

University of Montana

ScholarWorks at University of Montana

Graduate Student Theses, Dissertations, &
Professional Papers

Graduate School

2000

Literacy development within multiage and single grade age cohorts: The impact of organizational structure

Leslie J. Ferrell

The University of Montana

Follow this and additional works at: <https://scholarworks.umt.edu/etd>

Let us know how access to this document benefits you.

Recommended Citation

Ferrell, Leslie J., "Literacy development within multiage and single grade age cohorts: The impact of organizational structure" (2000). *Graduate Student Theses, Dissertations, & Professional Papers*. 10603. <https://scholarworks.umt.edu/etd/10603>

This Dissertation is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

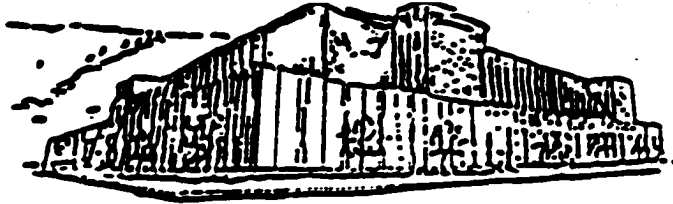
In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]



Maureen and Mike
MANSFIELD LIBRARY

The University of **MONTANA**

Permission is granted by the author to reproduce this material in its entirety, provided that this material is used for scholarly purposes and is properly cited in published works and reports.

*** Please check "Yes" or "No" and provide signature ***

Yes, I grant permission Yes
No, I do not grant permission

Author's Signature Leslie J. Ferrell

Date 17 June 2000

Any copying for commercial purposes or financial gain may be undertaken only with the author's explicit consent.

LITERACY DEVELOPMENT
WITHIN MULTIAGE AND SINGLE GRADE AGE COHORTS:
THE IMPACT OF ORGANIZATIONAL STRUCTURE

by

Leslie J. Ferrell

B.A. The University of Montana, 1971

B.A. The University of Montana, 1983

M.Ed. The University of Montana, 1993

presented in partial fulfillment of the requirements

for the degree of

Doctor of Education

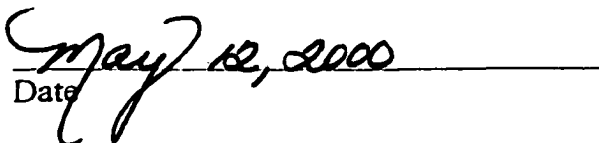
The University of Montana

Spring 2000

Approved by:


Chairperson


Dean, Graduate School


Date

UMI Number: 9971867

**Copyright 2000 by
Ferrell, Leslie Jayne**

All rights reserved.

UMI[®]

UMI Microform9971867

Copyright 2000 by Bell & Howell Information and Learning Company.

**All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.**

**Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346**

Literacy Development Within Multiage and Single Grade Age Cohorts: The Impact of Organizational Structure (235 pp.)

Chair: Dr. Marian J. McKenna



This combined design quasi-experimental study used multiple methods of assessment to compare and explore the impact of multiage and single grade organizational structure upon the literacy development of upper elementary students. Disaggregated in age cohorts of 8-, 9-, 10-, and 11-year olds, 235 students' test data from 10 grades 3-5 classrooms from two Title I schools within the same public school district were analyzed for statistically significant differences in literacy achievement. The control school was single grade only; the experimental multiage only, with the exception of kindergarten and one fifth grade. Two standardized test scores for reading and language from a Spring achievement test were analyzed using a t test. Two standardized test scores in reading and language from a criterion-referenced test were analyzed using an analysis of covariance, with students' Fall pretest score as covariate. Fluency and conventions from 244 writing samples were assessed by two trained 3-rater teams using a modified holistic scale, with a t test analysis.

Out of 28 separate statistical tests by age cohorts, 5 indicated a statistically significant difference at the .05 level. Two favored multiage cohorts: Cohort 8 in reading on the national standardized test, and Cohort 9 in reading on the criterion test. Three favored single grade cohorts in writing: Cohort 9 in fluency and conventions, and Cohort 10 in fluency. All experimental differences on the indirect measures were less than 5%. No consistent pattern emerged favoring either structure.

Qualitative observations regarding the instructional policies and programs of each school were made from interviews and documents. Emergent themes dealt with (a) historical origins: changes in structure to ameliorate behavior and academic problems; (b) leadership: collaboration among principal and teachers necessary for success; (c) meeting students' needs: assessment-driven instruction from goals; and (d) commonalities of experience: policies, programs, and practices were more similar than different, including early intervention in reading, homogeneous grouping by ability for skills' instruction, and no differentiated teacher training.

Overall, comparable literacy development was indicated. Thus, the classroom's organizational structure may be an inconsequential variable when structuring classrooms for improved academic achievement, but 12 out of 28 effect sizes $> .33$ warrant further study. Specific instructional policies and practices may account more strongly for literacy development among students with characteristics similar to this nonrandom sample.

ACKNOWLEDGMENTS

To the five University of Montana members of my dissertation committee:

Dr. Rhea Ashmore, Dr. Beverly Chin, Dr. Marian McKenna, Dr. Kelly Ward, and Dr. Stephanie Wasta, thank you for your unwavering support and mentorship. To Dr. Merle Farrier, thank you for your expert teaching and counsel in statistics. The individual guidance and expertise that I received from each of you will remain as testimony to the professional collegiality so necessary for learning. A special thank you to my advisor, Dr. McKenna, whose generosity toward her students is a truly unique experience for which I will be always grateful.

To those public school teachers across the United States who believe that each child is due a free and equal opportunity in education, thank you. To those of you who seek to question and research and improve us, I join you. Accepting the challenges daily, we make this democratic ideal happen.

To my husband, Clark, and children, Matthew and Jessica, I give my love and gratitude for your help in this endeavor. You three are all that make the rest of it worthwhile.

TABLE OF CONTENTS

ABSTRACT.....	ii
TABLE OF CONTENTS.....	iv
LIST OF TABLES.....	vi
LIST OF FIGURES.....	viii
I. INTRODUCTION.....	1
Statement of the Problem.....	4
Purpose of the Study.....	6
Research Questions.....	7
Significance of the Study.....	10
Definition of Terms.....	14
Limitations of the Study.....	17
Delimitations of the Study.....	19
II. REVIEW OF THE LITERATURE.....	21
Historical Background.....	21
Empirical Research 1960 - 1998.....	28
Theories of Learning.....	39
III. METHODOLOGY.....	46
Research Design.....	46
Quantitative Components.....	52
Qualitative Components.....	67
Standards for Quality of Conclusions.....	69

IV. DATA ANALYSIS.....	92
Quantitative Components.....	92
Qualitative Components.....	137
Summary.....	167
V. DISCUSSION.....	170
Findings.....	172
Conclusions.....	174
Implications.....	176
Recommendations.....	177
For Practice.....	177
For Future Research.....	179
REFERENCES.....	182
APPENDIXES.....	197

LIST OF TABLES

Table	Page
1. T test Results for TerraNova Reading by Age Cohort and Level Test.....	94
2. T test Results for TerraNova Language by Age Cohort and Level Test.....	98
3. Descriptive Statistics of Fall MALT Scores in Reading by Age Cohort.....	102
4. ANCOVA Summary for Cohort 9.....	106
5. ANCOVA Summary for Cohort 10.....	107
6. ANCOVA Summary for Cohort 11.....	108
7. Descriptive Statistics of Fall MALT Scores in Language by Age Cohort.....	111
8. ANCOVA Summary for Cohort 8.....	112
9. ANCOVA Summary for Cohort 9.....	113
10. ANCOVA Summary for Cohort 11.....	114
11. T test Results for Fluency in Writing by Age Cohort.....	117
12. Pre- to Post Fluency Scores Within Age Cohorts as Reported by Number of Students.....	120
13. T test Results for Conventions in Writing by Age Cohort.....	122
14. Pre- to Post Conventions Scores Within Age Cohorts Reported by Number of Students.....	124
15. Individual Students That Received an Increase or Decrease in BOTH Fluency and Conventions.....	125
16. Teacher Responses to Writing Assessment Questionnaire.....	131
17. Summary of Quantitative Results for Each Measure in Each Cohort.....	135

18. Summary of Length of Time of First Multiage Grouping Combined with Schoolwide Title I Interventions Up to Date of Research.....	162
19. CTBS Fourth Grade Summary of Building Data from 1991-1998.....	165
20. Comparison of Percentage of Free and Reduced Lunch from 1992-1998.....	165

LIST OF FIGURES

Figure	Page
1. Options for School Organizational Structure.....	16
2. Schemata of Research Design.....	205

CHAPTER 1

INTRODUCTION

Background of the Problem

Educational Reform

A nation at risk, schools in crisis, Johnny can't read...from over backyard fences to the Internet, we constantly scrutinize public education. In this nation that promises equal opportunity for each child, educational reform is ongoing. For some parents, retreat rather than reform is their solution. An increasing number are considering private or home schools for their children. For example, over 6,000 Cleveland families applied for vouchers which would allow their children to attend private schools rather than public schools (Gergen, 1996). In addition, as the estimate of K-12 homeschoolers has passed the one million mark, homeschooling is now recognized as a growing mainstream alternative (Archer, 1999; Pearson, 1996; Pulliam & Van Patten, 1995; Ray, 1996).

Why this retreat when America's public school system is replete with success stories? The American economy continues to be the strongest in the world. As a pluralistic society, U.S. immigration continues and has risen rapidly since 1980 (Bracey, 1996). The United States has educated the most diverse population in history. In 1993-94, one in three K-12 students were of minority racial-ethnic descent. With more Americans completing more years of schooling than ever before, the United States leads industrialized nations in terms of educational opportunity (Robinson, 1997). Public education has been recognized as a vital factor in these achievements.

Yet public education is in the midst of crisis, and many criticize present practice and policy. Is withdrawal to private, charter, and homeschools symptomatic of the failure of public education to answer reform demands? Pulliam and Van Patten (1995) state that private education, which is increasing in popularity, is "very traditional [with] few radical or innovation programs as of 1986" (p. 212). Could it be the type of reform, not the lack of it, which causes retreat from the public school system? For whichever reason, such withdrawals undermine public education in several ways. Immediately, it results in a monetary loss to public education which is funded according to number of students enrolled. In addition to loss of income, Comer (1997) argues there is a loss of diversity and thus, a loss of opportunity to gain understanding and mutual respect. If these losses continue, the effect upon public education and its promise for each child will be dramatic. If the factors contributing to this flight cannot be changed, we will compromise the American ideal of free and equal opportunity for all children.

Compounding the above issues, Berliner and Biddle (1995) declare that much of what is presented as evidence about education is misleading, inconclusive, or inaccurate. This type of evidence may lead to movements for poor, or unnecessary, reforms. When reform ideas are raised, to whom does the system listen? Which type of reform? How far should it be carried? In which direction? According to Drucker (1994), the "performance of schools...will be of increasing concern to society as a whole, rather than being considered professional matters that can be safely left to 'educators' " (p. 66). Goodlad (1984) concurs that "education is too important...to be left to the schools" (p. 46). In addition, Comer (1997) argues that demand for reforms through vouchers,

charter, and magnet schools is due, in part, because "children of the socially marginal are being denied even minimal learning conditions" (p. 295). So demands stem not just from criticisms of educational practice and policy, but according to Tanner (1993), from the "deteriorating social and economic conditions on the physical, mental and emotional well-being of children" (p. 295). Drucker, Goodlad, Comer, and Tanner agree that schools alone cannot solve these problems.

At the same time, Shannon (1994) asserts that "the school board, once the epitome of representative governance in our democracy, is undergoing profound change" (p. 387). For example, even in large school districts where bureaucratic central authority exists, parents and business stakeholders demand and bring about change. When diverse groups come together, collaboration provides a way to reach a common direction. To facilitate decisionmaking, schools need to be accountable through a variety of data. This study was predicated upon the idea that "our educational policies and practices must be based on the fullest available evidence so as to serve our deepest, widest, and highest social ideals" (Tanner, 1993, p. 297). Free and equal opportunity of education is a democratic ideal. Democracy cannot function without effective public schools. Without effective public schools we are truly then a nation at risk.

The Nature of Change in Education

Foundational to research in education is the question of how children learn best. While many schools have improvement goals and have begun to promote partnerships that increase parental and community involvement, Gipe (1992) reports that of 211 schools in the Northwest, approximately 50% have no current formal assessment of

curricular practices. In 1979 Goodlad stated that "we lack the base of knowledge required for comparing current school practices with alternatives... and for determining the precise changes that might prove helpful" (p.102).

As an educational researcher, Goodlad has investigated and promoted alternatives within organizational structure since 1959. In 1987 he stated "studies comparing graded and nongraded schools, taken as a group, are inconclusive" (p. 218). In terms of school structure, where does this leave parents who want the best for their child? Where does this leave teachers who want to instruct students in a way that will effect the greatest individual achievement for each student? Where does this leave administrators and school board members who must make a myriad of decisions regarding school practices while beset with financial limitations? Goodlad (1979) believes that "collaboration within the profession and between the school and community may be necessary for school improvement...and gathering data could be a good place to begin the necessary collaboration" (p. 103).

Dewey's (1916) "habits of mind which secure social change without introducing disorder" (p. 115) demand such collaboration. To consider change without disorder means that information must be available early and ongoing. Access to timely and understandable data must provide stakeholders time to review, collaborate, and make informed decisions about their issue.

Statement of the Problem

All of the challenges of educational reform and change were present in the issue of organizational structure of classrooms. Glickman (1998) states "there is no single

issue more controversial in public schools than how students are placed and grouped in schools and classrooms...homogeneous or heterogeneous? Horizontal or vertical?" (p. 46). Of these options, grouping children by the same age is called the graded classroom. This structure has been predominant for 150 years (Goodlad & Anderson, 1987). One of the alternatives is the mixed-age grouping called the multiage classroom. According to Davis (1992), the nongraded, or multiage, classroom has become a key element in reform, particularly for primary students, but increasingly for older students. Glickman says that this issue of systems of grouping children for learning polarizes people and has been met by "vehement resistance" (p. 47) from different stakeholders.

Similarly, requests for change in organizational structure from single grade to multiage classrooms had created uncertainty and dissension within the local district ("Committee reports," 1997). While there were ardent, sincere proponents on both sides of the issue, what we knew seemed confused. As the literature review shows, research on organizational structure exists, but contains equivocal findings, was dated, and provided little information above primary level (see Appendix A). Proponents of alternatives stated that the relevance of past research to today's nongraded or multiage classroom was questionable (Goodlad & Anderson, 1987; Gutiérrez & Slavin, 1992; Kasten & Clarke, 1993). Also, the terms nongraded, multigrade, multiage, and others have been used interchangeably which causes further confusion because they are not the same (see Definition of Terms, Literature Review, and Appendix B). Data were needed on academic achievement from multiple sources within clearly defined organizational structures to understand what makes a difference in literacy development.

Current information on brain development and learning further complicated this question. Research from several fields suggested children may have cognitive needs that were different from those of previous decades. Healy (1990) states that "subtle, but significant changes" [in the brain affect learning and that these] "fundamental shifts put children in direct conflict with traditional academic standards and methods...particularly at risk are abilities for language-related learning" (p. 46). She argues that alternatives to old school structures have "potential merit and potential problems. If what children get in school is ineffective or even damaging, simply adding more of the same will only exacerbate the problems" (p. 282). Therefore, information was needed as to how each organizational structure best supports learning and its impact on literacy development.

When educators do not or cannot satisfy parents' requests, reactions range from indifference to withdrawing their children to private or homeschools. When educators cannot agree, collegiality and school efficacy are threatened. When administrators and school board members face a controversial issue, they risk polarization that could impede action in the best interest of students. To address diverse concerns, all stakeholders must be able to compare and contrast through multiple types of data. An in-depth investigation of how classroom structure supports student literacy learning provides a broader basis for decisionmaking regarding organizational structure.

Purpose of the Study

The purpose of this combined design study was to delineate the impact of two different organizational structures-multiage and single grade classrooms-upon the literacy development of upper elementary students. In this study, literacy was defined

as "the capacity to accomplish a wide range of reading, writing, speaking and other language tasks associated with everyday life" (National Council of Teachers of English [NCTE] & International Reading Association [IRA], 1996, p. 139) and "requires active, autonomous engagement with print" (Venezky, 1995, p. 19). Through separate and distinct quantitative data sources, reading and language achievement were analyzed, with differences among the test measures integral to the analysis. Through interviews and document analysis, qualitative data were explored. Through triangulation of data, this study's combined design investigated how each structure supports literacy as reported by multiple methods of assessment. This study analyzed all available evidence in order to understand the nature of and make informed choices about the impact of organizational structure upon students' literacy growth.

The fundamental assumption in the purpose of this study was that:

collection, analysis and utilization of data...[is] the heart of professionalism. When schools embrace data-based decisionmaking as a school-improvement tool, they make measurable progress in attaining their objectives. They are able to plan next steps in such critical areas as creating small communities for learning, strengthening the core academic program, and reconnecting schools and communities based upon verified performance. (Lipsitz, Mizell, Jackson, & Austin, 1997, p. 536)

Overarching Research Questions

In the quantitative component, this study addressed three questions regarding students' growth in literacy, specifically reading comprehension and language composition:

1. To what degree does organizational structure impact student academic achievement on a standardized, norm-referenced achievement quantitative measure?

2. To what degree does organizational structure impact student academic achievement on a standardized, criterion-referenced district quantitative measure?

3. To what degree does organizational structure impact student writing development as demonstrated by a performance assessment of pre- and post writing samples?

Based on the first three broad research questions, specific research questions were narrowed to the following four questions. Because age configuration is an integral difference, disaggregation by age provided equity and specific focus. The questions were specific to age cohorts of 8-, 9-, 10-, and 11-year olds:

1. Will students who have completed one academic year within the experimental multiage structure demonstrate greater reading comprehension and language mean scores than students within the control single-grade structure as measured by the TerraNova?

2. Will students who have completed one academic year within the experimental multiage structure demonstrate greater reading and language mean scores than students within the single-grade structure as demonstrated by the pretest/post test (Fall and Spring) scores on the Missoula Achievement Level Tests?

3. Will students who have completed one academic year within the experimental multiage structure demonstrate greater literacy development than the students within the single-grade structure as demonstrated by writing samples?

4. Will there be a significant practical difference (effect size) between the pretest and post test scores of students in the experimental and control groups of age cohorts as measured by each of the three different types of assessments?

Thus, the null hypotheses were:

1. H_0 . There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohorts and the control (single grade) cohorts as measured by the TerraNova/CTB April 1999 Reading and Language tests.
2. H_0 . There is no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohorts and the control (single grade) cohorts as measured by the Missoula Achievement Level Tests in Reading and in Language.
3. H_0 . There is no statistically significant difference between the group mean scores in fluency or conventions of subjects' writing samples in the experimental (multiage) cohorts and in the control (single grade) cohorts.

The alternative hypotheses to each of the null hypotheses were nondirectional.

In the qualitative component, this study addressed two major questions about organizational structure:

1. What are the instructional programs and practices within the single grade and multiage organizational structures?
2. Does literacy growth differ within the age configurations of the two types of organizational structure?

According to Wolcott (1982), it is "impossible to embark upon research without some idea of what one is looking for and foolish not to make that quest explicit" (as cited by Miles & Huberman, 1994, p. 17). To prevent overlooking relevant or unanticipated information, specific questions were part of the protocol, but data collection was open

for discovery. With this assumption, the general direction of the two qualitative research questions was not limited to, but included: How are the schools' instructional programs and practices similar or different in curriculum delivery, teacher training, and activities? For example, are practices and strategies evident according to the current knowledge of best practices? Do upper elementary students receive different instruction? What part does assessment play in instruction? What information about school population is most important for this study? (see Appendixes C and R for general protocol).

Significance of the Study

The questions of this study have implications for all school districts that recognize educational and/or financial accountability. As more interest in alternative organizational structures arise, so do questions on how they may or may not provide academic opportunity, fiscal efficiency, or both. Although every choice made within the public school system regarding educational accountability has financial ramifications, this study addressed academic accountability only.

Interest in Multiage Classrooms

As of 1997, few multiage classrooms existed in Montana. A multiage program at primary levels existed in one rural city school and in two schools in two urban cities, but organization is primarily single-grade with some combination classrooms [D. Neilson, Montana Office of Public Instruction (OPI), personal communication, April 1997]. However, interest in a multiage alternative has been expressed locally, and in other Montana districts as well [L. Peterson, OPI, personal communication, June 15, 1998; D. Neilson, OPI, personal communication, August 23, 1998]. A local private school began

in 1998-99 "placing first through fifth graders in the same classroom" (Jahrig, 1998, B1).

Furthermore, to date, only one study of organizational structure had been conducted in the Northwest. Pawluk (1992) compared the academic achievement of middle school students in multigrade classrooms in private, parochial schools in Oregon and Washington. Therefore, a need existed in this geographical area for a relevant, current study of upper elementary students in a public school system. Implementation and performance records needed to be considered.

Implementation Considerations

Organizational change that requires teacher training, reassignment, or both, and either additional monies or reallocation of extant dollars, creates problems for districts whose general fund budgets have grown more slowly than inflation. Other issues include management of class size and hiring of additional teachers for multiage classrooms.

According to Montana accreditation standard 10.555.712:

In single grade rooms, the maximum class size shall be not more than 20 in grades K through 2; 28 in grades 3 through 4; 30 in grades 5 through 8. In multigrade classrooms, the maximum class size shall be no more than 20 in grades K through 3; 24 in grades 4 through 6; and 26 in grades 7 through 8. Multigrade classrooms that cross grade-level boundaries (e.g. 3-4, 6-7) shall use the maximum of the lower grade. In one-teacher schools, maximum class size shall be 18 students. Alternatives need approval from the board of education. (Administrative Rules, 1997)

Therefore, equity of size among structure of classrooms is an issue. Multiage classrooms have not been defined, nor their maximum class size addressed in standard terms.

Currently, even major proponents of the nongraded or multiage classroom such as Gutiérrez and Slavin (1992) question the relevance of past research of nongraded or

multiage grouping as it applies to today's educational problems. They state that we need "assessments of current forms...to understand what really changes...in schools and what differences these changes make in student achievement" (p. 24). Objective measures were part of their recommended research criteria. This research began to address these concerns.

Performance Considerations

When school districts consider reform proposals, past performance of achievement must be considered. In 1997 Montana had the 5th highest high school completion rate in the nation (Ludwick, 1998). In addition:

The 1990 and 1994 National Assessment for Educational Progress (NAEP) math and reading tests placed Montana students first among the states. College readiness scores (ACT and SAT) are significantly higher than the national average...despite the fact that more students are taking the exams...high school graduates in the armed services have the highest average qualification test scores in the nation. (Keenan, 1997)

One of the factors to be considered in educational performance is that our schools have been and are presently predominantly graded classrooms. Thus, the request for an alternative structure in several school districts presents administrators with a dilemma. As the state's elementary age population declines, funds decrease proportionately. New requests cost money. As the literature review presents, some research suggests that organizational structure of classrooms affects student learning. But its equivocal nature and limited data are not sufficient for school districts faced with substantive resource reallocation.

As each school district has unique needs, so do students. What is most appropriate for both must be decided by those near to the issues. Regarding organizational structure, little research had been conducted in this geographical region on upper elementary students, and what existed was limited in scope. This study attempted to fill this gap in the research. A school community that may be considering an alternative organizational structure will have research particular to this study which may help in its own decisionmaking.

The purpose of this study was to provide an in-depth, rigorous investigation of the impact on literacy development within two organizational structures. Slavin (1983) and Slavin et al. (1994) advocate that component-building research on practical issues can make a substantial contribution to school reform. According to Fisher (1997), who examined only instructional practices within four multiage classrooms, questions must be addressed regarding academic progress within multiage and graded environments that reflect "best" practices (p. 126). This research extends previous research by its specific focus on separate, older age groups, and its use of both quantitative and qualitative methods. Therefore, the significance of this combined design study was that it addressed the impact of organizational structure upon literacy development of upper elementary students within one public school district within one geographical region during one school year.

Definition of Terms

This study used the following definitions:

Alternative assessment is the term given to nonstandardized assessment processes such as writing samples and scales (Allington & Cunningham, 1996, p. 132) and may approach authentic assessment: tasks that evoke demonstrations of knowledge and skills in ways that they are applied naturally.

Cohorts are groups separated, or disaggregated, from the whole group for analysis. In this study age cohorts were determined by the student's age as of the date of the first assessment: October 5, 1998. To maintain confidentiality, one district coordinator compiled this data.

Combination grade is the grouping of more than one grade level in a classroom. Other terms are split, blended, multigrade, or double year classrooms. Each respective grade level receives a separate curriculum. These terms have been confused with multiage and nongraded.

Continuous progress "lets children progress according to their individual rates of learning and development without being compelled to meet age-related achievement expectations" (Katz, 1992). It can be a component of the nongraded and multiage structures.

Family grouping is the term used to describe multiage grouping today. Begun in Britain during World War II for children sent away from their families, the model divided children in three-year blocks in primary schools (Kasten & Clarke, 1993).

Formal assessment is the collection of data using standardized tests or procedures under controlled conditions rather than informal by casual observation or nonstandardized procedures.

Graded structure is the use of chronological age as the "primary, if not the only, determiner of entry" into school (Shepherd & Ragan, 1982, p. 44). Unit level grouping or single-grade grouping are equivalent terms for this organizational structure.

Holistic evaluation of writing is a "guided procedure for sorting or ranking pieces...quickly, impressionistically...guided by a holistic scoring guide which describes each feature and identifies high, middle, and low quality levels" (Cooper & Odell, 1977, p. 3).

Horizontal grouping determines instructional groups or classes of students, as well as allocation of teachers at various grades on the vertical axis. Common patterns include self-contained, departmentalized and team teaching classrooms (Shepherd & Ragan, 1982).

Independent measure indicates separation in time and topic for writing (Deiderich, 1974).

Literacy is defined as "the capacity to accomplish a wide range of reading, writing, speaking and other language tasks associated with everyday life" (NCTE & IRA, 1996, p. 139) and "requires active, autonomous engagement with print" (Venezky, 1995, p. 19).

Literacy outcomes are active, independent demonstrations of learning that pertain directly to competence in reading, writing, speaking and listening.

Multiage structure is a classroom grouping of students of an age span of at least two or three years. A basic construct is that heterogeneous groups form for instruction (Stone, 1997). Katz (1992) uses this term interchangeably with mixed-age grouping, but says that mixed-age classes use temporary, homogeneous subgroupings of children. The terms vertically grouped, vertical streaming, and family grouping have been used to define this configuration.

Multigrade structure is the grouping of students from two or more grades in one class, retaining grade-level assignments and respective grade-specific curricula.

Nongraded grouping designates a vertical organization that groups students of different ages without grade designations such as first grade through twelfth grade. This rejects the promotion-retention system and is differentiated from multiage in its homogeneous groupings by ability within the heterogeneous age group (Anderson, 1992).

Organizational structure is the control of the placement of students in vertical and horizontal directions within schools or classrooms according to age, ability, or both (Glickman, 1998; Shepherd & Ragan, 1982). Four combinations are possible (see Fig. 1).

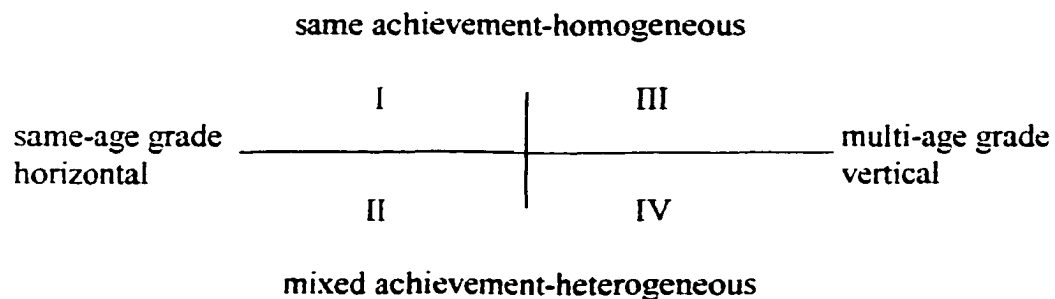


Figure 1. Options for School Organizational Structure. From Revolutionizing America's Schools by C.D. Glickman, Copyright (1998, Jossey-Bass, Inc.). Reprinted by permission of Jossey-Bass, Inc., a subsidiary of John Wiley & Sons, Inc.

Performance assessment is a "process which uses various strategies to provide students with opportunities to demonstrate their knowledge and skills in structured and unstructured situations" (Missoula County Public Schools [MCPS] Communication Arts, 1997). Writing samples are one example.

Retention is the act of nonpromotion so a student will repeat a grade level. A retained child repeats the previous curriculum during the year of retention (Gutiérrez & Slavin, 1992).

Rubric is a set of general criteria used to evaluate a student's performance in a given outcome area. Rubrics consist of a fixed measurement scale, a list of criteria that describe characteristics of products or performances for each score point, and sample responses which illustrate various score points on a scale (Maryland School Performance Assessment Program, 1993).

Stakeholder is a person who holds a share or interest in an institutional organization.

Vertical organization is a plan for the school "for identifying when and who is ready to enter, as well as the procedures for regulating pupil progress through the elementary school to a completion point" (Shepherd & Ragan, 1982, p. 43).

