ED 213 725 TM 810 943

AUTHOR Apling, Richard; Bryk, Anthony
TITLE Policy Paper: The Predictive Validity of Early Childhood Variables.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY Department of Education, Washington, D.C.
PUB DATE 80
NOTE 84p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Admission Criteria; Disadvantaged Youth; *Early Childhood Education; Educational Diagnosis; *Educationally Disadvantaged; *Federal Programs; *Predictive Validity; *Predictor Variables; Program Evaluation
IDENTIFIERS *Elementary Secondary Education Act Title I

ABSTRACT

Some early childhood variables are examined to evaluate their predictive validity. The selection of children needing early childhood Title I services is complicated by the lack of criteria for defining who is educationally disadvantaged and the special problems of early childhood testing and measurement. The study used re-analysis of longitudinal data on children in Head Start Planned Variation and Follow Through programs. The second approach used meta-analysis to synthesize results of studies that examined relationships between early childhood predictors and later outcomes. The strengths and weaknesses of these approaches complemented each other. Methods of selection and their predictive validity were the main focus of the paper. Another factor to be considered included costs of selection procedure. Special problems exist in assessing young children because tests for this age group are often of lower technical quality. Preschool children often lack the physical, intellectual and emotional prerequisites necessary for systematic assessment. Selection bias may result from the use of tests or variables which have different predictive validity for different groups. The importance of prediction stems from the goal of most ECT-I programs: the prevention of educational problems in later schooling. (DWH)

The Huron Institute

Policy Paper: The Predictive Validity of Early Childhood Variables

by

Richard Apling and Anthony S. Bryk

THE HURON INSTITUTE 123 MOUNT AUBURN STREET, CAMBRIDGE, MASSACHUSETTS 02138

Policy Paper: The Predictive Validity of Early Childhood Variables

By

Richard Apling
&
Anthony Bryk

The Huron Institute
123 Mt. Auburn Street
Cambridge, Mass. 02138

Fall 1980

Draft Not for Citation

FOREWORD

This policy paper has been prepared as part of a United States Education Department (USED) sponsored project on the evaluation of early childhood Title I (ECT-I) programs. Unlike the reports and resource books which are other products of this endeavor, this paper is intended for a limited audience, namely, USED staff concerned with ECT-I programs and the evaluation of those programs. It is not intended as a practical guide to states and local school districts on how to improve their ECT-I selection procedures. In fact, the paper deals only with some technical issues surrounding the selection of ECT-I children.

Deciding who receives ECT-I services is a complex multi-stage process that involves designating Title I attendance areas, identifying children in need of ECT-I services, and selecting those most in need for ECT-I program. This paper deals with the selection phase of the process by examining some early childhood variables that could be included in a selection strategy with regard to their predictive validity -- their accuracy in predicting later educational outcomes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

OVERVIEW OF THE PROBLEM

Because of our field work (Yurchak, Gelberg, & Darman, 1979; Yurchak & Bryk, 1980) and continuing conversations with USED staff, it became increasingly clear to us that the selection of children in need of ECT-I services presents special problems. These include the lack of criteria for defining who is educationally disadvantaged, disagreement on what constitutes disadvantage before school entry, and the special problems of early childhood testing and measurement. Despite these complications, the Huron descriptive study of ECT-I programs (Yurchak & Bryk, 1980) found that most Local Education Agencies (LEAs) are making a genuine attempt to fulfill not only the letter but also the intent of the law regarding ECT-I selection. The LEAs visited expressed "strong interest . . . in the need to find better ways to conduct . . . selection" (p. 6-15).

The Huron study found that school districts used a wide variety of indicators to select ECT-I children, including:

- A low score on a test or series of tests
- Teacher judgment
- A sibling who is or was a Title I student
- Parents with less than a high school education
- A child's inability to understand the language of instruction
- Parent judgment.

Although in almost every district tests were used in the ECT-I selection process, their importance varied enormously (Yurchak & Bryk, 1980). At one extreme, test scores were the sole determinant of who received ECT-I services. Less extreme was the practice of considering tests results together with teacher judgment.* At the other extreme, tests were given to comply with regulations but were not taken into account in selection decisions. In addition to the different ways in which tests were used, the Huron study found that many different tests were used in the districts studied. In all, we found 26 tests used for ECT-I selection in the 29 LEAs we visited. Only a few of these were used in more than one LEA.

The Huron study revealed widespread dissatisfaction with ECT-I selection practices. Some local and state staff were especially concerned about the inadequate quality of measures used to select children. Others expressed dismay about the inability to measure important attributes such as social and emotional development, task persistence, and the attention span of young children. Those interviewed generally agreed that ECT-I programs are aimed at the long-term goal of promoting general school competence in the early elementary grades, and thus must provide the necessary precursor skills. Unfortunately, however, there is little agreement on what those skills are; therefore there can be little agreement on what areas should be covered in an assessment battery.

The study reported here is an attempt to inform discussion of ECT-I selection procedures. Since a major goal of most ECT-I programs is to prevent problems from occurring when a child reaches elementary school, it follows that an adequate ECT-I selection procedure must be able to predict

* Such a combination of tests scores and teacher judgment was recommended in the evaluation of the Washington, D.C., Title I program (Stenner, Feifs, Gabriel, & Davis, 1976). The evaluators found "that a substantial number of eligible students are not being identified, . . . [and] a number of students not needing Title I services are, on the basis of faulty test scores, being placed in the Title I program" (p. 5). They therefore recommended that "the exclusive reliance on standardized tests should be discontinued in favor of a 'need index' computed from a weighted composite of teacher judgment and criterion-referenced test scores" (p. 7).

which children are most likely to experience later difficulty so that they may receive ECT-I services. An important criterion for assessing any ECT-I selection procedure is thus its predictive validity.

An ideal study comparing the predictive validity of possible ECT-I selection procedures would have several attributes. It would assess a large number of children at an early age using diverse predictors such as early childhood tests, socio-economic variables (for example, income and mothers' education), home characteristics such as how much parents read to their children, and teacher judgment. The study would follow these children until they reached early elementary school. They would then be assessed on general school competence in terms of school grades, achievement test scores, teacher judgment, attitudes toward school, and so forth. Alternative selection procedures, consisting of different combinations of these predictor variables, could then be compared for their relative predictive validity for later achievement test scores, future school grades, etc.

Unfortunately the ideal study for our purposes does not exist, nor is it likely to be done. Thus we have resorted to two imperfect but useful approaches. The first is a re-analysis of longitudinal data on children in Head Start Planned Variation and Follow Through programs, which approximate some characteristics of an ideal study. This reanalysis allows us to look at several combinations of variables for predicting later achievement. The data set has the advantage of including longitudinal data on a substantial number of children. It is limited, however, in not including potentially important variables such as teacher judgment and in having limited information on family characteristics.

The second approach uses meta-analysis to synthesize findings from studies that examine relationships between early childhood predictors and later outcomes. The meta-analysis combines a wider variety of predictor variables and outcomes; but, because these data come from scores of studies, it is impossible to examine different sets of predictors simultaneously. Thus the strengths and weaknesses of our two approaches complement each other.

SECONDARY ANALYSIS OF THE HSPV AND FT LONGITUDINAL DATA

The data on children in Head Start Planned Variation (HSPV) and Follow Through (FT) programs that we re-analyzed were originally assembled by Weisberg and Haney (1977) to evaluate the cumulative effects of these programs. Because this data set contains background variables, prekindergarten and kindergarten test scores, and later achievement test scores for several hundred children, it is useful for assessing the predictive power of multiple variables. In the remainder of this section we will describe this data set,* discuss how we analyzed the data, and report our results.

The HSPV/FT Data Set

The data on the two programs were merged to investigate "whether Follow Through helps maintain the benefits of Head Start in the early elementary grades; [and] the way in which Head Start experience of children may have confounded efforts in the national evaluation of Follow Through to calculate program effects" (Weisberg & Haney, 1977, p. i). As Weisberg and Haney point out, this data set is probably unique.

> . To our knowledge, these files represent the only data set with information on the experience and development of children from HS entry through the end of third grade. While it is in many respects painfully limited, it represents a unique source which required a considerable effort to create and may be of interest for purposes of secondary analysis. (P. 11)

* For a more comprehensive discussion of the data and of the original study, see Weisberg and Haney (1977).

Like many longitudinal data sets, the HSPV/FT data is "painfully limited" in several ways. For one, variables are inconsistent across groups: two cohorts of children were followed from prekindergarten through early elementary school,* but they received few tests in common. There are also inconsistencies within cohorts; for example, different versions of the Caldwell Preschool Inventory (PSI) were used by the two programs for cohort III. In addition, these are not data from random samples of children. As Weisberg and Haney (1977) point out, this is a special sample produced by a complex selection process:

> The flow of children into, through, and out of Head Start and Follow Through constitutes a vast and complex process. Children were selected for Head Start on the basis of general criteria applicable nationally, but local circumstances determined the specific make-up of program groups. Thus groups of Head Start children in different places vary widely on numerous dimensions. In Follow Through, too, the likelihood of participation depends on children's characteristics and local circumstances. Moreover, Head Start experience is one of the factors taken into account in the selection process. (p. 24)

As with all longitudinal studies, attrition creates problems with the data. Some children, although they remain in the sample throughout the study, inevitably are absent when some tests are given, and those data are lost. Similarly, other children leave the program, move to other schools, or for other reasons are unavailable for subsequent data collection. And children leave as they entered the study: in non-random patterns that make generalization to large groups difficult. As Table 1 shows, the usable samples were about half of the original cohorts.

* Figure 1 shows the years and seasons of the years when tests were administered to the two cohorts of children.

	1970-71		1971-72		1972-73		1973-74		1974-75	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Cohort	HSVP		K		1		2		3	
III[+]	*	*	*	*		*		*		*
Cohort			HSVP		K		1		2	
IV[+]			*	*		*		*		*

* Test administration times

[+] Follow Through cohorts. No Head Start data were available for children in Cohorts I and II; therefore these groups are excluded from the analysis.

Figure 1: Test Administration for Cohorts III and IV. (Adapted from Weisberg and Haney, 1977, p. 6).

Table 1: Sample Background Variables for HSPV/FT and NFT Children

	Cohort III	Cohort IV	Non Follow Through *
Sex	57% boys	52% boys	51.2% boys
Ethnicity	45% nonwhite	49% nonwhite	64% nonwhite
Average Family Income	$3700 (1970)	$3700 (1971)	$5900
Median Father's Education	Grade 10	Grade 10.3	-----
Median Mother's Education	Grade 10.4	Grade 10.5	Grade 11.6
Father's Occupational Status	12% unemployed	19% unemployed	-----
Family Receives Aid	50% yes	40% yes	-----
First Language	95% English	96% English	94% English
Sample Size	396	725	8676
Approximate Usable Sample Size	200	400	-----

* Data from Molitor, Watkins, and Napior, 1977, p. 12.

Finally, the children in the HSPV/FT sample are probably more disadvantaged than the pool of children from which ECT-I participants are chosen. Table 1 summarizes several background variables for the HSPV/FT children and the Non Follow Through (NFT) control group, which was made up mainly of children from Title I schools (Haney, 1977, pp. 165-166). Clearly the two samples differ significantly in minority enrollment, family income, and mother's education.

Despite these problems, the data are a unique resource for examining the predictive validity of ECT-I variables. They are valuable for our purposes because they follow children from preschool through second and third grade. For the subsample of children for whom complete records are available, we can easily examine the comparative predictive power of different groups of variables. In addition, children in the sample do not come from just one area but are drawn from 13 FT sites in 11 states (Minnesota, Utah, Washington, New Jersey, Nebraska, Delaware, Missouri, Illinois, Colorado, Florida, and Pennsylvania) that represent a geographical diversity.

