross motor domain on the Galileo for all of the students attending its site. There was
also a lot of variation in terms of the overall mean scores for the four domains when sites
were compared. One site reported that on the early math, language/literacy, and social
development domains, the average score was 1.8, 1.6, and 1.9 standard deviations below
the mean respectively. Another site reported that the average early math, fine/gross
motor, and language/literacy domains were 1.5, 1.79, and 1.78 above the mean. These
mean scores are significant when compared to both the national average and the Head
Start average. This may be the result of how the students were taught, implementation of
interventions, what each individual child brought in with her/him into the learning
environment, or a result of examiner bias. Without observing each examiner assessing with the Galileo, it is too difficult to determine why these results occurred. Finally, some of the student’s Galileo results were missing. For example, one student did not have any information regarding all four domains while another student at a different site only had one domain assessed (Early Math). Again, the reason for this occurring would be speculative.

A third concern with the CDC Head Starts usage of the Galileo is that the Galileo follows a dynamic assessment approach in that it evaluates how the child learns by engaging him/her in the learning process. The theory behind conducting a dynamic assessment is for the examiner to gather in-depth information regarding the learner and to create hypotheses about ineffective and effective interventions (Lidz, 1997). The inclusion of interventions is unique to dynamic assessment. Yet, whether or not individually tailored interventions are occurring was not documented and may differ from site to site. Again, this may be due to a non-standardized documentation practice at the CDC or due to individual sites choosing not to implement an intervention. According to Reschly and Grimes (1995) dynamic assessment procedures have limited standardization and it has not been proven that the interventions are effective. Another concern pertaining to dynamic assessment and thus the Galileo is the use of gain scores when evaluating progress. “The attempt to measure complex processes and change presents special challenges. The availability of pre- and post test scores offers the temptation to use gain scores, which have been shown to be unreliable and subject to regression effects” (Lidz, 1997, p. 286).
Denver II

The Denver II implementation was similar to the Galileo and hence the possibility of examiner bias must also be ruled out. Although the Denver II is a standardized assessment, there appears to be some variation in terms of how well the Denver II was given. At the CDC Head Start, the Denver II was given either by an educational specialist or by the classroom teacher. At one site where the Denver II was given by the classroom teachers, approximately four more children per classroom should have been referred to the CDC mental health consultants based upon the child’s Denver II results (Personal Communication, September 3, 2003). This indicates that there may be many more children at-risk for a developmental delay than initially indicated. Children were not referred even when the results of a developmental screener recommend for them to be referred. This poses the question of why were they not being referred.

The Denver II is a standardized screening instrument that may be given by the classroom teacher or by an educational specialist. The assessment is ideally to take place in a secluded area away from other children yet this was not always the case. Children were often tested within the classroom setting while their peers were engaged in activities. This may have had a negative impact on their performance. Also, the CDC does not have a process of checking on sites to ensure that children are being referred based upon the Denver II scores. If this was the case, many more children may have been referred for a further evaluation. Although the Denver II is a statistically sound instrument, which tends to have a high rate of false positives, how it was used at the CDC Head Start of Franklin County during the 2002-2003 school year indicated that children
may have be referred based more upon teacher in put rather than upon their scores on a developmental screening measurement.

*Examiner Bias and Instrumentation Threat*

The results of this current evaluation were greatly impacted by how the data were collected. This study used archival data. The analysis has concluded that there was a lot of variation in terms of how the Galileo and Denver II were given to students. This variation occurred across sites and was found throughout the study. Some researchers have indicated that examiners do differ on how they score responses even when the responses are the same or similar (Sattler, 1988). Examiner bias and instrumentation threat impacts internal validity, which means that the results of the Galileo and Denver II developmental screenings may be due to how the tests were given and/or examiner interpretation not based on the performance of the child. Therefore, no conclusions can be made regarding the results of each student’s Galileo and Denver II scores due to other potential variables that can not be currently ruled out.

*Effectiveness of the CDC Multiple Gating Program*

Since many of the preschool screening instruments have technical inadequacies, an argument has been made that norm-based, standardized assessment methods be replaced (Bagnato & Neisworth, 1994). Multiple gating is a technique that allows for the continuation of standardized assessment while combining it with more ecological based assessments. In this current evaluation, multiple gating by itself can not be evaluated for its effectiveness without an appropriate criterion measure. The measurements are important when implementing a multiple gating technique. Although the separate measurements chosen by the CDC are all valid instruments, in terms of
assessing preschool-age children who are at-risk for a developmental delay, these measurements do not appear to correlate with the criterion measure.

How an assessment is chosen and why may impact its usefulness. The Galileo is a state mandated criterion assessment that the CDC is required to use. It is unknown whether or not the State is aware that the Galileo does not correspond with the developmental screening instrument currently used by the CDC. The CDC did state, however that they believed the Galileo was a good criterion measure and that it correlated nicely with the Denver II (personal communication, June 4, 2003). Choosing an instrument that is meant to measure how well students are doing in the classroom is one that should be carefully done because one purpose for using a dynamic assessment to measure students performance is to create interventions for those areas that are in need of remediation. Without the proper instruments, the interventions may not target all of the right areas. The Galileo implementation is not standardized thus lending itself to possible examiner bias.