Limitations of the Study

This study was limited as follows:

1. Each classroom had a different teacher. Teacher demographics to include age, course training and workshops, educational level, years of experience, and choice of teaching position are stated. Since organizational structure does not "totally prescribe the

methods that a teacher may create, select, and adapt" (Shepherd & Ragan, 1982), possible extraneous variables included differences within classrooms of instructional practice and quality of delivery. However, as students from all 10 classrooms were disaggregated into age cohorts within both experimental and control schools, more than one teacher's influence resided within each cohort.

2. Student placement was not a random process. Some parents choose for their child to be in a particular classroom or school which may influence the child's attitude and may affect student learning. Student placement was also determined by teacher or principal recommendations. Therefore, the reality of a school setting prohibited a true experiment's randomization. Generalizability was limited by the quasi-experimental nature of this study, and so caution should be exercised in generalizing the results.

3. Students in this study were from two K-5 schools of similar demographic composition. The experimental school had all multiage classrooms except for self-contained single-grade kindergarten classrooms and one Grade 5. The control school had all single-grade classrooms. Within the district during this 1998-99 school year, only one other school had multiage classrooms, at first and second grade levels only.

4. Interviews regarding curriculum and instructional practice were limited to two school staffs: each principal, and any classroom teachers who would answer interview questions voluntarily. Letters were sent to each teacher requesting an interview. In the member check, the interviewee was asked to "nominate a person who, in his opinion, feels the same as he does about the evaluand" (Guba & Lincoln, 1981, p. 316) for an interview. No nominations occurred.

5. Midway into the research year, the control school population was informed that it would be closed the following year due to district budgetary factors. This could be considered an extraneous variable when considering student performance on test measures.

Delimitations

This study was delimited as follows:

1. This study focused on students within two K-5 public schools of similar demographic composition within the same district. Curriculum standards, objectives, and materials were presumed equal as well as district inservice training in literacy instruction.

2. The experimental school had components of its program in place for nine years, and so met the recommendation that programs have from three to five years of implementation before evaluation (Goodlad & Anderson, 1984). In addition, implementation through experience and teacher choice was stated to be part of its current delivery.

3. The subjects were upper elementary students in grades 3 through 5 and between 8 and 11 years of age. Literacy ability of these grade and age groups is usually more developed than primary groups, the extant research on this issue has been minimal at these older ages, and district norm-referenced and criterion-referenced standardized testing begins at third grade.

4. Students were disaggregated into age cohorts of 8-, 9-, 10-, and 11-year olds because age configuration is an integral factor of organizational structure. A grade-level

only study of multiage students would be inadequate. In addition, while the single-grade classrooms contain students closer in chronological age, this study recognized that same-age students may be at different developmental levels. Each structure was particular to each school in the study. Therefore, this disaggregation attempted to delimit by chronological age in order to provide a framework for study of literacy development which was most equitable for both organizational structures.

5. A common practice in schools has been nonrandom placement of students in classrooms according to ability, past academic achievement, and special needs. The disaggregation into age cohorts delimited the possible homogeneous placement as an extraneous variable and provided a more equitable comparison for both organizational structures.

CHAPTER 2

REVIEW OF THE LITERATURE

This literature review is organized in three sections. The first section reviews the historical development in America of the nongraded and the graded classroom from the colonial period of the 1600s to the 1960s. The second section provides a comprehensive account of research on multiage classrooms from the 1960s to the present. The third section reviews theories of learning for their relationship to instructional programs and practices.

Historical Background

The Oldest Organizational Structure

Imagine children seated on school benches according to chronological age. Brown (1970) documents a first instance of grouping of students in this manner as early as 1537 by Herr Sturm in Strassburg, Germany (p. 23). In most American schools today, classrooms replace these benches. Organizational structure by age within classrooms seems natural and customary to Americans. As predominant and permanent as it seems, this method of grouping children of the same age and different abilities was not America's first way to educate its children.

Before the 1800s, the family, religion, and a class system guided education. Private tutors, Latin preparatory schools, and theological colleges existed for the privileged in one-to-one teaching, or small groups of various ages. Parents, parishes, neighbors, and dame schools taught the rest of society (Pulliam & Van Patten, 1995). Within dame schools, "children as young as three associated with children as old as ten"

(Goodlad & Anderson, 1987, p. 44) and received instruction in a nongraded form (Miller, 1967). The belief that education was the parents' responsibility continued through the American colonial period and persists today, especially in homeschool families (parents, personal communication through informal survey conducted during a local book sale for homeschoolers, June 15, 1998). Yet, not unlike today, some parents in the 17th century did not fulfill this responsibility. In New England, the "Old Deluder Satan Act" of 1647 established the precedent that towns assume the responsibility for schools. A room full of children of various ages and abilities led by a poorly prepared teacher with meager equipment comprised many such schools. Often with few windows, and frequently with flogging to maintain discipline, it was "not a pleasant place, either physically or psychologically" (Pulliam & Van Patten, p .33).

This organizational structure continued through the Revolutionary War (Goodlad & Anderson, 1987; Pulliam & Van Patten, 1995). Children were taught in either privileged, private settings, or various-sized groups of children of various ages and abilities, with various instructors ranging from a widowed neighbor, to a schoolmaster, to an older student. Soon political, social, and economic changes would completely transform education from the responsibility of the family to that of the society.

Beginnings of the Graded System

Goodlad and Anderson (1987) state that five developments after the American Revolution were primarily responsible for emergence of the graded system: (a) public, state-supported education; (b) an effective monitorial system; (c) graded textbooks; (d) teacher training; and (e) German educational practices promoted by American educators.

First, separation of church and state disallowed use of public funds for church-supported schools. The selectmen of Boston, encountering increasing numbers of students to educate, began reading and writing schools separated by gender. Early in the 19th century, monitorial schools arose. Within a classroom as large as 300, "one teacher trained the older, brighter students to each teach... the same lesson to their groups of ten children" (Keliher, 1931, p. 3). Meyer (1957) wrote that a single classroom monitored by "junior henchmen" cost the public no more than \$1.06 per pupil per year (as cited by Goodlad & Anderson, 1987). Thus, cost-effective large group instruction, made possible through what could be called a type of multiage instruction, facilitated free, public education for many. The early results of educational evolution caused Alexis de Tocqueville to write in 1835:

I do not believe that there is a country in the world where, in proportion to the population, there are so few ignorant and at the same time so few learned individuals. Primary instruction is within the reach of everybody; superior instruction is scarcely to be obtained by any (p.54)...in no country in the world do the citizens make such exertions for the common weal. I know of no people who have established schools so numerous and efficacious...(p. 95)

The third development, publication of graded texts such as spellers, readers, grammar, and geography books began in the late 1700s, with Colburn's arithmetic text added by 1821. From 1836-57 the publication of McGuffey's Eclectic Reader with its graded levels changed everything (Parker, 1993). Parker asserts that "the 125 million copies sold are said to have influenced the American mind more than any other book except the Bible" (p. 2)..

A fourth development, the establishment of normal schools to train teachers, became a "powerful instrument for unifying educational practices [and] ordering the content of instruction" (Beggs & Buffie, 1967). Organization of subject matter, plus the graded textbooks, made it easier to handle large numbers of students (Keliher, 1931; Pulliam & Van Patten, 1995).

These large numbers of children were especially evident in the urban areas where immigrant populations grew rapidly. New school attendance laws for minimum ages added more students. Within this fifth development, administrators would reorganize classroom structure to meet the Industrial Revolution demand. Horace Mann and other influential educational leaders promoted the practice of graded structure they had observed in German schools. Academic achievement in the Prussian model that grouped by ages within separate grades impressed them. To them, this structure seemed to offer more educational opportunity.

During this era, grouping pupils according to their age became "familiar" (Miller, 1967, p. 48). In 1848 the first completely graded school opened in Boston. Principal John D. Philbrick instituted the Quincy Grammar School with new ideas of efficiency and organization (Case, 1931; Cuban, 1984; Goodlad & Anderson, 1987; Rollins, 1968). For example, separate classrooms for children at each age level had a separate teacher for each age group. With graded textbooks and course syllabi, graded classrooms could accommodate opportunity for more students in a structured, cost-effective manner (Goodlad & Anderson; Tewksbury, 1967).

The fact that only 45% of all school age children, urban and rural, attended any type of school emphasized this need (Pulliam & Van Patten, 1995). Growing numbers of children still had no school opportunities. Jacob Riis (1890) documented the "thousands of poor children crowded out of the schools year by year for want of room" (p.136). The graded system appeared to ameliorate this problem serving as an educational reform that provided equality of education. The graded system became firmly established (Beggs & Buffie, 1967). Mann, Philbrick, and others had instituted an organizational structure which would continue for the next 150 years to stand dominant today. Yet, as the next section relates, other organizational structures survived.

One-Room Schoolhouses Remain

After the Civil War in nonurban areas of the East, the typical school was still the one-room schoolhouse. It was often crowded, with bad ventilation, poor lighting, untrained teachers, and sporadic attendance. In the emerging West, the one-room schoolhouse existed for pioneer children as the alternative choice to homeschooling.

My Folks and the One-Room Schoolhouse (Webb, 1993) contains first-person accounts from people who attended one-room schoolhouses. Some excerpts include:

The teacher was a miracle worker....she had all eight grades....most of the time, however, not more than six of the grades would be represented, with probably two or three students in each grade. She gave us our lesson and from then on we were responsible for it. She did make use of older students in helping the younger ones which was good for all of us....we both feared and respected the big boys who could scare the smaller pupils and I learned to keep my mouth shut while sharing a desk with my sister. Whispering was strictly forbidden.

Classes could last about 10 minutes each [and] there were usually only 1 to 5 pupils in a grade so it was easy to help each other and still have time to help the younger ones. Much memorization was required in each grade....'background noise' was a geography lesson about the giant pyramids, the explanation of long-

division, or how to diagram a simple sentence. Slower learners profited from the repetition, quick learners absorbed material far beyond their years....much of the lessons were learned by rote.

According to Pulliam and Van Patten (1995), about 70% of the public school buildings in the United States were one-room schoolhouses until just after World War I. Muse, Smith, and Barker (1987) put the number at 196,037 in 1918, with about 1,000 remaining in 1980. In 1997 in Montana, 80 one-teacher schools remained [D. Neilson, personal communication, August 23, 1998]. Note one-teacher, not one-room schoolhouse, is the contemporary definition.

Reactions to the Graded System

Criticism of the graded system began almost at its inception. Shearer in 1899 complained that the pendulum had swung from no system to nothing but system (Goodlad & Anderson, 1987). First exceptions included W.T. Harris, St. Louis school superintendent in 1868, and later commissioner of education for the United States. His St. Louis plan refuted retention and recognized different abilities of children by instituting more frequent promotion and reassignment (Goodlad & Anderson, 1987; Keliher, 1931; Tewksbury, 1967). With ten-week intervals that assessed the progress of the child, a student did not have to struggle through an entire year of an inappropriate curriculum. Superintendent Harris said in 1900, "Like the current of a river there will be everywhere forward motion" (as cited in Keliher, 1931, p.13).

Documentation of early, and brief, efforts across the country to remedy the graded system exists (Case, 1931; Keliher, 1931; Miller, 1967; Otto, 1969). Some prominent attempts include the Pueblo Plan (1888), Cambridge Plan (1893), Batavia

Plan (1898), Wirt's Platoon Plan (1915), Dalton Plan (1919), and Winnetka Plan (1919). Although each had a different focus, of interest is how familiar each focus sounds today within most schools: ability grouping, tracking, theme units, team teaching, specialized teachers, mixing age groups, and individualized instruction. Each purported to recognize individual differences in children and to differentiate instruction.

Within the 20th century "practice in school organization [was] viewed against four sweeping movements" (Goodlad & Anderson, 1987, p. 51). First was the significant influence of John Dewey. Dewey's child-centered curriculum at the University of Chicago "eliminated arbitrary classification of grades, textbooks and subject matter" (Goodlad & Anderson, p. 50). He challenged " 'the lock-step' [where] the same subjects were taught in the same way using the same methods and same textbooks in every public school" (Pulliam & Van Patten, 1995, p.103). Second, research in human development suggested physical, emotional, social, and intellectual differences among children. Third, research on retention showed negative effects on cognitive and emotional development. Fourth, learning theories provided impetus for innovations in curriculum and instruction that moved teaching from a model of transmission to facilitation.

While the terms nongraded or ungraded did not become part of educational vocabulary until the 1940s (Tewksbury, 1967), plans that implemented all or part of a nongraded philosophy arose in the 1930s (Goodlad & Anderson, 1987; Miller, 1967; Otto, 1969). Some of the most frequently mentioned plans are Western Spring, Illinois (1934), Richmond, Virginia, (1936), Athens, Georgia (1936), and the Milwaukee Schools' Plan (1941). All eventually ended, but influenced subsequent revivals. With the Soviet

Union's launch of Sputnik (1957), the educational race was on. Reform received new interest and included alternative organizational structures. The next section presents research on organizational structure from the 1960s to the present.

Empirical Literature Since the 1960s

Organizational structure of classrooms and how it affects student learning has been addressed by a prodigious amount of research. This section discusses (a) two separate yet related revivals of interest in alternative organizational structures during recent decades, and (b) the confusing state of the research during this time.

The First Revival

During the 1960s the national response to the Soviet Union's Sputnik resulted in demands for accountability in education. The United States had to somehow increase student achievement, especially in math and science. "The beginning of massive public discontent...triggered ...increased emphasis on educational evaluation" (Popham, 1978, p. 3). Norm-referenced testing increased, and criterion-referenced testing emerged. One result was more grade retention of students, especially in urban areas. According to Gutiérrez and Slavin (1992), retaining more students improved test scores that reported by grade, not age. Therefore, schools appeared to be doing a better job. In The Nongraded Elementary School (1959), Goodlad and Anderson asserted that retention was harmful and applied inconsistently. Educators took note (Carbone, 1961; Gutiérrez & Slavin, 1992; McLoughlin, 1970; Shepherd & Ragan, 1982). According to Shepherd and Ragan, nongraded organization with its vertical and horizontal movement "based on ability...without regard for number of years" (p. 47) addressed retention concerns as it

provided a "successful experience...with no failure or retention" (p. 48).

During 1957-58 Goodlad and Anderson found fifty communities that were using some form of nongraded organization. However, information of actual implementation was "meager and somewhat confusing" (Shepherd & Ragan, 1982, p. 46). By the end of the sixties, less than 2% of American schools had nongraded programs (Slavin, 1986). From a national survey of elementary principals in 1968, Shepherd and Ragan found "that a little more than 10 percent of the schools were nongraded in the primary years" (p. 46) and by 1978, only 5.3% reported any organization other than graded. The movement is said to have "waxed and waned" through the 1970s (Pavan, 1992b). Yet, in 1983

A Nation at Risk renewed interest in alternative reforms.

The Second Revival

Mason and Stimson's (1996) study of twelve randomly selected states found that 95% of classes consisted of a single grade with the remaining four percent 2-and 3-grade combinations and less than 1% nongraded. Nevertheless, across the nation today, a return to nongraded or multiage programs is documented (Fogarty, 1993; Mason & Stimson, 1996; Nye, 1995). In 1990 Kentucky mandated ungraded primary schools and implemented multiage classrooms. Other states such as Tennessee, Mississippi, and Oregon had similar reforms. However, in 1996 the Kentucky legislature recalled the mandate, which returned decisionmaking about classroom structure to the local districts (KERA, 1997; Viadero, 1996).

Major reasons cited for organizational change to nongraded, or multiage, are (a) retention and (b) child development issues. Retention has continued through the

years "with a recent increase in incidence, without ever having been proven to be an effective practice" (Walters & Borgers, 1995, p. 300). Stronger is a Harvard Graduate School of Education research statement: "...we have no persuasive evidence that retention helps students to learn" (1986, p. 3). Other studies suggested long-term negative effects of retention (Holmes & Matthews, 1984; Shepard & Smith, 1990) and "the psychological ramifications of retaining young children" (Tanner & Decotis, 1995, p.135). Holmes' (1983) meta-analysis looked at 61 studies of academic achievement of promoted and retained students. According to Borg, Gall, and Gall (1993), meta-analysis has become the most widely used method for quantitatively combining research results from multiple studies. Borg et al. state that most meta-analyses use procedures developed by Glass (1976) that involve:

translating findings of a set of related studies into effect sizes. The studies typically are experiments that test the effectiveness of a particular program or method. The 'effect size' indicates how well the group that received the experimental method does relative to a comparison group that receives either no treatment or an alternative. (p. 171)

Holmes concluded that retention could not be supported. Students fall behind during the year they are retained and never catch up. Holmes and Matthews' (1984) second meta-analysis of attitudes, behavior, attendance, and academic achievement found no support for retention, with promoted students doing significantly better in every area. In addition, Holmes and Matthews declare "...cumulative research evidence [shows] that the potential for negative effects consistently outweighs positive outcomes...the burden of proof legitimately falls on proponents of retention to show there is compelling logic indicating success of their plans" (1984, p. 232). Shepard and Smith's (1990) study of 44

kindergartens and later studies of older students concluded that neither academic nor affective benefits were gained by retaining students. Their research is often cited. Johnson, Merrell, and Stover's (1990) study of fourth graders retained as first graders found that early grade retention was not "effective as an academic intervention" (p.337), and advised educators to look at other alternatives including "...strategic grouping of students within grades based on their academic needs" (p. 338).

In spite of this evidence, teachers and administrators continue to practice retention for various reasons. According to Tanner and Galis (1997), teachers' decisions are dependent on practical or tacit knowledge. They question whether teachers are aware of the research and disregard it, or just do not read the research. One major reason stated by teachers in support of retention is that one more year increases maturity. Mantzicopoulos' study of kindergarten children concluded that the "gift of time" did not contribute to school adjustment (1997, p. 126). Moreover, Roderick (1995) found that overage was a strong predictor of dropping out of school. However, Tanner and Galis included studies that suggest retention serves some purposes and concluded that:

there is no clear and consistent message for practitioners to use in guiding decisions because there exists sound evidence, although not in abundance, that supports retention. Therefore, there is enough published information to confuse decisionmakers and leave them to their own biases. (p. 108)

Another factor to consider in teachers' decisions regarding retention is the national standards movement. Called "Educate America 2000," this federal proposal, and thus monetary support and involvement in curriculum, wants states to use national standards and assessments for subject and grade levels. Glickman (1998) argues that

while this purports "to ensure a set threshold of academic outcomes for all students...[it] reinforces the very structures of subjects and grade levels" (p. 44). It furthers the use of standardized tests, letter grades, graded texts, exit exams, and retention.

Partially in response to this issue of retention, the National Association for the Education for Young Children (NAEYC) has suggested alternatives. Recommendations include nongraded primary and continuous progress programs with flexible groupings. Mixed-aged classrooms can facilitate both approaches. In nongraded or multiage, the practice of looping, a two- or three-year stay in one classroom, may forestall retention (Goodlad & Anderson, 1987; Stone, 1997; Tanner & Decotis, 1995). How teachers handle those students who are not developmentally ready to move after more than one year in one classroom is not apparent in the current literature. In addition, Bracey (1999) states that in the United States there is "little research backing" (p.169) the strategy of looping. Whether or not there is a difference in students' academic achievement associated with the number of years with one teacher is not established.

A second important factor in current reform is the research in early child development. While Goodlad and Anderson wrote about child development, there was still "little evidence to demonstrate the effects of developmentally appropriate practices...that allow young children to develop skills at their own pace" (Gutiérrez & Slavin, 1992, p. 339). Nongraded research simply did not define classroom practices in detail.

As stated earlier, multiage proponents maintain that multiage classrooms address not only retention, but also child development (Katz, 1992; Tanner & Decotis, 1995).

According to Katz, Evangelou, and Hartman (1990), ideal multiage grouping does not group by performance or ability within the classroom as nongraded does. Multiage classrooms are grouped initially by different ages. From there, heterogeneous, flexible groups are formed within the classroom with different grade level curriculum.

According to Katz (1996), this structure provides opportunities for nurturing found in Britain's family grouping, as well as differentiated learning. This idea follows the NAEYC's recommendations for appropriate school practices that meet developmental needs of children instead of children having to meet graded curriculum (Bredekamp, 1997). These beliefs parallel Goodlad and Anderson's (1987) concerns about curriculum and the wide range of abilities of children of similar ages.

Confusion in the Research Then and Now

Research from the 1960s to the present suggests that organizational structure differentially affects teaching and learning, but there is still little agreement on which structures significantly affect student success in terms of academic achievement, self-concept, or both (Brown & Martin, 1987; Gutiérrez & Slavin, 1992; Sheperd & Ragan, 1982; Slavin, 1986). The research on organizational structure has been confused in part by the different terms defining structures over the course of the decades (see Appendix B). Gutiérrez and Slavin (1992) state that the mixture of program types makes it difficult to single out benefits specific to the structure. Veenman (1995) adds that there is an "apples-and-oranges problem at the level of the independent variable" (p. 325). Gutiérrez and Slavin (1992) discuss two often-cited studies, McLoughlin (1967) and Pavan (1977), which reached opposite conclusions on graded and nongraded structures. Gutiérrez and

Slavin state that both studies were quite limited "...paying little attention to particular forms of nongrading used, the methodological quality of the studies, or the size of the effects" (p. 335). For example, in Pavan's 1977 study which summarized 64 studies between 1968 and 1976, she included nongraded, continuous progress, multiunit, individually guided education, multiage, ungraded, and mixed-age classrooms. Only 17 studies lasted more than a year, and differences within each program may have affected research results (Gutiérrez & Slavin, 1992).

To counter this problem, Gutiérrez and Slavin (1992) and Veenman (1995) both offer meta-analyses. As such, these two studies provide comprehensive information to date and a check to "distinguish good reviews from bad reviews" (Bickman & Rog, 1998, p. 315). Gutiérrez and Slavin's meta-analysis used a best evidence synthesis. Each study included had to have (a) an objective measure of achievement, (b) initial comparability of the two groups, and (c) programs in place for at least a semester. From the 57 studies that met these criteria, four different categories of nongraded programs emerged. Mixed conclusions emerged. Those nongraded programs that involved teacher-directed instruction showed positive effects. Students were grouped across age lines for a single subject, usually reading. Effects of those nongraded programs with individualized instruction appeared inconsistent and did not seem to enhance learning. Gutiérrez and Slavin (1992) state:

one interesting trend in the data on nongraded programs using individualized instruction: More positive effects were obtained with older rather than with younger children. It may be that students need a certain level of maturity or self-organizational skill to profit from a continuous progress program that includes a good deal of independent work. (p. 357)

They concluded that "there is a need for research combining qualitative and quantitative methods" (p. 369). For this research, three areas important to Gutiérrez and Slavin were included: (a) objective measures of achievement, (b) both programs in place for at least a semester and (c) initial comparability of the two groups, which was achieved with one measure, and similar demographic characteristics.

Veenman's (1995) meta-analysis synthesized research on the cognitive and noncognitive effects of (a) multigrade and single grade and (b) multiage and single-age elementary classrooms from several countries. His criteria were the same as Gutiérrez and Slavin's, with one exception: Veenman excluded nongraded programs, including only descriptors of multigrade, multiage, combination class, or vertical grouping.

Even though they may be distinct in curricular practices, Veenman's research of both multigrade and multiage follow because age configuration is the primary focus of this research.

For the multigrade versus single-grade, research findings for cognitive and noncognitive effects are similar. Multigrade students did not do better or worse than the single-grade classes. From 34 studies from which effect sizes could be estimated,

Veenman (1995) concluded:

that multigrade classes learn as much as their counterparts in single-grade classes. Across a number of studies, the number of years spent in multigrade was also not found to be associated with differences in achievement [and] of the 17 studies on noncognitive effects, five reported significant differences in favor of multigrade...but were so small they did not translate into higher achievement scores. (p. 357)

For the multiage versus single-age classes, Veenman's findings for cognitive and noncognitive effects from 11 studies were slightly different. His summary of cognitive effects states that "the findings do not favor multi-age classrooms...in most studies, no significant differences were found [and] multi-age classes appear to be generally equivalent to single-age classes" (1995, p. 362). Veenman states that the largest significant differences in achievement were found in favor of the single-grade classes, but with significant pretest differences. Only 2 of the 11 multiage studies provided evidence of initial equality. The summary of noncognitive effects found "a small positive effect for students in multi-age classes" (p. 366).

Veenman concluded that students in multigrade or multiage classes do not appear to learn more or less than their counterparts, though student attitudes are sometimes "better" in multigrade or multiage classes. It is important to note that where the differences exist, they "proved to be very small" (1995, p. 367) and cut across socioeconomic and grade level lines. Veenman listed four factors that may explain why no differences were found: (a) information on instructional practices in each of these four settings was not provided, (b) differential student selection criteria affected class composition, (c) absence of teacher training, and (d) time constraints for teachers. He recommended research on each of these areas. In this research, instructional practices, selection of students, and the types and degree of teacher training were part of the demographic description when possible.

In response to Veenman's research, Mason and Burns (1996) stated that multigrade classes have a slightly negative effect on achievement and a selection bias

toward quality of teachers and students. Veenman (1996) countered that he “suspects their conclusions are mainly based on studies in the United States and Canada [and that] a very small negative effect has been found only for the studies conducted in Europe” (p. 334).

In the only recent study in the Northwest, Pawluk (1992) found no statistically significant differences between the achievement of private, parochial school students in multigrade and single-grade classrooms. Grades 5 through 8 were measured in four subject areas through one standardized test. Muse et al. (1988) found in one-teacher schools in Montana, Nebraska, and South Dakota, students were “neither better nor less prepared” (p.19) than students from larger schools. However, since the tests varied from state to state, and school to school, no direct comparison could be made. In this study’s comparison, public school upper elementary students in grades 3 through 5 in Montana were the participants, and three separate, distinct, standardized measures were administered to all the upper elementary students in each school.

Studies that look at the noncognitive, or affective, dimension of this issue have shown positive benefits from heterogeneous age groups. Katz et al. (1990), Miller (1991), Pavan (1992a), and Pratt (1986) suggest that multigrade/multiage/nongraded grouping provide social gains. Miller concludes that “being a student in a multigrade classroom does not negatively affect academic performance, social relationships, or attitudes” (1991, p. 12).

Other affective studies suggest other considerations. Smith (1993) concludes that attitudes change toward structure as students get older, preferring same-age peers after

the fourth grade. Bergen's (1995) interviews found that while parents and students were supportive of multiage, the older students (8-year-olds) felt unchallenged, and parents felt they were learning less. Young and Boyle (1994) stated that fifth graders perceived third graders as incapable and instead of assisting, simply completed tasks for the third graders. Thus, since attitude is considered a factor in motivation and academic achievement, this is an area of concern. Moreover, in industrialized societies, puberty is beginning at even younger ages (Goodlad, 1984; "Onset," 1997). Wiles contends puberty is the time of the greatest developmental changes (1976). All these changes in child development speak to Tanner's (1993) statement of concern for the physical, mental, and emotional well-being of children, and need to be considered in the grouping of children of different ages. At present we do not know which combination of ages is most effective (Katz et al., 1990; Veenman, 1995), or the "advantages or risks associated from age ranges" (Katz et al., 1990, p. 56).

In summary, while benefits of alternative organizational structures have been found in studies, academic differences have really yet to be established particular to each specific type of organization (Brown & Martin, 1987; "Committee reports," 1997; Daily Report, 1995; Gutiérrez & Slavin, 1992; Katz, 1992; Miller, 1990; Nye, 1995; Pratt, 1986; Veenman, 1995). Goodlad and Anderson (1987) state that "the most serious problem afflicting all of the research on nongradedness...is researchers seem to accept the labels that are attached, without bothering to confirm that what is happening within the class or school is consistent with the label" (p.xxii). Assumptions were being made about classroom practices and attributed to one or the other structure without evidence to

support them, emphasizing the need to clearly define similarities or differences.

Slavin, Karweit, and Wasik (1993) agree in that "research from the first wave of nongraded primary schools supports [heterogeneous age grouping], but there is little consensus on its effects...we need to understand the conditions under which achievement was or was not enhanced "(p. 22).

This empirical review indicated a need for further study within this geographical region on the impact of organizational structures and how each supports all students' learning within the older age configurations. Therefore, theories about how children learn in the social environment of the classroom were critical to this study's framework.

Theories of Learning

Development Across Time

Current research in cognition draws upon the work of Lev Semenovich Vygotsky (1896-1924), a developmental psychologist whose work cuts across disciplines (Wertsch, 1985). Vygotsky's learning theory has been a part of American research since the 1962 publication/translation of his 1934 monograph Thought and Language, and in 1978 Mind in Society (1935). According to Jacob (1998), Vygotsky's work provides a theoretical and methodological framework to address the issues of how context affects learning. If learning can be understood only by considering how and where it occurs in growth, concentration on the process of development, not just the product is needed. A basic assumption is that "no single factor and corresponding set of explanatory principles" (Wertsch, p. 22) explains how students learn. Addressing the nature/nurture question, Vygotsky suggests that "multiple forces of development, each with its own set of

explanatory principles [are] the very nature of change" (Wertsch, p. 22). Vygotsky says that thinking, learning, and language occur through social interaction, and primarily through language. Therefore, our social/cultural groups affect our linguistic abilities. Vygotsky continues that "social relations or relations among people genetically underlie all higher functions and their relationships"(as cited in Wertsch, p. 61). This is the transition from the outside social influence to the point at which learning is internalized.