Study Variables

Table 2 lists the independent and dependent variables included in the analysis of the HSPV/FT data set. Outcomes are total reading and total math scores of the Metropolitan Achievement Test (MAT). This test was given in the spring to the first, second, and third grades of Cohort III and to the first and second grades of Cohort IV. All outcomes are raw scores. The same background variables were collected for the two cohorts. The original data set had several other background measures, excluded here because of large numbers of missing cases or high correlations with other background

Table 2: Prediction and Outcome Variables for Cohorts III and IV*

Cohort III	Cohort IV
Background Variables	
Sex	Sex
Age	Age
Ethnicity	Ethnicity
Total Family Income	Total Family Income
Mother's Education	Mother's Education
Family Receives Aid	Family Receives Aid
Number in Household	Number in Household
First Language in Home	First Language in Home
Prekindergarten Tests	
PSI (Fall)	PPV (Fall)
PSI (Spring)	PPV (Spring)
NYU Booklet 3D (Fall)	PSI (Fall)
NYU Booklet 3D (Spring)	PSI (Spring)
NYU Booklet 4A (Fall)	WRAT Reading (Fall)
NYU Booklet 4A (Spring)	WRAT Math (Spring)
	WRAT Numbers (Fall)
	WRAT Numbers (Spring)
Kindergarten Tests	
PSI (Fall)	MAT Primer Reading (Spring)
WRAT Reading (Fall)	MAT Primer Numbers (Spring)
WRAT Reading (Spring)	
WRAT Spelling (Fall)	
WRAT Spelling (Spring)	
WRAT Math (Fall)	
WRAT Math (Spring)	
PPV (Fall)	
PPV (Spring)	
Lee-Clark Reading Readiness (Fall)	
MAT Primer Reading (Fall)	
MAT Primer Numbers (Fall)	
Outcome Variables	
MAT Primary I Total Reading (1st grade)	MAT Primary I Total Reading (1st gr)
MAT Primary I Total Math (1st gr)	MAT Primary I Total Math (1st gr)
MAT Primary II Total Reading (2nd gr)	MAT Primary II Total Reading (2nd gr)
MAT Primary II Total Math (2nd gr)	MAT Primary II Total Math (2nd gr)
MAT Elementary Total Reading (3rd gr)	
MAT Elementary Total Math (3rd gr)	

*MAT = Metropolitan Achievement Test
PSI = Caldwell Preschool Inventory
WRAT = Wide Range Achievement Test
PPV = Peabody Picture Vocabulary

variables.* As Table 2 shows, there is little similarity between prekindergarten and kindergarten tests in the two cohorts. Only the fall PSI is common to both among the prekindergarten measures. Subtests of the MAT Primer are included in both sets of kindergarten predictors, but were given in different seasons in the two cohorts. Although this variability in the two sets of predictors makes it difficult to compare the two cohorts, such comparisons, when feasible, bring additional information to our study.

Data Analysis Strategy

The central question for analyzing the HSPV/FT data is whether combinations of early childhood variables do better than single variables in predicting problems in later school experience. In examining this question we looked at three procedures for predicting later achievement:

- Using an individual test or subtest
- Using a set of tests or subtests
- Using a set of tests and background variables.

We examined each procedure in two ways. First, we used multiple regression to generate R^2s (the percentage of variation in outcome variables explained by individual variables and by sets of variables). Then we determined which individual test or subtest accounted for most variation in an outcome. Next we added other tests or subtests that contributed significantly to the prediction of later achievement scores. Finally we added a set of background variables to the set of tests. By measuring increments to the

* In Cohort III these include Present Family Income (high correlation with Family Income), Father's Education (199 missing cases), Father's Occupation (156 missing), Father's Employment Status (164 missing), Mother's Employment Status (164 missing). In Cohort IV, variables dropped were Father's Schooling (358 missing), Father's Employment Status (312 missing), and Second Language (675 missing).

R^2s as we added successive sets of variables, we could compare the predictive power of the three procedures.

In addition to examining increments to R^2, we also analyzed the number of misclassifications produced by procedures predicting third-grade reading scores. Misclassification results when a procedure predicts either that a child will experience educational disadvantage and he does not, or that a child will not experience disadvantage when in fact he does. Thus we have a second way to compare prediction procedures: what are the rates of misclassification that result from each? In this subsection, we will discuss the multiple regression analysis and report our findings; in the next, we will explain our analysis of error rates and examine those results.

In designing the multiple regression analysis of the HSPV/FT data, we decided to analyze cohorts III and IV separately because different early childhood tests are used with the two cohorts. We chose to do separate analyses for MAT reading scores and MAT math scores because later Title I programs are often aimed at ameliorating either reading or math problems. We also decided to analyze separately predictors measured in the fall and in the spring. This resembles ECT-I procedures in that a program might use either spring or fall data to select children.

Benchmark R^2s

To compare the predictive power of single tests and sets of variables, we first determined benchmark R^2s by seeing how well background variables alone predict reading and math scores in grades 1, 2, and 3, and by examining how well all available variables predicted the same outcomes. These two groups of R^2s provide reference points by which to judge how well other combinations of variables predict achievement test scores.

Table 3 shows the results of our analyses with background variables and with all variables for each outcome measure. For each outcome in the two cohorts, we entered all background variables listed in Table 2 as independent variables in a stepwise regression. We stopped at the step for which all variables entered with $F \geq 1.00$. This cutoff rule ensures that random variation was not added to the prediction equation. The first row for background variables in Table 3 reports that R^2 at the last step for which F was greater than or equal to one. Each R^2 (in this and other tables from the HSPV/FT analysis) is adjusted for sample size and for the number of variables in the prediction equation. (See Cohen & Cohen, 1975, pp. 106-107, for a discussion of adjusted R^2.)

We followed a similar procedure for examining the predictive power of all variables. The background variables, fall prekindergarten tests, and fall kindergarten tests listed in Table 2 were entered as independent variables in stepwise regressions. The outcome variables in Table 2 were the dependent variables. The same cutoff rule ($F \geq 1.00$) was used to decide when to stop adding variables. The analysis was then repeated using spring prekindergarten and kindergarten tests. The result can be seen in the bottom half of Table 3.

We conclude several things from Table 3. First, the overall R^2s with all variables in the equations are substantial. For fall predictions, R^2s range from 0.32 to 0.62, and for spring prediction, from 0.38 to 0.61. Apparently, a set of background variables and early childhood tests account for significant amounts of the variation in later scores.

Looking again at Table 3, we see that background variables account for some, but not a great deal, of later test variation. This is not surprising

Table 3: Adjusted R^2s for Background Variables Alone and for All Variables Predicting Later-Grade Test Scores

	Cohort III						Cohort IV			
	MAT Read. 1	MAT Math 1	MAT Read. 2	MAT Math 2	MAT Read. 3	MAT Math 3	MAT Read. 1	MAT Math 1	MAT Read. 2	MAT Math 2
Background Variables	.11	.17	.14	.13	.15	.12	.18	.11	.17	.12
n	202	200	169	166	136	129	473	469	415	411
All Variables*										
Fall	.44	.62	.37	.49	.36	.32	.39	.32	.40	.35
n	141	139	122	121	97	94	433	432	383	382
Spring	.49	.59	.39	.45	.38	.40	.61	.57	.54	.54
n	137	137	117	115	93	91	384	381	347	343

* Includes all prekindergarten and kindergarten tests and background variables that entered the prediction equations at F=1.00 or more.

since the background variables in the HSPV/FT data are fairly crude measures. Prediction might have been improved if we had also had measures of home environment and parent-child interaction. Truncation of these background variables may also explain their modest predictive power. The children in our sample were selected for HSPV and FT programs on the basis of socio-economic measures. Thus this sample is more uniform than children in general on variables such as family income and mother's education, and this contributes to lower correlations and lower R^2s.

Table 3 presents two extremes against which to compare prediction procedures. Using only simple background variables we explain roughly 14 percent of the variation in later scores. Using all the variables at our disposal, which we would not expect any ECT-I program to have available, between 35 and 45 percent is a reasonable expectation. Other combinations of predictor variables, considered below, fall between these two extremes.

Three Sets of Prediction Variables

We used procedures similar to our analyses of background variables and all variables to examine the predictive validity of three alternative sets of variables -- a test or subtest used alone, a group of tests or subtests added to the single instrument, and background variables added to the set of tests or subtests. We performed separate analyses on tests given in prekindergarten and in kindergarten. We separately analyzed tests given in the fall and in the spring. We also analyzed the outcomes separately -- first-, second-, and third-grade reading, and first-, second-, and third-grade math. Finally we analyzed data from the two cohorts separately.

For each combination of predictor test time (e.g., kindergarten, fall), outcome measure and time (e.g., first-grade reading), and cohort, we followed the same analytic procedures. First we entered all appropriate tests and subtests (e.g., fall kindergarten measures) as independent measures in a stepwise regression. We used the $F \geq 1.00$ rule to determine the best set of tests or subtests. We then performed several regressions, in turn entering each test or subtest from the best set first. The test or subtest that produced the highest R^2 was designated as the best individual measure. We then added the remaining tests or subtests to the best measure. Finally we added all background variables stepwise after entering the best set of tests into the prediction equation. Once again, we stopped adding background variables just before F dropped below 1.00.

Tables 4, 5, 6, and 7 contain the results from these analyses. Tables 4 and 5 show results for each childhood test administered to prekindergarteners. Tables 6 and 7 contain test results for kindergarten children. Tables 4 and 6 are taken from cohort III; 5 and 7 are from cohort IV. Each table is read in the same way. The first row of numbers presents the adjusted R^2s for the single prekindergarten or kindergarten test that best predicts the outcome shown at the top of each column. The next row shows the R^2s and increments to R^2 when other prekindergarten or kindergarten tests are added to the prediction equation. The final row shows R^2s and increments when background variables are added to the best set of tests or subtests.

There are several patterns in Tables 4 through 7 that are partially masked because of the amount of data presented. To help clarify these patterns we have calculated median R^2s. These medians are presented in

Table 4: Predicting First-, Second-, and Third-Grade Reading and Math Scores From Prekindergarten Tests and Background Variables (Cohort III)

Outcome Tests

	MAT 1 Reading	MAT 1 Reading	MAT 1 Math	MAT 1 Math	MAT 2 Reading	MAT 2 Reading	MAT 2 Math	MAT 2 Math	MAT 3 Reading	MAT 3 Reading	MAT 3 Math	MAT 3 Math
					Best Test or Subtest							
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Test	PSI	NYU Booklet 4A	PSI	NYU Booklet 4A	PSI	NYU Booklet 4A	PSI	PSI	PSI	NYU Booklet 4A	PSI	NYU Booklet 4A
Adj. R^2	.15	.25	.25	.35	.13	.15	.14	.19	.14	.19	.09	.17
n	144	144	142	142	124	126	123	126	98	101	95	97
					Best Set of Tests or Subtests							
Tests	PSI	PSI NYU 4A NYU 3D	PSI	PSI NYU 4A NYU 3D	PSI	PSI NYU 4A NYU 3D	PSI	PSI NYU 4A NYU 3D	PSI	NYU 4A	PSI	PSI NYU 4A NYU 3D
Adj. R^2	.15	.26	.25	.38	.13	.16	.14	.23	.14	.19	.09	.18
Inc. R^2	--	.01	--	.03	--	.01	--	.04	--	--	--	.01
n	144	144	142	142	124	126	123	125	98	101	95	97
					Background Variables Added to Best Tests							
Variables Added	Sex Ethnicity Mom Ed Age Income Fam. Size	Sex Income Fam. Size Mom Ed. Age	Ethnicity Receives Aid Mom Ed. 1st Lang.	Ethnicity Mom Ed. Income 1st Lang.	Sex Ethnicity Receives Aid Fam. Size Income Age	Sex Ethnicity 1st Lang. Income Fam. Size Receives Aid	Ethnicity Receives Aid Mom Ed. Mom Occ.	Ethnicity Mom Ed. Receives Aid 1st Lang.	Income Fam. Size Age Receives Aid Ethnicity Mom Ed Mom Occ.	Income Fam. Size Mom Occ. Age Mom Ed.	Rec. Aid Ethnicity Mom Ed.	Rec. Mom Ed Ethnic. Mom Occ
Adj. R^2	.20	.29	.37	.44	.21	.21	.24	.29	.21	.24	.17	.23
Inc. R^2	.05	.03	.12	.06	.08	.05	.10	.06	.07	.05	.06	.05
n	144	144	142	142	124	126	123		98	101	95	97

Tab[redacted]5: [redacted]dic[redacted] Fi[redacted]- an[redacted]econ[redacted]rade [redacted]eading[redacted]and[redacted]ath Sc[redacted]es [redacted]
Prekindergarten Tests and Background Variables (Cohort IV)

Outcome Tests

	MAT Reading 1st Grade	MAT Reading 1st Grade	MAT Math 1st Grade	MAT Math 1st Grade	MAT Reading 2nd Grade	MAT Reading 2nd Grade	MAT Math 2nd Grade PK Fall	MAT Math 2nd Grade PK Spring
				Best Test or Subtest				
	PK Fall	PK Spring	PK Fall	PK Spring	PK Fall	PK Spring	PK Fall	PK Spring
Test	WRAT Reading	WRAT Reading	PSI	PSI	PSI	WRAT Reading	PSI	WRAT Reading
Adj.$\underline{R}^2$	.28	.49	.26	.36	.28	.42	.27	.35
$\underline{n}$	451	423	432	407	383	376	382	374
				Best Set of Tests and Subtests				
Tests	PSI Peabody WRAT Read WRAT Num.	PSI WRAT Read WRAT Num.	Peabody PSI WRAT Read	PSI WRAT Read WRAT Num.	Peabody PSI WRAT Read WRAT Num.	PSI WRAT Read WRAT Num.	PSI WRAT Read WRAT Num.	Peabody PSI WRAT Read WRAT Num.
Adj.$\underline{R}^2$	.35	.53	.31	.44	.36	.48	.32	.43
Inc.$\underline{R}^2$	.07	.04	.05	.08	.08	.06	.05	.06
$\underline{n}$	433	408	432	407	383	360	382	348
				Background Variables Added to Best Set of Tests				
Variables Added	Rec. Aid Sex 1st Lang. Income Fam. Size	Sex Receives Aid Ethnicity Mom Ed. Age Income 1st Lang.	Income Ethnicity 1st Lang.	Ethnicity Income	1st Lang. Income Family Size Age Sex Receives Aid	Mom Ed. Age Mom Occ. Sex Income Family Size Ethnicity	Income 1st Lang. Sex	Income Ethnicity Mom Occ.
Adj.$\underline{R}^2$	.39	.56	.32	.46	.40	.51	.35	.45
Inc.$\underline{R}^2$	.02	.03	.01	.02	.04	.03	.03	.02
$\underline{n}$	433	408	432	407	383	360	382	348