The CDC is required to assess all of its students with the Galileo at least two times a year with the Early Math, Language/Literacy, Fine/Gross motor, and Social Development domains. The purpose of the Galileo is to link assessment to instruction by evaluating each child using multiple means. The child’s development is assessed by the classroom teacher through direct observation during instructional activities and by conducting a one-on-one evaluation. This information is then used to provide corresponding learning opportunities to support development. The Galileo is aligned with the Head Start Framework. The Galileo is a measure that allows for the creation of individualized instruction yet as an assessment measure, it did not differentiate between
those CDC Head Start students who were in need of further evaluation from those children who were deemed as developmentally appropriate.

The purpose of assessment is to provide information regarding children to be used by teachers in creating interventions. “Ultimately, the validity of assessment is determined in terms of its applications” (Meisels & Atkins-Burnett, 2000, p. 233). The Galileo is an instrument whose purpose is to individually tailor each child’s curriculum to match those Galileo domains that need remediation. As a criterion measure the way the Galileo is currently implemented does not appear to be appropriate because it does not distinguish between those children who are identified as being at-risk for a developmental delay from those children who have been identified as developmentally appropriate.

Early intervention is essential for addressing the needs of those children who are at-risk for a developmental delay. It is imperative that there is a system set-up to identify those who are in need (valid positives) while simultaneously ensuring that we do not miss any students (false negative). Multiple gating is a technique that can help decrease the number of false positives while increasing the number of valid positives yet, there must be a way to assess that the number of false negatives are kept to an absolute minimum or at best, does not exist. Wolery and Bailey (2002) stated in their testimony to The President’s Commission on Excellence in Special Education for the “identification of assessment methods that prevent disabilities and identify needs early and accurately” (p.2). Although only a few of the CDC Head Start children may have a developmental delay, early detection is important. Finding as many children as possible is a goal of the CDC Head Start and they have even implemented a three-tiered system to achieve this goal. Yet, there is currently no measure to test the effectiveness of this system. “A field
with an over reliance on measures that are insensitive or invalid for many of its participants cannot ascertain precisely when interventions are working (Carta, 2002, p. 103).

*Program Evaluation and Accountability*

Accountability and program evaluation within education has received a lot of attention over the years (Bailey, 2001). Program evaluation is conducted to ensure that a particular practice is effective (Bailey). The question that was being asked during this current evaluation was how effective is the multiple gating program implemented by the CDC Head Start of Franklin county in locating those children at-risk for a developmental delay in the areas of self-help, socialization skills, gross motor, fine motor, or cognition. In addition, there was an attempt to answer the following questions:

1. Do the number of false positives decrease while the number of gates increase?
2. Does multiple gating improve on the rate of valid positives?
3. How many of the valid positive children were initially identified by the first screener?
4. How many of the valid positive children were identified by the combination of the first two screeners?
5. What are the error rates of the false positives and false negatives?

Although the questions can be answered, due to the lack of an appropriate criterion measure, the answers may not be meaningful. Though the CDC has followed the Federal guidelines in attempting to locate, evaluate, and serve those students who qualify for Part B of IDEA (PL 105-17, 1997) there currently is no way of knowing or
identifying how many children are not being identified when they should be (false negatives) with the current measurements that are being used.

**Recommendations**

First, it is recommended that the CDC Head Start of Franklin County implement a more standardized procedure for conducting criterion referenced assessments. Teaching training is very important and will have an impact on how well the teachers assess children, especially on a dynamic assessment. The CDC Head Start can not require teachers to correctly assess children without appropriately training them on how to conduct the assessment.

For the CDC to continue using the Galileo each teacher should be required to assess in the same way. This may entail the creation of a work sheet that involves sitting down with each child individually and assessing him/her on the current curriculum. With the passage of the No Child Left Behind act, early assessment of student’s performance in pre-school academic areas will become a lot more common. This act is requiring schools and educational programs to be accountable for educating children and it has recently begun to evaluate the national Head Start program. Having an instrument that requires all examiners to test and observe children for the same amount of time during the same or similar activity may result in less examiner bias. How the Galileo is implemented can impact the scores.

Third, the CDC may choose another criterion referenced assessment besides the Galileo. Again, the effectiveness of the CDC Head Start’s multiple gating program was unable to be evaluated possibly due to how the Galileo evaluation of children was implemented. When conducting an evaluation, the results may be skewed due to
examiner bias. In the case of the CDC Head Start implementation of the Galileo, the results did not differentiate between those were identified as developmentally appropriate from those who were identified as at-risk for a developmental delay. Also, the sites varied in terms of the children’s mean scores on the four domains, whether or not each domain was evaluated, and some data were missing for some of the children.