Vygotsky's construct, the zone of proximal development (ZPD), provides this transition. Vygotsky defines this as the "discrepancy between a child's actual mental age and the level he reaches in solving problems with assistance" (Vygotsky, 1934/ 1962, p. 187). By assistance he means social interaction with others is what facilitates the child's learning. This is done by "providing some slight assistance: the first step in a solution, a leading question, or some other form of help" (p. 187). He continues:

the development of a spontaneous concept must have reached a certain level for the child to be able to absorb a related concept [and this is found] within the zone of proximal development, in cooperation of the child with adults. (p. 194)

Again, the developmental process follows the learning process. Later, in Mind in Society (1978), Vygotsky states that this expert guidance can be not only from an adult, but also in "collaboration with more capable peers" (p. 86). This theoretical construct of interdependent learning provided an assumption upon which to question whether the age configuration of capable peers makes a difference. In this research, focus upon literacy development explored this factor of capable peers.

In addition, the construct holds two major points. One has to do with relationship to IQ, and the second to instructional practice. Vygotsky maintained, and studies by

Ferrara, Brown, and Campione (1983), and Campione, Brown, Ferrara, and Bryant (1984) suggest, that the actual level of development as measured by IQ is different from the potential level of development (as cited by Wertsch, 1985). In other words, different learning rates ("speed and/or degree of transfer") exist within students of similar IQ ranges (Wertsch, p. 71). From this, instruction appears most effective preceding development. Whether or not one organizational structure facilitates this cognitive development more than another within the context of academic achievement was a focus of this study.

Gardner's (1983) theory of multiple intelligences (MI) provides an even broader definition of diverse learning. He extends beyond just linguistic intelligence and incorporates at least seven more intelligences that emphasize the different ways people think and learn within social context. Gardner shares Vygotsky's assumptions as he asserts "constraints, both by epigenetic factors and by the operations of institutions" (1991, p. 264) and suggests alternative educational approaches. For example, Gardner's (1991) apprenticeship models for learning resemble Vygotsky's learning through collaboration with adults within the ZPD.

Cognitive studies emphasize the need for both assisted learning and accommodations for diverse abilities. For example, Shaughnessy (1993) suggests mentors for gifted students, and Falk-Ross (1997) for learning disabled students. Wood, Bruner, and Ross (1976) first used the term scaffolding to define the support that assists students (as cited by Graves & Avery, 1997). Support from a partner facilitates problem-solving. "When collaborators assume complementary roles, they begin to resemble peer

tutors" (Forman & Cazden, 1994, p. 155). Other educational researchers in the area of literacy have used Vygotsky as a framework in school (Baumann, Jones, & Seifert-Kessell, 1993; Heald-Taylor, 1996; Indrisano & Chall, 1995; Lehman & Scharer, 1996; McCarthy, 1994).

Research on language capacity of elementary children estimates an "exponential" increase in vocabulary at this stage (Bredenkamp, 1997). In addition, Goodlad and Anderson (1987) found that children enter first grade with a "range of from three to four years in their readiness"...[and] the "initial spread in abilities increases over the years so that it is approximately double this amount by...the end of elementary school" (p. 27). According to Heuston (as cited in Van Horn, 1999), as "classes get older, a class spread phenomena begins...rule of thumb is that there are as many years of difference in students' ability in a class as the grade level of the class...and the increase continues as students get older" (p. 296). This presumes a challenging environment for students as well as their teachers. Particular to this study was the focus of children's language development within each school and the potential for mentoring. One question was whether or not one organizational structure accommodates ZPD more than another. Germane to this issue were current recognized best practices for instruction, and whether or not they were implemented in either or both structures.

From Research to Practice

Through research, approaches such as collaborative and cooperative learning, heterogeneously grouped classrooms, learning styles, literature-based learning, reader responses, and literacy across the curriculum have become recognized as best practice

(Zemelman & Daniels, 1993). For example, using Vygotskian theory, Slavin (1986) states:

collaborative activity among children promotes growth because children of *similar ages* [emphasis mine] are likely to be operating within one another's proximal zones of development, modeling in the collaborating group behaviors more advanced than those they could perform as individuals. (as cited by Katz et al., 1990, p. 24)

Collaborative and cooperative learning are recognized strategies today. In a 10-year study of reading experts, Flippo (1997) found general agreement on appropriate practices across the curriculum. These included opportunities for integrating reading, writing, talking, and listening in cross-disciplinary instruction. NCTE and IRA (1996) added "viewing and visually representing" to language arts skills to make a total of six integrated literacy components. In addition, best practices includes making literacy functional and purposeful with authentic materials, and providing literature of quality in a variety of forms.

Harste (1989) asserts that the socio-psycholinguistic process of brain development relates directly to meaningful literacy activities. Thus, the social nature of learning and specific facilitative practices and contexts enables the student to become an active learner, and not merely a passive recipient (Harste, 1989; Healy, 1990; Smith, 1983). Hiebert (1994) states that these shifts in literacy practices result in different accomplishments which she calls authentic tasks. Authentic literacy tasks "are ones in which reading and writing serve a function... for...communication" (p. 391) [and] these "literacy processes... that rely on authentic tasks contrast with those that stress skills"

(p.393). Literacy definitions and standards involve authentic tasks, with outcomes that demonstrate competence, as in “the clear, rapid, and easy expression of ideas in writing or speaking” (NCTE & IRA, 1996, p. 72) defined as fluency.

Contemporary brain research explores the social/cultural concept that physical experience shapes brain development. Neurobiologists suggest “two broad stages of brain wiring: an early period, when experience is not required, and a later one, when it is” (Begley, 1996, p.55). Challenging the traditional view of predetermined brain development, these scientists also challenge the way some schools operate. For example, researchers found that early music training translated later into increased spatial intelligence and then math and reasoning skills (Begley, 1996). Healy (1990) believes we are rearing a generation of “different brains” at every socio-economic level and argues the neural plasticity of the brain in that:

a brain's organization, its proficiency with language... and its very patterns of thinking may be physically changed to a significant degree by early language environments (p.133)...there is as yet no substitute for language, used in tandem with visual reasoning, to hone precision of expression and analysis. In the schools to which we consign youngsters for so many hours of their lives...language is the coin of the realm. (p.107)

Healy maintains that students are less attuned to both spoken and written language, and thus, they are harder to teach. A visual, fast-paced lifestyle and a lack of physical, intellectual, and emotional nurturance are among hypothetical reasons. Her research suggests that children's brains are no less intelligent today, but learn differently, both temporally and topically. If so, then educational practices must give attention to the new research, in the area of language as well as the organizational structure of schools

(Begley, 1996; Gardner, 1991; Healy, 1990).

In summary, the influence across time of the ideas of Vygotsky, Gardner, and other cognitive scientists upon educational policies, programs, and practices is evident. The social-cultural theory of language acquisition was a framework for this study. The research as it relates to best practices in literacy instruction for development combined with the new concerns about cognitive development, developmental levels, and language learning. How all of this comes together within the organizational structures of classrooms and age configurations was the focus of this study.

CHAPTER 3
METHODOLOGY
Research Design

To investigate how organizational structure impacts literacy development, this combined design study used multiple methods of data collection and analysis. According to Reichardt and Cook (1979), research with multiple methods “can build upon each other to offer insights that neither one alone could provide” (p. 21). Similarly, Jick (1979) recommends multiple methods as “complementary” (p. 602).

Of Creswell’s three models of combined design, this study followed the dominant-less dominant design (1994, p. 177). The dominant paradigm, the quantitative method, used three different quantitative data sources. The less dominant paradigm explored qualitative data from two different categories to “probe in detail another aspect” (Creswell, p. 177). As a complementary component, the qualitative method attempted to provide a “more complete portrayal of the unit(s) under study” (Jick, 1979, p. 603). Because both quantitative and qualitative data collection procedures and analyses were used, this combined design involved a “between methods” approach (Creswell, 1994).

Merging various data is called triangulation (Denzin, 1970). This study used two of the four ways to triangulate data (Tierney, 1992): (a) a variety of data sources, and (b) the use of multiple methods. Triangulation may “uncover some variance which otherwise may have been neglected” (Jick, 1979, p. 603). In addition, triangulation

attempts to neutralize bias within the researcher or methods (Creswell, 1994; Reichardt & Cook, 1979; Yin, 1984). A fundamental assumption in this study was that multiple methods of data collection are necessary for decisionmaking. Thus, "the decisionmaker may need to utilize an alternative lens to understand" (Tierney, 1992, p. 1) and to answer different questions about one issue (see Appendix D for schemata).

As both quantitative and qualitative methods were used in a combined design, assumptions of both paradigms are presented. Each paradigm addresses (a) the meaning of reality, (b) relationship of the researcher to the setting, and (c) the process of research.

Assumptions of the Quantitative Paradigm

1. Reality is objective and singular. The quasi-experimental design used is "one of the most widespread experimental designs in educational research (Campbell & Stanley, 1963, p. 47). Reality is apart from the researcher.

2. The researcher is independent from data collection, being distant and circumscribed. The quantitative measures were administered without the researcher present.

3. Research is context-free. However, Campbell and Stanley (1963) state that "there are many natural social settings in which the research person can introduce something like experimental design into scheduling of data collection procedures" (p. 34), and they encourage the use of quasi-experimental design situations. Glass and Stanley (1970) state this offers "a middle ground between the controlled experiment of the laboratory and the uncontrolled experiment of nature" (p. 501).

4. The research is accurate and reliable through validity and reliability.

Assumptions of the Qualitative Paradigm

1. Reality is subjective and multiple. Qualitative data were collected from two different schools. Interviews and document analysis contributed to a more complete understanding of organizational structure within the complex mix of academic policy, program, and practice. The insider's perspective "illuminates the inner dynamics of situations – dynamics that are often invisible to the outsider" (Bogdan & Biklen, 1992, p. 32).

2. The researcher interacts with that being researched. In this study interviews were conducted on the natural site when possible. The researcher was an instrument of data collection (Bogdan & Biklen, 1992) of interviews and documents.

3. Research is context-bound with a natural setting paramount. The school was the only setting with which this inquiry was concerned. "Qualitative researchers believe that human behavior is significantly influenced by the setting in which it occurs" (Bogdan & Biklen, 1992, p.30). All interviews were conducted within the school or district buildings and in the context of school requirements and procedure.

4. The research is accurate and reliable through verification.

These assumptions provided direction for the combined design. It is important to note that data collection of test scores as a quantitative component is not dichotomous with the qualitative paradigm (Creswell, 1994; Jick, 1979; Yin, 1984). Schoolchildren take tests and write in the classroom, not the laboratory, and so this is part of the reality of the classroom. Objective data, that are well-established parts of the reporting of student progress, can be useful to different stakeholders (Anderson, Hiebert, Scott, &

Wilkinson, 1985). In addition, the use of more than just one assessment sought diversity in a critical examination of students' products. The complementary qualitative data, providing an alternative lens, expanded the breadth and scope of this study, and thus, "makes the most efficient use of both paradigms" (Creswell, p. 176). In the complex nature of a school setting, it seemed logical to use combined methods in order to "counteract discrepancies or biases" that may arise from only one method (Reichardt & Cook, 1979).

The Setting and Its Participants

Sites for investigation were two K-5 schools within one urban public school district in one northwestern Rocky Mountain community of approximately 87,000. Within this district, 3 out of its 12 elementary schools offered some form of multiage structure as of Fall 1998. The school selected as the experimental school had 13 classrooms: three kindergartens, one fifth grade single-grade classroom, four 1-2, one 2-3, three 3-4, and one 4-5. This configuration of multiage from grade 1 through grade 5 had been in place since 1995-96, beginning in 1990-91 with multiage in first and second grade only. Thus, the configuration, the length of time the structure had been in place, and its singularity in the community accounted for its selection. Its development has been with the principal as advocate, first as a teacher, and then as principal for six years.

Of the other two possible sites with multiage configurations as of the beginning of this research, one school had only one multiage classroom that had been in place for only one semester, and so was not considered. The third district school with multiage classrooms had only grades 1-2 multiage classrooms. No school within the district

offered both multiage and single grade options for all age and grade levels.

The control school had 14 K-5 single-grade only classrooms. The control school was selected because (a) its students' demographic composition was similar to the experimental school, and (b) it also had Title I schoolwide status. The numbers of students within each classroom were similar. In the single-grade school, classroom sizes were 26, 28, 20, 22, 19, and 25; in the multiage school 23, 25, 23, 24, and 28 as of September 1998. In December the enrollment was 23, 23, 24, 25 in each multiage class, with 30 in the single-grade fifth. At the control school, enrollment was 19, 20, 22, 25, 26, and 28 with the larger class sizes in the third grades. Division by gender was equal at both. The schools' enrollments were 299 and 274 respectively. Both qualified for Title I services, a federal K-12 remedial program for disadvantaged students authorized through the 1965 Elementary and Secondary Education Act, with comparable socio-economic (SES) numbers, adopting schoolwide status the same school year. Free and reduced lunch percentages had been high in relation to other district schools: the control school had ranged from 49 to 66% over the past six years; the experimental school had been from 61 to 76% during the same time (MCPS, 1998a).

This urban school district espoused open enrollment, but enrollment was usually limited to neighborhood boundaries. Students may attend a school outside their home boundary if classroom enrollment limits have not been reached. In 1998, 12 elementary schools, 4 middle schools, and four 4-year high schools made up the building units. As of September 8, 1998, the school district reported 9,507 K-12 students including 3,533 K-5; 1,990 middle school; and 3,984 high school students.

Special education services were provided districtwide under Public Law 105-17, the Individuals with Disabilities Education Act (IDEA) guidelines. Additional special education services, accommodations, or both are provided through Section 504 of the Rehabilitation Act of 1973. The English as a Second Language (ESL) program served students from several national and ethnic backgrounds including Native American, Russian, Asian, and Latino populations. Minorities comprised close to 8% of the district's population (B. Williams, ESL supervisor, personal communication, November 1997). Eight of 12 elementary schools qualified for Title I services. Both schools in this study have diverse populations of students, with the control school having the largest cultural diversity in the district with 24% bilingual students. However, this district's student composition does not approach the composition of other urban areas. It has what Comer (1997) characterizes as an "untraumatic social history" (p. 168) which he would argue may account for some of its academic achievements.

Data Collection Procedures

Access to participants and data was obtained by this researcher through the overt approach (Bogdan & Biklen, 1992), with entry from the superintendent, principal, and teacher, in that order of authority positions (Dean, Eichhorn & Dean, 1969, p. 68). Permission from the superintendent to entertain this project was obtained in the spring of 1997 following her reading of a first draft proposal. Meetings with school principals, and then teachers followed. In November 1998, following a need for change in the original design, this researcher met with the superintendent and obtained direction and permission for the present study. In a June 1999 telephone conversation, the

superintendent authorized access to the standardized test scores. On June 30, 1999 the superintendent, curriculum director, and this researcher met to plan procedures to access student scores in a manner that protected confidentiality.

Standardized test scores were identified by a code number and disaggregated by birthdate into age cohorts of 8-, 9-, 10-, and 11-year olds by the district coordinator who processes testing. During the following week, writing samples identified only by a code number were matched to birthdates or ages to disaggregate into age cohorts, just as the test scores had been. It was understood that the superintendent, as well as my dissertation chair, would be apprised of the study's direction during the course of this research.

It is important to note that the utmost confidentiality and anonymity was observed during this research. Because of past discussions within the community regarding differences of opinions on this issue, during the entire process no information was shared by this researcher with any persons within or outside the school other than the required gatekeepers in their order of authority. In addition, no classroom, teacher, or individual student was singled out at any time. Anonymity was a priority before, during, and after the course of this study.

Quantitative Components

Standardized data were collected from 11 classrooms over a period of one school year (see Appendix D for timeline). The small number of multiage classrooms in this community necessitated "convenience, or purposive sampling, of data collection [to] exhibit the phenomena of interest" (Borg et al., 1993, p.101). Purposive sampling "must

select a sample from which the most can be learned" (Merriam, 1998, p. 61), thus producing "information-rich" cases. This was a total of five multiage classrooms from the experimental school, and six single-grade classrooms, two at each grade at the control school.

As all third through fifth grade classrooms in each school participated in each quantitative measure, no preference could be indicated. The 263 participants were between 8 and 11 years old. Data from one fifth grade classroom at the experimental school were not included in the analysis as it was not a multiage classroom, resulting in a total of 235 students in 10 classrooms. Measures were administered at different times in the school year from October through June. Along with mortality, sample sizes per age cohort per measure vary also because one of the test measures, the TerraNova, is administered only by grade level, not age. One measure, the MALT, provided both pre- and post data.

Since student placement was not random, the classrooms were nonrandom "naturally assembled collectives...as similar as availability permits" (Campbell & Stanley, 1963, p. 47). For the writing assessment, stratified random sampling was used for samples to be read. Thus, all classrooms and grade levels were represented in an equal manner. This procedure also provided an additional check for student confidentiality, and attempted to equalize sample sizes (Borg et al., 1993).

Measures

To provide triangulation, three different quantitative measurements included two indirect and one direct assessment. The two indirect assessments in use in the district

were the standardized norm-referenced TerraNova/CTB, and the criterion-referenced Missoula Achievement Level Tests (MALT). The third measure was the standardized direct assessment of student pre- and post writing. Recognition of each measure's different characteristics (Farr, 1992) is purposeful and part of the analysis.

In Becoming a Nation of Readers (1985) Anderson and others recommended that the "attitude toward standardized tests is one of balance" (p. 101). They further suggested that reading comprehension subtest scores are the most significant. Allington and Cunningham (1996) suggested that "standardized achievement test data work well when comparing performances of groups of children" (p. 124) in classes or similar schools, and are "best used to monitor basic reading achievement patterns in a school" (p. 127). They went on to say that standardized achievement tests "do not measure everything that children might know or be able to achieve...[assessing] only a narrow range" (p. 126), but that data can be a valid assessment of "development of groups of children" and used for a broad program evaluation (1997).

The TerraNova/CTB

This district introduced TerraNova/CTB as its norm-referenced, standardized achievement test for the school year 1998-99, after 15 years use of the Comprehensive Test of Basic Skills (CTBS). It is the newest edition from the same company, McGraw-Hill. "While designed to provide continuity with previous editions of CTB tests, aspects of TerraNova... reflect new directions in today's curriculum" (CTB/McGraw, 1996, p. 9). Major strands in the reading test are basic understanding, analyze text, evaluate and extend meaning, and identify reading strategies. The major strands in the language test

are sentence structure, writing strategies, and editing skills (see Appendixes E and F for subdimensions). All questions are multiple-choice format.

CTB/McGraw (1997c) states that:

primary inferences from test results include measurement of the achievement of individual students relative to a current nationwide normative group and relative program effectiveness based on results of groups of students...results can also be used as one factor in making administrative decisions about program effectiveness, class grouping, and needs assessment. (p. 29)

This research emphasizes the “one factor” in recognition of the limitations of this type of assessment, and the need for judicious use of data interpretation.

Test administration. During the week of April 19-23 each classroom teacher in Grades 3 through 11 administered the timed TerraNova/CTB Battery using standardized instructions for the students and the teacher. Materials provided were a preprinted answer sheet, a No. 2 pencil, and level tests: Level 13 (Grade 3), Level 14 (Grade 4), and Level 15 (Grade 5). Number of questions per section corresponding to levels were: for Reading 42, 50, and 46; and for Language 28, 30, and 34. Students took the level of test that corresponded to their grade, not age, in both the control and experimental classrooms. Students in grades 3-5 took only the reading, language arts, and math sections, except for Grade 4 which takes science and social studies as well. Students were exempt from testing if an Individual Education Plan (IEP) so indicated. This district included tests scores of special education students. The district advised teachers that morning is preferable for testing, and to administer only one section a day. Degree of adherence was not certain as administration was not monitored on a formal basis, nor was this researcher present during any testing.

Missoula Achievement Level Tests

The MALT is a standardized, normed, and criterion-referenced test, with multiple-choice items matched to the district curriculum by its local test construction. It has been used for the past four years. One of the seven stated purposes of the MALT most relevant to this research is to monitor individual student growth (MCPS, 1996a). The major strands in the reading test are word meaning, literal comprehension, interpretive comprehension, and critical analysis. Major strands for the language test are the composing/writing process, composition structure, basic grammar/usage, and conventions. All strands are composed of multiple-choice items (see Appendixes G and H for subdimensions).

Level tests systematically increase in difficulty. Each student has a level appropriate to his individual level of proficiency as indicated by a previous test or initial locator test. Student progress is reported in the form of scores on a Rasch Unit, or RIT scale, each with benchmarks for performance expected at each grade level. The Rasch model assumes “that all items are equally discriminating and that items cannot be answered correctly by guessing” (Lord, 1980, p.189). The National Assessment of Educational Progress (NAEP) tests used this model of item response theory and similar scales for reporting scores (Ralph, Keller & Crouse, 1994, p. 3). According to the district and the Northwest Evaluation Association (NWEA) which guided construction of the district test, this type of testing is also ideal for an ungraded instructional program (MCPS, 1996a). Thus, there was equity in using this test for comparison of both organizational structures. In this manner, the same-age cohorts were compared

according to their own degree of growth, not just whether a score was higher or lower than another student of the same age or grade.

Test administration. During the week of October 5-9 Fall MALT tests in Reading, Language, and Math were administered to students in grades 3 through 8 in all district classrooms. The MCPS MALT Administration Guide (1998) states that level tests are not timed, and students may be exempted by teacher decision. Materials provided to each student were a test booklet at the predetermined level as indicated by the level assignment report received from the curriculum department, a preprinted answer sheet, and No. 2 pencil. Teachers read standardized directions for each test. Although specified as not a timed test, the test instructions to the teacher included:

After 45 minutes of testing, alert students that 15 minutes remain in this testing period. This is not a timed test. The test period should be long enough for all students to finish. If even one student is still working, however, do not collect materials until the test period ends. When you determine it is time to stop, say: Stop! (MCPS, 1998b, p. 4)

Consequently, the length of time given to students between classrooms could be an extraneous variable. However, through three separate verifications, both control and experimental schools' teachers allowed all students as much time as each individual needed. Only when a student appeared to be struggling was the teacher then to discontinue the test. It was assumed that teachers followed instructions.

The teacher or the retest report determines the need for a retest. The retest report indicates students who scored above or below the valid range. Each student must then take a second test at a level "normally two levels higher or lower...to give them opportunity to do their best" (MCPS, 1998b, p.6). Retest scores are part of this data.

The Writing Assessment

Pre-and post writing samples. From assessment of student writing, writing scores can be “treated just like scores obtained from standardized tests, but they are more valid in that they are based on actual pieces of writing, on some writer’s real performance” (Cooper & Odell, 1977, p.ix). The writing evaluation documented (a) students’ growth over a specific period of time, and (b) described and measured group differences (Cooper & Odell). Allington and Cunningham (1996) viewed writing samples and scales as “high-quality information about the acquisition of literacy” (p. 133). This direct assessment of students’ writing triangulated as an alternative measure with the two indirect assessments of literacy, the TerraNova and the MALT.

For research Allington and Cunningham (1996) recommended (a) more than one writing sample from each student and (b) prompts about which “most children know a lot” (p. 132).

Collection of samples. One standardized writing sample from each student was collected by teachers in the morning during the first week in January at both schools in all 11 classrooms. A second was collected during the first week in June. Test administration was conducted within the time parameters suggested by the school principals. Instructions for this timed writing were directed toward a “typical” (or average) performance in contrast to a “best” performance (Arter, 1993; Brossell, 1986; Hawk & Cross, 1987). This “static procedure [sought] objective, neutral, impartial assessment” (Shaughnessy, 1993, p. 4) of how each student writes independently. It attempted to control for extraneous variables such as time and outside writing process

assistance in Venezky's "active, autonomous engagement with print" (1995, p. 19). Part of the assumption of independence necessary in hypothesis testing was met in that responses of one student did not affect the responses of other students, as would have occurred with peer editing or teacher assistance.

Students from all classrooms wrote in bluebooks provided by this researcher. Students were instructed to use additional paper if needed. However, upon investigation, no student in either pre- or post writing used more than eight pages total, writing on both front and back pages of the 16-page, wideline bluebook. As with the other measures, this researcher was not present during test administration.

Selection of prompts. To select prompts, an informal pilot study was conducted within two elementary classrooms from a third school over the course of one school year using different prompts to see which elicited typical writing within the time frame. This researcher analyzed these writing samples and selected a final prompt (see Appendix I). The prompt followed the criteria for effective writing prompts (Barry, 1997; Gray, 1982; Spandel & Culham, 1993) for students across a broad range of development. It provided a topic that spanned the students' diversity due to limitations of experience (Calkins, 1986). Rhetorical specification of prompts followed recommendations for a typical timed writing (Brossell, 1986; Brand, 1991; Hawk & Cross, 1987). The prompts and purpose of each pre- and post writing was standard across the 11 classrooms. Teacher feedback regarding the pre-and post writing was gathered through a questionnaire (see Appendix J).

Choice of method of scoring for writing assessment. "It is critical to keep in mind that there is not now, nor will there ever be, a single best way to assess writing skill. The

method of choice is tied to the specific writing skills one desires to assess and the purpose of the assessment” (Anderson, 1980, p. 20). In this study a modified holistic method of assessment, the Holistic Developmental Writing Scales (HDWS), was used for four main reasons.

The first of the four reasons involved the content/form issue. While holistic scoring is the most commonly used method for writing assessment in elementary schools (McLean, 1992, p.12) and a “valid way of scoring large sets of compositions” (Proett & Gill, 1986, p. 26), it has been criticized in that it either “glorifies content and ignores form” (Gregory, 1991, p. 20), or form over content. This denies full scores if either is not strong (Proett & Gill, 1986). HDWS is a modified system of holistic scoring that separates conventions from fluency so that one will not influence the other in assessing scores (Elser, 1997). The procedure for scoring prevents the bias for highly conventional writing in that the paper is first read aloud by one member of the rating team. This also lessens Remondino’s factor, the influence of handwriting and neatness (Diederich, 1974).

Secondly, an assessment of language development as a whole, rather than several separate traits, was desired in order to be equitable for both organizational structures. HDWS analyze student writing at developmental levels, providing a goodness-of-fit to the heart of this research. Since developmental levels present in these control and experimental samples were not known, then rather than use grade level training for raters, a broader developmental range was needed. HDWS provided equity to both control and experimental organizational structures by examining writing from a developmental mode, rather than grade level expectations.

Third, it was assumed that all teachers in their instruction addressed developmental levels of students within the writing process paradigm (Zemelman & Daniels, 1988; Zemelman, Daniels, & Hyde, 1993). But it was not known at what time in the year each of the six traits in the six-trait writing instruction used by the district had been introduced within each classroom, or to the degree. Therefore, to assess using the six-trait writing assessment would not be equitable for both structures or all classrooms. HDWS offered an assessment that would ameliorate time and degree as extraneous variables and provide more equity for both structures between and among classrooms.

Fourth, as a former rater using six-trait assessment, this researcher wanted an assessment that (a) would be more collaborative and less isolated, (b) would eliminate the “go for the middle” score tendency when two raters are not in agreement after a first reading, and (c) reduce the possibility of different opinions regarding subskills.

Procedures for Assessment of Writing Samples

The site. The writing assessment was completed in three 3-hour afternoon sessions on June 29, 30, and July 1 at a local high school. The site was centrally located with free parking. Sessions began promptly at 1:30 p.m. and ended promptly at 4:30 p.m. Initial training was conducted in a classroom. The scoring took place in the adjacent cafeteria which was quiet, pleasant, and cool. The cafeteria area had some natural lighting and sufficient space to spread out the samples as needed. During the nine hours, the raters were uninterrupted. Once or twice a day the custodian or his two helpers would walk through the cafeteria, but they did not disturb the raters. Care was taken to avoid fatigue

with frequent breaks encouraged. Free food and drink were provided. As three raters had small children at home, the researcher provided a cell phone. It was used once. The facilitator's two daughters were present for part of the second and third days, but stayed apart from the raters, playing quietly in an adjacent room. In addition to the \$20.00 per hour that each rater received, the working conditions were quite satisfactory.

The participants. Eight people were involved in the writing assessment and were present at all three sessions: Dr. Tammy Elser, the six raters, and this researcher who acted as coordinator and host, answering only logistical questions. As developer of the Holistic Developmental Writing Scales over the past ten years, Dr. Elser provided training, instruction, and guidance for the raters of this writing assessment. She trained the raters in the use of the scales, facilitated the scoring during the three sessions, and was available to clarify any points, answer questions, and address problem papers. The six raters were all known to this researcher through different avenues of professional experience. Each person had been recommended by at least one other educator. These people were solicited because each met the preset criteria for raters: (a) previous training in writing assessment and/or as full-time teachers, have had at least seven years' experience evaluating and assessing student writing (Myers, 1985); and (b) not employed at either the control or experimental school (see Appendix K). Two were employed by the district in the study. In addition, the raters needed to be naive raters, i.e. they were unaware of the focus of the study before and during the assessment. This researcher solicited each rater first by phone, and then sent a reconfirmation letter two weeks prior to the scheduled assessment (see Appendix L).

