A LONGITUDINAL STUDY OF METROPOLITAN ACHIEVEMENT TEST SCORES

FINAL THESIS

Submitted to the Graduate Committee of the
Department of Education and Human Development
State University of New York
College at Brockport
in Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Education

by
Eric W. Purdy
State University of New York
College at Brockport
Brockport, New York

November 6, 1989

SUBMITTED BY:

Eric Purdy — 11/6/89
Candidate — Date

APPROVED BY:

Robert C. Pibble — 11/7/89
Thesis Advisor — Date

Barbara M. Lorenz — 11/6/89
Second Faculty Reader — Date

Director of Graduate Studies — Date

TABLE OF CONTENTS

ABSTRACT

This longitudinal study compared how well students scored on the Metropolitan Achievement Test (sixth edition) (MAT6) in third grade, fourth grade, and fifth grade. Thus, the study examined the scaled scores of 50 randomly selected students over a three year period. The comparison of performance means between grades, and between genders at each grade, were drawn by conducting statistical t-tests. The three subtests of the MAT6 that were used in this study were; total reading, total math, and social studies. Of the 50 students selected, 28 were boys and 22 were girls. It was found that there was a substantial statistically significant increase of scores from third grade to fourth grade among the 50 students on all three subtests (reading t=9.021, math t=8.886, social studies t=7.979). However, there was no statistically significant difference between fourth grade scores and fifth grade scores on any of the subtests. It was also found that there was no statistically significant difference between the boys' scores and the girls' scores on any of the three subtests at any grade level. It was concluded that the statistically significant increase in fourth grade reading, math, and social studies scores from third grade could be due to a variety of factors.

The conclusion discusses such possibilities as "test coaching", a more closely aligned fourth grade curriculum with the MAT6 format than that of third or fifth grade, and cognitive developmental factors.

CHAPTER I. INTRODUCTION

The objective of this study was to compare how well students scored on the reading, math, and social studies subtests of the Metropolitan Achievement Test (sixth edition) (MAT6) over a three year period. The 50 randomly selected students proceeded through each grade together and had their scaled scores on the three subtests recorded at third grade, fourth grade, and fifth grade. The study investigated whether there would be any statistically significant increase or decrease in scores from one grade to the next. It also compared boys' scores to girls' scores at each grade level to determine if there was any statistically significant difference between genders. To compare performance on the subtests from grade to grade, and between boys and girls at each grade level, statistical t-tests were conducted. This was done in order to obtain a t-value which indicated whether or not there was any statistically significant difference between the propositions tested.

Due to the fact that the foundation of this study is the MAT6, it is necessary to discuss what some of the current issues are surrounding the use of standardized tests in our schools. Different groups have different things to say about the increasing reliance upon standardized tests to evaluate the status of students,

teachers, administrators, and schools in general.

Howards (1987) saw the accountability of teachers, schools, and administrators as necessary, although he argued that "a means of insuring accountability has not yet been discovered."(p.1) Howards proclaimed that, "testing often lacks fairness, objectivity, and appropriateness and is rarely apolitical"(p.1). He also suggested that, "Research on most standardized tests shows that they are incompatible and incomparable, that is they don't measure the same thing the same way"(p.8).

Farr and Carey are quoted from Howards' (1987) article in a summary of how incompatible the most commonly used standardized reading tests are. They compared the (MAT), the California Test of Basic Skills (CTBS), California Achievement Test (CAT), the Iowa Test of Basic Skills (ITBS), and the Stanford Achievement Test (SAT): They stated, "The MAT measures sight words and so does the SAT, the CTBS and others do not; MAT measures visual discrimination of letters and words, and so does the CAT, but none of the others does; CTBS, CAT, and SAT measure syllabication, the others do not; recognition of root words is measured only by the CTBS; rhyming words is included in the CTBS and ITBS, and so it goes. The lack of consistency across the tests does not stop with these differences. Often there is variation as to when certain

skills are tested"(p.20).