Table 6: Predicting First-, Second-, and Third-Grade Reading and Math Scores From Kindergarten Tests and Background Variables (Cohort III)

Outcome Tests

	MAT 1 Reading	MAT 1 Reading	MAT 1 Math	MAT 1 Math	MAT 2 Reading	MAT 2 Reading	MAT 2 Math	MAT 2 Math	MAT 3 Reading	MAT 3 Reading	MAT 3 Math	MAT 3 Math
						Best Test	or Subtest					
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Test	WRAT Reading	WRAT Reading	MAT Primer Numbers	WRAT Math	WRAT Reading	WRAT Reading	MAT Primer Numbers	WRAT Math	WRAT Reading	WRAT Reading	MAT Primer Numbers	WRAT Spelling
Adj. R^2	.33	.39	.45	.46	.23	.28	.32	.30	.22	.29	.25	.28
n	202	189	197	188	168	157	165	154	136	131	129	127
						Best Set of Tests or Subtests						
Tests	PSI WRAT Read WRAT Spell Peabody Lee-Clark MAT Read MAT Numbers	-- WRAT Read WRAT Sp WRAT Math -- -- --	-- WRAT Read WRAT Sp WRAT Math Lee-Clark MAT Read MAT N	-- WRAT Read WRAT Sp WRAT Math -- -- --	-- WRAT Read -- -- Peabody MAT Read MAT N	-- WRAT R WRAT Sp WRAT M Peabody -- --	PSI WRAT R -- Lee-Clark Peabody MAT R MAT N	WRAT R WRAT Sp WRAT M Peabody -- --	WRAT R -- -- Peabody MAT R MAT N	WRAT R WRAT Sp WRAT M -- -- --	WRAT R -- -- Lee-Clark MAT R MAT N	WRAT R WRAT Sp WRAT M -- -- --
Adj. R^2	.41	.46	.56	.54	.32	.36	.41	.40	.32	.35	.30	.39
Inc. $\bar{R}^2$	.08	.07	.09	.08	.09	.08	.09	.10	.10	.06	.05	.11
n	199	189	197	188	168	149	165	146	136	131	129	127
						Background Variables Added to Best Tests						
Variables Added	Sex Income Age Receives Aid 1st Lang. Ethnicity Mom Occ. Mom Ed.	Mom Ed. Receives Aid Income Age 1st Lang. Ethnicity Sex	Ethnicity Income Mom Occ. Mom Ed. Fam. Size	Ethnicity Mom Ed. Income Mom Occ.	Sex Income Fam. Size 1st Lang. Age	1st Lang. Sex Fam. Size Income Age	Ethnicity Mom Ed. 1st Lang. Mom Occ. Age	Ethnicity Mom Ed. Mom Occ. Sex 1st Lang.	Income Age Fam. Size Mom Occ. Mom Ed. Sex	Income Fam. Size Mom Occ. Age Mom Ed.	Income Mom Occ. Ethnicity Receives Aid	Mom Occ. Sex Receives Aid Mom Ed. Ethnicity
Adj. R^2	.45	.49	.62	.59	.38	.39	.47	.45	.38	.39	.33	.41
Inc. $\bar{R}^2$	.04	.03	.06	.05	.06	.03	.06	.05	.06	.04	.03	.02
n	199	189	197	188	168	149	165	146	136	131	129	127

Table 7: Predicting First- and Second-Grade Reading and Math Scores From Kindergarten Tests and Background Variables (Cohort IV)

Outcome Tests

	MAT Reading 1st Grade	MAT Math 1st Grade	MAT Reading 2nd Grade	MAT Mat 2nd Grade
Best Subtest				
	K Spring	K Spring	K Spring	K Spring
Subtest	MAT Primer Reading	MAT Primer Numbers	MAT Primer Reading	MAT Primer Math
Adj.R^2	.44	.49	.38	.45
n	437	424	391	377
Best Set of Subtests				
Subtest	MAT Primer Reading Math	MAT Primer Math Reading	MAT Primer Reading Math	MAT Primer Math Reading
Adj.R^2	.50	.52	.41	.49
Inc.R^2	.06	.03	.03	.04
n	424	424	381	377
Background Variables Added to Best Tests				
Variables	Receives Aid Family Size Sex Mom Ed. 1st Language	Mom Ed. Ethnicity	Mom Ed. Received Aid Family Size Income Ethnicity Mom Occ. Sex	Income Mom Ed. Sex
Adj.R^2	.53	.53	.45	.50
Inc.R^2	.03	.01	.04	.01
n	424	424	381	377

Tables 8 and 9, and graphed in Figures 2, 3, 4, and 5.* Table 8 and Figures 2 and 3 display two patterns: the comparative predictive power of the three procedures and the effects of predicting outcomes at later and later times. Thus medians for this table and these figures are calculated for each selection procedure and for each outcome time. These medians combine predictor test time (prekindergarten and kindergarten) and type of outcome measure (reading and math).

The patterns in Table 8 and Figures 2 and 3 are similar for the two cohorts. In both cases, background variables alone have some predictive power but not a great deal. Using just one test or subtest results in higher R^2s. Adding further tests and subtests increases the R^2s still more. And combinations of tests and background variables always do the best. In addition we see a consistent decline in R^2s as the time between prediction and outcome increases.

The relationship of predictive power to the time between measurement points is more thoroughly explored in Table 9 and Figures 4 and 5. To do this, medians were calculated for each prediction time and for each outcome time. These medians combine the three sets of predictor variables and the two outcome measures. In both cohorts we see that R^2s are highest when prediction of first-grade scores takes place in spring of kindergarten -- the shortest time span between prediction and outcome -- and declining as the time between prediction and outcomes grows longer. This phenomenon

* To further clarify what we are doing, we will reproduce one calculation from Table 8. The median R^2 (.34) in the upper left-hand corner was obtained as follows. The median was taken for R^2s of all first-grade outcome (reading and math) and for all prekindergarten and kindergarten single-test predictions for cohort III. Thus a median is obtained for the R^2s .15, .25, .25, .35 (from Table 4), and .33, .39, .45, and .46 (from Table 6).

Table 8: Median R^2s for Three Sets of Predictor Variables

	Cohort III			Cohort IV	
	1 Reading and Math	2 Reading and Math	3 Reading and Math	1 Reading and Math	2 Reading and Math
Individual Test (Both Pre K and K)	.34	.21	.20	.40	.36
Set of Tests (Both Pre K and K)	.40	.28	.24	.47	.42
Tests and Background Variables (Both Pre K and K)	.45	.34	.28	.49	.48

Table 9: Median R^2s for Three Prediction Times

	Cohort III			Cohort IV	
	1 Reading and Math	2 Reading and Math	3 Reading and Math	1 Reading and Math	2 Reading and Math
Pre K Fall 3 Sets Combined	.22	.14	.14	.32	.34
Pre K Spring 3 Sets Combined	.32	.20	.19	.48	.44
K Fall 3 Sets Combined	.45	.35	.31	—	—
K Spring 3 Sets Combined	.48	.38	.37	.51	.45

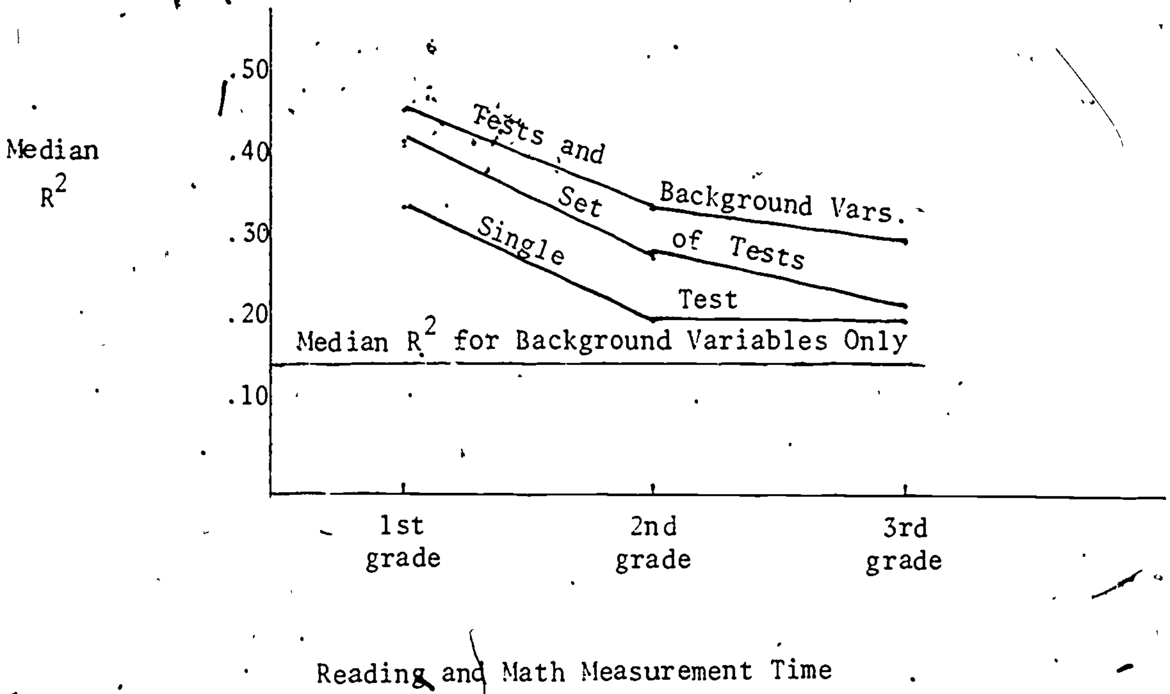

Figure 2: Median R^2s for Three Sets of Predictor Variables (Cohort III)

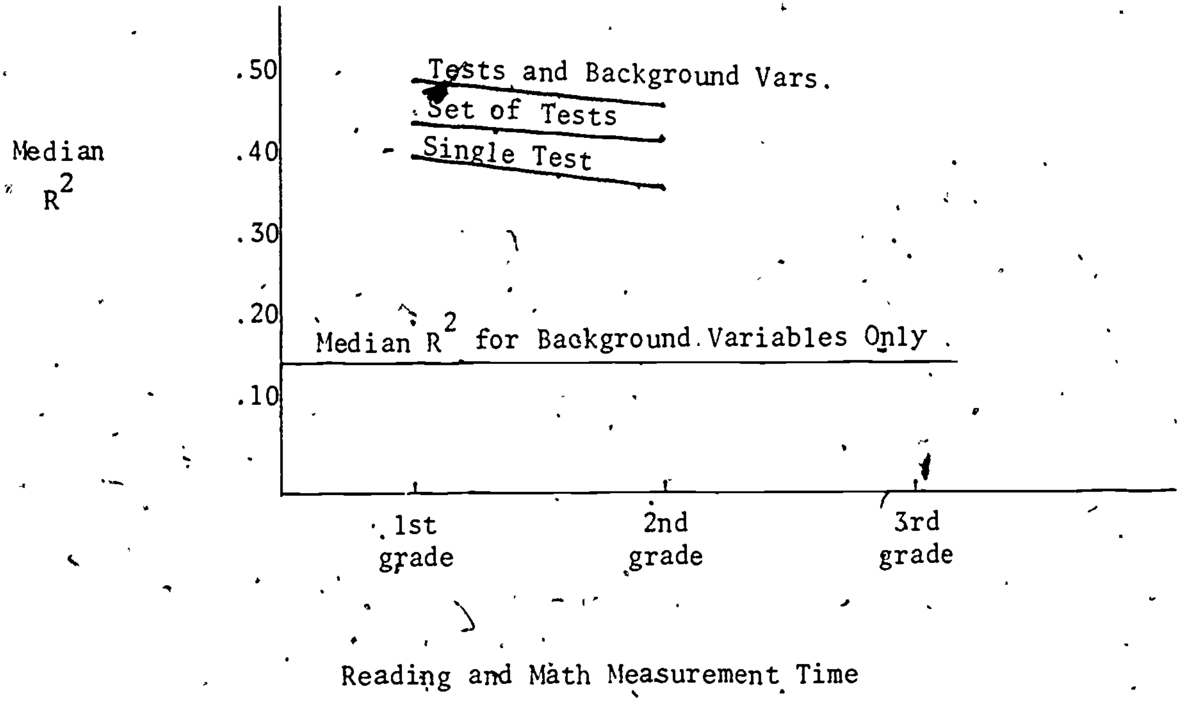

Figure 3: Median R^2s for Three Sets of Predictor Variables (Cohort IV)