The process. On the first day, introductions were made. A summary of the purpose of the assessment was given: this writing assessment is one component of research for a dissertation on literacy development among elementary students. No other details were given. Dr. Elser then gave a brief overview of the HDWS and proceeded to train the raters in a 90-minute session, providing samples of work that met each of the scales' criteria.

When all raters agreed that they were ready to begin reading papers, this researcher explained she would organize teams to provide diversity within each team. All were amenable to this arrangement which achieved equalization by gender; years of experience; public or private school employment; and primary, upper primary, or middle school experience. This last criterion placed on each team at least one person familiar with emergent writing. In addition, the nonrandom selection of teams provided another measure to facilitate a "focus beyond a set of grade level expectations" (HDWS, p. 17). A husband and wife were placed on opposite teams. No one person knew any of the other team members through any close relationship. Three were previously acquainted through workshops or university classes, but none of the members of each team were close social friends, relatives, or in positions of authority through employment.

Student sample selection. To control for mortality, only students who wrote both pre- and post essays were included in the total number of essays to be read. After the "lonely" samples were pulled, then all names and dates were removed and replaced with a coded number/letter written on the back of each sample. Then each coded paper was drawn according to a stratified random sampling. For this process in the control school,

all third grade papers were sorted together in the order of their pre-designated code number, then the same for fourth and fifth. In the experimental school, the same sorting took place as each student was designated by grade within each multiage classroom. The third graders were sorted together from the three classrooms; the fourth graders from the four classrooms; and the fifth graders, each in the order of their pre-designated code number. After this sorting, the student papers were then drawn according to a random sample table of numbers (Myers, 1985) and placed in ranked files in designated folders. As a result, every student with a pre- and post sample in each classroom had an equal chance of being selected within the total samples read.

Equal samples from each school were then placed into piles within each grade cohort in the order of each random sample number. This procedure was to equalize sample size within grade cohorts according to the least number of students within an age cohort. All student papers were then mixed into one group, so that raters did not know student names, ages, grades, classroom, teacher, or organizational structure. This process provided a measure against rating bias according to any of these factors, thus reducing, if not eliminating, the halo effect (Stanley & Hopkins, 1972). Additional samples were pulled and mixed in the same procedure on the third day because time was available to score more papers.

An additional check on confidentiality was provided by the fact that not all samples were read. The total number of samples scored were 244 (122 pre and 122 post).

Scoring. The procedure for scoring followed the HDWS (1998) instructions and the facilitator's directions. Samples were divided into an equal number for each of the

two teams. This first division included 90 samples for each team. This would meet the HDWS' estimated number possible within the nine hours. This researcher emphasized that raters were to take their time and there was no required number to complete. This verbal guide and division of papers avoided the "assembly-line" (Gregory, 1991) atmosphere of some writing assessments. During all sessions, team members were encouraged to take breaks whenever needed. For each team, a 90-in. x 11-in. laminated scale was placed on the long cafeteria tables. Each of the nine sections contained the 1-through 9-point fluency and convention rubrics. Before each scoring session, one member from each team read aloud the fluency scale criteria which gave the team a quick review of the criteria. To begin, each team member took a handful of writing samples and read them in relation to fluency, placing each sample below the number on the 9-point fluency scale where it fit best. Through this "quick-read" each member independently placed their samples along the continuum until all 90 had been placed.

Teamwork then began with a team assessment of each sample to determine if the sample fit the criteria as it had been initially placed. Members of each team took turns reading one sample aloud to the other two members. The listening two responded first with their judgment as to where it should be placed on the scale according to its content and development only. Since the listening two were not reading the paper, they were not influenced by the handwriting or conventions/mechanics of the paper. The reader gave a score last thus providing an additional measure against bias. In this first reading, the team is looking for "development of ideas, the creation of a story line, and other factors that indicate growing fluency using English for written expression" (Elser, 1997, p. 15).

Each sample is placed from a 1 (can't be read by anyone) to a 9 (indicating high engagement). Level one is the point at which no literate adult can decode any of the writing (Elser, 1997), thus indicating the total absence of the characteristic being measured (Christensen & Stoup, 1991; Elser, 1997).

According to team judgment, samples that didn't fit the criteria at the first reading were placed at the bottom of the stack one level ahead or one below. These samples were reassessed after all the papers had been read. Those that were judged to be properly placed initially remained in that level. Each time a team finished a stack of samples at one level, they moved to the next level. They usually reread the criteria, either silently or aloud. This collaborative process proceeded until all samples had been assessed. Then each sample was marked in the top right corner with the numeric score corresponding to its level on the scale and placed back into its stack.

With each atypical sample, the team followed the HDWS instructions to reread the criteria at that level, reread the sample, and then use their collective judgment and place the sample. The pool of papers previously read that collected under each level provided benchmark samples to which raters referred in this decision. If there was still a concern, the team members referred the paper to the facilitator. Discussion among the members and the facilitator then followed, with placement becoming a four-member decision.

In addition, teams had been instructed that any papers indicating a "crisis" were to be reported to the researcher who would refer the paper to the school principal. Crisis was defined as a reference indicating possible harm to the writer or others. Two crisis papers were reported by one team. The students' principal was notified by telephone

message later that day.

At this point the second stage of assessment began. Each member took a stack of writing samples at one level and sorted them based on the conventions scale of high, middle, low, emerging, or indiscriminate conventions. Samples were skimmed. Raters were instructed to not reread completely, as this might let fluency interfere with a conventions rating. A corresponding letter was placed by the numeric score. Thus, each paper then had a complete rating, e.g. 5-H, 6-L or other combinations. The conventions score later was converted to a numeric score for statistical analysis. This separation distinguishes these scales as modified holistic scoring that recognizes the different skills involved in fluency and conventions as separate but equal.

By the third session additional papers were added because of additional time and the desire to increase the size of the final sample. The process was repeated. Upon completion 16 papers were used to recalibrate individual scores among team members.

A total of 244 papers, 122 pre- and 122 post, were read. Upon conclusion, each member answered the rater questionnaire. One team finished earlier than the other and voluntarily stayed in its group discussing the students' writing. All raters left by 4: 40 p.m. on July 1.

Qualitative Components

Interviews

According to Bogdan and Biklen (1992), an interview is a "purposeful conversation...that varies in the degree to which it is structured" (p. 96). The semi-structured interview helps in collection of comparable data across samples of subjects. However, since this study took place during one school year and explored instructional

and academic components of each school, the format of interviews had temporal and topical considerations. I piloted each protocol with participants from a third school. Initial interview questions to develop rapport discussed research objectives and all questions attempted to "minimize the imposition of predetermined responses" (Patton, 1980, p. 211). Tierney (1992) states that this frees the researcher "to move in a direction that appears interesting and rich in data" (p. 4). In addition, I used probes and follow-up, and tried not to deter participants from digressing from the protocol:

the interviewer [needs] more flexibility in probing...and in determining when it is appropriate to explore certain subjects in greater depths or...undertake whole new areas of inquiry...not originally included in the interview instrument. (Patton, p. 204)

Interviews were taped only with participants' permission. Immediately following the interview, I filled out a cover sheet noting central topics. In addition, I reviewed my notes and wrote a summary within 24 hours of the interview for the audit trail and for later data analysis. I transcribed all interviews in order to retain confidentiality and to know my data more fully, consistent with Tierney's recommendations to "develop familiarity with notes" (1992, p. 23). Within the week of the interview, I mailed a transcription copy to each interviewee, with a cover letter of appreciation, and reexplaining and scheduling a member check (Tierney, 1992). Within this letter I also offered them the opportunity to nominate, or recommend, a person to be interviewed about this issue (Guba & Lincoln, 1981). No nominations were received.

Documents and Archival Records

In the collection of data, two main categories were considered and searched: official public documents and archival records (Bogdan & Biklen, 1992). Official

documents included the district curriculum guides, standards and benchmarks, district and school mission statements, district and school goals, district assessment reviews, and documents from the state Office of Public Instruction (OPI). Archival records used were newspapers and newsletters to the present date which provided an historical description of the alternative organizational structure from multiple perspectives. In addition, recent school developments that provided a thick, rich description of each neighborhood were provided by school newsletters, local newspapers, and federal program information.

Standards for Quality of Conclusions

Quantitative Components

The independent variable was the classroom organizational structure: the multiage and the single grade classroom. The effects were measured by the TerraNova/CTB, the Missoula Achievement Level Test (MALT), and writing samples. The dependent variable was growth as measured by mean scores by age cohorts in reading, language, and writing fluency and conventions.

Validity and Reliability of TerraNova/CTB

Until 1998, this school district used the CTBS/4 as its standardized achievement test. In The Eleventh Mental Measurements Yearbook (MMYB) reviewer Kenneth D. Hopkins (1992) states that the CTBS/4 continues to be “among the very best general achievement test batteries” (p. 216), although there are “major unanswered questions about the representativeness of the norming sample” (p. 217). According to the curriculum director, this is one of the reasons for the district’s 1998 adoption of the new standardized test, the TerraNova (R. McKean, personal communication, July 6, 1999).

Representativeness of the norming sample of TerraNova. Standardization procedures were based on a stratified national sample. Variables used were geographical region (Northeast, Southeast, Midwest, or West), community type (large urban, urban, suburban or rural), school size (small or large), socioeconomic status (high or low), and school type (public, Catholic, or private non-Catholic). The spring study involved 100,650 kindergarten through grade 12 students from 295 school districts. At least eight Montana schools participated. The exact number cannot be known as only 88% of participating schools agreed to be listed. Scores for students were weighted to represent national proportions based on national census data. CTB/McGraw-Hill obtained the schools' demographic data through a self-reported questionnaire.

Validity of TerraNova. Test validation "is not a quantifiable property but an ongoing process" (CTB/McGraw-Hill, 1997c, p. 29). Technical Bulletin I (1997) presents three types of validity: content, criterion, and construct-related. Under content validity, CTB developers state:

Content-related validity is evidenced by a correspondence between test content and instructional content. To ensure such correspondence, CTB developers conducted a comprehensive curriculum review and met with educational experts to determine common educational goals and the knowledge and skills emphasized in today's curricula...content is more thematically integrated...graphic design mirrors types of materials students read...minimized ethnic and gender bias....[it] accurately represents the important educational objectives set throughout the nation. (1997c, p. 29)

In addition, usability studies; student input regarding graphic design, background color, navigational items; and teacher surveys about test directions were conducted as part of evidence of content-related validity.

Hopkins (1992) states the districts must examine their own curriculum and determine for themselves content validity (p. 217), thus restating the “heavy reliance on human judgment [that] does not lend itself readily to quantification” (Popham, 1978, p. 35). TerraNova was selected because it best meets this district’s curriculum, standards, and benchmarks (R. McKean, personal communication, July 6, 1999).

Criterion-related validity tells us how well the test measures what we want it to by indicating how closely the test relates to some criterion (Lyman, 1971, p. 23). Evidence is presented through a validity coefficient. Data are not available as the studies have not been completed as of the latest technical bulletin publication. The bulletin also states that anticipated studies include links to the National Assessment of Educational Progress, Third International Mathematics and Science Study, Scholastic Assessment Test, and American College Testing Battery. However, CTB did equate TerraNova to the CAT/5 and CTBS/4 using equipercentile methods, and the results were mixed.

Technical Bulletin I (1997) states that construct validity, what test scores mean and what inferences they support, are evidenced by several components. First, a comprehensive description of skills, concepts, and processes, and expected growth in scale scores and raw scores is present. Secondly, “minimization of construct irrelevant variance and construct underrepresentation is addressed in the steps of the test development process of specification, item writing, review field testing, test construction and standardization” (p. 30). Third, guidelines for appropriate test administration and use for students, including special needs students have been reviewed (pp. 34-35). In addition, convergent and discriminant validity correlations with Test of Cognitive Skills/2 are

“consistent with how measures of academic performance should relate to measures of cognitive processing” (p.74).

Reliability of TerraNova. Content reliability is the consistency with which a test measures what it measures. This may be estimated by a reliability coefficient based on split halves, alternate forms, or internal consistency. CTB/McGraw-Hill states that “on the average the test difficulties are well targeted to student performance and show appropriate growth from fall to spring [as reflected] in p-values, the Kuder-Richardson Formula 20 coefficient, and standard errors of measurements...and indicate that the tests are providing good measurement” (1997c, p. 112). Articulation studies indicated that for any given test and level comparable results are attained.

According to Lyman (1971) if the test is not “highly speeded, evidence on content reliability can be obtained by Kuder-Richardson or split-half formulas... but neither may be used when speed is an important factor” (p. 29). TerraNova is a timed test. CTB/McGraw states that “typically fewer than 4% fail to complete the tests as indicated by responding to the last item...TerraNova tests show little speededness” (1997c, p. 73).

Validity and Reliability of the MALT

MALT questions were drawn from the Northwest Evaluation Association (NWEA) item banks. NWEA researchers have calibrated each test item to a continuum of skill levels and tested for validity and reliability over the past 20 years (MCPS, 1996a, p. 4).

Representativeness of norming sample of the MALT. Currently 21 states and 150 school districts across the United States use achievement level tests through the NWEA (G. Kingsbury, NWEA, personal communication, July 1998). Initial norming samples

were drawn in 1995 for grades 3-8 from 14 participating districts. Approximate sample size for reading was 65,000 students; for language, 18,000. Ethnic makeup was compared to 1994 U.S. census data (see Appendix M). In 1998, NWEA conducted a norming study of 104 school districts with over 500,000 students. Mean scores and average annual growth for grades 2-10 are available for 1998-1999. For grade level means, standard deviations, and annual learning growth from this norming sample see Appendix N.

Validity of the MALT. NWEA develops test items. Once test items have passed the bias review panel, each item is field tested. A minimum of 300 students in each grade takes a test on these calibrated and developmental items. Researchers revise test items that do not perform well. Each is field tested again, or discarded. The level tests depend on the difficulty of the questions to estimate student performance levels. A common scale of difficulty was conducted for each subject area using the Rasch model of Item Response Theory (Lord, 1980). Recalibration of test items whose difficulty may change over time are completed regularly. Each district constructs its own test particular to its curricula from this bank of thousands of multiple-choice questions.

Content validity is nonstatistical and refers to the extent that the curriculum is reflected in the test items (Lyman, 1971). District teachers constructed each level test five years ago. This researcher participated in both reading and language constructions. Test questions were selected according to district curriculum goals and objectives, with explicit efforts to align the test with the curriculum (see Appendixes G and H). This is the district's test blueprint and is the first step toward insuring content validity (NWEA, 1996, p. 11). All tests were piloted in several local schools before districtwide testing began.

In addition, face validity has not been an issue to date.

Criterion-related validity is empirical. It tells us how well the test measures what we want it to measure by indicating how closely the test relates to some other criterion (Lyman, 1971, p. 23). An equivalence study that relates performance of the MALT to a nationally normed test would be necessary. NWEA does not conduct these tests due to the diverse nature of tests among districts. According to Gage Kingsbury at NWEA (personal communication, July 1998), one school district of 2500 students obtained grade level validity coefficients between .80 and .75 for the reading test and the Comprehensive Test of Basic Skills (CTBS). These samples did not include special education or ESL students.

The district under study used the CTBS as a second standardized achievement measure until 1998 and was to begin comparisons with the MALT in July 1998. These data were not available according to the district curriculum director (R. McKean, personal communication, July 6, 1999).

Reliability of the MALT. A test with high reliability is one that will "yield very much the same relative magnitude of scores for a group of people under different conditions or situations" (Lyman, 1971, p. 24). "Consistently high reliabilities" have been found by NWEA research (1996, p. 14). In 1995 reading achievement level tests scores were calculated according to marginal reliability statistics based on a norming sample of 9,000 students in five states. For language, the sample size was approximately 3000. Marginal reliabilities were obtained for grades 3-8 by subject area (see Appendix O). NWEA (1996) states that reliability estimates should be accurate provided the distribution of achievement in the local district is similar to the norming sample.

Validity and Reliability of Writing Samples

Validity. Deiderich (1974) states student writing samples are “direct measures of the ability we wish to measure and hence are valid by definition” (p. 102). However, variability of student writing is affected by several factors including subject matter, rhetorical specification of topic, testing time, and audience (Brossell, 1986; Graves, 1983; Gregory, 1991; Myers, 1985; Proett & Gill, 1986). This study’s carefully developed prompt addressed equity of writing between both organizational structures, across three grade levels, and individual students’ differences. Its content validity is subject to the same conditions as the study’s other two measures: Content validity is nonstatistical and refers to the extent that the curriculum is reflected in the test (Lyman, 1971). Miller and Crocker (1990) state that content validity is strengthened by pre- and post prompts that are specific, structured, within the general experience of all students, and in the same mode of discourse (as cited by McLean, 1992, p. 28). Both pre- and post prompts did not require of students any writing skill beyond the district curriculum goals, objectives, or training teachers received about the writing process (MCPS, 1997).

Reliability. Reliability of writing samples is “achieved by asking for more than one piece of writing on more than one occasion and then involving two or more people in...rating each piece” (Cooper & Odell, 1977, p. xi). Pre- and post samples of writing were assessed by groups of three trained raters using an agreed upon criteria of judgment (McLean, 1992, p. 29). Reliability was enhanced by the prompt which was “fair to [all] writers” (Cooper & Odell, p. xi), written on different days, and “written under controlled conditions to insure the student actually does the writing” (Cooper & Odell, p. 19).

Validity and Reliability of HDWS

The Holistic Developmental Writing Scales (HDWS) were in use in 22 school districts across several regions as of 1998. The rubric scoring criteria provided a metric measure by which to assess and evaluate student writing (Arter, Culham, Pollard, & Spandel, 1994; Elser, 1997; Nye, 1995), and the scales met construct validity as well as reliability tests, through the original study and a replication study (Elser, 1997).

Validity. Construct validity was met by through analysis of HDWS and its relation to theories of writing assessment, process, language acquisition, and cognitive development. Content validity, which is “nonstatistical” (Lyman, 1971) was also addressed by this researcher’s participation in the district curriculum, inservice training, and assessments, as well as study of the emphasis upon the qualities of writing measured by these scales. Agreement among the district curriculum, writing instruction, and the scales is demonstrated through comparison of criteria (see Appendixes P and Q).

Reliability. The HDWS Fluency Scale has an inter-rater reliability coefficient of .9941 for all levels 1-9. The Conventions Scale has an inter-rater reliability coefficient of .9830 for its five levels (Elser, 1997, p. 46). A reliability coefficient of .80 for program evaluation and .90 for individual growth measurement is considered “high enough” (Cooper & Odell, 1977, p. 18). Reliability of raters for the sample was “achieved by...involving two or more people in...rating each piece” (Cooper & Odell, p. ix) [and] “when raters are from similar backgrounds and when they are trained with a holistic scoring guide...they can achieve...scoring reliabilities in the high eighties and low nineties on their summed scores from multiple pieces of a student’s writing” (p.19).

Threats to Internal Validity for All Measures

1. History, selection-maturation, and maturation were not threats, but worked with the research question of developmental growth within one academic school year. For the MALT, each child received a level test appropriate to his/her last functional level test in order to analyze growth. Disaggregation into age cohorts eliminated selection-maturation (age differences) as an extraneous variable. Maturation, or developmental differences among students of the same chronological age, was recognized and is part of the narrative of this study.

2. Testing effect was minimal due to the length of time, one academic year, between criterion tests. The intent of each test was to measure growth over the year. For the writing samples, the time interval was five months. Any reactive effect should have been countered by the adequate length of time between testing combined with the maintenance of normal routine.

3. The threat to instrumentation validity was minimal due to the constructed forms of the entire level series of the TerraNova, MALT (NWEA, 1996, p. 10), and writing samples. Teachers received standardized directions for administration. Complete information on each measure was provided. Time constraint and a structured writing topic were extraneous variables necessary within the parameters set by principals for the writing sample. The samples offer an accepted measure of a student's independent writing, i.e. without peer or teacher editing (Arter et al, 1994; Brossell, 1986). Instrument decay was controlled by the "shuffling" (Campbell & Stanley, 1963, p. 9) of samples to eliminate raters' knowledge of age, school, teacher, and organizational structure.

4. Differential selection was present in this quasi-experimental design, so due to nonrandom assignment, generalizability of any effect of the treatment should be viewed with caution. However, the pre- and post tests, the similar demographic composition of each school (Borg et al., 1993), and the inclusion of all accessible classrooms within each school were attempts to control for this threat within this quasi-experiment. Bias due to selection of schools by parents was considered minimal as both are neighborhood schools.

5. To control for mortality threats, students who took the pre- and the post MALT, or the writing samples, had scores included for each analysis. Mortality rates reduced the initial expected size of the samples that was based on Fall enrollment, but were not due to characteristics of the treatment. Attrition due to the loss of the single-grade classroom from the experimental school was unavoidable. Loss of students due to these reasons did not distort the post test results in any type of systematic bias. It simply reduced the sample size. For the TerraNova measure, students' entry was not controlled, so data include students enrolled any time before the spring test.

6. Groups were not selected on the basis of extreme scores, nor were the tests being analyzed over more than one academic year, so the threat of statistical regression was minimal.

7. The threat of the Hawthorne effect, that is knowledge of the experiment affecting participants' behavior, was controlled by the research design in that all measures were part of the regular routine of the school, and were administered by each classroom's teacher. However, the possibility of different emphases placed by individual teachers upon any of the measures was present and considered an extraneous variable.

Threats to External Validity for All Measures

1. For population validity, it is acknowledged that the subjects are from two schools in one district. While it could be argued that the experimental group was the only accessible population of upper elementary multiage and that all accessible upper elementary classrooms from both schools were used, both groups were not part of a true random selection. Where other districts have similar characteristics, results could be generalized, but only with caution. Due to the nonrandom selection of schools, generalizability is possible only if it is "reframed to reflect the assumptions underlying qualitative inquiry" (Merriam, p. 208), and user or reader generalizability is practical only from the quasi-experimental design and its controls. This research includes complete descriptive statistics of each group with which to compare initial group scores, and pre- and post data for the most rigorous statistical tests.

It is necessary to remember Campbell and Stanley's (1963) support of quasi-experimentation with reference to Design 10, the nonequivalent control group design on which this qualitative design is modeled:

...naturally assembled collectives such as classrooms as similar as availability permits, but yet not so similar that one can dispense with the pretest...Design 10 should be recognized as well worth using in many instances in which Designs 4, 5 or 6 are impossible...in particular it should be recognized that [this design] reduces greatly the equivocality of interpretation over what is obtained in the experimental One-Group Pretest-Post test design. The more similar the experimental and the control groups are in their recruitment and the more this similarity is confirmed by the scores on the pretest, the more effective the control becomes." (pp. 47-48)

According to Miles Myers (1985), the "drawbacks of nonrandom assignment of students and teachers" can be ameliorated by "obtaining pretest and post test data, employing multiple treatments for comparison with the traditional treatment, and using

the class rather than the individual as the unit of study” (p. 134). This research accomplished the first two. However, because the students were disaggregated by age from classrooms into age cohorts for test score analysis, the unit of statistical analysis is the individual student:

The units of statistical analysis are the data (the actual numbers) that we consider to be the outcomes of independent replications of our experiment. If you will, the units of statistical analysis are the numbers that we count when we count up degrees of freedom “within” or “for replications.” (Glass & Stanley, 1970, p. 505)

2. Personological items - "An interaction is present if the experimental results apply to subjects with certain characteristics, but not to subjects with other characteristics" (Borg et al., 1993, p. 304). Demographics were defined as completely as possible for comparison, including socio-economic status, ethnicity, gender, age, grade, and other data per school, but were not available per age cohort.

3. Ecological validity of students was addressed by similar age and ability configurations that would be found in most public elementary schools in this geographical region. Ecological validity of teachers was addressed by description of teachers' background, including training, workshops, and experience. Since instructional practices within each school are extraneous variables, qualitative data from interviews attempted to investigate these variables within each school (see Appendixes C and R). Self-reported teacher opinions of the writing assessment procedures were summarized from the questionnaire (see Appendix J).

Procedures for Quantitative Analysis

Descriptive statistics for the three separate measures for each age cohort include group's mean, standard deviation, variance, skewness, and kurtosis of distributions.

Experimental differences (difference between two post test means) is expressed in a percentage. A difference of over 5% would warrant consideration. Statistical tests, analyses, and concomitant inferences were made based on these conditions, as well as discussion of practical significance by effect sizes. This reporting attempts to clarify questions of internal validity.

Choice of statistical tests was determined by the nature of each of the individual measures. The TerraNova, MALT, and writing sample scores are equal interval scales. To investigate the difference between the means of each cohort on the TerraNova Spring test in Reading and in Language, and with the writing post scores in fluency and conventions, the independent sample t test was used. The t test is a robust technique for small as well as large size groups (Christensen & Stoup, 1991). With samples of unequal size between control and experimental groups, the t statistic was calculated using the pooled variance estimate. "The sample variances are weighted by their degrees of freedom" (Howell, 1997, p. 192), and this weighted average corrects for the difference in sample sizes. Variances were reported within all mean averages.

For the TerraNova, conversion from the raw score to an equal interval standard score specific to each level was necessary (CTB/McGraw-Hill, 1997b). Students take the level test that meets their designated grade level. Thus, different-aged students took the same level test. Then, disaggregation by age according to each corresponding level test taken was done, i.e. 8-year olds that took Level 13 were separated from 8-year olds who took Level 14. This reduced sample sizes and created unequal sample sizes.

Because there were pre- and post test scores for each student on the MALT, the F -

test with the analysis of covariance (ANCOVA) was considered the most appropriate and powerful analysis to determine statistical significance. Many studies reviewed in Chapter 2 used gain score analyses. However, Campbell and Stanley (1963) state that “simple gain scores are applicable but usually less desirable than analysis of covariance” (p. 49). Hopkins and Glass (1978), Howell (1997), and Keppel (1973) also state that gain scores are not preferred. It is necessary to account for differences that may exist between the groups prior to the treatment. In the ANCOVA, “each student’s post test score is adjusted up or down to take into account the pretest performance” (Borg et al., 1993, p. 162), and thus is a “method of statistically controlling variables” (Hopkins & Glass, 1978, p. 153). Wildt and Ahtola (1978) recommend the ANCOVA in nonrandom assignments in order to remove bias among intact groups, and “increase the precision of the experiment by reducing the error variance” (p. 14) which is an increase in the statistical power of the analysis (Freed, Hess, & Ryan, 1989, p. 438; Huitema, 1980, p. 25).

Wildt and Ahtola (1978) also state that the ANCOVA is appropriate when the “observations on the covariate are obtained after the presentation of treatment but before the treatment has had an opportunity to affect the covariate...”(p. 15). The MALT pretest was administered one month after school started. It would be imprudent to suggest one month of treatment would affect pretest scores to the extent they would account for statistically significant post test differences. However, it is for this same reason that an analysis of covariance test was not used for writing scores since the pretest writing was administered almost five months after school had begun. Thus, the covariate in writing would not be independent from the treatment. However, pre- and post tests were obtained

to include students enrolled at least since January and presented as descriptive statistics.

For both parametric tests, the independent t test and the ANCOVA, assumptions are reported. When assumptions are not met, a nonparametric test was used. If a combination of both unequal sample size and heterogeneity of variance was present within groups, then the Mann-Whitney U-test was used as a follow-up. When the assumptions underlying the t-statistic or analysis of variance cannot be met, the Mann-Whitney U-test is one of the most powerful nonparametric tests for independent samples, is especially sensitive to differences in distributions, and does not require equal group sizes (Christensen & Stoup, 1991, p. 387).

For all measures, the alpha level of probability was set a priori at .05 to define significance for all statistical tests. This minimized the danger of both Type I and Type II errors. A Type I error is made when the null hypothesis is true, but an alternative hypothesis is accepted. A Type II error is made when the null hypothesis is retained and the alternative hypothesis is true (Christensen & Stoup, 1991). A nondirectional (two-tailed) test was indicated because a direction of difference between means was not specified a priori (Hopkins & Glass, 1978).

An effect size for each test was computed for the control and experimental groups within each age cohort. An effect size was “computed by taking the difference between the mean score of the experimental treatment and the mean score of the control treatment on the criterion measure and dividing this difference by the standard deviation of the scores for the control group”(Borg et al., 1993, p.171). For the MALT score, the adjusted mean score was used. For each measure, this provided a numerical expression of how

well the experimental group performed relative to the control group. An effect size greater than .33 was considered of practical significance and part of analysis (p. 164).

Qualitative Components

Does a study do what it says it is doing? Is it believable? In qualitative research, Lincoln and Guba (1985) state these are questions of trustworthiness. To insure rigor in trustworthiness, the four areas of credibility, transferability, dependability, and confirmability were addressed. The following methods were used to insure rigor and establish trustworthiness:

Credibility

Credibility asks if there is truth in the findings. The truth, or credibility, derived from this study is from the analysis of the perspectives of the participants. A qualitative description of the site, interviews, and documents attempts Geertz's (1973) "thick, rich description" which is aided by triangulation of data (Jick, 1979). In this triangulation, two different categories of data sources were explored: interviews (see Appendix R) and public documents.