Yatvin (1987) elaborated on the care that goes into standardized tests by stating, "It is clear to most school people that the tests are seriously flawed. This is hardly surprising since test makers face at least two insurmountable obstacles: (1) what each school teaches is different from what any other school teaches; and (2) ways of learning cannot be reproduced on a machine scored test"(p.86).

Another major reason for why the use of standardized tests to try and measure academic performance is criticized is because of cultural bias. Neill (1989) claimed that, "Because of class and cultural bias in the tests, children who are not white and middle or upper-class are disproportionately judged as 'not ready' or 'not gifted' or they are otherwise penalized for their background. The essential danger of decisions based on test scores is that too often racial minority and low-income children are placed in programs that virtually guarantee that they will never obtain even a decent education"(p.11).

Sledd (1986) added to this by stating, "With an increasingly larger percentage of the population becoming foreign speaking, it is no wonder that standardized test scores have fallen over the years. Quite simply, it is a

gross injustice to command a mastery of standard English from students who, through no fault of their own, have had no chance to master it"(p.28).

Many standardized tests are often used in conjunction with, or are supposedly reflective of what is included in, a given school district's curriculum. The original Metropolitan Achievement Test (MAT) was developed by the Psychological Corporation and dates back to the early 1930s when the test was designed to meet the curriculum needs of New York City. Later editions of the test were expanded to better reflect a more national curriculum (Balow, et al. 1985). According to Balow, "The MAT provides an overall measure of the whole content in a curriculum area, but in these reviewers opinion, a stronger tie to the specific curriculum should be taken into consideration. Although the MAT has been developed on a representative, general curriculum, there is no reason to assume that it includes the principal program that is to be evaluated"(p.430).

The statement by Balow seems to indicate that the MAT and other tests are often used to evaluate a system of education whose curriculum is not tied to that of the MAT. Yatvin (1987) sees this exemplified in many ways. She stated, "The tests under consideration in our district emphasized fractions at fifth grade, our

curriculum emphasizes them at sixth. At the heart of our science program are observation, recording data, making hypotheses, and drawing conclusions; the tests care only for the facts. In reading we think it important to predict, summarize, and interpret, none of those skills are included in the tests"(p.88).

Neill (1989) elaborated on this further by indicating, "The standard nature of standardized multiple-choice exams precludes their use as appropriate tools for shaping the curriculum. Only the basic, the simple and the trivial can be measured with these instruments-and even then, only narrowly"(p.11)

It is claimed by many experts that the increasing use of standardized testing in schools is adversely effecting the way students are taught by teachers. More and more instructional time seems to be spent on how to take a test rather than actually learning something.

Prell (1987) pointed out that, "The relation between time and test results is not a direct linear relationship, but a logarithmic one. One will observe diminishing returns with increasing investments of time spent on preparing for a test." Prell also stated, "Coaching is a technique which means training children how to answer specific types of questions and providing them with information about a test"(p.2). Quoted from Prell's (1987) article is

James Guines, Associate Superintendent for Instruction in Washington D.C.. He stated, "We want all our children to do well. We'll even begin to tell them some of the things on the test"(p.2).

Neill (1989) criticized test coaching by stating, "In preparing students to do well on a test, teachers divert educational time and energy from the higher order curriculum, as well as from non-academic efforts. Several major reports issued during the past year all concluded that students in the U.S. are not developing higher order thinking skills. Research shows that the methods commonly used to raise standardized test scores-drill, memorization, learning by rote, and repitition-are counterproductive to teaching higher order thinking skills"(p.11).

Richards (1989), a former teacher who retired early because of the increasing emphasis being placed on standardized testing in his school stated, "A change came when my district's competition for high standardized test scores reached maniacal proportions. Teachers told about how teaching to the test left little time or energy for real teaching. Consequently they said, their students write less, think less, read fewer books, and are passive rather than active participants in learning" (p.65).