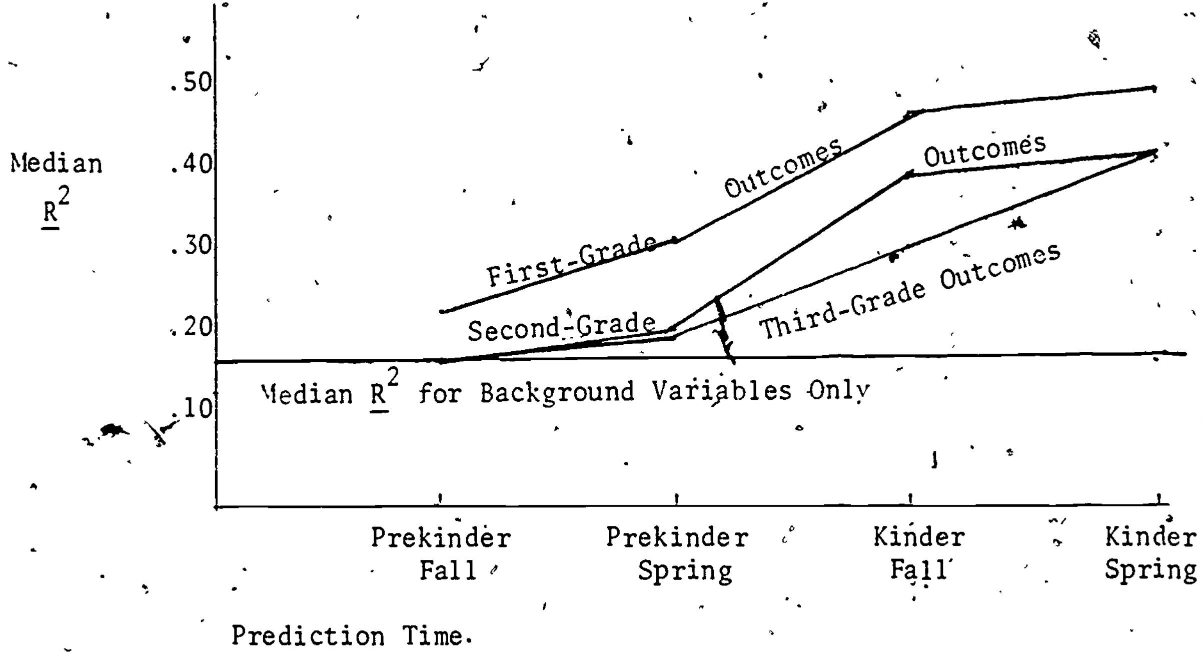

Figure 4: Median R^2s for Reading and Math Outcome Measures and All Predictor Variables* (Cohort III)

* Outcome measurements took place in the spring of first, second, and third grade.

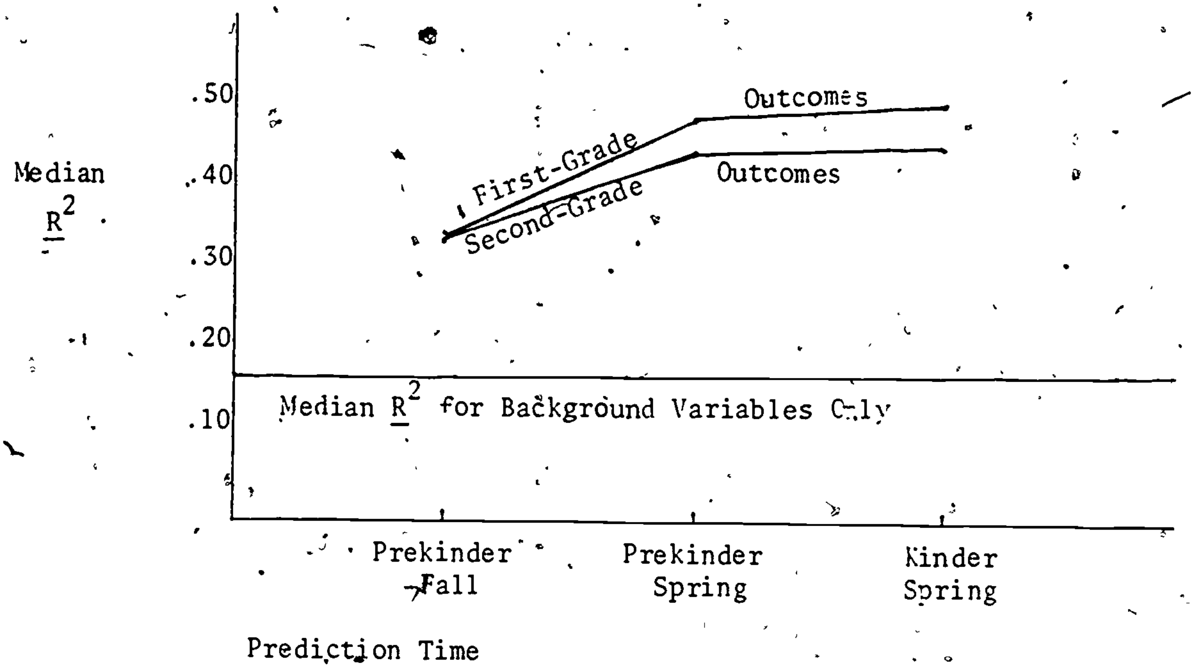

Figure 5: Median R^2s for Reading and Math Outcome Measures and All Predictor Variables* (Cohort IV)

* Outcome measurements took place in the spring of first, second, and third grade.

is apparent in three ways. First, R^2s for second- and third-grade outcomes are generally lower than R^2s for first-grade outcomes. Second, prediction improves as the test time is moved from fall to spring for both prekindergarten and kindergarten. Third, prekindergarten predictions do not do as well as kindergarten predictions.

Regarding this last point, the higher predictive validity of kindergarten tests over prekindergarten tests probably cannot be explained just as a factor of different durations between prediction and outcome. The results here are consistent with the view that kindergarten tests are more reliable and that kindergarteners are developmentally better prepared to take tests. We definitely see the "better test" effect in cohort III (Figure 4). The prekindergarten tests used in cohort III (the PSI and the NYU booklets) do not do much better than background variables at predicting later outcomes. When we examine the R^2s of the kindergarten tests in cohort III (the WRAT and the MAT), we see a substantial increase in predictive power over the prekindergarten tests. The nonlinear "jumps" in the data graphed in Figure 4 suggest that "better tests" as well as the passing of time may contribute to increased R^2s.

We do not see the same substantial increases in R^2s of cohort IV (Figure 5). The R^2s for the spring prekindergarten tests are nearly as large as those for the kindergarten tests. These results may be due in part to the use of better prekindergarten tests in cohort IV than in cohort III. Specifically, the NYU booklets* are replaced in cohort IV by the WRAT. Thus,

* According to Walker, Bane, and Bryk (1973), these booklets "are shortened versions of six Early Childhood Inventories which are being developed...at the New York University School of Education" (p. 271). These authors make the following evaluation of the NYU tests: "Neither Booklet 3D nor 4A is an adequate achievement estimate alone since they both have low internal reliability and the 3D has definite floor and ceiling effects" (p. 299).

if we compare Figures 4 and 5, we see some indication that the large increases in R^2s for cohort III might have been reduced if the "better" tests, used in cohort IV, had been used also in cohort III,

Additional Data on Early Childhood Prediction

Shipman, McKee, and Bridgeman (1976), in their study of stability and change in disadvantaged children's family variables, report findings that parallel some of what we found in our re-analysis of the HSPV/FT data. In part of the ETS Head Start Longitudinal Study, these authors examined how well measures of family status, mothers' direct and indirect influence on children, and one prekindergarten test predicted third-grade reading and math achievement test scores.

Shipman et al. measured background variables such as number of possessions in the home and mother's education, together with direct and indirect process variables such as whether the mother reads to her child. The authors also tested children two years before first grade with the PSI and again in third grade with the Cooperative Primary Test. Thus Shipman's study resembles our reanalysis of the HSPV/FT data in several respects. Both use data from Head Start children. Both have a measurement time period from prekindergarten to third grade. Both use background variables and a preschool test to predict third-grade outcomes.

Table 10 shows the relevant results from the Shipman et al. report. Overall their findings are similar to ours. Background variables account for some of the variation in third-grade scores, and a single test adds an appreciable amount to the R^2s. In some respects the results of the two studies differ, however. Shipman et al. included family process variables as well as status variables, but the process variables added little to the

Table 10: Predicting Third-Grade Reading and Math From A Wide Range of Early Childhood Variables

	Cooperative Primary Test	
Additional Variables	Reading Third Grade R^2	Math Third Grade R^2
Status/Situational		
# Possessions	.17	.14
Crowding Index	.24	.21
Head of Household Occupation	.31	.30
Race	.35	.38
Mother's Education	.37	.39
Mother/Child Interactions		
Reads to Children	.38	.39
Rational Punishment	.39	.40
Responds to Child's Questions	.39	.40
Physical vs. Verbal Punishment	.39	.40
Expectation	.39	.41
Mother's Behavior		
Reads Magazines	.39	.41
Votes	.39	.41
No. of Groups a Member of	.39	.41
Preschool Inventory (PSI)	.46	.48

Adapted from Shipman, McKee, and Bridgeman, 1976, pp. 150-155.

accuracy of prediction of later scores. The main difference between their data and ours is their finding of substantial R^2s when using simple background variables to predict third-grade test scores. Their R^2s are more than twice the analogous HSPV/FT results. Part of this could be due to differences in the sets of variables. The Shipman study included information on the home environment, such as number of possessions and a crowding index, which was not available from the HSPV/FT data. These additional variables may have added to the power of their predictor variables. Moreover, the sample of children that Shipman and her co-authors studied differed in several ways from the HSPV/FT sample, and there is some indication that it was more varied in terms of background variables than the latter. For example, mother's education averaged about 11 years with a standard deviation of about three years for the last year of data that the Shipman group analyzed. For both cohorts of the HSPV/FT data, mother's education averaged about 10.5 years with approximate standard deviations of two years. That HSPV/FT children are more homogeneous in their background variables reflects the fact that they were in part selected on the basis of economic criteria. It is well understood that the resulting restrictions in range reduce the predictive power of the variables. Thus the amount of variance accounted for by background variables in the HSPV/FT data may be relatively small because the range of some of these variables has been restricted. Because their background variables have wider range, Shipman, McGee, and Bridgeman's data better estimate the predictive power of such variables for a somewhat more diversified Head Start population.

Measuring Misclassification to Assess ECT-I Prediction Strategies

The use of R^2s and increments to R^2 is one way to evaluate the predictive validity of ECT-I variables. If one set of variables results in a higher R^2 than another, it may make sense to include those variables in an ECT-I selection strategy. But R^2s provide only one measure of prediction effectiveness. Since ECT-I selection involves identifying the most educational disadvantaged children to receive Title I services, an alternative assessment of potential selection variables is to examine how well individual variables and sets of variables classify children. In this section, then, we will illustrate how analysis of misclassification rates can be used to evaluate potential selection variables.

The identification of educationally disadvantaged children to receive Title I services may be viewed as a problem of categorical classification. Based on test results, teacher judgment, or other information, school systems try to identify children who are educationally disadvantaged from those who are not. At the early childhood level, especially before children enter first grade, educational disadvantage is often hard to define. If this identification process is viewed in a predictive manner, the goal is to identify children who will be educationally disadvantaged after they enter school, so that they can receive the benefit of ECT-I services.

There are four possible results from such an attempt to identify future educationally disadvantaged children. First, there are two ways in which prediction can be consistent with subsequent performance: a child predicted to experience future disadvantage actually does show it in future performance (a "true positive" identification of disadvantage), or a child predicted not to show later disadvantage does not in fact show it in later school performance (a "true negative" identification). Second, there are two errors or misclassifications in such an identification process: a child predicted to show disadvantage in later performance does not in fact show it (a "false positive" identification), or a child not predicted to show later disadvantage does (a "false negative" identification). From this perspective, one way to assess the ECT-I selection process is by examining the misclassification associated with different selection information.

We were able to estimate rates of these two misclassifications for several combinations of variables using the HSPV/FT data. We illustrate this approach using third-grade reading scores as a criterion of later performance. As a rough indicator of later educational disadvantage, we may define children scoring at or below the 25th percentile of the national norms for MAT Total Reading as being educationally disadvantaged in reading.* Table 11 shows the four possible results from using this criterion. By third grade about 35 percent of the children in our HSPV/FT sample scored at or below the 25th percentile.