Including triangulation, six other ways to insure rigor in the credibility of the study were prolonged engagement, member check, literature check, peer debriefer, negative case analysis, and an audit trail (Lincoln & Guba, 1985). A brief description of how each was conducted follows:

1. Prolonged engagement means the longer the study, the more rigor it will have. This study began in September 1998 and continued with interviews through August 1999. This length of time was longer than many of the studies cited in the literature review.

However, the interviews were not as frequent nor varied as desired due to unforeseen circumstances. Students and staff at one school learned in March that their school would be closed the following year. Both principals were notified mid-year that they would be transferred to different schools at the close of the present school year. These events were considered extraneous variables.

2. Member check is the most crucial technique for establishing credibility (Lincoln & Guba, 1985, p. 314). It requires asking during the interview such questions as, "I think I heard you saying this. Did I get it right?" It is also the systematic process of checking back with interview participants to verify transcriptions and any other record. This was accomplished with each person interviewed in a formal interview. I telephoned each person a week after the typed transcription had been sent to them to see if they had any concerns. Two follow-up interviews with each interviewee were scheduled, and the same procedure followed. In addition, one rater participant in the writing assessment was asked to read and review the description of the three-day process as an additional member check. He confirmed that this analysis documented his experience and that the writing assessment was conducted in a professional manner.

3. Literature check involves reviewing previous literature and keeping current with new literature to verify and develop new ideas. I did this on a regularly scheduled basis, continuing to use computer searches through Educational Resources Information Center, the Thesaurus of ERIC Descriptors reference, Dissertation Abstracts International, and Newsline Online through March 2000. Priority was given to search the terms multiage, nongraded, and multigrade classrooms. I used Merriam's (1998)

selection criteria: author's authority, date of work with both early and most recent research considered, relevancy of characteristics, quality of overall study, with an abbreviated annotated bibliography of all references.

4. Peer debriefer involves asking a peer to question and check the research (Lincoln & Guba, 1985). As a University of Montana student, I had a UM doctoral graduate student who met Lincoln and Guba's (1985) criteria: (a) neither junior nor senior in authority, (b) familiar with the substantive area of inquiry and methods of research, and (c) serious enough to play the "devil's advocate" (p. 308). She read independently all transcriptions and field notes from the audit trail.

5. Negative case analysis means that the researcher looks for data that might not fit with previous hypotheses. I examined "both the supporting and discrepant evidence to determine whether the conclusion in question is more plausible than the potential alternatives" (Bickman & Rog, 1998, p. 93). Miles and Huberman (1994) state that "when a preliminary conclusion is in hand, the tactic is to say, 'Do any data oppose this conclusion, or are any inconsistent with this conclusion?'" (p. 271). Because the nature of the organizational structure as conducted by the experimental school was apparent only after interviews, I needed repeated follow-up contacts to principals and teachers through telephone, voicemail, and letters.

6. An audit trail determines confirmability and dependability. An audit trail is the organization of data so that another researcher could examine methods and procedures, taped interviews, and transcriptions and thus replicate the study. This researcher's audit trail was contained in four three-ring binder notebooks; tapes; color-coded, categorized

file folders and inspected by the peer debriefer. Prior to inspection, I blocked out personal names to retain confidentiality that was promised. In addition, items that participants asked to remain confidential, I deleted from the transcript and marked with an asterisk.

Transferability

Transferability asks if the findings can be applied, or transferred, to other contexts or subjects. This transferability, or generalizability, relies on comprehensive description and analysis. All data were described in as thick and rich a description as possible to enable the reader to make connections to other settings. I described all demographics as fully as was possible with limited information access and member check confidentiality. I discussed findings congruent, or contradictory, to prior theory or research. As Lincoln and Guba (1985) state:

Transferability... must be reassessed in each and every case in which transfer is proposed...an investigator can make no statements about transferability for his other findings based solely on data from the studied context alone. At best the investigator can supply only that information about the studied site that may make possible a judgment of transferability to some other site; the final judgment on that matter is, however, vested in the person seeking to make the transfer. (p. 217)

Dependability

Whether or not this study could be replicated in a subsequent study addresses the issue of dependability. One of the major concerns with research on this subject as stated in the literature review was lack of comprehensiveness, as well as scope. This combined design study attempted to present a triangulation of data. However, the subjects' characteristics were particular to this community's district and geographical region. Also, it is acknowledged that this study takes into account "factors of instability, factors of

phenomenal change" (Lincoln & Guba, 1985, p. 299), and limited accessibility. The audit trail facilitated dependability (Lincoln & Guba, 1985) as much as this particular set of time and circumstances could be replicated.

Confirmability

Confirmability addresses neutrality and bias. Triangulation, discussed earlier, addressed this issue. For example, the use of a standards-based performance writing assessment through naive but trained raters, as well as the separate quantitative measures of the norm-referenced tests attempted to investigate neutral, objective measures for data analysis. In addition, an audit trail, or chain of evidence, and a reflexive journal, also called a field diary, contribute to the neutrality and confirmability of this study. I used Halpern's (1983) audit trail categories, file types, and evidence such as tapes of interviews (as cited by Lincoln & Guba, 1985, pp. 382-384). A field diary of personal reflections on the process was another way to self-check and be checked (Bogdan & Biklen, 1992; Lincoln & Guba, 1985). The field diary attempted to keep an "accurate record of methods, procedures and evolving analysis" (Bogdan & Biklen, p. 121), and contained handwritten notes, copies of transcripts, as well as documents received from participants.

Procedures for Qualitative Analysis

Bogdan and Biklen (1992) suggest that qualitative research leaves "the formal analysis until most of the data are in" (p. 154). Informal analysis took place during the study to facilitate direction of data collection and ensure substantial data. Therefore, data collection and formal analysis were not a simultaneous process. The following points as recommended by Bogdan and Biklen (1992) and Merriam (1998) considered in this

study's ongoing analysis:

1. Plan data collection in light of what is found in previous observation;
2. Try out ideas and themes;
3. Use visual devices to summarize thinking and complexities.

Following Babbie's (1998) advice, typed notes are a "stimulus to recreate as many details of the day's experiences as possible...comprehensive and detailed" (p. 295) with two copies made for backup and for later mechanical steps. Organizing and filing of notes was the first step to "finding the underlying meaning" (p. 295) of all this data. The type of files began according to the components of the research questions and was a continuous process.

Additional mechanics of working with the data as described by Bogdan and Biklen (1992) were used at the onset: wide margins and all data numbered sequentially. Data such as interviews and fieldnotes were numbered to be kept separate. Reading of the material was paramount, and during this time a preliminary list of coding categories that seemed relevant to each research question was kept and contained in the audit trail notebooks. From this, coding categories were abbreviated and assigned units of data- "pieces of fieldnotes, transcripts, documents that fall under the particular topic represented by the coding category" (Bogdan & Biklen, p. 176). I labeled the evidence related to each question and then entered this into data summaries. This allowed examination of any trends within the data. As I typed my own notes, I always cross checked the data to gain more perspective and organized the data according to Bogdan and Biklen (pp. 177-179).

Role of the Researcher

Within the assumptions of the qualitative paradigm, Creswell (1994) states that the role of the researcher is an integral part of a qualitative study and perceptions must be stated explicitly. My professional experience as a public school elementary teacher for the past 13 years shapes my perceptions of education. I have taught at three grade levels, all of which were single grade classrooms. At each of my schools, options of organizational structure were not present. During one year, the possibility of a multigrade classroom arose. Though I was reluctant to volunteer because of my lack of experience and knowledge of a combination structure, I was open to the assignment. However, the option did not materialize at this school at my grade level.

For the past 13 years, I have served the district on two language arts curriculum review and selection committees, as well as other content areas. My interest in literacy has been longstanding. My interest in organizational structures began more than seven years ago when a group of parents requested of the school district an opportunity to have alternative classrooms within two district schools. It was intriguing to me that strong opinions on both sides of the issue formed so quickly. Negative discourse occurred with some discussions. Everyone seemed to have an opinion, but opinions, including my own, seemed based on generalizations, grounded in a natural skepticism. While I felt my contextual awareness would help understand the challenge of this issue, I did not begin to realize its complexity.

While I feel that I am open to new ideas, a principle from the Hippocratic oath to “first do no harm” has appeal for me in the advocacy of classroom practice. I knew that I

wanted answers to questions. I needed evidence that was more than selective evidence. I have continually revised, changed, and challenged my instructional strategies and practices within my own classroom, in combination of knowledge gained from research, continuing coursework, professional workshops, visits to other classrooms, and most importantly, daily experiences with children and parents. This study, at the very least, offered an opportunity to improve my own teaching with the insight gained from extensive study, a comprehensive review of assessment measures, and the perspectives of others outside my own school about this issue. At the very most, this research may contribute to component-building research. Adherence to the rigorous requirements for access to data within well-established methods of a combined design research has been foremost in my mind and upheld at all times during the course of this research.

CHAPTER 4

ANALYSIS

This study investigated multiple evidence regarding the impact of organizational structure upon upper elementary students' literacy development. Its fundamental assumption is that multiple methods of collection and analysis of data provide a diverse body of verifiable information necessary for a comprehensive evaluation. First, three quantitative measures' data, and analyses are presented. Second, qualitative data within an analytic narrative follow.

Quantitative Components

The quantitative measures are the TerraNova, the Missoula Achievement Level Tests (MALT), and pre-and post writing samples. Each is a dependent variable to measure the effects of the independent variable, organizational structure. Results of each quantitative measure are reported by age cohorts in the following order: TerraNova Reading, TerraNova Language, MALT Reading, MALT Language, and post writing assessments. No emphasis or preference is indicated for any measure by the order of presentation. Each measure has properties unique to its type of assessment and analysis. Data were collected and analyzed within the context of those properties. All data entries were triple-checked for coding errors. Rounding was performed only in final answers, and then according to standard rules for rounding (Christensen & Stoup, 1991, p. 22).

Descriptive statistics are presented first. Experimental differences are reported for each comparison. Discussion of assumptions of inferential statistical tests are presented prior to data tables. Effect sizes are reported for each comparison. Cohort summaries are

in the text with corresponding tables. This reporting accomplishes an analysis from which inferential statistics should not be confusing or subject to misinterpretation by the reader (Borg et al., 1993; Howell, 1997; Mallows, 1983, pp.135-36).

TerraNova

TerraNova Reading

TerraNova reading scores are reported within the district in percentiles and raw scores. For this research, using the TerraNova conversion tables, raw scores were converted to standard scores to provide an equal interval measure for statistical analysis using the independent sample t test. Each cohort met all assumptions of the t test, with some exceptions. Because this school district administers tests according to a student's grade level, an 9-year old in a third grade class takes a Level 13 test, while an 9-year old in a fourth grade class takes a Level 14 test. Within the multiage classrooms, the school and the school district designate students by grade levels and tests are taken accordingly. Consequently, partitioning by grade and level created smaller sample sizes, and in all cases, unequal sample size among cohorts.

In addition, if a combination of unequal sample size and heterogeneity of variance existed, the nonparametric test, the Mann-Whitney U , was used. In this measure, Cohorts 9 - Level 13, and Cohorts 10 - Level 15 required this additional analysis. Also, Cohort 8 - Level 14 for Reading and Language were not entered into statistical analysis due to both sample sizes of 3. The control group reported one 8-year old, and the experimental group reported two 8-year olds, with all three designated as grade 4 students. Table 1 summarizes the results of the statistical tests.

The null hypothesis, H_0 . There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohorts and the control (single grade) cohorts as measured by the TerraNova/CTB April 1999 Reading test.

Table 1

T test Results for TerraNova Reading by Age Cohort and Level Test

<u>Group:</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>t-value</u>	<u>*p < .05</u>	<u>Effect size</u>
<u>Cohort 8 - Level 13</u>		(651)				
Control	36	634	32.24	-2.68	.0093*	0.68
Experimental	34	656	36.46			
<u>Cohort 9 - Level 13</u> (alternative test follows)						
<u>Cohort 9 - Level 14</u>		(660)				
Control	24	660.79	39.01	-.96	.3426	0.28
Experimental	19	649.84	34.60			
<u>Cohort 10 - Level 14</u>						
Control	9	648.77	38.24	1.07	.3009	0.64
Experimental	8	624.25	55.54			
<u>Cohort 10 - Level 15</u> (alternative test follows)						
<u>Cohort 11 - Level 15</u>		(675)				
Control	17	654	36.24	-1.50	.1355	0.64
Experimental	7	677.29	24.07			

Note. Cohort 9-Level 13 and Cohort 10-Level 15 are not included due to the combination of heterogeneous variance and unequal sample size. Instead, the alternative test, the Mann-Whitney U , was used and results follow in text. District 1999 averages per level test are noted within parentheses.

*p < .05.

Alternative Tests

Cohort 9 - Level 13. The results of the Mann-Whitney U test for Cohort 9-Level 13 reports a U-value of 44 which reveals a p value of $.0987 > .05$, indicating no statistical significance. The null hypothesis fails to be rejected.

Cohort 10 - Level 15. The results of the Mann-Whitney U test for Cohort 10-Level 15 reports a U-value of 64.5 which reveals a p value of $.5653 > .05$, indicating no statistical significance. The null hypothesis fails to be rejected.

Analysis of Hypotheses for TerraNova Reading by Cohort

Cohort 8. There is a statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort as measured by the TerraNova/CTB April 1999 Level 13 Reading test. Therefore, the null hypothesis is rejected. The direction of difference is indicated by the experimental group's greater mean of 656 (SD 36) as compared to the control group mean of 634 (SD 32).

Cohort 9. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort as measured by the TerraNova/CTB April 1999 Reading test at Level 13 or 14. The null hypothesis fails to be rejected.

Cohort 10. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort as measured by the TerraNova/CTB April 1999 Reading test at Level 14 or 15. The null hypothesis fails to be rejected.

Cohort 11. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort as measured by the TerraNova/CTB April 1999 Reading test. The null hypothesis fails to be rejected.

From this analysis, the only cohort which indicated a statistically significant difference was Cohort 8 which took the Level 13 (Grade 3) test. Practical significance is indicated in the difference of 22 points between the means in favor of the experimental group, with an effect size of $.68 > .33$. The experimental difference is only 3%. This cohort of students has had the longest exposure to schoolwide interventions beginning at kindergarten. These 8-year olds would be the younger students in the multiage classroom, designated as third graders within a 3/4 classroom. It should be noted that this is the first standardized testing experience for this age cohort of students in third grade.

None of the other cohorts which completed one academic year indicated a statistically significant difference between the reading comprehension mean scores of each structure as measured by the TerraNova Reading test. Interpretation of results of Cohort 10 and 11 should be cautious due to either small Ns or unequal sample size.

Within each age cohort and between the two groups, literacy growth appeared comparable. Experimental differences ranged from 0% to 3%. Both 9- and 10-year olds designated as fourth grade indicated greater mean scores in the control groups, a pattern demonstrated within another measure as well.

Overall, these results suggest that students who have completed one academic year within the experimental multiage structure did not demonstrate any pattern of statistically

significant greater reading mean scores than the students within the control single-grade structure as measured by the TerraNova.

TerraNova Language

Student scores were reported within the district in percentiles and raw scores. Raw scores were converted to standard scores (Norms, 1998) to provide an equal interval measure for statistical analysis within the independent sample t test. Level tests were administered to students by grade level designations. Each cohort met all assumptions unless otherwise noted. Due to its small sample size ($n = 3$), Cohort 8 - Level 14 was not entered into statistical analysis. The control group reported one 8-year old, and the experimental group reported two 8-year olds designated as grade 4 students. Table 2 summarizes the results of the statistical tests.

The null hypothesis. H_0 . There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohorts and the control (single grade) cohorts as measured by the TerraNova/CTB April 1999 Language tests.

Table 2

T test Results for TerraNova Language by Age Cohort and Level Test

<u>Group:</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>t-value</u>	<u>*p < .05</u>	<u>Effect size</u>
<u>Cohort 8-Level 13</u>		(644)				
Control	36	632.39	29.41	-.83	.4084	0.20
Experimental	34	638.15	28.44			
<u>Cohort 9-Level 13</u>						
Control	13	641.54	36.90	.73	.4742	0.27
Experimental	11	631.54	28.90			
<u>Cohort 9-Level 14</u>		(660)				
Control	24	666.83	46.07	2.90	.3754	0.27
Experimental	19	654.32	44.72			
<u>Cohort 10-Level 14</u>						
Control	9	655.22	32.23	1.33	.2023	0.76
Experimental	8	630.63	43.61			
<u>Cohort 10-Level 15</u>		(670)				
Control	25	659.12	32.32	-.53	.5976	0.25
Experimental	6	667.17	36.93			
<u>Cohort 11-Level 15</u>						
Control	17	658.24	26.52	.11	.9171	0.05
Experimental	7	657	25.03			

Note. District 1999 averages per level test are within parentheses.

*p < .05.

Analysis of Hypotheses for TerraNova Language by Cohort

Cohort 8. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort of 8-year olds as measured by the TerraNova/CTB April 1999 Language test at Level 13. The null hypothesis fails to be rejected.

Cohort 9. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort of 9-year olds as measured by the TerraNova/CTB April 1999 Language test at Level 13 or Level 14. The null hypothesis fails to be rejected.

Cohort 10. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort of 10-year olds as measured by the TerraNova/CTB April 1999 Language test at Level 14 or Level 15. The null hypothesis fails to be rejected.

Cohort 11. There is no statistically significant difference between the group mean scores of subjects in the experimental (multiage) cohort and the control (single grade) cohort of 11-year olds as measured by the TerraNova/CTB April 1999 Language test at Level 15. The null hypothesis fails to be rejected.

These results suggest that students who have completed one academic year within the experimental multiage structure did not demonstrate statistically significant greater language mean scores than the students within the control single-grade structure as measured by the TerraNova at any age cohort, or level test. Therefore, for each cohort, the null hypothesis fails to be rejected.

Within each age cohort and between these two groups, literacy growth in language appeared comparable. The experimental differences ranged from 0% to one 4%. No groups indicated effect size $> .33$ except for Cohort 10-Level 14 (fourth grade) which indicated an effect size of 0.76. The control group ($n = 9$) achieved a greater mean difference of 24 points over the experimental group ($n = 8$). In Cohort 9-Level 14 (fourth grade), the control group ($n = 24$) achieved a greater mean difference of 13 points over the experimental group ($n = 19$). Thus, both the older and younger students within the multiage classes who were designated as fourth graders achieved lower mean scores than the single grade fourth grade students, following the previous pattern indicated in Reading. Other results were mixed. Results from unequal sample size or small samples should be viewed with caution.

Missoula Achievement Level Tests

MALT Reading

Student scores were reported within the district in RIT scores, percentiles, and goal performance. The RIT score, an equal interval score, reports the test's composite reading score and quantifies growth. It was the only unit of measurement used in this analysis of student pre-and post tests. The "scale of difficulty for all the items in a subject area transcends grade levels, test forms and school years" (NWEA, 1996, p. 6).

Due to their ordinal scale of measurement, the individual percentiles and goal performances were not part of the analyses. As a note of interest, the percentiles form the basis for the reported goal performance for each strand in reporting to teachers and parents. Student goal performance within each strand is reported only as high, average, or

low. High indicates that a student performed above the 66th percentile, average indicates between the 66th and 33rd percentile, and low indicates below the 33rd percentile (see Appendix S). Individual student longitudinal reports present test results over three years indicating student growth in comparison to district and norm group averages, and ranks students as basic, proficient, or advanced along the RIT continuum (see Appendix T).

The null hypothesis. H_0 . There is no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohorts and the control (single grade) cohorts as measured by the Missoula Achievement Level Tests in Reading and in Language.

Descriptive statistics. Reading pretest data for each age cohort were summarized in Table 3 and followed by assumptions necessary to the statistical tests and analysis. Assumptions not met are noted. Also, within the Fall to Spring raw score data, some individual scores reflected no gain or a decrease. Because the MALT is constructed to provide the appropriate level of test, not too easy and not too hard, this was of interest. In consult with the MALT coordinator, it was learned that this does happen, even with retests and is not infrequent [L. Curry, personal communication, August 2, 1999]. Percentage of students that exhibit this phenomenon within each cohort is reported and discussed.

Table 3
Descriptive Statistics of Fall MALT Scores in Reading by Age Cohort

Group	<u>n</u>	<u>M</u>	Variance	<u>SD</u>	Skewness	Kurtosis
<u>Cohort 8</u>						
Control	37	194.11	94.82	9.74	-.75	2.19
Experimental	36	192.28	241.92	15.55	-.68	.73
<u>Cohort 9</u>						
Control	35	200.37	198.95	14.10	-1.27	3.07
Experimental	32	197.38	235.27	15.34	-.64	.21
<u>Cohort 10</u>						
Control	35	206.2	152.99	12.37	-.53	1.61
Experimental	15	202.6	362.11	19.03	-1.15	2.56
<u>Cohort 11</u>						
Control	17	206.12	122.74	11.08	-.31	-.91
Experimental	7	210.71	143.90	11.00	-1.01	-.082

From these data it is apparent that pretest scores are similar between the two groups. Means' differences are no larger than 4 points on the scale and all well within the expected Fall range of 186 to 203.8 (NWEA, 1999). Standard deviations are also below the expected Fall range (16 to 17) except EC10.

Assumptions Necessary to the ANCOVA

Assumptions of normality. The distribution of the scores around the mean within each group was normally distributed. Tools used to assess normality included the histogram, central tendency measures, skewness, and kurtosis. None of the above groups' scores deviated substantially from the normal curve, either from a "rough estimate" (Keppel, 1973, p. 74) which noted scores and estimated the general shape of the distribution, or by a histogram generated by GB-Stat. The standard deviations, which are

sensitive to extremes (i.e. outliers), indicated little variation between control and experimental groups within each cohort, another indication of the normality of distribution of each group. Outliers were left in as they are part of the reality of a classroom. The major concern would have been if more than 5% of the scores were beyond two standard deviations from the mean, but this did not occur within any cohort. The underlying distribution of each of the cohorts was consistent with all values.

The degree of symmetry of the distribution of scores around the mean is its skewness. Skewness generally ranges between -3 and +3, with 0 indicating exact symmetry of a distribution (Glass & Stanley, 1970, p. 90). Each pair of cohort pretest scores indicated a negative skewness, i.e., a tendency for a greater frequency of high scores than low. More important was the fact that each group's fall test is skewed in the same direction, indicating more similarity among initial scores.

Kurtosis is the property that describes the "peakedness" of the normal curve. A normal curve is mesokurtic with a value of 3. Of interest is that two groups approach this normal distribution: CC 9 (3.0775) and EC10 (2.55637). All other groups were less than 3 indicating platykurtic curves, or degrees of broader distributions, with scores that move from the center and tails into the shoulders, and well within a normal distribution.

The assumption of homogeneity of variance. This assumption requires that the variances of each group be the same, in that the "precision of result...is greatest when both groups are equal" (Hopkins & Glass, 1978, p. 257). But since the normal curve is theoretical, variances vary. Variance scores within each cohort pair were reasonable, as they are within the accepted standards' limit of the larger variance no more than four

times the smaller variance (Howell, 1997, p. 321). In addition, the analysis of variance and covariance are robust (Box, 1953, as cited by Christensen & Stoup, 1991; and Keppel, 1973; Wildt & Ahtola, 1978) with unequal sample sizes as long as the assumption of homogeneity of variance is met. GB-Stat formulates tests for homogeneity of variance. Homogeneity of variance was met by all cohorts unless otherwise indicated.

The assumption of homogeneity of regression. “Aside from the usual analysis of variance assumptions of normality and homogeneity of variance, we must add two more assumptions for the analysis of covariance” (Howell, 1997, p. 587). The first is that the covariate and post test relationship is linear. The second is homogeneity of regression, i.e., the incremental impact of the covariate is the same for all treatment groups. GB-Stat formulates the test for homogeneity of regression. It must be tested prior to interpreting results of the ANCOVA. If not met, an alternative parametric test, the Mann-Whitney U, is recommended. All cohorts met the assumption of homogeneity of regression, except for Reading 8 and Language 10. Alternative analyses were presented for these two cohorts.

Reporting of unadjusted and adjusted means. Both are reported within Tables 4-6 for Reading and Tables 8-10 for Language so that the reader is informed of the difference between the two pretest means in comparison to the difference between the two post test means that have been adjusted for differences in the covariate (pretests). If the F-ratio is not significant, this means that the adjusted post test means are much closer to each other than the original unadjusted means, and that most of the differences can be attributed to pretest differences. The adjusted means answers the question, “What if the covariate

means were the same?" If significant, it indicates only that a difference exists. Further investigation would be needed to determine whether the independent variable only had an effect upon the dependent measure (Huitema, 1980).

Results of the ANCOVA for MALT Reading by Age Cohort

Cohort 8. Cohort 8 met the assumption of homogeneity of variance. However, the homogeneity of regression data for Cohort 8 indicated an observed F -ratio of 4.79, $p = .0319$. This indicates significance, and thus heterogeneity of regression.

A plot of the data clearly revealed the similar nature of the slopes. The computation within this formula was running a test on the same differences and essentially divided by 0 because the two groups' scores were so alike. "In the covariance model the coefficient of the covariate is assumed to be nonzero. If this were not the case, there would be no benefit to complicating the analysis by the inclusion of the covariate" (Wildt & Ahtola, 1978, p. 28). The ANCOVA is not the appropriate statistical test for this group of scores.

The follow-up nonparametric test, the Mann-Whitney U -test reports a U -value of 541 which reveals a p value of $.1645 > .05$. Conclusions from these data warrant the null hypothesis fails to be rejected.

Continued analysis looked at the group mean scores and the practical significance of each. Both groups' Spring scores were above the expected spring RIT average (196) for this age/grade. More importantly, the expected learning growth from Fall to Spring for grade 3 was 9.8 points (NWEA, 1999). The control group gained 6; the experimental gained 12. The raw scores of these two groups revealed that in the control group, 22%

did not show growth on their tests. Out of 37 students, 8 achieved post test scores that were the same as or lower than their pretest scores. These scores included retest scores as well. Discussion of this phenomenon, evident in other control and experimental groups, is presented within this chapter. The effect size was .44.

Cohort 9. The ANCOVA reports an F value of 4.42 which reveals a p value of $.0392 < .05$. This indicates a statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohort and the control (single grade) cohort as measured by the MALT in Reading for the 9-year olds. Thus, the null hypothesis is rejected. The direction of difference is indicated by the experimental group's greater adjusted mean of 207.79 as compared to the control group's adjusted mean of 204.20 (see Table 4).

Table 4

ANCOVA Summary for Cohort 9

Unadjusted Mean Y 1 = 205.37		Adjusted Mean Y 1 = 204.20			
Unadjusted Mean Y 2 = 206.62		Adjusted Mean Y 2 = 207.79			
<u>Source</u>	<u>Sum Sqres</u>	<u>Df</u>	<u>Mean Squares</u>	<u>F-Ratio</u>	<u>Probability</u>
Between	223.30	1	223.30	4.42	.0392*
Covariate	8365.77	1	8365.77	165.73	<.0001
Error	382.11	67	50.48		
Total	11971.17	69			

Note. Effect size = .23

* $p < .05$.

The difference of one point in the unadjusted means and 3 points in the adjusted is minimal. The average expected learning growth for grade 4 of 6.5 points. In unadjusted means, the control group gained 5 points, and the experimental gained 9 points. Standard

deviations were comparable. This point spread is important to both groups, as the gain for the experimental group is almost twice that of the control group. Both groups scored above the expected Spring average (203) for grade 4. In this cohort, 20% (7 out of 35) of the control group, and 13% (4 out of 32) of the experimental group scores showed no gain or a decrease in test scores.

Cohort 10. The ANCOVA reports an F value of .25 which reveals a p value of .6189 > .05. This indicates no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohort and the control (single grade) cohort as measured by the MALT in Reading for the 10-year olds. The null hypothesis fails to be rejected (see Table 5).

Table 5

ANCOVA Summary for Cohort 10

Unadjusted Mean Y 1 = 210.29		Adjusted Mean Y 1 = 208.67			
Unadjusted Mean Y 2 = 207.67		Adjusted Mean Y 2 = 209.29			
<u>Source</u>	<u>Sum Sqres</u>	<u>df</u>	<u>Mean Sqres</u>	<u>F-Ratio</u>	<u>Probability</u>
Between	6.59	1	6.59	.25	.6189
Covariate	8429.38	1	8429.38	319.70	<.0001
Error	1766.55	67	26.37		
Total	10202.56	69			

Note. Effect size = .05

The difference of three points on the unadjusted means and a difference of one point on the adjusted means is minimal. The expected spring average for grade 5 is 210 with the expected growth of 5.4 points (NWEA, 1999). The control group gained 4 points, and the experimental group gained 5 points. In the control group 26% (6 out of 35) and in the experimental 20% (3 out of 15) showed no gain or a decrease.

Cohort 11. The ANCOVA reports an F value of 1.26 which reveals a p value of .2717 > .05. This indicates no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohort and the control (single grade) cohort as measured by the MALT in Reading for the 11-year olds. The null hypothesis fails to be rejected (see Table 6).