Nirth (1988) summed things up very well when he

indicated, "In the final analysis, institutions like schools cannot become good work places until they support the power of the people to learn through 'conjoint communicated living'. The prevalence of standardized testing does not support this philosophy"(p.7).

Despite the cry from teachers, administrators, and educational intellectuals to put standardized testing in the shadows rather than the spotlight of education, there remains a group of people in favor of more standardized testing. That group of people is the politicians. One public official was quoted by Schecter (1981) at a symposium on education as stating, "What if we could some how move away from competency testing, even now? Who then would be disadvantaged? Lawyers and testmakers would surely suffer, but who else? Students, especially minorities and the handicapped would surely suffer because they wouldn't be as aware of their problems. Educators would be at a disadvantage because they would still be feeling public pressure and not be able to give so clear a response. And taxpayers would be left with out being able to judge value received"(p.50).

Power (1986) indicated, "Most of the testing controversy these days centers on standardized tests for practicing teachers, as reform minded governors and legislators sound the call for teachers to prove their

competency. Instead of focusing on substantive issues - say, lowering class size- they're coming up with these quick-fix schemes that appeal to the media. Too many politicians still don't understand that a paper-and-pencil test can measure only knowledge, not competence. Never Competence"(p.3).

On a more positive note, it seems that the testmakers are beginning to become more aware that better tests are necessary, and less emphasis should be placed upon them. The President of the Educational Testing Service (ETS), Gregory Arnig, in an interview with U.S. News & World Report (1986) stated, "We are focusing too much on standardized testing. We forget that the most common exam is the one prepared by the teacher and given in the classroom every week or two. Teachers are always trying to see how much their students are learning. But that does not fully satisfy everyone because teachers tests are not standardized"(p.84).

In an effort to make better tests, the ETS recently joined forces with the National Education Association (NEA) to work on the problem. Jenkins (1986) outlined the coalition's plans to improve currently used standardized tests by stating, "Their approach to testing incorporates a range of 10 types of test items that attempt to measure -more accurately than do the

traditional essay, multiple-choice, and true/false questions- what information and skills the students have learned. Models of new tests include questions asking students to arrange items or events in rank order, select the best choice from several correct answers, complete a matrix showing the relationship between two variables, and so on"(p.25).

In his interview with U.S. News & World Report (1986), Arnig elaborated on how computers will help to make better tests. He indicated, "We have created the prototype of diagnostic tests where a pupil sits down at a console and a computer program helps the youngster find out how well he or she is doing. Tests are individualized so he or she proceeds until a question is answered wrong, at which point the tests show the child what is being done wrong and helps them to correct their mistakes." When asked if we'll see an increased use of testing in the future, Arnig responded by stating, "If we do, I hope we see more testing that helps people and not the kind that is used to make decisions about them. A lot of time and money is being spent on testing, and we must be sure that investment is paying off with improved learning opportunities. People want to raise standards and performance, but in using tests to reform schools, I think they're using a very dull meat-ax to do it"(p.84).

Hopefully, better standardized tests will be administered in our schools in the not-to-distant future. However, until testmakers do come out with better tests, it is probably wise not to read too much into the scores that are recorded or the trends that they may indicate.

As far as this study is concerned, it was expected that the 50 students scaled scores on the reading, math, and social studies subtests of the MAT6 would improve as the students became older. In other words, it was expected that scores would increase from third to fourth grade and again from fourth to fifth grade. The logic behind this expectation revolves around the assumption that each year students would become more familiar with the types of questions that the MAT6 asked and the overall format of the test. It was also expected that the girls would do significantly better than the boys on all three subtests at each grade level. This can be explained by the fact that there was a general consensus among teachers who have taught these 50 students that the girls are more academically inclined than the boys.

CHAPTER II. METHODOLOGY

SUBJECTS:

The 50 students (28 boys and 22 girls) selected for this longitudinal study were chosen randomly. The subjects, in general, shared relatively homogeneous cultural, demographic, and socio-economic characteristics. They all attended a small Central School (composed of high school students and elementary students in the same building) in the rural Southern Tier of Upstate New York. The population is primarily white middle-class and the major industry is agriculture. The average age of the students in third grade was eight, in fourth grade the average age was nine, and in fifth grade the average age of the students was ten.