These issues lead to a fourth recommendation, which is that it is recommended for the CDC to have some type of method for ensuring that each site is completing all of the Galileo assessments and that they are being completed correctly. Agencies are ultimately accountable for whether or not its sites are completing what they are required to complete. In the case of the CDC, some of the sites were not completing the requirements as mandated by the state of Ohio. Of course, how the assessments are being implemented is very important and is also one that must be monitored to ensure that each site is assessing all of the kids appropriately. This may be a somewhat difficult task due to the large number of children who attend the CDC Head Start. Yet, this is important because without the proper measurement, the agency is ultimately responsible for making sure that each child who is at-risk for a developmental delay is identified. The Galileo is the assessment chosen by the CDC and therefore should be one that is implemented correctly.

Fifth, the CDC Head Start of Franklin County should evaluate each site to ensure that all children who score within the refer range on the Denver II are being referred for an additional evaluation.

Finally, it is recommended that the education specialists conduct all of the Denver II developmental screenings. Educational specialists are employees of the CDC
and they have been extensively trained on how to correctly administer the Denver II. Having educational specialists conduct all of the Denver II developmental screenings may decrease the amount of examiner bias. If teachers are to continue to conduct the Denver II screenings, then they must receive extensive training, including role-playing and, sites must be accountable for referring children in a timely manner.

Limitations

This study used archival data collected by numerous employees from the CDC head start of Franklin County during the 2002-2003 school year. Therefore, the results of this study may be limited only to other agencies with similar characteristics as the CDC Head Start. Another limitation pertains to the use of a convenient sample. The demographics were not readily obtained for this convenient sample and therefore, the results may be limited to only those children and sites included in this current study.

Directions for Future Research

The effectiveness of a multiple gating program was unable to be conducted. The reasons for this may be many. One recommendation given was for an increase in program evaluation and accountability. An area of future research would be to evaluate how the CDC Head Start evaluates each site and in what ways it holds the sites accountable for serving children. The results in this study indicate that program evaluation is necessary in each site, especially when it comes to identifying children who are at-risk for a developmental delay.

Another possibility for future research pertains to standardizing the implementation of the Galileo for the CDC Head Start sites. This may entail the creation of a work sheet for each teacher to follow. This work sheet could be implemented and
observed and modified where needed. The standardization of a dynamic assessment does not occur often yet in the case of the CDC Head Start this may be necessary.

Finally, the teachers may be observed in how they implement both the Galileo and Denver II for purposes of obtaining inter-observer agreement. Observing to see whether or not examiners are scoring children in a similar manner will help with internal validity, resulting in stronger more valid multiple gating method.

Conclusion

Although the CDC Head Start of Franklin County has chosen the Galileo instrument due to its correlation with the Denver II developmental scales (personal communication, June 4, 2003), upon analysis though, there appears to be very little correlation between the Denver II and the Galileo. The *Galileo Alignment with the Head Start Framework* (2002) indicates that the Galileo corresponds with academic areas addressed by the CDC Head Start, yet the Galileo does not correlate with the majority of the questions asked on the developmental screening instrument used by the CDC, the Denver II. For example, the Language component of the Denver II only correlates with two Early Math questions, and one Literacy/language question on the Galileo (Child Development Council of Franklin County, 2002). A potential consequence is that if a child scores within the Ok range on the Denver II, then there is no other measure which to compare the results lending itself to the possibility of false negatives. This lack of correlation between the Denver II developmental screening instrument and the Galileo is reflected in the performances of the student’s on these two different instruments. There appears to be no pattern between those students who were referred for further evaluation based upon their screening results compared to those students who scored within the ok
range. This lack of differentiation between the two groups of students indicates that no definitive conclusions can be drawn regarding the effectiveness of a multiple gating program currently being used by the CDC Head Start.

Multiple gating may be an effective methodology yet the tests utilized must be statistically sound instruments and implemented correctly. Although the CDC uses a multiple gating program, the Galileo may not be implemented the same way for each child and thus lend itself to possible examiner bias. Therefore the results of this analysis indicate that the CDC multiple gating program can not be currently evaluated for its effectiveness due to the lack of a criterion referenced measurement.

Teacher training must address how to appropriately conduct assessments and how to report the results. The quality of teacher training can have an impact on how well they conduct their job. If the CDC requires teachers to conduct assessments then the CDC must ensure that teachers are trained appropriately and verify the results or the results may be meaningless. It is recommended that the CDC Head Start of Franklin County conduct periodic checks on each site to ensure that all children who are at-risk for a developmental delay are being referred on for further evaluation. Early intervention is imperative and must begin as soon as possible.
REFERENCES


Galileo alignment with head start. (2002). The Child Development Council of Franklin County. Columbus, OH: Authors.


*Individuals with Disabilities Education Act Amendments of 1997 (PL 105-17).* 20 USC Chapter 33, Sections 1400 et seq.