Table 6

ANCOVA Summary for Cohort 11

Unadjusted Mean Y 1 = 210.65	Adjusted Mean Y 1 = 212.56
Unadjusted Mean Y 2 = 216.49	Adjusted Mean Y 2 = 214.52

<u>Source</u>	<u>Sum Sqres</u>	<u>df</u>	<u>Mean Sqres</u>	<u>F-Ratio</u>	<u>Probability</u>
Between	30.75	1	30.75	1.26	.2717
Covariate	2206.00	1	2206.00	89.87	<.0001
Error	760.97	31	24.55		
Total	2997.72	33			

Note. Effect size = .18

The difference of 4 points on the unadjusted means and the difference of 2 points on the adjusted means were not differences of practical significance. Both groups met or exceeded the expected Spring average of 210. The expected learning growth was 5.4 points. The control group achieved 4 points; the experimental achieved 6 points. In the control group, 18% (3 out of 17) indicated no gain or a decrease; in the experimental group, all 7 students showed gain.

Analysis of Hypotheses for MALT Reading by Cohort

Cohort 8. The difference between the means of the 8-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis

fails to be rejected.

Cohort 9. The difference between the means of the 9-year old cohort control and experimental group scores is statistically significant ($p > .05$). Therefore, the null hypothesis is rejected. The direction of difference is indicated by the experimental group's greater adjusted mean of 207.79 as compared to the control group's adjusted mean of 204.20. The adjusted means' difference answers the question "what if" the groups had initial comparability of achievement.

The inference to be drawn from this is that differences do exist for these two groups after covariate adjustments. However, due to the nonrandom assignment of groups, causality cannot be inferred from this result. Rather, the results are observational and need further research to suggest causality (Huitema, 1980). The practical significance of these results indicate that the experimental group started lower and finished higher than the control. On an individual student level, the experimental group had fewer students (13% compared to 20%) reporting a no gain or decrease in pre- to post scores. This warrants consideration from both a classroom and district perspective.

Cohort 10. The difference between the means of the 10-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis fails to be rejected.

Cohort 11. The difference between the means of the 11-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis fails to be rejected.

MALT Language

Scores are reported in a manner identical to the MALT Reading test.

The null hypothesis. H_0 . There is no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohorts and the control (single grade) cohorts as measured by the Missoula Achievement Level Tests in Reading and in Language.

Descriptive statistics. Language pretest scores for each cohort are summarized in Table 7. Note comparisons between the standard deviations and variances, skewness, and kurtosis. Fall average mean scores were within only a few points of each other within each cohort. Assumptions not met were noted and warranted separate analysis. In addition, experimental differences ranged from 1 to 4%.

Table 7

Descriptive Statistics of Fall MALT Scores in Language by Age Cohort

Group	n	M	Variance	SD	Skewness	Kurtosis
<u>Cohort 8</u>						
Control	36	195.75	121.22	11.01	-.41	.60
Experimental	36	193.42	139.45	11.81	.85	.85
<u>Cohort 9</u>						
Control	34	201.44	225.47	15.02	-.51	.02
Experimental	32	196.94	230.55	15.1	-.56	-.31
<u>Cohort 10</u>						
Control	35	208.71	120.56	10.98	-.80	1.13
Experimental	15	203.07	240.35	15.50	-.09	-1.26
<u>Cohort 11</u>						
Control	17	209.35	85.62	9.2	-.03	.57
Experimental	7	208	159.95	12.65	-1.55	.82

Most of the Fall Language RIT scores for the four cohorts fall at or between the expected RIT averages of 188 and 205 (SDs between 14.95 and 15.24) for grades 3 through 5 (NWEA, 1999, p. 11). The exceptions are: control group Cohort 10's scored 208, and both control and experimental groups Cohort 11 scored above 205. This indicates development in language at or above MALT 1998 Fall norms (see Appendix N).

Results of the ANCOVA for MALT Language by Age Cohort

Cohort 8. The ANCOVA reports an F value of 1.72022 which reveals a p value of $.194 > .05$. This indicates no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohort and the control

(single grade) cohort as measured by the Missoula Achievement Level Test in Language for the 8-year olds (see Table 8).

Table 8

ANCOVA Summary for Cohort 8

Unadjusted Mean Y 1 = 201.36		Adjusted Mean Y 1 = 200.43			
Unadjusted Mean Y 2 = 201.42		Adjusted Mean Y 2 = 202.34			
<u>Source</u>	<u>SS</u>	<u>df</u>	<u>MS</u>	<u>F-Ratio</u>	<u>Probability</u>
Between	65.11	1	65.11	1.72	.194
Covariate	5710.48	1	5710.48	150.89	<.0001
Error	2611.52	69	37.85		
Total	8387.11	71			

Note. Effect size = .18

The expected Spring mean for grade 3 was 196.69 which both groups surpassed.

The expected learning growth was 8.9 points. Using unadjusted scores, the control group achieved 6 points; the experimental 8. Five out of 36 students indicated no growth or decrease in both groups.

Cohort 9. The ANCOVA reports an F value of .001 which reveals a p value of .9647 > .05. This indicates no statistically significant difference between the group mean scores from pretest to post test of the experimental (multiage) cohort and the control (single grade) cohort as measured by the Missoula Achievement Level Test in Language for 9-year olds (see Table 9).

Table 9

ANCOVA Summary for Cohort 9

Unadjusted Mean Y1 = 208.91		Adjusted Mean Y1 = 207.10			
Unadjusted Mean Y2 = 205.24		Adjusted Mean Y2 = 207.04			
<u>Source</u>	<u>SS</u>	<u>df</u>	<u>MS</u>	<u>F-Ratio</u>	<u>Probability</u>
Between	.06	1	.06	.001	.9647
Covariate	9743.26	1	9743.26	300.66571	<.0001
Error	2106.36	65	32.41		
Total	11849.69	67			

Note. Effect size = .00

In terms of practical significance, the expected Spring mean for grade 4 was 204 which both groups surpassed; the expected learning growth was 5.7 points (NWEA, 1999). The control group achieved 7; the experimental 9. An increase in growth occurred in all student scores except for two in the control and one in the experimental.

Cohort 10. The homogeneity of variance assumption was met. The homogeneity of regression for C10 indicated an observed F -ratio of 4.49, $p = .0378$. This indicates significance, and thus heterogeneity of regression. Therefore, the analysis of covariance could not be used. The alternative nonparametric test, the Mann-Whitney U -test for Cohort 10 reports a U -value of 217 which reveals a p value of $.3249 > .05$, indicating no statistical significance.

Cohort 11. The ANCOVA reports an F value of 3.96 which reveals a p value of $.0554 > .05$. Thus, there is no statistically significant difference between the group mean scores from pre- to post test of the experimental (multiage) cohort and the control (single grade) cohort as measured by the MALT for the 11-year olds (see Table 10). With an $n < 20$ and unequal sample size, interpretation must be cautious. In addition, a follow-up

test, the Mann-Whitney U , was conducted.

Table 10

ANCOVA Summary of Cohort 11

Unadjusted Mean Y 1 = 212.59 Adjusted Mean Y 1 = 212.29
 Unadjusted Mean Y 2 = 215.29 Adjusted Mean Y 2 = 215.58

<u>Source</u>	<u>Sum Sqres</u>	<u>df</u>	<u>Mean Sqres</u>	<u>F-Ratio</u>	<u>Probability</u>
Between	92.04	1	92.04	3.96	.0554
Covariate	945.47	1	945.47	40.71	<.0001
Error	719.88	31	23.22		
Total	1757.4	33			

Note. Effect size = .41

The results of the follow-up test, the Mann-Whitney U test, for Cohort 11 reports a U value of 43 which reveals a p value of $.2664 > .05$, indicating no statistical significance.

For practical significance, the expected Spring mean for fifth grade was 210, which both groups surpassed. The expected learning growth was 4.8 points. The control group gained 3; the experimental 7 points.

Analysis of Hypotheses for MALT in Language by Age Cohort

Cohort 8. The difference between the means of the 8-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis fails to be rejected.

Cohort 9. The difference between the means of the 9-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis fails to be rejected.

Cohort 10. The difference between the means of the 10-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis

fails to be rejected.

Cohort 11. The difference between the means of the 11-year old control and experimental group scores is not statistically significant ($p > .05$). The null hypothesis fails to be rejected.

The Spring expected means for grades 3-5 range from 196.69 to 210.71 (NWEA, 1999). All cohorts' unadjusted Spring means were at or above these ranges for each of their respective grade levels, indicating acceptable growth in language development. However, the individual scores that did not demonstrate growth need to be part of further analysis. In terms of practical significance, only one effect size warranted consideration, the 11-year old cohort, $.41 > .33$. However, due to unequal sample size, results should be viewed with caution. The experimental differences were from 1 to 4%.

The Writing Assessment

In the writing assessment, pre- and post test writing samples were gathered, first in January and then in June. Pretest scores were obtained five months after the introduction of the experimental treatment and therefore, an analysis of covariance was not appropriate. Only post test writing scores were used in statistical measurement. However, for issues of practical significance, descriptive statistics of pre- and post test scores are provided for the readers' information regarding numbers of students who increased or decreased their writing scores, as well as to indicate the range of writing scores within each cohort. Only students with both pre-and post scores were included. All student scores were four or above in fluency; all student scores were above emergent in conventions.

Fluency

Literacy definitions and standards involve authentic literacy tasks, with outcomes that demonstrate competence, as in “the clear, rapid, and easy expression of ideas in writing or speaking” (NCTE & IRA, 1996, p. 72) defined as fluency.

Description of Scales

Fluency. The criteria used in the raters’ assessment include distinct, related sentences; apparent story line; use of time words; sequence of events; writing with a beginning, middle, and end; and other features (see Appendix Q).

The null hypothesis. H_0 There is no statistically significant difference between the group mean post test scores in fluency of subjects’ writing samples in the experimental (multiage) cohorts and in the control (single grade) cohorts.

All cohorts from both control and experimental groups met all assumptions underlying the use of the t -ratio. Each t test was conducted using the pooled variance due to unequal sample sizes in each cohort. Cohorts of small sample sizes need to be interpreted with caution. Table 11 summarizes the fluency results by age cohort.

Table 11

T-test Results for Fluency in Writing by Age Cohort

<u>Group:</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>t-value</u>	<u>*p< .05</u>	<u>ES</u>
<u>Cohort 8</u>						
Control	19	5.18	.90	.14	.8866	0.08
Experimental	18	5.11	1.08			
<u>Cohort 9</u>						
Control	19	6.05	.97	2.45	.0191*	0.83
Experimental	20	5.25	1.07			
<u>Cohort 10</u>						
Control	21	6.67	1.32	2.32	.0265*	0.78
Experimental	14	5.64	1.26			
<u>Cohort 11</u>						
Control	5	7	1.58	1.24	.2468	0.63
Experimental	6	6	1.10			

*p < .05.

Analyses of Hypotheses for Writing Assessment for Fluency by Age Cohort

Cohort 8. There is no statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the fluency scores for the 8-year olds. The null hypothesis fails to be rejected.

Cohort 9. There is a statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the fluency scores for the 9-year olds. Therefore, the null hypothesis is rejected. The direction of difference is indicated by the control group's greater mean of 6.05 in comparison to the experimental mean of 5.25.

Cohort 10. There is a statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the fluency scores for the 10-year olds. The null hypothesis is rejected. The difference of direction is indicated by the control group's greater mean of 6.67 in comparison to the experimental mean of 5.63. Results should be viewed with caution due to the unequal sample size.

Cohort 11. There is no statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the fluency scores for the 11-year olds. The null hypothesis fails to be rejected. The results of this cohort should be viewed with caution due to the N of 11.

Summary of Fluency

Cohorts 9 and 10 each indicate a statistically significant difference. The difference in direction of both groups was indicated by greater mean scores of the control group. Enough cases were observed to provide a reasonable assurance that a difference exists. The results do not tell us why the difference exists. The effect sizes in three of the four cohorts are considered large enough to warrant consideration of practical significance. Experimental differences were larger than 5% in all but Cohort 8.

However, the difference in mean scores in both Cohorts 9 and 10 is only one point on the rating scale. The growth in fluency in writing is important to note, but both groups are still within the developing fluency phase. Both groups would appear to have made comparable gains of practical significance. Also to be noted is that there is no initial comparability of groups to compare beginning achievement levels. The focus of the

control school upon writing across all grade levels could be considered an extraneous variable, and will be discussed in Chapter 5. Combined with a nonrandom sample and small sample size of cohort 10, results should be viewed with caution. Of consideration are the descriptive statistics of individual pre- and post scores which demonstrate the numbers of students within each level and their development over the 5-month interval (see Table 12).

Table 12

Pre to Post Fluency Scores Within Age Cohorts Reported by Number of Students
Fluency Score

	<u>Fluency Scores</u>											
	4		5		6		7		8		9	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
<u>Cohort 8</u>												
Control	4	4	8	10	6	3	1	2				
Experimental	5	5	12	7	0	2	1	3				
<u>Cohort 9</u>												
Control			7	6	9	8	1	3	2	2		
Experimental	6	4	7	11	4	2	2	2	1	1		
<u>Cohort 10</u>												
Control			7	3	8	10	3	2	3	3	0	3
Experimental	4	2	5	6	1	2	3	3	0	1	1	0
<u>Cohort 11</u>												
Control			0	1	2	1	3	1	0	1	0	1
Experimental			2	2	2	3	1	0	1	1		

Another way to analyze the writing development is to compare the number of students who within each individual comparison either increased, decreased, or stayed the same in the ratings given. The following comparisons of individual growth include:

In fluency scores of control Cohort 8, four students increased by one level, nine stayed the same, and six student scores decreased by one level. In experimental Cohort 8, 4 students increased by one level, 13 stayed the same, and one decreased by one level.

In control Cohort 9, 5 increased, 11 stayed the same, and 3 decreased. In experimental Cohort 9, 5 increased, 10 stayed the same, and 5 decreased.

In control Cohort 10, 11 increased, 5 stayed the same, and 5 decreased. In

experimental Cohort 10, seven increased, one stayed the same, and six decreased.

In control Cohort 11, two increased, one stayed the same, and two decreased. In experimental Cohort 11, two increased, two stayed the same, and two decreased.

Conventions. The criteria used in the raters' assessment include conventions that interfere with readability such as punctuation, sentence fragments, run-ons and other features (see Appendix Q).

The null hypothesis. H_0 There is no statistically significant difference between the group mean post test scores in conventions of subjects' writing samples in the experimental (multiage) cohorts and in the control (single grade) cohorts.

All cohorts from both control and experimental groups met all assumptions underlying the use of the t ratio unless noted. Each t test was conducted using the pooled variance due to unequal sample sizes in each cohort. Results of conventions are summarized in Table 13.

Table 13

T-test Results for Conventions in Writing by Age Cohort

<u>Group:</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>t-value</u>	<u>*p < .05</u>	<u>Effect size</u>
<u>Cohort 8</u>						
Control	19	3.89	.81	-.64	.5254	0.21
Experimental	18	3.72	.83			
<u>Cohort 9</u>						
Control	19	4.26	.65	-2.14	.039*	0.71
Experimental	20	3.8	.70			
<u>Cohort 10</u>						
Control	21	4.48	.60	-.99	.2775	0.44
Experimental	14	4.21	.80			
<u>Cohort 11</u>						
Control	5	4.4	.55	-.58	.579	0.43
Experimental	6	4.17	.75			

*p < .05.

Analysis of Hypotheses for Writing Assessment of Conventions by Age Cohort

Cohort 8. There is no statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the conventions scores for the 8-year olds. Therefore, the null hypothesis fails to be rejected.

Cohort 9. There is a statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the conventions scores for the 9-year olds. The null hypothesis is rejected. The direction of difference is indicated by the control group's greater mean of 4.26 in comparison to the experimental mean of 3.8.

Cohort 10. There is no statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the conventions scores for the 10-year olds. The null hypothesis fails to be rejected.

Cohort 11. There is no statistically significant difference between the group mean post test scores of the experimental (multiage) cohort and the control (single grade) cohort as measured by the conventions scores for the 11-year olds. The null hypothesis fails to be rejected.

In terms of practical significance, three out of four cohorts, all but the youngest group, indicated an effect size $> .33$. In each cohort, the greater mean score was indicated by the control group. Experimental differences were larger than 5% in all but Cohort 8. However, the difference in mean group scores was less than one point on the rating scale. A reasonable analysis would be that this is not a difference large enough to warrant any conclusions toward preference for either structure. Also to be noted is that there was no initial comparability of groups. Descriptive statistics of individual pre- and post scores are summarized to demonstrate the numbers of individual students that demonstrated an increase in scores from pre- to post writing (see Table 14).

Table 14

Pre- to Post Conventions' Scores Within Age Cohorts Reported by Number of Students

Cohort	Low		<u>Scores</u> Medium		High	
	Pre	Post	Pre	Post	Pre	Post
<u>Cohort 8</u>						
Control	11	7	7	7	1	5
Experimental	9	9	6	5	3	4
<u>Cohort 9</u>						
Control	2	2	12	10	5	7
Experimental	12	7	7	10	1	3
<u>Cohort 10</u>						
Control	4	1	9	9	8	11
Experimental	4	3	9	5	1	6
<u>Cohort 11</u>						
Control			4	4	2	2
Experimental	0	1	4	3	2	2

In conventions, the following comparisons of individual growth are:

In control Cohort 8 scores by individuals, 10 increased, 7 stayed the same, and 2 decreased. In the experimental Cohort 8 by individuals: six increased, seven stayed the same, and five decreased.

In control Cohort 9, eight increased, six stayed the same, and five decreased. In the experimental Cohort 9, 10 increased, eight stayed the same, and two decreased.

In control Cohort 10, 5 increased, 15 stayed the same, and 1 decreased. In experimental Cohort 10, seven increased, six stayed the same, and one decreased.

In control Cohort 11, one increased, three stayed the same, and one decreased. In experimental Cohort 11, one increased, three stayed the same, and two decreased.

In analyzing writing development of students, five months is usually not considered a sufficient length of time to demonstrate large differences in writing, and results indicate this. Table 15 reports individual students that achieved a difference in both fluency and conventions, and the direction of difference.

Table 15

Individual Students That Received an Increase or Decrease in BOTH Fluency and Conventions

Cohort	Control			Experimental		
	Increase	Decrease	<u>n</u>	Increase	Decrease	<u>n</u>
Cohort 8	2	0	19	2	1	18
Cohort 9	0	0	19	4	1	20
Cohort 10	1	1	21	4	1	14
Cohort 11	1	0	5	0	0	6

This may suggest that it is difficult to increase writing skills in both areas at the same time. It also follows the literature that it is difficult to find an appreciable difference within a 5-month interval. Yet 4 students in the control cohorts and 10 students in the experimental cohorts demonstrated an increase in scores. This development is noted.

Analysis of the Procedures of the Writing Assessment

The opinions of each rater regarding the process were important to this researcher and solicited verbally during the 3-day rating session, and from a questionnaire filled out at the end of the session (see Appendix K). Raters were assured their responses would be

kept anonymous. In addition, as coordinator, I was free to listen and reflect on the process and comments throughout the three days. From handwritten notes and the questionnaire, the following narrative attempts to recapture the atmosphere of the 3-day event within two related categories, the collaborative process, and the raters' opinion of the scales. Direct quotes from the session are indicated by quotation marks. Quotations from the questionnaire are followed by a 'RQ' to indicate Response to Questionnaire. New paragraphs indicate a different speaker. Comments from discussion contained within researcher notes are indicated by 'RN'. Comments from observations begin with 'OC'.

Process of collaboration necessary to the HDWS:

"I enjoyed working with the people in my group, I felt that we were very compatible...Having three people in a group was perfect. I'm glad I was not doing this alone." (RQ)

"It was enjoyable working with a team."(RQ)

"It was nice to be able to read aloud and discuss the writing with two other people." (RQ)

"I enjoyed it. The process will be helpful in assessing my students' writing. I liked the process better than 6 trait. The team approach is an improvement over individual scoring"(RQ)

"I wouldn't have wanted to do this longer than the 3-hour time period, because my brain would get too tired to be effective." (RQ)

OC: During each session the raters worked in a professional manner. They were all always on time and in fact, some came early. One commented that she really "wanted

to learn something from this.” (RN) Except for pleasantries, little conversation exchanged on arrival: each was ready to work. This was not to say the raters were not congenial. In fact, they were all congenial and polite people. As each person seemed comfortable, with the most quiet persons able to express their views, the collegiality was evident. While no less was expected, in rating sessions with other groups, this researcher had experienced different levels of congeniality and collegiality that was dependent on the participants. These six raters were exemplary in their professional and personal conduct.

OC: On arrival the second day, one rater said that she really looked forward to coming again and that she felt very enthusiastic about the whole process. She commented that she was learning a lot (RN). Another indicated the same and brought a notebook the second day in which she kept notes of the procedures. She said it helped her in the scoring, and she wanted to be able to remember it and use it “all” in her teaching (RN).

Before each session, the researcher asked the raters if they had any comments or concerns from the previous day. This was to allay any problems with rating, and to see if the participants needed anything to facilitate their comfort zone. One team had remained in the classroom for one hour of rating on the first day. They commented that the fluorescent light was flickering. Even though the custodian fixed this by the second day, the team moved to the cafeteria on the second day to give them more space. The teams were across from one another in the large cafeteria and did not disturb one another. One rater commented that “once or twice a conversation was held right next to us.” (RQ). This could only have been this researcher and the facilitator. While conversations were held at a “library” level, it was not known at that time that this was a disturbance. Again

this comment indicated the level of professionalism of the raters; they were here to work. No tension was evident during any of the sessions between or among any raters. Both groups of raters demonstrated “time on task.” If any negative criticism can be made perhaps they were too polite, and should have spoken up when conversations near to them disturbed their work.

Comments about the modified holistic scoring criteria:

“It was a great way to remember/reinforce the fact that writing is developmental in nature. When we score children’s work, we need to remember that and devise or utilize a means to analyze the writing appropriately.” (RQ)

“This was a wonderful way to think about the writing process within the classroom. I have always tried to separate mechanics from the actual process and this experience gave me additional ideas.” (RQ)

“I’m glad I had the opportunity to look at this evaluation tool again because it reminded me of how affected I am by conventions. It also helped me see how a little coaching can help a student excel quickly.” (RQ)

“I was really impressed and excited by using this method. We learnt a lot about the process by going through the evaluations. I looked forward to continuing it each day.”(RQ)

OC: In the process considerable reflection occurred on the 4-, 5-, and 6-point level of the scales. A free flow of dialogue continued, and no one seemed rushed to move on to the next paper. With intermittent laughter and enthusiastic comments about some examples of writing, sincere enjoyment of the reading appeared evident. Comments

included “he’s very precise,” “ I like that intentional thing,” “ I think there’s more reflection than a critical 5,” “ I was thinking an 8 because of the reflection. What do you think?” “ This is so much easier than 6-trait,” “How about these adjectives!” (RN).

Raters referred to the rubric on a regular basis to check assumptions and decisions. Evidence of reflection was demonstrated by these comments: “What makes this a 6? It’s more fluid; it holds together,” “This would have been rated higher on fluency due to the very neat handwriting,” “I’m not sure. Let’s reread as a 4,” “It’s higher than a 5; it doesn’t ramble.” All through the assessment raters’ comments were to the point, demonstrated clear distinctions between rating criteria, indicated a collaboration among peers, and expressed not only knowledge about, but an enjoyment of reading children’s writing.

Additional comments recorded by raters on their questionnaires:

“Writing is developmental...Given this is the case, do our curriculum, the institutional constraints imposed by the district, and our effective teaching practices support the development nature of writing?” (RQ)

“I definitely learned as much as I could have in a university class for credit.

Thanks!” (RQ)

Raters’ concerns:

Raters at one point wanted more clarification on the terms “extensive, frequent, and occasional.” Other than referral for two crisis papers, several problem papers, and several comments on exemplary student writing, the teams worked separately. This researcher made every attempt not to interfere or make comments on any aspect of writing

or the assessment unless asked.

Out of all samples read, one rater indicated that one student wrote that the prompt was not very interesting (RN). This student's pre- and post writing in full context is included later in this discussion in connection with teacher comments.

Analysis of the Writing Assessment from Questionnaire Responses of Teachers

Student motivation was a primary concern, so the opinions of participating teachers were solicited. The writing assessment was conducted in each classroom by individual teachers at the request of the principal and after authorization from the superintendent. Standardized written directions for the teachers for both pre- and post tests were delivered to the principal, along with student directions and prompts, and bluebooks (see Appendix I). The distribution of these packets was conducted by the principal within the suggested time frame.

Since this district advocates process writing for students, and its own Spring writing assessment at the fifth grade level is conducted over a period of four days, it was anticipated that teachers would prefer a similar assessment. However, this "typical" performance as compared to a "best" performance attempted to control for outside assistance, indicating a measure of a student's independent writing, and also was necessitated by time constraints. Therefore, as part of the analysis, a questionnaire was included to be completed after the post writing samples were administered (see Appendix J) to gauge the degree of acceptance or nonacceptance of this writing sample by the classroom teachers. Results of teachers' comments are varied (see Table 16), but the majority were favorable. From the 11 classrooms, 11 questionnaires were returned.

Of the 11, 3 were blank, and 8 were completed, all anonymously as was indicated as a choice on the questionnaire. Of note, while all 11 classrooms participated in all measures, only 10 classrooms were included for statistical analysis as the 11th classroom was the only single grade in the experimental school.

Table 16

Teacher Responses to Writing Assessment Questionnaire

Question	Yes	No	No Response
1. Do you feel the writing topic was suitable for your students?	6	2	0
2. Did the student direction give them enough guidance?	7	1	0
3. Did the directions give you enough information?	8	0	0
4. Do you feel from observing the class that enough time was given for this prompt?	4	3	1
5. Has your class had experience writing on topics similar to these two prompts?	2	6	0
6. Have you had workshops on the teaching of writing according to the six-trait writing analysis?	7	1	0
7. Do you use the six-trait writing language in your writing instruction?	6	0	2 said "some"

Analysis of comments voluntarily added to questionnaires falls into two groups: those that seemed to accept the assessment and those that did not. The majority were favorable. Within the group that checked 'yes' more often came these comments:

My class wanted more time.

[Topic] could have been one with more interest to them.

[Regarding guidance] But we wondered if it all had to happen in [our city] -could we go to Flathead or Lolo Hot Springs?

We haven't written on seasons.

Self-editing and redoing is hard for them to do. Revising is hard to do without adult supervision.

They do write better on self-selected topics, though. You'll see that their spring stories take "birdwalks" into what they want to write about [smiley face inserted].

I suppose a higher skilled writer would be more able to stick to the subject.

Of the 11 questionnaires returned, it was clear that one teacher was dissatisfied with the assessment. She wrote:

Pretty difficult topic to get them interested in...This was not enough to give them ideas to write about. They were just not intrigued enough with the idea to develop ideas like they have for other topics...Very frustrating for students to get into a project and not necessarily have time to finish...But we write a lot on different kinds of topics...the timing of this project was very poor. If we had written this 2 weeks ago, the results would have been different. Students are very hard to motivate this late in the year. I also resent having a project dumped on me on a Monday morning and told I have to complete it by Friday. While I respect your project, to give me no earlier notice or a longer time frame to complete your project does not respect my time as a teacher. We have things we are trying to finish also, you know. (QE1)

However, all perspectives were welcome and solicited in this research. Overall, the response seemed favorable and the student writing demonstrated interest and growth. It is noted that the direct assessment, the writing sample, is bound by parameters just as the indirect measures were, and its limitations recognized.

Whereas writing assessments have struggled and will continue to struggle concerning the effectiveness of prompts, I selected this prompt as well as an independent, structured assessment as an attempt to provide equity for all ages and both organizational structures. Also, that some students may not be as eager on one particular day is a

constant challenge for teachers. As was mentioned earlier in the analysis of the writing assessment and one rater's comment regarding the student who said he was bored, this student's writing needed to be noted, as a fine example of the challenges we face. The student in his pre-writing sample wrote (the following text is exactly as he wrote it):

There is nothing in the winter that is to much fun. I do not like this topic. I would rather set on my bum. But sence you are making me, I guess this is a start, to a very boring story, so here is the first part. It was a day with lots of snow, it look like a lot of fun ya' know until on came the news. It said it was 20 below and there there were boohool (exsspelly me!) Thats why I don't like this topic you see. I told you it was a very boring story. So the next time you ask us to, please make it fiction, the only thing I can say now is the Spanish word Fin (end).

Consideration for an interesting topic was recognized in the criteria necessary for a standardized prompt. In the directions for the prompt the stated option between fiction or non-fiction apparently wasn't made clear enough for this student through teacher instructions. However, it is interesting to note that in the post writing, essentially the same topic but a new season, the writer appears to be in a better humor. Even if he was bored, he demonstrated more enthusiasm in his writing. Transcribed here just as he wrote:

It was wonderful spring day and I had to spend en school. But luckely my mom and dad decided to drag me home to go camping. Unfortunatly it would took us an hour to get there. We were going to flathead lake but today it took only an hour because we held up a sign telling people that there tires were flat. When we got there the lak was shimmering and the place was peaceful. But the best thing was was that there were no girls. The End (not). Then a girl happen to pass by. "Nooo"! I thought. "My vacation ruined!! The only thing I could do was ruin her. So I went down to the lake and I happen to find a snake. I grabbed the snake and snuck behind the girl and dropped it down her skirt. The luches sensation of screening. I it wasn't so bad after all. The End really.