PROCEEDURE:

The first step of this study involved a collection of the data. The scaled scores of the 50 subjects on the total reading, total math, and social studies subtests were recorded from a computer spreadsheet. The spreadsheet indicated the scaled scores of all 50 subjects from when they were in first grade up through

the time the subjects completed fifth grade. The third grade scores, fourth grade scores, and fifth grade scores were recorded for this study.

After the data was collected it was entered into a computer. Statistical t-tests were conducted which produced a t-value. This t-value indicated whether or not there was any statistical significance between the propositions tested and allowed the researcher to draw conclusions.

CHAPTER III. RESULTS

The primary results of this study indicated that, when comparing third grade scores to fourth grade scores, there was a fairly large statistically significant increase in the fourth grade scores on all three subtests. However, there was no statistically significant difference between fourth grade scores and fifth grade scores on any of the subtests. When comparing boys to girls in the same grade, there was no statistically significant difference between the two genders on any of the subtests at any of the three grade levels.

The following t-tests highlight in detail the results of this study.

COMPARISON OF GRADE THREE READING ACHIEVEMENT MEANS
TO GRADE FOUR READING ACHIEVEMENT MEANS

2-SAMPLE T-TEST

GROUP:	READ3	READ4
SIZE:	50	50
MEAN:	607.36	639.8201
SD:	55.935	45.616

T-VALUE: 9.021
DF: 49
2-TAIL PROB. <.000

ETA SQUARED: .624

COMPARISON OF GRADE FOUR READING ACHIEVEMENT MEANS
TO GRADE FIVE READING ACHIEVEMENT MEANS

2-SAMPLE T-TEST

GROUP:	READ4	READ5
SIZE:	50	50
MEAN:	639.821	646.999
SD:	45.616	52.829

T-VALUE: 1.983
DF: 49
2-TAIL PROB .053

ETA SQUARED: .074

COMPARISON OF GRADE THREE MATH ACHIEVEMENT MEANS
TO GRADE FOUR MATH ACHIEVEMENT MEANS

2-SAMPLE T-TEST

GROUP:	MATH3	MATH4
SIZE:	50	50
MEAN:	600.900	630.560
SD:	34.450	33.649

T-VALUE:	**8.866**
DF:	49
2-TAIL PROB:	<.000
ETA SQUARED:	.616

COMPARISON OF GRADE FOUR MATH ACHIEVEMENT MEANS
TO GRADE FIVE MATH ACHIEVEMENT MEANS

2-SAMPLE T-TEST

GROUP:	MATH4	MATH5
SIZE:	50	50
MEAN:	630.560	635.68
SD:	33.649	32.299

T-VALUE	**1.905**
DF:	49
ETA SQUARED:	.068

COMPARISON OF GRADE THREE SOCIAL STUDIES ACHIEVEMENT MEANS
TO GRADE FOUR SOCIAL STUDIES ACHIEVEMENT MEANS

2-SAMPLE T-TEST

GROUP:	SOC3	SOC4
SIZE:	50	50
MEAN:	591.000	617.259
SD:	35.242	33.391

T-VALUE	**7.979**
DF:	49
2-TAIL PROB:	<.0001
ETA SQUARED:	.565

COMPARISON OF GRADE FOUR SOCIAL STUDIES ACHIEVEMENT MEANS
TO GRADE FIVE SOCIAL STUDIES ACHIEVEMENT MEANS

2-SAMPLE T-TEST

GROUP:	SOC4	SOC5
SIZE:	50	50
MEAN:	617.259	616.199
SD:	33.391	41.890

T-VALUE	**.296**
DF:	49
2-TAIL PROB:	.768
ETA SQUARED:	.001

COMPARISON OF MALE/FEMALE ACHIEVEMENT MEANS FOR GRADE THREE
MALE=1 AND FEMALE=2
READING

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	600.107	616.590
SD:	61.713	47.380