We used predicted scores to forecast which children would score at or below the 25th percentile in the third grade. These scores were calculated from the regression equations from our previous analysis. Thus we were able to calculate predicted scores using only background variables, using one prekindergarten test, using one kindergarten test, and using a combination of background variables and tests. We would then predict that children would show later educational disadvantage if their predicted third-grade MAT reading score fell at or below the 25th percentile.

* Although this criterion is not hard and fast, it has some precedent. For instance, Becker (1977) used the 25th percentile on the MAT to estimate entry-level performance of Follow Through students (pp. 526-528).

Table 11: Four Possible Results from Comparing Predicted and Actual Performance

	Predicted Performance	
Actual Performance	Predicted Score Above 25th Percentile (No Disadvantage Predicted)	Predicted Score at or Below 25th Percentile (Disadvantage Predicted)
Children Score Above 25th Percentile (No Disadvantage Develops)	True Negative	False Positive
Children Score Below 25th Percentile (Disadvantage Develops)	False Negative	True Positive

By combining information on predicted scores with information on who actually fell below our criterion score (the 25th percentile), we were able to evaluate several prediction strategies in terms of two misclassification rates, which correspond to the upper right corner and lower left corner of Table 11. The other cells in Table 11 represent correct or consistent predictions.

When we first carried out this analysis, we used the 25th percentile score as our criterion. We found that this approach resulted in many more false negative errors than false positives. We therefore decided to try the 34th and 40th percentile scores as prediction criteria while keeping the performance criterion at the 25th percentile.

Table 12 shows the results of the analysis for several prediction strategies. Each row presents results for a different strategy -- using only one prekindergarten test in the fall, using a prekindergarten test and background variables in the fall, etc. The three sets of three columns present the results obtained when the 25th, 34th, and 40th percentiles were used as the criterion. The last column shows the R^2 from the regression analysis for each prediction strategy.

Strategies can be compared in three ways: by examining the rate of error 1 (false positives), the rate of error 2 (false negatives), and the uncertainty coefficient.* The last indicates "the proportion by which

* There are other statistics for measuring misclassification rates. Subkoviak (1980), for example, in a discussion of the reliability of mastery classification decisions, recommends Cohen's kappa when scores from two forms of a criterion-referenced test are available. This coefficient measures the reliability of the two forms in classifying children as either "masters" or "nonmasters" of the items tested. Another approach is asymmetric lambda, which "measures the percentage of improvement in our ability to predict the value of the dependent variable once we know the value of the independent variable" (Nie et al., 1975, p. 225). Of course, results will differ somewhat depending on the statistic used.

Table 12: Misclassification Rates for Strategies Using Three Cut-off Scores to Predict Third-Grade Reading Scores

	25th Percentile			34th Percentile			40th Percentile			
	Error* 1	Error+ 2	Uncert. Coeff.	Error* 1	Error+ 2	Uncert. Coeff.	Error* 1	Error+ 2	Uncert. Coeff.	R^2
Background Variables	7.1	27.7	.072	20.6	14.2	.071	29.8	5.7	.094	.15
Best Test PK Fall	7.5	30.8	.037	23.4	18.7	.0187	34.6	7.5	.037	.14
Best Test PK Spring	8.0	31.0	.033	22.1	13.3	.0675	31.0	5.3	.091	.19
Best Test K Fall	13.3	19.9	.077	26.5	7.8	.097	32.5	4.8	.090	.22
Best Test K Spring	6.5	23.2	.114	12.3	14.8	.148	25.8	7.7	.107	.29
BT & BV PK Fall	7.1	31.0	.068	19.7	7.9	.154	22.4	5.8	.165	.14
BT & BV PK Spring	10.1	29.2	.045	18.3	11.3	.127	25.2	4.3	.159	.19
BT & BV K Fall	9.6	23.0	.089	15.5	17.9	.081	25.0	8.3	.078	.32
BT & BV K Spring	7.1	25.4	.092	19.1	14.6	.076	22.5	6.7	.137	.35
All Variables Fall	6.6	18.4	.209	20.0	9.6	.134	22.2	5.2	.188	.36
All Variables Spring	8.7	21.7	.122	15.9	14.3	.116	24.6	6.3	.138	.38

*Error 1 = False positive error

+Error 2 = False negative error

'uncertainty' in the dependent variable [here, whether or not the child scored below our cut-off score] is reduced by knowledge of the independent variable' [whether or not a low score for that child was predicted]" (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975, p. 226). The uncertainty coefficient ranges from 0.0, which indicates no improvement with knowledge of the independent variable, to 1.0, which indicates complete elimination of uncertainty about the dependent variable given knowledge of the independent variable, to 1.0, which indicates complete elimination of uncertainty about the dependent variable given knowledge of the independent variable.

In most respects, the results of the error rate analysis parallel the findings from the regression analysis. We see a fairly consistent improvement in the strategies from the use of only background variables. We again see prediction improving as the prediction time is moved closer to the time when outcomes are measured. Overall, however, the prediction results viewed in terms of error rates and uncertainty coefficients seem less impressive than the R^2s. For example, use of a test and background variables from the spring of kindergarten to predict third-grade reading scores results in a combined misclassification rate of 32% and a reduction in uncertainty of only 9%, whereas the R^2 is 35%.

It is important to note that in using an error rate analysis to assess the predictive validity of early childhood variables, the rates of misclassification are influenced by choosing different criterion scores for prediction. As shown in Table 12, using the 25th percentile produced many more false negatives than false positives, the 34th percentile resulted in roughly the same percentages of both errors, and the 40th percentile produced more false positives.

More generally, note that when deciding on variables to include in a selection strategy, one needs to consider the economic and social costs associated with these different errors. If one believes the costs are similar, a criterion score that equalized error rates is indicated. But if one thinks that missing a child who needs help is worse than helping one that does not, a score that minimizes false negative errors is preferable. If, however, erroneous prediction of disadvantage is seen as worse, more weight should be given to reducing false positive misclassification.

Summary of the Findings from the HSPV/FT Data

We have learned and confirmed several things that bear on the discussion of predicting later educational outcomes from measures of early childhood variables:

- Although the predictive power of background variables in the HSPV/FT data was modest at best, such variables seem to have a place in selecting children for ECT-I programs. One reason for including background variables in a selection strategy is that their predictive power does not seem to decrease over time. In addition, we have evidence from the Shipman et al. study that background variables may have greater predictive power for populations that are more diverse than the HSPV/FT groups.

- In some cases, one test or subtest does fairly well in predicting later outcomes, especially when prediction occurs during kindergarten. Moreover, some tests do much better than others. The WRAT and the shorter version of the PSI did best in the HSPV/FT data set.

- Time between test points influences the predictive power of early childhood tests. The longer the time, the less accurate the prediction.

- In addition to the influence on time, there is some evidence suggesting that the poor quality of prekindergarten tests and the difficulty of testing very young children reduce the predictive power of tests given during prekindergarten. This, in turn, may argue for relying more on other indicators such as background variables or teacher judgment for selecting children for pre-K ECT-I programs.

- The misclassification analysis also illuminated the importance of two different errors any prediction strategy can make -- the false positive classification and the false negative classification. The relative costs of these errors should be considered when assessing variables for an ECT-I selection strategy.

META-ANALYSIS OF PREDICTIVE STUDIES

Meta-analysis, a term coined by Gene Glass (1976, 1977), is a strategy for quantitatively combining the results of similar studies. It involves examining as many published and unpublished studies as possible in the area of interest. The analysis then proceeds by determining summary statistics (such as effect sizes or correlation coefficients) from each study, aggregating these statistics, and obtaining a distribution of study statistics for which a mean, median, standard deviation, and other descriptors are calculated. Meta-analysis rests on the assumption that each study in an area of inquiry is analogous to sampling from a population of interest and estimating the population parameter. Thus, Glass argues, averaging study results produces an accurate estimate of the parameter in question. Glass acknowledges that some studies are better than others and should be weighted more heavily. To determine whether studies should be weighted according to study characteristics, he advises computing correlations between characteristics of interest and the magnitude of correlations. If correlations are substantial, these characteristics should be taken into account in combining studies. If there is little relationship, Glass maintains that the characteristics can be discounted.

Meta-analysis is a plausible approach for looking at the predictive validity of individual variables that might be used in selecting ECT-I children. As we stated earlier, no comprehensive study has been done on the predictive validity of ECT-I selection strategies, but there are hundreds of studies that contain bits of relevant information. For example, scores of studies over half a century have examined the predictive validity of readiness tests. Meta-analysis can be used to assess the overall predictive power of such tests. Likewise, many authors and practitioners argue that teacher judgment is as good a selection mechanism as readiness scores. Meta-analysis can be used to combine studies of teacher judgment and the results can then be compared to the predictive validity of readiness tests.

Scope of the Analysis

While planning this meta-analysis, we decided to focus on reading, math, and language arts outcomes, (as measured by both standardized test scores and school grades) since these are the primary areas of interest in early elementary Title I programs. We next made a list of possible predictor variables such as sex, race, test scores, teacher judgment, measures of socio-economic status (SES), and family variables. (For the initial list of predictors and outcomes, see Table 13.) We decided to look at studies that examined relationships (usually simple correlations) between one or more of these outcomes and one or more predictors.

The list of predictor and outcome variables needs some further explanation. As one can readily see, we included a wide variety of predictor and outcome variables in our initial list. Later we found it necessary to eliminate some predictors and outcomes because too few studies containing those variables could be located. A few of the variable labels in Table 13 require some description. Items in the home refer to family possessions such as vacuum cleaners and television sets, which are often used as indicators of social class. Other SES measures include scales used to assess social

Table 13: Predictor and Outcome Variables Sought in Studies Assessed by the Meta-Analysis

Predictor Variables	
Sex	
Age	
Race	
Income	
Father's Education	
Mother's Education	
Father's Occupation	
Mother's Occupation	
Items in the Home	
Other SES Measures	
Sibling Variables	
Family Variables	
Teacher Judgement:	PK II
	PK I
	K
	1
Reading Readiness:	PK II
	PK I
	K
	1
Other Readiness:	PK II
	PK I
	K
	1
IQ Tests:	PK II
	PK I
	K
	1
Other Tests:	PK II
	PK I
	K
	1
Parents' Desires	
Prior School Experience	

Outcome Variables	
Reading Achievement	1*
	2
	3
	4
	5
	6
Math Achievement	1
	2
	3
	4
	5
	6
Language Arts Ach.	1
	2
	3
	4
	5
	6
IQ Test	1
	2
	3
	4
	5
	6
Composite Achievement	1
	2
	3
	4
	5
	6
Reading Grades	1
	2
	3
	4
	5
	6
Composite Grades	1
	2
	3
	4
	5
	6
Other Measures	1
	2
	3
	4
	5
	6

* Arabic numerals are grade levels: 1 = first grade, 2 = second grade, etc.

class. Sibling variables refer to such things as number of brothers and sisters, birth order, and siblings' eligibility for compensatory education. Family variables include measures such as assessment of parent-child interaction.

Teacher judgment* and all early childhood tests were grouped according to when assessment took place. PK II refers to a test time two years prior to kindergarten; PK I, to one year prior. First grade (1) refers to fall of first grade for teacher judgment and early childhood tests. The other readiness tests include composite readiness scores and subtest scores other than reading readiness subtests. Other tests include socio-emotional and psycho-perceptual tests such as the Bender and the Wepman. If we were unsure where a test fit, we consulted Buros (1972), and followed his categorization.

Initially, we categorized study outcomes according to achievement test scores, IQ test scores, school grades, and other measures, and sought studies that reported these outcomes measures for the first to the sixth grade. We categorized a first-grade measure as an outcome only if it was obtained in the spring of first grade. For other grades, we made no distinction between fall and spring.

Locating Studies

We used several methods to find studies. We made an ERIC search of

* Teacher judgment was assessed in a variety of ways, from 5-point scales to elaborate questionnaires.

journals and ERIC documents.* We consulted literature reviews (e.g., Bryant, Claser, Hansen & Kirsch, 1974) and other meta-analyses (for example, White, 1976); and searched through dissertation abstracts and reviewed the indices of relevant journals for the last ten years.** Finally, we examined the bibliographies of articles, books, and reports that we reviewed. In all, approximately 300 studies were read. These are listed in part II of the bibliography.