Once again, it is important to recognize that writing, as in the other two measures, is just one sample of performance at a one given time on one particular day. With regard

to student motivation, the control school's principal was asked about the effect of the impending school closure upon the test results of the control school's students, and if she felt it would be a factor in analysis. She responded: "Absolutely...that the students were "pretty disconnected. Lost. A lot of unknowns. Even though they knew the schools they would be going to...it is a definitely [a factor]." (3IP17) The notification of the control school's closure and possible effects on motivation during end-of-the-year testing are considered extraneous variables.

Summary

In total, the inferential statistics indicated no pattern of statistically significant differences in academic achievement between the students within the multiage and single-grade cohorts as evidenced through results of each of the three types of quantitative measures. Table 17 presents a summary of these results.

Table 17

Summary of Quantitative Results for Each Measure in Each Cohort

Cohort	<u>TerraNova</u>		<u>MALT</u>		<u>Writing</u>	
	Reading	Language	Reading	Language	Fluency	Conventions
Cohort 8 Level 13	Experimental*	No	No	No	No	No
Cohort 9 Level 13 Level 14	No No	No No	Experimental*	No	Control*	Control*
Cohort 10 Level 14 Level 15	No No	No No	No	No	Control*	No
Cohort 11 Level 15	No No	No No	No	No	No	No

* indicates statistical significance with $p < .05$ in the direction of the group stated.

Differences of statistical significance were reported among 5 of the 28 possible cohort measures. Twenty-three did not indicate statistical significance. Among the five, statistical significance on the indirect measures were indicated by the TerraNova reading scores of the Level 13 test for the experimental group of eight-year olds, and MALT reading scores for the experimental group of nine-year olds. Among the five, statistical significance was indicated on the direct measure with the control group of 10-year olds in fluency, and the control group of nine-year olds in both fluency and conventions. In addition, it should be noted that three out of the five tests with statistical significance were within the 9-year olds' cohort, two of which favored the control group; one the

experimental. Neither indirect measure in language indicated statistically significant differences in any cohort. In none of the other tests did the oldest students demonstrate statistically significant differences, a fact that would follow much of the literature. Twelve out of 28 effect sizes $>.33$ suggests that further study is warranted. None of the experimental differences within the indirect measures were above 4%; however, within the direct measure 5 out of 8 were 5% or above in favor of the control group.

One group that indicated a pattern of differences in practical significance was the fourth grade level group. The 9-year olds in fourth grade and the 10-year olds in fourth grade in the control single-grade groups indicated greater mean scores than their experimental multiage counterparts in both TerraNova results, and half of MALT results.

An important reminder for inferences is that with larger samples, i.e. > 30 , the smaller the observed result required for statistically significant differences; conversely, the smaller the sample, the larger the observed result required (Borg et al., p. 164). Small N s are to be regarded with caution.

Also, it is important to note (a) inferences can be drawn that differences exist, but not the cause of the differences, and (b) that “failure to reject the null hypothesis often means that we have not collected enough data” (Howell, 1997, p. 93). This is one group of students from one school year. Evidence of their literacy development over a period of time would ferret out extraneous variables present in this particular research and is a recommendation for future research.

What was of most interest and will be more clear with the following qualitative analysis were the areas in which the statistically significant differences occurred. From

this combined design study, the data from interviews and documents yielded information that posed a unique consideration for the focus of literacy development in this study.

Findings and conclusions are discussed in Chapter 5.

Qualitative Components

In this combined design study, the less dominant paradigm explored the impact of organizational structure upon literacy development of upper elementary students through interviews and document analysis. Research was directed toward, but not limited to, these questions: What are the instructional programs and practices within the single grade and multiage organizational structures? Does literacy growth differ within the age configurations of the two types of organizational structure? From the triangulation of data, four major categories emerged: historical origins, leadership, meeting students' needs, and commonality of experiences.

Overlap exists within the categories as they connect and support each other. Subcategories within two categories emerged as well. Quotations from interviews and documents substantiate each category. In parentheses, following each quotation, is a code denoting type of data, number, page number in transcript, and type of source.

Historical Origins of Alternative Organizational Structure Within District

Prior to 1990, single-grade classrooms constituted this district's organization. In the early '80s, the district schools made a transition from K-8 schools, to K-5 and middle schools 6-8. The experimental school in the study remained a K-8 school for two years after other district schools implemented the elementary and middle school transition.

Eventually its sixth, seventh, and eighth graders were bused to middle schools. For five years, a K-5 structure with kindergarten, transitional kindergarten, and single-grade classes was in place similar to the other district schools.

In 1990 this school initiated an alternative organizational structure for its primary students in the form of multiage grouping. Within this category of historical origin, a statement made in 1990 by one teacher reflected what had been a growing opinion of the staff since their transition that: "Education as it exists today was not working at our school as well as it was in other parts of the city."(3N90T) This was substantiated in the statement of another staff member: "Nobody was satisfied with how things were going." (3N90T)

Statements of the condition of the school climate during this time reveal a frustration with the fit of the system to the population of students. A clear example is a statement by a staff member who had taught at the school since 1988, and who is now the principal:

It was clear by the end of the first quarter these kids had needs that I had never seen before...it was culture shock...we had conversation about doing all these things, what we weren't doing...what do these kids need. (11P)

The following excerpt emphasizes the needs as the staff perceived them and presented them to the visiting state governor in 1989:

Teachers often have to instruct the students in how to set alarm clocks and make their own breakfasts and what to do when they're home alone at night. Often the students are tired because they didn't go to bed until 11:30 p.m., or were afraid someone would come into their room, or were wondering if their parent would even come home at night. (1N89T)

This teacher added to the staff theme that survival had to be taught first before the basics: "...counselors [and nurses] are needed for the children, some of them 5- to 6- year olds who already have problems." (1N89T) Personal and social needs due to one-parent families, transiency, and attendance were stated as factors contributing to behavior and discipline problems at this school. These concerns seemed precipitating factors for teachers in their request for change. However, the fact that academic achievement, as reflected by test scores, was also a concern is reflected in this teacher's statement: "Our scores are constantly lower in comparison to other districts." (1N89T) She concluded: "Many students improved from the 10-20 percent levels they came in at...[and] test scores are but one measure." (N89T)

The governor proposed more faculty involvement in the curriculum. One year after the governor's visit, the school again made the news in 1990 with its advocacy of an alternative approach. That intervening summer, two teachers and the principal attended the 20th National Alternative Education Conference in California. Supported by research from this conference, the staff began discussing change directed toward an alternative organizational structure. An attending teacher stated: "[The research] showed that 50 percent of the population doesn't learn well in a traditional learning style, such as sitting at desks." (3N90T)

That the administration agreed with the teachers that reform was needed is reflected in later statements by the school's principal in that:

Factors such as the school's transient student population, behavior problems, attendance and the many 'non-traditional' or single-parent families in the area [as well as]...the poor performance [of this school] on a national achievement test taken yearly by all District students also contributed to the staff's decision to

investigate alternative schooling...scores 20 percent on the average lower than the district average. (3N90A)

In the earlier quote from a teacher that nobody was happy with the way things were going, if the “nobody” included the administration and the current school board, it appeared from public comments made by the 1990-91 administration that the responsibility for decisionmaking was in the hands of the faculty. The assistant superintendent said:

School board members and administrators are eagerly watching the developments...they’ve put the focus for change at the staff level...the staff needed freedom and flexibility to make responsible changes. The main learning goals will still be taught, but the way it’s taught may vary. (3N90A)

An editorial in the local newspaper in September 1990 may have provided impetus to change. Declaring that high test scores aren’t the same as a good education, it went on to praise the district for its ranking in the top 10 percent of schools nationwide on a standardized test. One-third of the district schools achieved this rank. The editorial continued: “By and large, schools that produce students who score well on tests are probably doing a better job than those whose students consistently perform at lesser levels.” (4N90E) With such a public statement, schools that did not meet these standards must certainly have felt pressure, even though the editorial continued with:

Credit for students’ good test scores goes beyond the classroom...[along with administrators and educators] don’t overlook the contribution of people who play important roles outside the classroom - the district’s parents and taxpayers. They, too, are essential [to a] winning team. (4N90E)

This is an interesting point that speaks to the issue of accountability as well as leadership. The staff accepted the responsibility and implemented alternative measures.

Later, as controversy arose, the superintendent in 1991 was reported to have said: “The

decision on the program's future [is] in the hands of the...staff. It's a local school matter initiated at the local school level. We'll support it and we don't intend to get involved."

(5N91A)

A newspaper article in 1991 reiterated the factors of "lowest test scores in the District, a raft of disciplinary problems and [the school's] highly transient population" (N910) as the reasons for change. In spite of some controversy, this organizational structure has continued to the present. Yet not until the 1999-2000 school year have all classes except kindergarten been multiage classrooms. The change has evolved over nine years with minimal staff replacement.

From examples of the origin of the multiage structure of this school emerged a second category. While the administrative support was present, a sense of responsibility from the teachers precipitated change to meet the needs of this population of students. This set of circumstances leads us to the second emergent category of leadership.

Leadership

Within both schools today, leadership has promulgated programs, training, and schoolwide practices. But evidence indicates it is a limited collaboration of leadership. Notwithstanding the responsibility of an administrator, it is clear that staff at both schools were encouraged and provided with the opportunity to explore what they felt would work best with their population of students.

At the experimental school, it was evident that, with principal support, some members of the staff were instrumental in implementing change as reform. With the 1991 superintendent's statement above, the staff was given license to continue. In the

ensuing years, one teacher who had been part of the change became the school principal and continued through 1998-99. She remembers that, as a teacher from 1988-92, the staff made decisions through collaboration:

We had conversation about doing all these things, what we weren't doing, what do these kids need? We decided, and the principal was very supportive that as a building we would participate in the Onward to Excellence and focus not necessarily on the academic portion but more on other needs first. (1IP6)

She recalls that they decided to begin with first and second grades only: "The staff talked with a variety of people, what research was out there. The principal who was visiting from Australia talked with us for a day...how some schools are structurally different...."(1IP6) Even though she gives credit to the "gains made initially with multiage" to the other two principals involved, she alludes to the need for stronger leadership in the ensuing years:

T. was the principal for one year, E. for two, and myself for six. Looking back, somebody needed to draw some lines in the sand. Those lines must say that this is expected of you as a teacher. I talk with teachers. Let's look at the evolution. Every time I am in a class, I look for a piece of what I asked for. We have a meeting, and if it's word meaning, we talk about. We get it squared out...This is a good place for kids. (1IP9)

The collaborative leadership between principal and teachers in one school is in evidence at the other building. From the control school principal comes the statement: "The staff decides what is our most important goal here." (3IP5). During the six years as principal, she has facilitated a district curriculum arts adoption as well as elementary grade level meetings, particularly at the primary levels, leading to policies adopted by the staff. Stating that she regularly attends literacy workshops she adds:

Usually when I go somewhere I take people with me...we can share our learning afterward, and I can be supportive...I do a tremendous amount of reading. (3IP3)

This is supported by one teacher's comment regarding this principal's leadership: "She was very supportive. We always went to workshops and discussed them. She was very concerned about the students and how we would work with their needs." (II99ct) Her commitment for providing staff training and introducing them to new ideas is illustrated in several statements:

I facilitated the training of at least, the very least, five people in my building for Reading Recovery training...When I took a couple of staff members with me...to the at-risk conference in Phoenix...it was a whole new concept for them...they seemed quite taken aback that they should be thinking in terms of accelerating students learning. (3IP4)

Frequently this principal spoke in terms of the staff's recognition of the needs of the students and is reflected in a statement regarding curriculum:

It was never our intent to arrive at one best way for everyone to use in the district, but rather for us to arrive at one best way that we continually refined to use with students at [our school] tailored specifically for our population...teachers need that latitude, that decision ability. I certainly if I were still in the classroom I would be one teacher standing right on my back legs and saying...I know what my kids need, and how best to deliver that and I'm continually adjusting that. Please give me the respect that I'm due as a teacher...even though I haven't been in the classroom for a long time, I remember how that felt. (3IP9)

While these statements make a strong case for collaborative leadership, the similarities between the principals and within their staffs contributed to making this possible at each location. From interviews and documents, a common picture emerges. The principals' length of leadership within each of their schools is six years, both had extensive experience as teachers, and both express intense satisfaction with their careers. Both displayed enthusiasm toward their job in all interviews. Both were supportive of this research. With 29 years in education, the control school principal had taught K-12 at every grade, including art. With an advanced degree, she had a total of 17 years'

experience as a principal, including within three other states: "I've been in four states in four separate school districts and that has been very enriching to me overall to see how there are such similarities. Yes, I've made an abbreviated loop." (I399p) Much of her discussion involved the use of "we" and "us."

With 20 years in education, the experimental school principal had taught grades 3 through 8. She has an advanced degree, and prior to becoming principal at the experimental school, was employed as a teacher at this school since 1988, with one year of experience outside of this district. She states:

Education has been my life. My dad was a teacher, and I have followed in his footsteps. I love education, and I love the classroom...I regret that I don't get to teach...If I were teaching again it would be 3/4/5. It did that much to change my inner core. (I199p)

She also relates her feelings about teachers' choice: "For any school that wanted to do it [multiage] to feel it is a better education for kids you have to have people who want to do it....I would never force anyone to do multiage." (I17p).

Differences exist between and among the principals' staffs. Between the two schools, the range of teaching experience extended from 35 years to seven years, with the control school having more teachers with seniority. At the control school, each teacher's years in teaching were: 35, 29, 27, 21, 20, and 8.5 (average 23.4; median 24); at the experimental school, years in teaching were: 19.5, 17.9, 13.2, and 7 (average 14.4, median 15.5). Three of the six at the control school, and two of the four at the experimental school had completed either an M.A. or M.E. degree. Each group had one teacher with an endorsement in Special Education.

Due to contractual procedures between the district administration and the teachers' union, not all teachers had a choice of placement at each school. In addition, involuntary transfers to and from each of the schools have occurred. However, voluntary transfers are an option if openings at other schools exist, or if teachers at other buildings wish to switch building placement. All teachers in each building had been in their school for at least three years.

From events from 1990 to the present, the term 'choice' was prevalent, not only in principal and teacher perspectives, but parents'. Even though one teacher stated in 1991 that by the second quarter most of the parents were proponents of the new system, not everyone was satisfied. A subcategory of parental choice emerged and is integral to historical origin and the leadership that guided the reform.

Parental choice. During the first year of implementation, several statements make clear the dissatisfaction of some parents with the change to an alternative structure. Elimination of choice seems to be a predominant theme as one parent is quoted as saying: "The school reneged on a promise to offer them a choice between the traditional classroom setting and the new system." (5N91Pa) Another parent complained that:

We thought we'd be given the choice of traditional or the new learning group. But three days before school started they said we were out of luck. They said, "if you don't like it, take your kids to another district. (N91p)

Indeed, the original statements from the school staff indicated that parents would be consulted first. Consider the statements, first from a teacher: "The [Stanford] conference showed the importance of offering choice." (3N90T) And second, from the administrator: "Nothing will be changed until parents have been consulted...parents will

be notified about any proposed changes through the mail...a key element of alternative education is to encourage more parental involvement.” (3N90A)

Another strategy to meet the school’s needs was an advisory council formed during the summer before implementation in 1990. This advisory council became one of the first in the district and consisted of: “Five parents, three teachers, school counselor, principal, two university education professors, a citizen-at-large and a home-school coordinator [who] met on a biweekly basis to discuss problems and talk about goals.” (3N91O) The first objective of the advisory council bylaws was stated to be “the education of parents and community about school programs and operations” [and] “mobilizing and coordinating all community resources in a concerted attack on the problems of children who are at risk.”(D6sd)

As a forerunner to a later school board policy that adopted school centered decisionmaking, “these councils represent the views and direction of a school community over time providing a dimension of stability for the school....[role is to] positively affect student achievement in a collaborative manner” (D96BP/AR 6001). The school board was changing, and seemed to relinquish some of its central authority for the collaboration of diverse groups.

Yet some parents felt left out. According to the principal, he sent every parent a letter in August, inviting them to a meeting. The principal is reported to have said: “Parents were given the option of requesting an “attendance-area exemption” allowing them to take their children to another District school at no charge. However parents [were] responsible for providing their own transportation.” (5N91P) One month earlier

in February 1991 the teachers were reported as saying that because they felt the system may not be not suitable for all schools or even every child: "You're looking at a school within a school...if a parent requests a traditional program we will provide it." (4N91T)

In March came the statement from the principal that the school would consider: "Offering traditional classroom settings in the primary grades if there [was] enough demand. A minimum of 30 to 35 children would be needed before the school would consider offering traditional first and second grade classes." (5N91P)

The issue of choice of program is one that surfaces in more recent discussions, but with a reverse focus. While an open enrollment district, students outside of the neighborhood boundaries are admitted to a classroom only if the student number does not exceed state standards. Therefore, if class sizes are below state standards, any student from any part of town can attend a school of their choice. However, in the past several years, numbers have not afforded that option at all schools. It is reasonable to assume that transportation would affect whether or not parents have a real choice.

In the experimental school, options between single-grade and multiage classes have not existed since 1990-91. In other parts of town, the same non-option occurs, only in reverse with only single-grade classes available except in one school. The issue of choice re-emerged in 1995 with the Committee for a Magnet School whose purpose was: "...to convince the local district to open a K-5 school that would offer an alternative curriculum that allows students to learn at variable speeds...to include multiage options." (16N95P) Note the specific delineation of "alternative curriculum...to include multiage options." The committee sent out surveys to parents and teachers, meetings were held,

and options were discussed. The survey received a 10 percent response from teachers, with 22 expressing interest (Dms96s). Nine percent of the district parents responded. It was evident that some parents felt some students' needs were not being met by the present curriculum or organizational structure. The issue of equity of choice was evident.

District administrators and teachers were faced with: When reform ideas are raised, to whom does the system listen? Which reform? How far should it be carried? In which direction? Administrators responded to parental requests through dialogue and public meetings. Several options were presented. A review of the issue and analysis of the options with fiscal projections were discussed at meetings and a school board special work session, with an application for magnet schools provided to the committee. At the same time a position paper submitted by the principals indicated that if the board was going to consider this change, the principals needed more planning time for research, citing "undefined and unresolved issues" (FD96p.01). Academic and fiscal accountability combined during a time of severe budget constraints. From these categories of historical origins and leadership evolved a third category integral to both: Meeting students' needs.

Meeting Students' Needs

The connection between the first two categories and this third one is made clear from previous statements that one population of students needed something different. The teacher who later became the principal believes that the change implemented was in the best interest of the students. This is demonstrated in her comments:

I made a conscientious choice to be in multiage. I felt kids would benefit. It was one more thing we could do. The data shows that academically they hold their own, are the same. The other thing which is striking is the ability to have life skills that employers want such as initiative, flexibility, cooperation, empathy, courage to do something else...it's tough to test. Those are the things these kids have more of...[our school] will struggle forever on standardized tests...Until other basic needs are met, the child is not ready to learn...there are so many things like divorce, etc. (1118,7p)

During the school year 1990-91, the experimental school implemented three major changes: a method of organizational structure that “scrapped age divisions,” assigned students to one teacher for up to three years, and eliminated report cards with grades. Consistent with meeting the needs of the student one teacher said: “We could implement fast or slow, depending on where you were at the time. We decided the first year that we would begin with first and second and aim for intermediate later.” (11p) At the time other statements made about the change included:

The most radical changes to date have taken place in the primary grades which up until last fall consisted of first and second grade and transitional kindergarten, for students who have completed kindergarten but are deemed not quite ready to advance to grade 1. Now, instead of dividing students by age, all 154 6- to 9-year old pupils are divided equally into seven primary classes. (N91T)

Upper elementary students remained in single-grade classrooms until 1997, and not until the 1999-2000 school year were all classes multiage. The school's web page states the school strives

to provide a developmentally appropriate education for each student in a safe, stimulating supportive environment...multi-age/Grade classes are based on flexible grouping which allows children to participate at their own level of learning and social interaction. (5S99)

This recent technological document reiterates the early goals of the staff, evidenced by the following teacher's comment made during the first year of change: “Kids work at

their own speed and with each other. Sometimes the older ones will do reading for the younger ones.” (5N91T) Another teacher repeated the “work at their own speed” and added that: “By mixing them up you no longer have older kids saying ‘I’m the dumbest kid in the class.’”(N91T)

Other conversations addressed the affective needs of the current population.

“Dependent on their age, Myers [the gym teacher] said students can only climb to certain levels. However [I] let children experiment to some degree on the wall...we’re trying to do developmental things that are good for kids (Jahrig, 1997, p. B2). It appears that age is still considered a factor in some departments. Connecting with this issue of self-esteem, the principal stated:

And so guess how, they’re fourth graders, you’re in the dumb group... When we group kids and say you are all fifth graders and for some reason little Joey can’t do math at fifth grade, then everybody says he’s dumb. Multiage, I’m sure, and I’m not going to be naive enough to say that doesn’t exist in there, but I think that is far less a factor. (2IP11)

Age spans from the target beginning date of October 15 were the following: the experimental school ranged from eight years old through 10 in the 3/4, and from nine years old through 11 in the 4/5 (difference by months was not available). At the control school, the age spans in months within each of the single grade classrooms varied from 12, 13, 16, 16, 18, and 26.

One teacher stated that the new structure met the challenge of transient rates during the first years of the change and helped students’ affective concerns: “New students coming in can fit in where they belong. You don’t have to hold back the rest of the class while they catch up.” (2N91T) While transiency rates remained high, in the

later years it had been decreasing. The principal stated: “[But] the transient population is not what it used to be [in this school area]. The neighborhood is changing.” (11P6)

One major project demonstrated the burgeoning community aspect of the neighborhood: Project Playground, constructed on the school’s adjoining park area by the area parents and neighborhood association, used schoolchildren’s ideas. Other projects include the Neighborhood Tool Library, where tools are loaned out for home maintenance for the growing number of home buyers and more permanent renters. Fundraising by the parents, in this area whose 1990 median income was reported to be \$14,750 compared to the city’s \$21,033 helped construct the climbing wall within the school (Chaney, 2000).

The control school has garnered much parental support through many different programs emanating from the school, and federal grants obtained largely through the efforts of the principal. In addition to having the largest multicultural diversity within one school in the district, almost 50 percent of this school’s population was bussed from other areas of the city. Through the grant writing activity of the principal, several literacy programs existed. One tutoring program taught through RSVP volunteers, ranging from one 85-year old to high school honor students. The coordinator reported: “Because 24% of the population are bilingual, tutors focus on reading skills. When they are learning English in school, they go home and their parents don’t speak English. It’s a big problem.” (22N970)

A summer school program offering classes for bilingual students was housed at the control school. Available to all district students, many of the area students’ progress

in literacy may have been affected by this extended learning time. Yet not all students approached grade level expectations, and so arose the question of retention.

Retention. When asked how retention was handled, principals responded in a similar vein. Witness these remarks from the experimental school principal:

Often times there are kindergarten kids who aren't, you know maybe would benefit from another year of kindergarten...So we'll talk about where they are and look at some of their assessment results, and if we believe that it is in the best interest of the child to stay in kindergarten, then we will make that recommendation. There probably aren't very many, maybe one or two a year...if we still see an issue, we say to the parents...we're recommending to you that your child spends three years in the primary...three years in either a 1-2, or three years which would be 1-2, two years then and then move to a 2-3. (2IP8)

Although she has said there probably aren't very many, she continues within the same discussion:

The most interesting thing that's happened though is we make lots of those recommendations and they don't need to stay three years because they're in a situation where maybe they needed longer with math and suddenly their math comes around and they don't need it. There's a couple kids I can think of right now that did three years in the primary and it was the best thing for them. There were some...it wasn't necessary. (2IP8)

From the other school the principal addressed this issue in that they were always looking at other viable options first, and to her recollection, no students had been retained during her six years:

We want to give the gift of time, to have students be able to accomplish what they need to accomplish. We are always scrambling for strategies and always discussing it...We agonized over some students. We would discuss at length. The teachers often felt that it was too late in a social sense. If they kept them another year, what would be the advantage? And these are strong teachers, it is easy to listen to their opinion. They know the families and their history. It was never an easy decision...due to the transiency rate, and the history with the population, retention would probably not do a bit of good. All the factors the child has to deal with are considered in such a discussion. (3IPtc)

The same feeling is part of these remarks from the experimental school principal:

We haven't done retention with kids that are in the intermediate or 2-3...I remember retaining some kids myself early on in my teaching as sixth graders and wondering why I even thought I should do that. There has to be another way and the better way is to diagnose and say these are the gaps. (2IP)

Although the term was never brought up by any interviewee, the availability of continuous progress at the multiage level was demonstrated in this example:

We had a girl who was an older kid when she went to kindergarten. She was age appropriate but for some reason she was older than the other kindergartners. Then in the 1-2 the first year she was a superstar and at the beginning of second the teacher could see she was way beyond any of the other kids. So at the end of the quarter we moved her into the 2-3...we haven't skipped her, but so she will be in there one more year as a third grader...we had two boys already old when they entered kindergarten. They didn't enter kindergarten with us, but then were retained in kindergarten because of absentee issues. They stayed in a 3-4 combination one year and then were moved to fifth grade. They would have both started their senior year of high school as 19-year olds, and we knew they would not be around then. (1IP8)

Both principals indicated aversion to retention and used different ways to avoid it, saying that kids need every opportunity to be successful. The control principal seemed to suggest that retention was avoided when the possibility arose, as in this case:

You just find other ways around it...we did some limited kind of, well, parents would request for example when a teacher moved from kindergarten to first grade a lot of parents requested that the child be with that teacher again. so it was a parent request kind of thing. (3IP15)

Assessment and evaluation. The experimental staff had eliminated report cards with grades, using instead a written narrative of progress for every grade level. The narratives were different at primary, intermediate, and fifth grade level:

The third and fourth grade report is two pages, and essentially it's a checklist with some of those categories...the fifth grade is like long legal paper, two pages and essentially lists the skills from the district curriculum and then we use...acquire, practicing, mastery...you know the interesting thing about it is people can actually

if they wanted to associate a letter grade with those things, too. (2IP7)

The principal remarked that she had one parent write at the end of the year they never got a report card all year long, "because I guess because it wasn't a card like probably they thought it should be." (1IP7)

At the same time, the control school principal uses the district report card with letter grades because: "Teachers were unable to choose due to the district mandates. So they did what the district required and then they did narratives above and beyond...they actually did two sets at the lower elementary...tremendous amount of work." (3IP13)

So while one school dispensed with grades concomitant with the alternative organizational structure and used narratives, the other continued to use letter grades due to district mandates, and supplemented with narratives. An exception to the mandate seems to have been allowed to exist.

At each school, the profiles of the student body prior to this year of research indicated similar educational challenges, documented by both schools meeting Title I requirements each year for the past six years. In 1996 both schools submitted plans for schoolwide Title I programs which would enable them to address the needs of more students. Each had investigated schoolwide status in 1994 and begun assessment in 1995. The following statistics from the schoolwide program plan descriptions for 1996-97 for both schools indicate the two schools' comparability in SES, transiency, and ethnic diversity during the 1995-96 school year:

Demographics of students receiving free and reduced lunch based on parental income was 56.16% for the control school. The principal stated that 68% was the

maximum qualification during the recent history. For the experimental school, 76% qualified with 34.4% of its families living below poverty level.

Transiency rate for the control school was 37.62%, but ranged from 30-50% over the 3-year period. The experimental school stated no percentage, but mobility/stability raw data: "in August and September 1995, there were 120 transfers to and from [this school]. During the period of October 1, 1995 to February 27, 1996, students transferring in and out totaled 102, bringing the total number of student transfers in and out to 222" (MCPS, 1996, p. 5). Challenges created by transient students seem compounded at the experimental school, contributing to behavior and discipline problems in earlier documentation. As stated by the principal, the degree of student transfers had been a significant problem for the experimental school, and one of the major reasons for changes in organizational structure. However, over the past few years, the rate had been declining, and as both principals commented, students entered, moved, and returned.

The ethnic diversity of the experimental school was stated to be 12 Russian, 4 Hmong, and 22 Native American students in May 1996, which would be 10% of the stated 389 total population. At the control school, the number of minority students was 23.19%, with predominantly Hmong, Native American, and Russian populations, but was as high as 28% within the last few years, and the largest within the district. Literacy challenges inherent in teaching children for whom English was a second language was evident at both schools, but compounded at the control school.

About the control school closing, one teacher remarked: "This has been a wonderful place for kids. Kids who need more deserve more." (II99ct) Yet another

commented that even after all her 20-plus years of teaching: “I’m still not sure that I’m doing the right things in my classroom. These kids have so many problems.” (II98Tfn)

Therefore, it is evident that challenges related to lower socio-economic status and concomitant academic and affective needs of students received leadership directed toward these ends. More analysis on how these needs were met culminates within the fourth category: Commonalities of experiences.