F-RATIO (VAR):	1.695
DF:	27, 21
2-TAIL PROB:	.217
T-VALUE:	**-1.696**
DF:	48
2-TAIL PROB:	.305
OMEGA SQUARED:	.001
ETA SQUARED:	.021

MATH

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	598.464	604
SD:	36.716	31.907

F-RATIO (VAR):	1.324
DF:	27, 21
2-TAIL PROB:	.513
T-VALUE:	**-.560**
DF:	48
2-TAIL PROB:	.578
OMEGA SQUARED:	-.OI3
ETA SQUARED:	.006

SOCIAL STUDIES

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	586.607	596.590
SD:	36.480	33.593
F-RATIO (VAR):	1.179	
DF:	27	21
2-TAIL PROB:	.704	
T-VALUE:	**-.994**	
DF:	48	
2-TAIL PROB:	.325	
OMEGA SQUARED:	.000	
ETA SQUARED:	.020	

COMPARISON OF MALE/FEMALE ACHIEVEMENT MEANS FOR GRADE FOUR
MALE=1 AND FEMALE=2
READING

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	637.214	643.136
SD:	50.631	39.226
F-RATIO (VAR):	1.666	
DF:	27	21
2-TAIL PROB:	.233	
T-VALUE	**-.451**	
DF:	48	
2-TAIL PROB:	.653	
OMEGA SQUARED:	-.016	
ETA SQUARED:	.004	

MATH

2-SAMPLE T-TEST

SUBSET #	1		2
SIZE:	28		22
MEAN:	624.857		637.818
SD:	35.558		30.294
F-RATIO (VAR):		1.377	
DF:	27		21
2-TAIL PROB:		.454	
T-VALUE:		**-1.363**	
DF:		48	
2-TAIL PROB:		.179	
OMEGA SQUARED:		.016	
ETA SQUARED:		.037	

SOCIAL STUDIES

2-SAMPLE T-TEST

SUBSET #	1		2
SIZE:	28		22
MEAN:	613.392		622.181
SD:	37.556		27.256
F-RATIO (VAR):		1.898	
DF:	27		21
2-TAIL PROB:		.135	
T-VALUE:		**-.922**	
DF:		48	
2-TAIL PROB:		.360	
OMEGA SQUARED:		-.002	
ETA SQUARED:		.017	

COMPARISON OF MALE/FEMALE ACHIEVEMENT MEANS FOR GRADE FIVE
MALE=1 AND FEMALE=2
READING

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	650.607	642.409
SD:	63.710	35.411
F-RATIO (VAR):	3.236	
DF:	27	21
2-TAIL PROB:	.008	
T-VALUE:	**.540**	
DF:	48	
2-TAIL PROB:	.591	
OMEGA SQUARED:	-.014	
ETA SQUARED:	.006	

MATH

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	631.785	640.634
SD:	36.783	25.471
F-RATIO (VAR):	2.085	
DF:	27	21
2-TAIL PROB:	.341	
T-VALUE:	**-.961**	
DF:	48	
2-TAIL PROB:	.341	
OMEGA SQUARED:	-.001	
ETA SQUARED:	.018	

SOCIAL STUDIES

2-SAMPLE T-TEST

SUBSET #	1	2
SIZE:	28	22
MEAN:	612.249	621.227
SD:	48.383	32.209

F-RATIO (VAR):	2.256	
DF:	27	21
2-TAIL PROB:	.059	
T-VALUE:	-.748	
DF:	48	
2-TAIL PROB:	.457	
OMEGA SQUARED:	-.008	
ETA SQUARED:	.011	

CHAPTER IV. CONCLUSIONS

The results of this study indicated that there was not a statistically significant increase in scores from grade to grade. Nor was there a statistically significant difference between boys and girls at any grade level. Thus, both original hypotheses were not supported by the findings.