Criteria for Including Studies in the Analysis

To be included, a study had to report at least one measure of a relationship between an early childhood predictor and a later outcome. Most of the studies we included reported simple correlations. Others reported statistics that could be converted into correlation coefficients. (See Glass, 1977, for details on converting various statistics to Pearson r's.) Studies that reported only multiple regression analyses without correlation matrices were excluded, since simple r's could not be retrieved. Because children develop rapidly during early childhood, we discarded any study that did not report at least approximate indications of children's ages for the times when predictor and outcome variables were measured. Some articles reported ages in months, others in years and fractions of years; and still others reported the grades and seasons when tests were given. To standardize our coding of ages, we decided to record grades and seasons when measurements were made. Table 14 shows how we converted ages into grades and seasons. Finally, we omitted any study

* We first selected all studies with the keywords Early Childhood. Then from all early childhood studies we selected those with the keywords Predictive Validity, Siblings, Achievement, Failure-Success Prediction, Reading Readiness, SES, or Parent-Child Relations.

** The journals were American Educational Research Journal, Child Development, Educational and Psychological Measurement, Harvard Educational Review, Journal of Educational Psychology, Journal of Learning Disabilities, School Review, and Teachers College Record.

Table 14: Children's Ages and Corresponding Grades and Seasons of the Year

Age (years)	Grade and Season
2.5 to 3.0	Fall Pre-K II
3.1 to 3.5	Spring Pre-K II
3.6 to 4.0	Fall Pre-K I
4.1 to 4.5	Spring Pre-K I
4.6 to 5.0	Fall K
5.1 to 5.5	Spring K
5.6 to 6.0	Fall First Grade (1)
6.1 to 6.6	Spring First Grade (1)
7	Second Grade (2)
8	Third Grade (3)
9	Fourth Grade (4)
10	Fifth Grade (5)
11	Sixth Grade (6)

that did not report sample size. Based on these criteria of acceptability, 119 of the 300 studies initially identified were included in the analysis. (These are identified with asterisks in part II of the bibliography.)

Recording Study Characteristics

As noted earlier, the magnitude of correlations can vary with study characteristics. (For example, White [1976] found that a somewhat stronger relationship between socio-economic class and achievement test scores was reported in published than in unpublished studies.) Therefore we decided to record a wide range of such characteristics. For our analysis, we attempted to record the following information:

Date of the study
Author's affiliation
Source of the study (e.g., journal)
Number of subjects
Study population (local, regional, etc.)
Percent minority in sample
Attrition rate
Whether children received some special program
Independent variables
Initial measurement time
Means of independent variables
Standard deviation of variables
Reliability of independent variables
Evidence of truncation in independent variables
Outcome measures
Outcome measurement time
Means of outcome measurements
Standard deviations of measurements
Reliability of outcome measures
Evidence of truncation in outcome measures
Correlation between independent variable and outcome measure

Of course, few studies reported all of these characteristics.

Analyzing the Results

We had three main questions in mind when we examined the results of the meta-analysis:

- On average, how does each early childhood variable correlate with various outcome meaures?

- Do some variables have significantly higher correlations, indicating that they could be better predictors of later outcomes than other early childhood variables?

- To what extent do the results of the meta-analysis substantiate or differ from our findings from the HSPV/FT study?

Our first step was to calculate the average correlations between predictors and outcomes.* This resulted in a matrix of 40 predictors by 48 outcomes and a total of 1291 correlations. Despite the large number of correlations, many cells contain few or no correlations. Some predictors, such as PK II tests, had no correlations. Some outcomes, including most from the later elementary grades, had as few as one correlation. We decided to eliminate those predictors and outcomes that had just a few cases or no cases at all. By so doing we reduced the meta-analysis so that it would more nearly parallel the data available from the HSPV/FT study. We therefore concentrated on reading, math, and language arts achievement test outcomes at the first, second, and third grade levels. This is much like the HSPV/FT study, which has reading and math outcomes in the first three grades. We also pared down the number of predictors. Still included are the 12 background variables (sex through family variables in Table 13), teacher judgment, and the four types of early childhood tests. Again, this set of predictors

* Our unit of analysis in calculating averages and other statistics was the correlation coefficient, not the study; thus some studies contributed one correlation to the analysis, others contributed 20 or 30.

parallels the HSPV/FT study; but it also permitted us to look at a broader range of preschool tests than was included in the HSPV/FT data, and at teacher judgment, which was not included in the HSPV/FT data. Appendix A contains the matrix of average correlations (821 in all) between all predictors and the three outcomes reading, math, and language arts achievement test scores.* Appendix B contains correlations (624) between early childhood tests and the three outcomes.

Some questions about the predictive validity of ECT-I selection variables can be addressed by examining average correlations. Other questions require further analysis. When we ask whether some variables do better than others in predicting later outcomes, we are asking an inferential question, for we want to know whether differences in correlations are just chance variations or are statistically and educationally significant. To answer such questions we prefer to use parametric statistics, which assume normally distributed variables. Correlation coefficients, however, are not normally distributed. (See Fisher, 1915; McNemar, 1969; and Cohen & Cohen, 1975). Fortunately, Fisher devised a method for transforming correlation into Fisher's z, which approximate normal distributions (Cohen & Cohen, 1975, pp. 50-52).

Transforming correlations into z's makes it sensible to use analysis of variance (ANOVA) and other parametric procedures. The next subsection illustrates how we used ANOVA to compare the average correlations of different early childhood tests, different outcome measures, and different measurement points.

* These correlations are based on a combined sample of size of 147,780. Studies on average had an n of 180.

The Findings of the Meta-Analysis

Discussing the findings of the meta-analysis is not easy. We are examining average correlations from a large number of studies, which vary in several ways. They vary in sample size, attrition rates, and other study characteristics. They vary regarding the predictor variables (X's) employed. They use different outcome measures (Y's). Prediction times (t_x) also vary, as do outcome times (t_y). In principle, we could analyze simultaneously all the ways in which studies vary. But such an analysis (e.g., a four-way ANOVA with several covariates) would be complex and difficult to interpret if there were significant interactions. Instead, we have decided to examine smaller pieces of the puzzle. We leave to the last a look at the influence of study characteristics such as sample size, turning first to differences resulting from varying predictors, outcome measures, predictor times, and outcome times. At most we will discuss two of these four dimensions at one time. To do this we will have to "average across" the other two dimensions. For example, when we examine the effects of different outcome measures (reading, math, and language arts tests) and different outcome times (first grade, second grade, and third grade), we average across different predictors (such as types of early childhood tests) and predictor times (pre-K'I, K, and first grade).

In examining variations in X's, Y's, t_x's and t_y's, we concentrated on five areas:

- Background variables as predictors
- Differences in predicting reading, math, and language arts outcomes
- Effect of time between prediction and outcome on predictive validity
- Predictive validity of different categories of early childhood tests
- Predictive validity of teachers' judgment.

We will discuss each of these in turn.

Background Variables. Tables 15 and 16 summarize the results for a set of background variables. We are interested in whether the predictive validity of background variables varies over outcome time (t_y) and with different outcome measures (Y's). Thus we are looking for effects in these two dimensions, and we are averaging across the different background variables (X's). The number in the center of each cell in Table 15 is an average correlation* computed across 12 background variables.** This table and the accompanying two-way analysis of variance show two things about the predictive validity of background variables. First, we see no difference in the overall ability of background variables to predict different outcome tests: the mean correlations between background variables and reading tests (.20), math tests (.17), and language arts tests (.18) are not significantly different ($F = 1.65$, $df = 2, 91$ $p > .05$). Second, we find no decrease in the predictive power of background variables as they are used to predict later and later outcomes (first through third grade). The average correlation between background variables and first-grade outcomes is actually the lowest, but the differences among means are not significant ($F = 2.41$, $df = 2, 91$, $p > .05$). These two results parallel what are found in the HSPV/FT data.

Table 16 presents a further comparison between the meta-analysis and the HSPV/FT results. The predictive power of the background variables represented in both data sets is similar. Table 16 shows that, with a few exceptions,

* All average correlations are weighted by the number of cases (i.e., number of correlations) per cell.

** Sex, age, race, income, father's education, mother's education, father's occupation, mother's occupation, items in the home, other SES measures, sibling variables, and family variables.

Table 15: Average Correlations Between a Set of Background Variables and Reading, Math, and Language Arts Scores for Three Outcome Times*

	Outcome Time			Row Averages
Outcome Tests	Grade 1	Grade 2	Grade 3	
Reading	.17 27	.23 13	.22 20	.20
Math	0.0 1	.26 5	.16 20	.17
Language Arts	.11 1	.21 6	.11 1	.18
Column Averages	.16	.23	.19	

*The number in the lower right corner indicates the number of correlations per cell.

Analysis of Variance:

Fisher's z by Type of Outcome Test and Time of Outcome Test

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	0.139	4	0.035	1.652	0.169
- Outcome Test	0.070	2	0.035	1.649	0.199
- Outcome Time	0.102	2	0.051	2.411	0.096
2-Way Interactions	0.049	4	0.012	0.576	0.680
	0.049	4	0.012	0.576	0.680
Explained	0.188	8	0.023	1.114	0.362
Residual	1.751	83	0.021		
Total	1.939	91	0.021		

Table 16: Comparison of Correlations Between Background Variables and Outcomes from the Meta-Analysis and HSPV/FT Data Sets (Cohort III)

Background Variables	Outcome Tests and Times Reading 1	Reading 2	Reading 3	Math 1	Math 2	Math 3
Sex						
Meta	.19	.17	.13	—	—	.01
HS/FT	.06	.16	.06	.20	.00	-.01
Age						
Meta	.05	.19	.01	—	—	.08
HS/FT	.15	.01	-.02	.01	.10	.05
Race						
Meta	.14	.23	.16	.27	.14	
HS/FT	.36	.21	.23	.19	.33	.28
Income						
Meta	—	—	.07	—	—	.12
HS/FT	.21	.17	.20	.14	.16	.20
Mother's Educ.						
Meta	—	.19	.34	—	—	.26
HS/FT	.09	.10	-.02	.02	-.02	-.02
Mother's Occ.						
Meta	—	.22	.10	—	—	—
HS/FT	-.06	-.07	-.08	-.07	-.10	-.11

common correlations from the two analyses are similar. Like the HSPV/FT data, the meta-analysis indicates that background variables are modestly correlated with achievement outcomes.

Predicting Different Outcomes. Table 17 reports average correlations between early childhood tests and reading, math, and language arts outcomes at grades 1, 2, and 3. The number in the center of each cell is an average correlation taken across all early childhood tests (reading readiness, other readiness, IQ, and other tests) at three prediction times (prekindergarten I, kindergarten, and grade 1). All averages are weighted by number of cases per cell.

Table 17 and the companion two-way ANOVA show some unexpected results. Looking at the row averages, we see that the mean correlations for reading and math outcomes are about the same but that the language arts mean is considerably lower, and the analysis of variance shows a significant difference among means (F = 8.18, df = 2, 615, p < .01). Instead of these findings, we would expect that early childhood tests would differ in how well they predict reading and math scores, since reading and solving math problems presumably involve quite different skills. We would expect smaller differences in the prediction of reading and language arts scores, since the underlying skills are probably more closely related than reading and math skills.

The analysis of variance also shows that the times of outcome measurements are significantly different (F = 30.07, df = 2, 615, p < .01). However, the column means indicate a pattern unlike what we would expect. Initially, correlations go down -- predictions of first-grade scores do better than those of second-grade scores. But third-grade predictions are higher than

Table 17: Tests as Predictors of Reading, Math, and Language Arts Outcomes*

	Outcome Times			
Outcome Tests	Grade 1	Grade 2	Grade 3	Row Averages
Reading	.44 193	.33 121	.38 43	.40
Math	.47 27	.41 12	.35 19	.42
Language Arts	.39 86	.27 88	.34 35	.33
Column Averages	.43	.31	.36	

* The number in the lower right corner is the number of cases per cell.

Analysis of Variance:

Fisher's z By Type of Outcome Test and Time of Outcome Test

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	3.550	4	0.888	21.150	0.000
- Outcome Tests	0.686	2	0.343	8.175	0.000
- Outcome Time	2.523	2	1.262	30.066	0.000
2-Way Interactions	0.125	4	0.031	0.746	0.561
	0.125	4	0.031	0.746	0.561
Explained	3.675	8	0.459	10.948	0.000
Residual	25.808	615	0.042		
Total	29.483	623	0.047		

second-grade. This is puzzling, since we expect prediction to decline monotonically. This seems to result from some complex confounding of study characteristics and outcome measurement times. In looking for a precise explanation, we explored the effects of several characteristics. We found, for example, that attrition is related to the magnitude of correlations. However, we were unable to find a completely satisfactory explanation for these results.

Types of Predictor Tests. Table 18 summarizes our findings for the correlations between types of predictive tests (reading readiness, other readiness, IQ, and other tests) given at three times (prekindergarten I, kindergarten, and first grade) and the three outcome measures (reading, math, and language arts). The numbers in the center of each cell are the average correlation across the three outcome measures (Y's) and the three outcome times (t_y's).