The only statistically significant finding indicated an important increase in scores on the subtests from third grade to fourth grade. There could be many reasons for why there was a such a significant increase in scores from third to fourth grade and no significant increase from fourth to fifth grade.

The age of the students may have something to do with it. Going from third to fourth grade could be a time period when most youngsters acquire and retain more information as opposed to the time period going from fourth to fifth grade.

Another possible reason that we see a statistically significant increase in scores from third to fourth grade, and not from fourth to fifth grade, could be because of curriculum alignment. In other words, the fourth grade teachers in the school may have taught material which paralleled the material on the MAT6, while

the third and fifth grade teachers may not have.

"Test coaching" could also be a an explanation. The fourth grade teachers may have taken the test very seriously, and in an attempt to better the students scores, gone over the types of questions that would be asked on the test shortly before administering it. The third and fifth grade teachers may have neglected to do this.

The reliability of the test itself may also be a factor contributing to this finding. Although the MAT6 is considered to be one of the better standardized achievement tests on the market, there seems to be a general consensus among those in education that all standardized tests need improvement in many areas. Thus, the ability of the MAT6 to yield consistent results should be questioned.

The second hypotheses predicted that the girls would do better than the boys on the subtests at all three grade levels. However, this was not supported by the findings which indicated that there was no statistically significant difference between boys and girls. There could be a number of reasons for why the boys and girls did about the same on the tests at every grade level.

One reason could be that there really is no significant difference between the knowledge level of the boys and

the girls in reading, math, and social studies. The teachers at this school may have said that the girls are more apt to do better on the tests because of certain things which the MAT6 does not measure. The girls may indeed be more academically inclined, but what is the definition of academically inclined? The teachers at this school may have thought that the girls would do better than the boys because they seem to be more active participants in class and more eager to learn. The girls may even do better on the classroom tests that the teachers administer. Howvever, the MAT6 may not measure these factors. This could contribute to the fact that there was no statistically significant difference between the two genders.

It would seem however, that if a group of teachers unanimously agreed that the girls were better, academically speaking, than the boys in school, then there should be a significant difference between the knowledge level of the two genders. The fact that this study does not indicate this could also have something to do with the construction of the MAT6. The test may not be measuring what students actually know or can do and what they don't know and can't do. Based on what many of the experts are saying about the validity of standardized tests, it would not be surprising if this were the case.

REFERENCES

Balow, I. H., Farr, R., Hogan, T. P., Prescott, G. A. (1985). Test critiques. Test Corporation of America. p.430.

Howards, M. (1987). Testing: The illusions of measurement. (ERIC Document Reproduction Service No. ED 300 393)

Jenkins, K. C. (1986, November). NEA, ETS join forces to help teachers write better tests. NEA Today, p. 25.

Neill, D. M. (1989). It's time to end the misuse and overuse of standardized tests. Journal of New York State School Boards Association, Incorporated. 11.

Power, J. (1986 January-February). The good news about testing. NEA Today, p. 3.

Prell, J. M. & Prell, P. A. (1986). Improving test scores--teaching test wiseness. A review of the literature. Bloomington, IN. Phi DElta Kappa, Center on Evaluation, Development, and Research. (ERIC Document Reproduction Service No. ED 280 900)

Richards, T. S. (March, 1989). Testmania: The school under siege. Learning, p. 65.

Schechter, J. (1981). Issues of competency and accountability Austin, TX. The Proceedings of an Invitational Symposium. (ERIC Document Reproduction Service No. ED 208 569)

Tests can't solve every problem. (May, 1986). U.S. News & World Report, p. 84.

Sledd, J. (1986). A basic incompetence in the defining of the basic competencies. English Journal, 26-28.

Wirth, A. G. (1988). Towards a post-industrial intelligence: Gadmer and Dewey as guides. New Orleans, LA. Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 298 613)

Yatvin, J. (1987). Playing the testing game. Educational Leadership, 88.