The unadjusted and adjusted means in Table 18 and the analyses of variance again show some unexpected findings. The unadjusted means and the ANOVA for type of predictor test indicate that reading readiness tests do slightly better than other readiness and IQ tests. Other tests seem to do worst. This finding supports our argument in the previous section that predictive validity should be an important criterion for choosing a selection test, since not all tests predict equally well. The results regarding the effects of prediction time on the correlation between early childhood tests and outcomes seems to contradict expectations. The first analysis of variance (with unadjusted means) shows no significant difference among the prediction times ($F = 1.29$, $df = 2, 612$, $p > .05$). Moreover, the pattern of the unadjusted means is unexpected: tests given in prekindergarten appear to be the best predictors of later outcomes.

At first we thought that there might be some confounding between the time at which the predictor test was given and the total time between

Table 18: Average Correlations Between Types of Predictor Tests and Later-Grade Outcomes Unadjusted and Adjusted for Time

	Unadjusted Means	Means Adjusted for Time Between Measurements
Predictor Test:		
Reading Readiness Tests	.47 (264)	.47 (264)
Other Readiness Tests	.41 (67)	.40 (67)
IQ Tests	.41 (80)	.40 (80)
Other Tests	.29 (213)	.30 (213)
Predictor Time:		
Prekindergarten	.42 (41)	.52 (41)
Kindergarten	.37 (348)	.40 (348)
First Grade	.42 (235)	.38 (235)

Analysis of Variance:

Fisher's Z by Predictor Test and Predictor Time

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	5.522	5	1.104	28.989	0.000
-Predictor Test	5.010	3	1.670	43.834	0.000
-Predictor Time	0.098	2	0.049	1.291	0.276
2-Way Interactions	0.646	6	0.108	2.824	0.010
		6	0.108	2.824	0.010
Explained	6.167	11	0.561	14.717	0.000
Residual	23.316	612	0.038		
Total	29.483	623	0.047		

Analysis of Variance:

Fisher's z by Predictor Test and Predictor Time
Controlling Time Between Measurement

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	1.069	1	1.069	29.081	0.000
-Total Time Between Measurement Points	1.069	1	1.069	29.081	0.000
Main Effects	5.436	5	1.087	29.580	0.000
-Predictor Test	4.430	3	1.477	40.176	0.~~000~~
-Predictor Time	0.717	2	0.358	9.752	0.000
2-Way Interactions	0.519	6	0.087	2.354	0.030
	0.519	6	0.087	2.354	0.030
Explained	7.024	12	0.585	15.925	0.000
Residual	22.459	611	0.037		
Total	29.483	623	0.047		

predictor test and outcome test. If the time between tests for studies using prekindergarten tests tended to be shorter then the time between tests when kindergarten and first-grade tests are used, the average correlation from the prekindergarten studies might be equal to or larger than the averages from kindergarten and first-grade studies. For example, if most prekindergarten tests were used to predict first-grade outcomes whereas most kindergarten and first-grade tests were used to predict third-grade outcomes, the prekindergarten tests might appear to do as well as or even better than the kindergarten and first-grade tests. To test this hypothesis, we entered the total time between tests as a covariate and then performed thw two-way ANOVA again. The column of adjusted means in Table 18 shows that the average correlation for prekindergarten tests increases by nearly 25 percent while the other means hardly change when time between tests is taken into account. Moreover, the differences among these means become statistically significant (F = 9.75, df = 2, 611, $p < 0.001$). This is a surprising result, quite at odds with our findings from the HSPV/FT data.

A closer examination of the studies that are represented in Table 18 helps explain these results. First, most of the correlations fall in the kindergarten and first-grade rows. One might expect that the prediction times for studies using kindergarten tests and those using first-grade tests would differ by about one year on average. But many of the kindergarten studies tested in the spring and all the first-grade studies tested in the fall because we classified any test given after fall of first grade as an outcome. Thus, in many cases, the prediction times for kindergarten and first-grade studies differed by only a few months. This short time difference may account for finding no difference between the means of kindergarten and first-grade prediction times.

The high average value for prekindergarten tests may be caused by an abundance of "good" tests in prekindergarten studies. The 41 correlations of prekindergarten tests with later outcomes come from four studies. These studies, upon closer examination, used prekindergarten tests that we found in the HSPV/FT study to be good predictors of later reading and math scores. For example, one study reported correlations of 0.60, 0.59, and 0.49 between the WRAT and reading scores at first, second, and third grade. These findings are comparable to our results from the HSPV/FT data, which showed correlations of 0.70 and 0.64 between the WRAT reading subtest and first- and second-grade reading scores. Thus we would expect the average correlation for prekindergarten tests to be lower if we had found studies with a wider range of prekindergarten tests.

The final result of note from Table 18 is the statistically significant interaction between test time and test type. We cannot explain this result. The interaction seems to be small in comparison to the main effects, and we suspect that it is not of substantive significance.

The Predictive Validity of Teacher Judgment

Our final results deal with teacher judgment as a predictor of later achievement. As the Huron field study (Yurchak & Bryk, 1980) reported, teacher judgment is often used explicitly or implicitly to select ECT-I participants. Indeed, it is often suggested as a complement to or substitute for test results. We were able to locate studies with a total of 75 correlations between some kind of teacher judgment and reading, math, and language arts scores in grades 1, 2, and 3.

Table 19 presents a summary of what we found. Each cell averages across measures of teacher judgment in kindergarten and in first grade.

Table 19: Teacher Judgment as a Predictor of Reading, Math and Language Arts Achievement*

	Outcome Time			
Outcome Tests	Grade 1	Grade 2	Grade 3	Row Averages
Reading	.41 27	.56 4	.46 6	.43
Math	.33 11	—	.51 3	.37
Language Arts	.37 23	—	.37 1	.37
Column Averages	.38	.56	.47	

* The number in the lower right corner of each cell is the number of cases.

(We found no studies relating teacher judgment during prekindergarten with later outcomes.) Teacher judgment in these studies encompases a range of activities. In some cases, teachers were asked to rate children's future achievement on a 5-point scale. In other studies, teachers were asked to judge children on the same criteria that the tests used. Still others used lengthy questionnaires for teachers to assess children. Studies also varied on how long teachers knew the children they assessed and on whether teachers had seen test results before making their assessment.

Table 19 shows that teachers seem to do well. Average correlations range from 0.33 to 0.56, which is not much different from the correlations for the best tests in the HSPV/FT data. We did not perfrom an analysis of variance on these data, but we can get some useful impressions from the row and column means. First, there seems to be little difference in predicting the three outcomes, although the average reading correlation is higher than the correlation for math or language arts. Second, we do not see a decline in average correlations as outcome time lengthens. In fact, the first-grade average correlation is lowest of the three. This indicates that teacher judgment may behave like background variables and unlike test scores; i.e., the predictive validity of background variables seems to be fairly stable over time whereas the predictive validity of early childhood tests declines over time.

We need to note an important caveat regarding the studies from which our results on teacher judgment were obtained. Almost all of the correlations (68) measured the relationship between teacher judgment in the spring of kindergarten and test scores in grades 1, 2, and 3. No study reported results from the fall of kindergarten, and there were only 7 correlations that resulted from fall first-grade teacher judgment. Thus one reason teacher judgment accurately predict later score may be because the teachers knew the children they were assessing for nearly a full school year. But teacher judgment used as part of an ECT-I selection procedure would most likely take place in the fall of the year, when the teachers have known the children they are assessing for only a short time. Thus we must be cautious about being overly enthusiastic about teacher judgment until more data are available.

Limitations and Caveats

Meta-analysis is a new and somewhat controversial analytic technique; therefore, one must be cautious in its use. As meta-analysis had been used to synthesize results in more and more areas, critics have raised some important concerns (see, for example, Eysenck, 1978; Gallo, 1978; and replies to Rosenthal and Rubin, 1978). This paper is not the place to examine the "virtues and vulnerabilities" of meta-analysis in general (see Hauser-Cram, Note 1, and Jackson, 1980), but it is appropriate to raise some caveats and limitations regarding the results from our application of the technique.

Our use of meta-analysis to examine potential ECT-I selection variables has been an exploratory process. We have sought to move beyond what Glass and his colleagues have tried. They attempt a single meta-analysis -- for example, synthesizing the effects of class size on achievement or the effects of psychotherapy. We, on the other hand, have tried to synthesize findings from several areas simultaneously -- analyzing studies employing one or more prediction variables from several sets of variables -- background measures, teacher judgment, and early childhood tests, for example. We have also attempted to examine a wide range of criterion measures and outcome times.

Because meta-analysis is a new approach and because we have applied it in several ways simultaneously, findings of our analysis must be viewed cautiously. One way in which we exercised caution was to compare and contrast the meta-analysis findings with the results from our re-analysis of the HSPV/FT data. Where results are similar (for example, those on the predictive validity of background variables), we are fairly confident of the meta-analysis findings. But where the meta-analysis produced results at odds with the HSPV/FT data, we are more skeptical. For example, when we found prekindergarten tests more highly correlated than kindergarten or first-grade tests with later outcomes, we began looking for alternative explanations. Likewise, although teacher judgment shows promise as a predictor of later achievement, our conclusions in this regard must be tentative because teacher judgment was not included in the HSPV/FT data.

Another reason for caution is the complex way in which study characteristics appear to influence study outcomes. Glass argues that the influence of such characteristics as sample size can be ignored if the correlation between the characteristic and the magnitude of the relationship is near zero; he does not discuss at length what to do if a relationship is not near zero. When we looked at the relationships between study characteristics and magnitude of r's, we found some large and some counterintuitive results. Table 20 presents correlations between four study characteristics -- attrition rate, predictor and outcome reliability, and sample size -- and the size of r's reported in these studies.

Three relationships are statistically significant and two are fairly substantial. Attrition rate (which is measured by percentage of subjects missing for later measurement) has a strong relationship (0.29) with correlation

Table 20: Correlations Between Magnitude of Pearson's r and Selected Study Characteristics for Studies Reporting Relationships Between Early Childhood Tests and Reading, Math, and Language Arts Achievement

Characteristic	Mean	Standard Deviation	Cases	Correlation
Attrition Rate	22.6	20.0	392	.29*
Predictor Reliability	80.3	12.4	404	.31*
Outcome Reliability	86.0	5.0	556	-.03
Number of Subjects	170.0	256.9	624	.10*

*p <.05

size, but it is in the unexpected direction (that is, attrition and the size of the correlation are directly rather than inversely related). Unless attrition is random, we would expect correlations to decrease with higher attrition rates because the sample becomes more homogeneous. Here we find the opposite relationship. Predictor reliability is in the expected direction, but its relationship with the size of the r's is surprisingly high (0.31). The relationship between outcome reliability and correlation size is essentially zero, probably because of the low varability in outcome reliability in our sample. Sample size and magnitude of r's are weakly related with studies having larger samples, with large samples tending to produce slightly higher correlations.

It is difficult to know what to make of these relationships. It seems likely that complex interactions among studies are at work. For example, it is possible that many studies of prekindergarteners are done by universities and research organizations; and, for that reason, perhaps are better controlled. Predictive validity studies of reading readiness tests given to kindergarteners and first-graders may be done more often by school districts; and may be less well controlled. With less well controlled studies, one would expect reduced correlations between predictors and outcome measures. To try to unravel these complexities is a substantial task requiring time and resources beyond those available to us. Hence our concern that the results of our meta-analysis be viewed with caution.

Summary of the Findings from the Meta-Analysis

Overall, our meta-analysis of studies supports the conclusions we reached from the re-analyzing of the HSPV/FT data set.

- We found that background variables correlate weakly with later educational outcomes, but these correlations do not seem to decrease as the time between prediction and outcome measures increases.

- Some individual early childhood tests predict later outcomes fairly well; and some tests do better than others. Reading readiness tests appear to do best; non-cognitive tests seem to do worst.

- Both analyses revealed significant relationships between predictive power and time between measurement points, but these relationships appeared to be more complex in the meta-analysis data than in the HSPV/FT data. Probably this is due to interactions between study characteristics and the relation of time to predictive power.

- The meta-analysis data permitted us to examine teacher judgment as a predictor of later educational outcomes. Our tentative findings are that teacher judgment does nearly as well as tests and that its predictive power seems not to decline over time to the extent test results do.

CONCLUSIONS

Our aim in this paper has been to inform discussion of ECT-I selection policy, and our main audience has been people at the federal level who think about, formulate, promulgate, and monitor such policy. Our study has been limited to a discussion of the predictive validity of early childhood variables that, in some combination, could make up part of local ECT-I selection strategies. The study has also been limited to the data at hand -- data which were collected originally for other purposes. In this last section, we will mention briefly some considerations besides predictive validity that should be taken into account in a complete examination of ECT-I selection policy. Then we will discuss the implications from our findings for ECT-I selection policy.

Some Other Considerations

This paper has assessed the predictive validity of some early childhood variables that could be part of an ECT-I selection process. Of course, there are other criteria for judging selection variables and for assessing the overall process by which young children are chosen to receive ECT-I services. This subsection will briefly discuss some important considerations that we have not discussed in this paper.

Other Aspects of Choosing ECT-I Children. Determining who receives ECT-I services is a three-staged process: Title I attendance areas are specified (on economic grounds), a pool of eligible children residing in the attendance area is identified (based on educational need), and the neediest children are selected from that pool. We have examined only aspects of the last stage, selection. But the other stages need to be considered in any overall discussion of ECT-I identification and selection. The second stage, identification, is particularly problemmatic for ECT-I. Identification for Title I programs aimed at children in grades 2-12 is made easier because almost all potentially eligible children are in school and available for identification. Children are not so readily available for ECT-I identification since ECT-I programs often provide the first school experiences for educationally disadvantaged children. Some work has been done on the problem of identifying young children in need of services (see Hauser-Cram, Note 2; Yurchak, Note 3), and further consideration seems warranted.

Costs of Selection Procedures. We have said little in this paper about the monetary and non-monetary costs of ECT-I selection, which are obvious concerns in assessing any selection procedure. We have seen that several variables together usually predict later outcomes more accurately than one variable, for example, a test score. But using multiple measures such as a combination of test scores, background variables, and teacher judgment to select ECT-I children may be expensive, using resources that might be better spent serving those children who are selected. Moreover, multiple measures can be a burden on teachers, children, and parents. Giving several tests, collecting background information, and judging children's readiness for school can be laborious for teachers and, worse, can take time from instruction. Providing detailed information about their children can be annoying or threatening to parents. Data collection, especially test taking, can be boring, confusing, or threatening for children. Such costs should be assessed in judging alternative selection procedures.

Problems in Assessing Young Children. In judging variables that could make up an ECT-I selection procedure, one must continually keep in mind the special problems in assessing young children. In another report from Huron's study of ECT-I evaluations, Haney and Gelberg (1980) not only point out that "tests and instruments for use with young children are generally of lower technical quality than those for use with older children" (p. 7) but also argue that preschool children often lack the physical, intellectual, and emotional prerequisits necessary for systematic assessment. Given these special difficulties it may make sense, especially when selecting prekindergarteners or children who have had no school or preschool experience, to emphasize variables that are not so dependent on obtaining direct information from young children in strange situations -- variables such as family characteristics, teacher judgment, and sibling information.

Selection Bias. Much of the discussion about bias against minority groups in the literature dwells on the misuse of standardized tests leading to the misclassification of children. (For example, see Mercer, 1975). But such discussions could be broadened to other variables of a selection strategy. Haney and Kinyanjui (1979), who aim their discussion at tests, provide a useful perspective on bias. They argue that tests are not usually biased but the use of tests may be. By extention, the components of an ECT-I selection procedure (which might include test scores, background variables, and teacher judgment, for example) probably would not be intensionally biased for or against minorities. However, the use of the strategy might be biased.

Haney and Kinyanjui argue that "a test is biased if, when it is used to make decisions or inferences about a person or group, those decisions or inferences are less valid than those made when it is used analagously with people generally" (p. 5). Their definition of bias rests on the validity of the decision or inferences resulting from the test. Similarly an ECT-I selection procedure might be said to be biased if the selection decisions for some groups are less valid than decisions for all children. Furthermore, the predictive accuracy of a procedure is one criterion in terms of which to judge its bias. That is, if a variable, a set of variables, or a selection procedure has lower predictive validity for some group than for others, it might be viewed as biased.

Implications for ECT-I Selection Policy

Importance of Predictive Validity. The importance of prediction stems from the central goal of most ECT-I programs: the prevention of educational problems in later schooling. This goal, together with the requirement of selecting the neediest children, suggests that selection be based at least in part on which children are most likely to experience later educational disadvantage. Thus, we have argued and tried to demonstrate that the predictive validity of background variables, teacher judgment, and test scores is an important consideration.

The implication here is that local program staff should examine the predictive validity of their selection method. This means more than just looking up a validity index for the tests they use. It means studying the predictive validity of their procedures in terms of their unique population of children, since similar procedures can produce varying results with different samples, as we saw in the HSPV/FT re-analysis and the Shipman data. Another reason for encouraging local staffs to examine the predictive validity of their selection methods is that definitions of educational disadvantage differ from community to community. Some LEAs are most concerned about preventing future reading problems; others want to help children at-risk become better prepared to achieve later school success in general. Moreover, some LEAs emphasize standardized test scores as measures of later school success; others are more interested in grades or students attitudes toward school. Although we did not find striking differences in how well early childhood variables predict different outcomes, in some cases, the composition of a set of variables or the weighting of the variables that go into a selection strategy may differ depending upon how a district chooses to define educational disadvantage in later grades.

Useful Statistical Procedures. Fortunately, some statistical tools exist to help local staffs assess the predictive validity of their selection procedures. We have described two in this paper: examining R^2 (and increments to R^2) and examining misclassification rates. Using the latter seems to be a particularly fruitful approach to assessing different strategies; it makes explicit the errors and the successes of any strategy and can help the staff focus their attention on the costs and benefits to the children they select and do not select.

Importance of Longitudinal Data. Clearly, data collected over time are needed to assess predictive validity. All the data we used were longitudinal. The HPSV/FT data set and the Shipman data followed some children from prekindergarten through third grade. Studies included in the meta-analysis measured children as early as two years before kindergarten and as late as sixth grade. The shortest time span of studies included in the meta-analysis was six months -- fall of first grade to spring of first grade.

To judge the predictive validity of a selection strategy, data are needed at the time of selection and at a future time when some important event will take place (such as success or failure in third-grade reading). Many districts have neither the capacity nor the expertise to collect and analyze longitudinal data. However, given the importance of these data not only for studying selection but for evaluating programs, help should be provided to LEAs to enhance their capacity in this respect. Some of this help might come from Title I Technical Assistance Centers. Additional help might come from consortia of LEAs having ECT-I programs. These districts could band together to share computer facilities, analytic strategies, and even data.

Some LEAs have the capability to collect and analyze longitudinal data but may lack the resources needed to apply the data to a topic like local ECT-I selection. One large district we visited had extensive longitudinal files tracing Follow Through children for several years after the program. Some files held test scores and background information; other files held student identification. The only problem was a lack of resources to merge these files and apply them to early childhood selection and evaluation questions. Given the importance of a predictive perspective and the difficulty in collecting and analyzing longitudinal data, encouragement and support from USED for collecting and analyzing such data seem warranted.

Selection Instruments and Variables. Both the re-analysis of the HSPV/FT data and the meta-analysis suggest instruments and variables that have roles to play in selecting ECT-I children. From the standpoint of prediction, early childhood tests seem to have a place in selection strategies. Some instruments, such as the WRAT and the PSI, appear to be fairly accurate predictors of later achievement. Others seem to have less accuracy. Thus the predictive validity of a test should be considered in deciding whether to use it.

Although some tests have reasonably high predictive validity, we have seen that adding other variables and assessment methods to test results can improve the accuracy of a selection procedure. One set of such variables is SES-related measures, which usually improve prediction of later achievement and do not seem to diminish in predictive power over time spans for which we had data. In addition to simple background variables such as income and parents' education, some ECT-I programs may have access to or the capacity to collect more sophisticated variables such as measures of parent-child interaction. We have little data on the usefulness of such variables for ECT-I selection, but local efforts to collect and assess such information warrant some encouragement and support.

Teacher judgment also has a place in ECT-I selection. Tentative findings from our meta-analysis show that teacher judgment may do as well as early childhood tests in predicting later achievement. As we noted earlier, teacher judgment may be particularly useful in selecting very young children, whose lack of skills and school experience reduce the reliability and validity of tests. Unfortunately, we had no data that allowed us to examine how much teacher judgment would add to the predictive power of tests and background variables. In addition, there seems to be a dearth of data on teachers' ability to assess prekindergarteners. Clearly, here is an area for further research, some of which could be carried on by LEAs with or without TAC assistance.

Finally, there are several potentially fruitful instruments and variables for which we have no information and for which further investigation is warranted. We had no data and found no studies that used early childhood criterion-referenced tests to predict later outcomes, although this use for

CRTs has been suggested (for example, Stenner, et al., 1976). Parental judgment is another area for which we have no data, but where some further investigation may be warranted (see Johansson, 1965, for a study of parental judgment in Sweden). In our descriptive study of ECT-I programs, we found several LEAs using information about siblings and about language proficiency to select children. Again we cannot comment on these approaches except to say that they deserve examination.

REFERENCE NOTES

1. Hauser-Cramm, P. Research synthesis: Virtues and vulnerabilities. Unpublished paper, Kennedy School of Government and Harvard Graduate School of Education, 1980.

2. Hauser-Cramm, P. Developmental screening: A review and analysis of key issues. Report in preparation for the Huron Institute, Cambridge, MA, 1979.

3. Yurchak, Mary Jane. Identifying and selecting children for early childhood Title I programs. Paper in preparation for the Huron Institute, Cambridge, MA, forthcoming.

REFERENCES

Becker, W.C. Teaching reading and language to the disadvantaged. Harvard Educational Review, November 1977, 47, 518-543.

Bryant, E.C., Glaser, E., Hansen, M.H., & Kirsch, A. Associations between educational outcomes and background variables: A review of selected literature: Denver, CO: National Assessment of Educational Progress, 1974.

Buros, O.K. (Ed.). The seventh mental measurements yearbook. Highland Park, N.J.: The Gryphon Press, 1972.

Cohen, J. & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers, 1975.

Eysenck, H.J. An exercise in mega-silliness. American Psychologist, 1978, 33, 517.

Fisher, R.A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 1915, 10, 507-521.

Gallo, P.S. Meta-analysis--A mixed meta-phor? American Psychologist, 1978, 33, 515-517.

Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Glass, G.V. Integrating findings: The meta-analysis of research. In L.S. Shulman (Ed.). Review of research in education (Vol. 5). Itasca, Ill.: F.E. Peacock Publishers, 1977.

Haney, W.M. The Follow Through planned variation experiment. Vol 5: A technical history of the national Follow Through evaluation. Cambridge, MA: The Huron Institute, 1977.

Haney, W.M., & Gelberg, J.W. Assessment in early childhood education. Cambridge, MA: The Huron Institute, 1980.

Haney, W.M., & Kinyanjui, K. Competency testing and equal educational opportunity. IRCD Bulletin, 1979, 14, 1-11.

Jackson, G.B. Methods for integrating reviews. Review of Educational Research, 1980, 50, 438-460.

Johansson, B.A. Criteria of school readiness: Factor structure, predictive value, and environmental influences. Uppsala: Almquist and Wiksells Boktryckeri AB, 1975.

McNemar, Q. Psychological statistics (4th ed.). New York: Wiley, 1969.

Mercer, J.R. Psychological assessment and the rights of children. In N. Hobbs (Ed.) Issues in the classification of children (Vol. 1). San Francisco: Jossey-Bass, Inc., Publishers, 1975.

Molitor, J.A., Watkins, M., & Napior, D. Education as experimentation: A planned variation model, The non-Follow Through study. Cambridge, MA: Abt Associates, Inc., 1977.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. Statistical package for the social sciences (2nd Edition), New York: McGraw-Hill, 1975.

Rosenthal, R., & Rubin, D.C. Interpersonal expectancy effects: The first 345 studies. The Behavioral and Brain Sciences, 1978, 3, 377-415.

Shipman, V.C., McKee, D., & Bridgeman, B. Stability and change in family status, situational, and process variables and their relationship to children's cognitive performance. Disadvantaged children and their first school experiences. ETS Head Start Longitudinal Study, September, 1976. 261 pages. ERIC Reproduction Number ED 138339.

Stenner, A.J., Feifs, H.A., Gabriel, B.R., & Davis, B.S. Evaluation of the ESEA Title I program of the public schools of the District of Columbia, 1975-76. Washington, D.C.: The Public Schools of the District of Columbia, no date.

Subkoviak, M.J. Decision-consistency approaches. In R.A. Berk (Ed.) Criterion-referenced measures: The state of the art. The Johns Hopkins University Press, 1980.

Walker, D.K., Bane, M.J., & Bryk, A.S. The quality of the Head Start planned variation data (2 Vols.). Cambridge, MA: The Huron Institute, 1973.

Weisberg, H.I., & Haney, W.M. Longitudinal evaluation of Head Start planned variation and Follow Through. Cambridge, MA: The Huron Institute, 1977.

White, K.R. The relationship between socioeconomic status and academic achievement. (Doctoral dissertation, Univ. of Colorado, 1976).

Yurchak, M.J., & Bryk, A.S. ESEA Title I early childhood education: A descriptive report. Cambridge, MA: The Huron Institute, 1980.

Yurchak, M.J., Gelberg, W., & Darman, L. Description of early childhood Title I programs. Cambridge, MA: The Huron Institute, 1979.