Commonalities of Experiences

Within this fourth category analysis revealed that more similarities than differences existed in instructional program and practices for upper elementary students. Within one school there was greater cultural diversity; the other higher transiency rate, but between the two, socio-economic status was comparable. Family structures were alike with both schools with a large number of single-parent families, excepting within the Hmong and Russian ethnic groups which were typically two-parent families. Within the approximate same number of years each principal had to work with virtually a non-changing staff, and each with school populations with special needs. The similarities of direction to meet student needs, each special and unique, were striking. Each school focused on early intervention in literacy in the primary grades.

At the experimental school, their schoolwide profile indicated a need for time-on-task and early intervention. Their needs were to be addressed by:

focusing on reading instruction in the primary grades, reducing teacher-student rations, and creating uninterrupted blocks of time....Title 1 staff collaborated with primary teachers to facilitate literacy instruction for small groups of primary students in an uninterrupted ninety-minute language arts block (LAB) each morning...specific criteria determined student placement in various groups. (1199p)

All first and second graders were assessed with running records. All scores were placed on a continuum from emergent to fully fluent. Resource students with IEP for reading were pulled out. According to the principal, flexible groups formed, according to student skill levels as needed: “Students moved in and out and back according to the teachers’ assessments...a lot of strategies we are using are commiserate with...Reading Recovery early literacy pieces....” (1IP5)

Specific strategies for upper elementary students. When asked what the strategies are with upper elementary students, the principal of the experimental school responded:

What we do the rest of the day with our Title people then is we kind of divvied them up in the schedule so that if you have a class of intermediate kids you have a Title person every day for either forty-five minutes, in some cases it’s an hour, but the idea is to be doing inclass teaming or breaking out for skill instruction...so kids are really getting some lower student-teacher ratio with some intensive help at that time. (1IP5)

The schoolwide profile added that to address the language arts instruction and needs of intermediate students a math lab for tutorial/remedial time was available. (SD196p)

When asked about specific upper elementary practices in relation to the goal of literacy, she replied:

[We] don’t do anything specific in multiage. Growth has taken place in the primary. Strategies that I do aren’t any different than any other district. The reasons: no more training, same place for training, same reading workshops, writing workshops. [What we do] is based on contacts outside these are vastly different...learning styles, multiple intelligence...exposure to that in multiage you see more quickly than a homogeneous grouping. Some [teachers] have started learning style workshops and we talk about this in staff meetings and discuss. I started this year with teachers presenting - to fit in with schoolwide goals. I began with teacher I thought would feel comfortable...there is a tendency at 3/4/5 to do more grade specific things, but directed at ability specific. (1IP7)

Later when asked what the upper elementary teachers are doing that may be different from other organizational structures she commented:

You know, I don't think necessarily different, but I think maybe because of some of the things that we've built into our schedule that they have more opportunity to do things...like the teaming. They have every day with their team a half hour that they could and some do [emphasized] use to meet together within the confines of their 8-4 school day...can be personal prep, or we could choose for the three of us to get together. (2IP5)

The only other specific feature directed toward specific strategies used by the upper elementary teachers that was mentioned by the experimental principal was:

You know I really think that everybody ought to be exposed to and become at least well enough versed in it and apply it if they feel comfortable with it, some of the literacy learning. I think that's a very great approach and it certainly applies to intermediate and to primary...and I think by the end of this particular summer session that there may be only one, or two people on the staff that haven't participated at least once. (2IP5)

That the same strategies and practices were used in both schools was connected in the conversation of the control principal. However, her perspective included the statement:

I can't separate lower and upper elementary. It's just one whole continuum to me. We used running records approach which is part of Reading Recovery and most of the teachers were trained, including the upper elementary...I wanted the third, fourth and fifth grade teachers to really tie into early intervention...we did provide service of Title I daily all the way through third through fifth...we had a flexible inclusion mode where again we pooled our human resources. Special Ed and Title and all classroom teachers would meet weekly and do planning, and be in the classroom on a daily basis. (3IP6)

The pullout for skills at the upper elementary level was also practiced:

As we got to grades three, four and five, it was obvious to us there would be times when students would need more of a pull-out model. They were missing specific skills, splinter skills if you will, so then one of the teachers would take a group out. They might do it for a week and half and then be back in. They might do it for a longer period of time. We did what the kids needed...they were flexible groups. (1IP7)

The only substantially different practice revealed in interviews and from documents seemed to be the instructional practice of writing across the curriculum and schoolwide writing assessment that was a schoolwide focus at the control school.

According to the control principal when asked about instructional practices:

By far the strongest instructional practice that we implemented was writing across the curriculum. By far the strongest measurement tool K-5 was the Holistic Development Writing Scale which meshes very nicely with six-trait writing instruction which is what the district uses. So we did not use it in exclusion to six-trait writing. It actually is very complementary to it. The difference is that there is a developmental component. (3Icp)

In stating that she believes “writing leads,” this principal asserted that: “the writing was the piece that I think helped our multicultural population the most, helped our at-risk population the most. We saw the greatest growth I think with those students.” (3Icp7)

The schoolwide writing assessment entailed all K-5 staff to collaborate in reading and assessing student writing according to the developmental scale. Under the principal’s direction, assessment teams comprised of the: “entire staff, speech and language clinician, librarian, Title, special ed, music, PE...everyone helped with assessing that writing so they all had an inservice...that drew the staff together in lots of ways.” (3Icp8)

And this brought to the staff, according to the principal:

a flex of emotions in a positive way so that people were respectful with each other and they could have a better understanding of what the fourth and fifth grade teachers go through if you will and vice versa...and that focused the entire school on writing. (3IP8)

When asked about specific approaches or programs she remarked:

I wanted the third, fourth and fifth grade teachers to tie into early intervention, and it’s not an easy thing for them to do...their look at it is, if you’re putting the money at the front, what’s left for us...we did provide services Title I daily all the way through third through fifth, with a flexible inclusion model. (I36cp)

Asked again about specific instructional practices for upper elementary students, she replied:

What worked particularly well at the fifth grade level...is that two teachers used a short diagnostic measure that they put together themselves...to determine the level the student was at and for reading time actually created two groups between two classrooms. Now I know that's an old, old kind of idea. (I312cp)

The grouping by ability through pull-out that was part of the experimental school's effort to teach skills, as well as the control school as was mentioned earlier.

The control school principal continued:

The special ed teacher stayed with the teacher who took basically a group that had lower students, but they weren't all low students. We made homogeneous groups, not homogeneous, heterogeneous groups. Then the other teacher who was probably a bit more interested in doing gifted education kinds of things took more of the upper elementary, sorry the upper ability students, but we did not ability group. I talked really long and hard to the staff about ability grouping. It's something that I feel is very damaging to students, but we saw wonderful gains. The students that were with the special ed teacher, most of the special ed kids were in there and they could do flexible groupings back and forth. They could also group between the other classroom. It was all very flexible...the Title I teacher was in the other fifth grade classroom so there were two teachers in each classroom for an hour. You know, as a fifth grade teacher when students come in to you and they're so low, they're reading at the second grade level perhaps, and most of those students were transient...That worked really well. (I312cp)

When asked about specific curriculum practices she responded:

When we tried to make some decision in that area, we went to best practice. We did a lot of reading together as a staff and discussing ...It also came from watching other teachers in the district, from going to other districts..curriculum's function is not how to do it...there are many ways of getting...we had one way that worked for us. (3IP9)

Schoolwide intervention. In the both schools, the schoolwide Title I program began in 1995-96. The number of years each of the students in this research may have been exposed to each organizational structure and intensive intervention strategies and

practices was a variable. Within the experimental school, the fifth grade students are the body of students that has had the longest educational exposure to multiage grouping, but the least years with interventions implemented by the schoolwide plans. It is essential to note that it is not known which individual students were enrolled continuously, and then would have had both exposures. Due to transient rates at each school, this is difficult to measure. Nevertheless it is a fact that the fifth graders, or 10- and 11-year olds, are the cohorts that did not have the intensive interventions during their primary years. Only the fourth and third graders had this additional educational experience for four years. It is at this level that 4 out of the 5 statistically significant differences occurred in the 28 total measures.

Only the third graders had their entire school career covered by the intensive intervention; the fourth graders had four years since first grade, and the fifth graders four years since second grade. Again, the numbers of individual students that were enrolled continuously is not part of this research data; the tracking of individual students through their school years is recommended in Chapter 5. However, both principals noted students return. The control principal stated that “because we were in a low rent part of town, we got a lot of the same students back...They would move away for a year or so, and then they would come back.” (3IP16) This seemed to be a pattern in schools in both lower socioeconomic neighborhoods. Table 18 summarizes intervention time per cohort.

Table 18

Summary of Length of Time of First Multiage Grouping Combined with Schoolwide Title I Interventions Up to Date of Year of Research

School Year				Experimental	Control
1990-91				Multiage (1 / 2 only) begins (excluding kindergarten)	
1991-92					
1992-93					
1993-94	K				
1994-95	1	K			
1995-96	2	1	K	Schoolwide intervention begins	Schoolwide intervention begins
1996-97	3	2	1	Interventions continue	Interventions continue
1997-98	4	3	2	Interventions continue	Interventions continue
				First year all classes multiage, with kindergarten and one Grade 5 exception	
<u>1998-1999</u>	5	4	3	Interventions continue	Interventions continue

Note. The school year for this research data collection was 1998-1999.

Another difference between the two schools relating to time of exposure to interventions and instructional programs is the fact that the experimental school may have used different communication arts materials up until the last two school years. Difference in materials would be considered an extraneous variable. The principal states:

The district doesn't prescribe anything but six-trait, and with Scholastic we are being bound to this reading curriculum. Initially [beginning multiage] we were not bound to adoption. Curriculum and adoption are different. With the new adoption [1997] we are bound to it and the curriculum. If the teachers are well organized and managed this will work. (I18ep)

From results of the Comprehensive Needs Assessment, the stated implementation plans for the school year for the experimental school was the improvement of reading/literacy. The experimental school decided to focus on reading instruction in the primary grades, uninterrupted blocks of time, and reduced teacher-student ratios. Each school received federal funds through Goals 2000, Eisenhower, Title VI, and Drug Free Schools' grants. A summary of the many variables affecting the literacy development of the students at each school documents the similar emphasis in over 50 programs and training within each of the two schools (see Appendix U). From this comprehensive summary, most instructional programs and practices were similar within the single grade classrooms and the multiage classrooms at the upper elementary levels. A primary exception was the time allowed to plan as a team. Provided for the experimental school, as the principal stated "they could and some do use to meet together [or it] can be personal prep." (21p5)

Assessment-driven instruction. To continue the analysis exploring whether literacy growth differs, the testing question emerged again. At both schools, assessment of primary students was made with running records with assessment for third grade and above with the MALT fall and spring scores, and standardized testing with CTBS and then TerraNova. The fact that test scores were part of both schools' instructional decisionmaking process was integral to the analysis.

The control principal stated that the standardized achievement tests:

"don't really reflect what it is we teach in the classroom day to day. They are a measurement tool that we need to be aware of and to use...we felt running records were...and MALT scores in the third, fourth and fifth grades to inform our

instruction, but we did a lot of sharing with the lower elementary. (I1p18)

Standardized test results were a component of district reporting, with test results periodically published in the local newspaper. These longitudinal scores correlate with the data from this study in its look at fourth grade students, who are a mix of 9- and 10-year olds. From a district document (1998), achievement test scores for a full battery given fourth graders were analyzed over an 8-year period. The experimental school test scores were summarized as follows:

Test results for [the school] have remained very consistent since 1992. However, 1991 was a year when the school's test profile was very much in line with the rest of the elementary schools in the district...The percentage of those scoring in the bottom quartile jumped dramatically in 1992. Many of the communication and math sub-test scores improved from 1992 to the present. Only math computation and language mechanics fell during that period (MCPS, 1998a, p. 18)

The district analysis document (1998) stated that the control school:

has done extremely well on the CTBS when the results are tracked from 1991 to the present. Overall Battery total improved five percentile points during that period. This is despite the increase in the percentage of students on Free and Reduced lunch. That figure rose from 54% in 1991 up to 66% in 1998 [note this summary does not match the table from the same page of the document]. Also, during that same time period, the percentage of students falling within the bottom quartile remained relatively low with the exception of 1996. [This school] is one of the elementary schools where the vocabulary drop has been minimal over time. Language Mechanics, Reading Comprehension, Math Concepts and Applications, and Math Computation all improved during the past eight years....(MCPS, 1998a, p.15)

Data reported by the district indicates that fourth grade mean scores for the complete battery at the experimental school were below the control school for a period of eight years with the exception of 1991 and 1996 (see Table 19).

Table 19

CTBS Fourth Grade Summary of Building Data from 1991-1998

	1991	1992	1993	1994	1995	1996	1997	1998
Battery Total								
Experimental	62	40	36	43	41	47	42	37
Control	51	69	75	69	58	38	61	56

Note. Scores reported in percentiles (MCPS, 1998a)

Note that the greatest mean score for the experimental school was the year that the multiage program began with the primary students, 1990-91. The 1998 fourth grade students enrolled continuously would have had four years of multiage, grades 1-4. Yet these scores must be analyzed in conjunction with data reported that indicated an increasing number of students from low income households within each school, with the larger percentage in the experimental school (see Table 20).

Table 20

Comparison of Percentage of Free and Reduced Lunch from 1992-1998

	1992	1993	1994	1995	1996	1997	1998
Experimental	61	62	64	74	73	76	73
Control	49	51	54	60	60	66	62

Note. Data from MCPS, 1998a.

That assessment in multiple forms has been a part of each school's focus in instruction is evident from previous comments, as well as this statement from the principal of the experimental school: "I get a rush when I look at data and disaggregation" (1IP6). She has data from the past six years, tracking student progress through running records, CTBS, and the district MALT scores, clearly demonstrating

literacy growth. She continued with:

So I took their running records when they entered first grade. 94% were emergent, 2 % early, 4% fluent. After two years and these are all the same kids, we moved to 0 percent emergent, 6% early and 27% fluent, and 67% had passed fluent. So at the end of second grade 6% were still not at grade level. (I215ep)

When asked again if these were all the same students she replied: “Yes, so you know you can see that we have kids coming and going. And so they are the same kids. They come back.” (I216ep)

Her assessment-driven instruction for the upper elementary teachers’ instructional practices has the benefit of instructional planning time, and according to this principal, seems to have future goals to meet:

[Upper elementary teachers] have a half hour built into their schedule where they can talk...if we in our MALT scores have seen that literal comprehension is a weakness for our students as a whole, then that’s what we can talk about...let’s put together some strategies and go out and do them for month and come back and talk about how they’re working....a recommendation would be that teachers have more skills in reading assessment, in assessing whether or not the child is not reading the material because they can’t call the words, they can’t decode, or they do not understand...as teachers we haven’t arrived yet where we can sit down and diagnose specifically...in order for intermediate teachers to be more successful they need a broader understanding of the reading process and how to identify where a child is...I recommend lots of families to check out Sylvan...they can diagnose specifically where the child is...we’re not skilled at that yet....Each year we’ve made new changes with the intermediate definitely based on what the assessment data shows us and how we need to apply that information to change and modify instruction. (I25ep)

That both schools achieved literacy gain for their students is evident in the data each collected and in the results of this study. That the staff at both schools felt their school was a good place for kids is also evident.

Summary

Both schools were responsible to meet the vision of the district to “provide a broad, effective education for each student in a safe, stimulating supporting learning environment” and the district mission “to provide a foundation for each student to become a lifelong learner, to promote development of the whole individual and to prepare each student to become a responsible, productive citizen of our community, state, nation and world” (MCPS, 1998-99, p.3). Both schools set major schoolwide goals in reading/literacy improvement to achieve this. All of the educational challenges were met by similar policies, programs, and practices. Academic achievement was comparable between the two groups and in most measures, well within expected norms.

Yet writing was the area in which two out of the four cohorts at the control school indicated a statistically significant positive difference. The professional development for instruction to students for writing through the six-trait writing was the same for all district teachers. Both schools were engaged through the same goals, objectives, and training to implement writing through the process paradigm. Improvement in writing, according to interviews, was one of the control school’s primary instructional goals. To achieve this goal cross-grade level collaborative training in assessment of writing occurred with the entire staff. This was one practice different from the experimental school.

The areas in which two cohorts at the experimental school showed a statistically significant positive difference were on the standardized achievement test, 8-year olds in reading; and on the criterion test, 9-year olds in reading. Improvement in reading,

according to interviews and documents, was this school's primary schoolwide goal. To achieve this intensive intervention in the primary years occurred.

Another salient point is that in the intermediate grades, the 8- and 9-year olds, not the 10- and 11-year olds which had been in the multiage structure the longest length of time, are the multiage cohorts that demonstrated two out of five statistically significant differences. They are the students that received the longest length of time with intensive interventions. It is the 9-year old cohort in the control group that demonstrated two out of five statistically significant differences in writing. It is the 9-year old cohort in the fourth grade and 10-year old cohort in the fourth grade that demonstrated practical significance in favor of the control group in a consistent pattern.

However both schools made literacy growth of a comparable gain, well within or above expected levels, and most importantly, within their particular focus of goals and objectives. This would suggest that academic achievement is possible when specific goals and objectives, guided by assessment-driven instruction, are implemented schoolwide for a concerted period of time.

From interviews with the principals, surveys from the teachers, and formal and informal documents, this picture emerged: Developmentally appropriate practices and assessment-driven instruction, with the principals' observations and leadership, were predominant in both settings. Instructional programs and practices were similar. Many were derivations of the same professional training choices, as well as programs implemented within the school setting. Differentiated training of teachers to supplement either organizational structure was not apparent. As for the implementation of the

training. this researcher will not make any assumptions except for the previously stated assumption that programs and policies for which teachers receive training would most likely be part of their regular classroom practice. One must generalize from one's own experiences in the classroom, an isolated experience particular to the individual.

Because of impending district school closures, each principal by April 1999 received new assignments for the next school year, 1999-2000. By May, teachers within the closed school were reassigned by seniority to available positions afforded by the school closures, which included two upper elementary multiage classrooms added to a third school. None of the control teachers involved in the research chose to be assigned to the multiage classrooms. Two teachers who had been at primary level made this choice.

CHAPTER 5

DISCUSSION

The purpose of this combined design study was to compare and explore the impact of organizational structure upon literacy development of upper elementary students within two organizational structures: multiage and single grade. Students were disaggregated by age as an integral focus and to provide equity within the statistical analyses. Three separate standardized test measures were used. Statistical tests of difference were conducted appropriate to each measure with an alpha level of .05 set a priori.

Quantitative questions were:

1. To what degree does organizational structure impact student academic achievement on a standardized, norm-referenced general achievement quantitative measure?
2. To what degree does organizational structure impact student academic achievement on a standardized, criterion-referenced district quantitative measure?
3. To what degree does organizational structure impact student writing development as demonstrated by a performance assessment of pre-and post writing?

Triangulation of data through combined methods included interviews and document analysis directed toward two qualitative questions:

1. What are the instructional policies, programs, and practices within the single grade and multiage organizational structures?
2. Does literacy growth differ within the age configurations of the two types of organizational structure?

Participants were upper elementary students in grades 3 through 5 in two public K-5 schools of similar demographic composition in an urban setting of 87,000 in the Northwest region of the United States. The student population was considered homogeneous in terms of ethnic diversity; yet the number of minority students within each school was well above other area schools. Other factors contributed to the similar needs of students at each school: lower socioeconomic status, transiency, and nontraditional families.

The literature review suggested that previous research was equivocal and dated in relation to needs of today's students, parents, teachers, administrators, school board members, and community. In addition, the many terms associated with organizational structure contributed to ambiguity that precluded a clear definition of what to expect within classrooms. A clear definition rather than prevalent assumptions was necessary to determine if there were actual differences. While affective benefits of multiage grouping of primary students have been reported, evidence of academic achievement has been equivocal. Seeking evidence of a component-building nature was a purpose of this study. Overarching questions in this combined design were (a) whether there are differences in academic achievement of students and (b) whether there are differences in the instructional policies, programs, and practices between structures, and if so, what they are.

Assumptions within the paradigms of this combined design research that were stated earlier are reemphasized by the following additional opinions:

Literacy is the degree to which someone is able to merge all the language systems - reading, writing, listening, speaking and even music, art and drama...[some tests] come closest, but they are not a true test of literacy. There is no such thing and probably never will be. All our tests are nothing but faint imitations of real literacy. (Farr, 1992, p. 27)

Even so, Farr constructs standardized assessment measures. He may agree tacitly with Mohr (1990) that:

Social science researchers...search for the sort of factors that make a difference in people's lives, with different aspects of life being salient...factors that seem to be important in one population or at one time have an annoying way of appearing inconsequential later on. But the identification of such factors at work in at least one setting is a strong beginning for much thought and research that must then go on at a deeper level (p.27)...one cannot be a slave to significance tests. But as a first approximation to what is going on in a mass of data, it is difficult to beat this particular metric for communication and versatility. (p. 74)

Findings

To come to a decision about growth in literacy development, all parts must be considered separately, and then as a whole.

1. With regard to the three major quantitative questions, 28 separate statistical comparisons within four age cohorts were conducted. Twenty-three indicated no statistically significant difference. Of the five cohorts in which a statistically significant difference was found, three favored the single-grade cohorts and two favored the multiage cohorts. Three out of the five cohorts were the 9-year old students. Specifically:

a. On the standardized, norm-referenced general achievement measures in Reading and Language, only one of the 12 analyses indicated a statistically significant difference: Cohort 8 on Level 13 of Reading in the direction of the experimental group.

b. On the standardized, criterion-referenced district measure in Reading and Language, only one of the eight analyses indicated a statistically significant difference: Cohort 9 in Reading in the direction of the experimental group.

c. On the writing assessment, three of the eight separate analyses indicated a statistically significant difference. Cohort 9 in both fluency and conventions indicated a direction in favor of the control group. Cohort 10 in fluency indicated a direction in favor of the control group.

d. Practical significance indicated 12 out of 28 effect sizes $> .33$. The determination of this significance is a subjective decision for the reader, but with 42% of the separate analyses $> .33$, further study would be warranted. A consistent pattern of greater mean scores was indicated by the control age cohorts 9 and 10 within the fourth grade. However, no experimental differences over 4% were reported within any age cohort on either indirect measure. Five out of 8 on the direct measures were 5% or over. However, all mean scores in writing were within one point of each other and within the same level of writing development.

2. Overall, a definitive pattern of differences in literacy development was not indicated for any cohort or within any of the three literacy areas. This finding reflects previous literature (Brown & Martin, 1987; "Committee reports," 1997; Daily Report, 1995; Gutiérrez & Slavin, 1992; Katz, 1992; Miller, 1990; Nye, 1995; Pratt, 1986; Shepherd & Ragan, 1982; Veenman, 1995).

3. Each statistically significant difference corresponded to a major focus of assessment-driven instruction within the school: (a) for the experimental group, the 8- and

9-year olds that had intensive intervention in reading at the primary level and (b) for the control group, assessment training for writing had been a schoolwide cross-grade level collaboration among staff.

4. From interview and document analysis, findings indicated that the instructional programs and practices within each school were more similar than different. Both schools engaged in schoolwide goals implemented through assessment-driven instruction. Collegial leadership to meet the needs of students was integral to the choices made within the historical origin of the choice of organizational structure of one school, and for instructional practice within both. The commonality of choices and experiences in instructional programs and practices makes it difficult to single out any pattern of difference. In addition, no differentiation in teacher training was evident.

5. Both schools demonstrated comparable gains in achievement. Findings indicated that homogeneous grouping of students by ability for skill instruction was a primary mode of instruction for both schools, and constituted a considerable portion of literacy instruction and planning time. Even though groups were flexible, this type of grouping contradicts one prevalent assumption of a “pure” multiage concept of heterogeneous grouping of students by ability, as well as age. So the question still remains: when there is a difference, what is truly making the difference.

Conclusions

The findings of the combined design study suggest the following conclusions:

1. When instructional programs and practices that are within the definition of best practice are implemented on a schoolwide basis, the effect of organizational structure may

be inconsequential. In other words, between these two schools, the variables associated with instruction and training were more similar than different. Also, a wide variety of programs were implemented to improve student achievement. That being the case, and with only five out of 28 statistically significant differences, it would be unwarranted to suggest an effect upon literacy development based on organizational structure alone.

2. The most compelling finding of this study was that differences were demonstrated within the literacy area that each school chose as a particular focus for instruction, and both used assessment to guide instruction. A reasonable conclusion is that this type of concerted effort brings about student success that is not only equitable success, but success that is achievable even within limited budgets. The nature of the leadership at both schools produced a collaborative focus. This practice holds promise for schools to be more effective.

3. From the available evidence, it is difficult to conclude the degree to which a “pure” multiage environment of heterogeneous grouping by age and ability existed in this study. Therefore no conclusion can be drawn as to the degree that this experimental sample represents the multiage concept as defined by the research literature. Prevalent assumptions about multiage beliefs guiding practices different from other structures are not evidenced by this study. This is not a simple dichotomy. Beliefs may or may not translate into practice. This study suggests that practice may not be particular to structure, or assumed unique to one structure.

4. This study indicates that practices are predicated by need. Each structure seemed to work for the similar special needs of the students. In order to reach any

conclusion about alternative structures, changes that were initiated, developed, and continued must be considered. Discipline problems, high transient rates, and low test scores were precipitating factors for one staff. The issue of choice for teachers, both collectively and individually, was found to be integral to the continuance of the program, and ultimately, the effectiveness of the school.

5. In the collection and analysis of this data, its utilization is limited in its generalizability. The conclusions from these findings are understood to be generated by one nonrandom study within one school population. Nevertheless, its components suggest implications for further research.

Implications

1. The implications of the findings and conclusions of this study indicate that decisions made regarding one organizational structure over the other require more complete descriptions before future implementation. The term single grade did not preclude use of instructional programs, practices, or strategies similar to those within the situated nature of the multiage classroom. If best practice as defined in the literature takes place within the classroom, the organizational structure may be a secondary factor for consideration.

2. The literature review suggested that additional training and other considerations need to be part of a multiage implementation. Yet, within this study, no additional monies were spent to prepare the multiage teachers for their structure. If further training were part of the program, would there be a difference in academic achievement?

3. Moreover, when training teachers to meet the needs of students, the increase in the spread of differences in ability needs to be addressed more specifically. The literature suggested that the increase in ability differences increases with age, and therefore is greater in the upper elementary classes. If this is true, how will teachers best meet the needs of all students? Homogeneous grouping by ability is counter to Vygotsky's "more capable peer," or Gardner's mentor, both of whom facilitate the learning. Whether or not age configuration makes a difference could not be determined when grouping by ability.

From this data it appears that multiage students were grouped within two of Glickman's (1998) quadrants for different instruction, as were also the single grade students. Therefore, are the community of learners that the late Ann Brown researched within the zone of proximal development just as effective within either structure? The descriptive data of this study would suggest this, but the implications are that we still do not know what is truly making a difference. Would that not be the heart of the matter when exploring benefits of one structure over another based upon different age groups?

4. Because this issue has generated controversy, the implications for open discussion are critical. We must extend generative understanding and dialogue. To understand that the similarities may be more important than assumed differences is an implication of this research.

Recommendations for Practice

1. Effective schools require a commonality of goals and objectives across staff levels. Administrators need to know their staff well before attempting to implement change. Similarly, teachers need to know their entire school population and its needs

thoroughly before considering change that affects the whole school. Dewey's "habits of mind which secure...change without disorder" (1916, p. 115) must be a guiding principle.

2. Rather than emphasizing differences between structures, emphasis upon what is best in policy, program, and practice within each needs to be an ongoing dialogue, with accompanying educational accountability of past and present performance. When change is proposed, reasons need to be available to all stakeholders with sufficient data, especially when faced with substantive resource reallocation.

3. Choice for both parents and teachers is a vital element to consider in public education. However, equity both for students and teachers is essential to implementation, success, and continuation of an effective core academic program.

4. Observation of discussions within the groups would illuminate if older students facilitate ZPD. Attention to sound evidence regarding earlier puberty, changing learning styles, and other physical, social, and emotional conditions of today's children must be part of any decisionmaking.

5. Districts adopt and report to the public their curriculum with accompanying goals, standards, and benchmarks, curricular issues, and assessment specific to each structure. If, for example, there are curricular differences in scope and sequence or materials, or types of grade expectations and reporting between structures, these differences need to be delineated specifically for parents. Mandates are understood, but if special dispensations exist, stakeholders should have this information.

6. Equity in achievement reporting needs to be realized. Academic achievement can be reported in standardized and nonstandardized ways with differences in population

by socioeconomic status, ethnicity, and other variables stated. In this way, true gains made by distinct populations can be recognized. The inequitable but continuing comparison of schools dissimilar in demographic factors needs to be rendered obsolete.

7. Decisionmakers need to implement programs and practices predicated on replicated studies, preferably those with random assignment and selection. Full quantitative and qualitative data as recommended must be provided in order to make decisions particular to the needs of individual students and groups of students.

Recommendations for Future Research

The data in this study did not answer all of the research questions. Therefore, recommendations for further research are that:

1. An in-depth observation and documentation of curricula and teaching strategies within both types of structure be conducted by an independent researcher to determine the extent and type of practices implemented within daily and ongoing instruction.

2. A formal exploration of the dialogue and interaction among children within multiage and single grade classrooms be conducted to explore the relationship of the more capable peer within the zone of proximal development. Multiple zones of proximal development within a community of learners should be explored within both organizational structures. That is to say, what types of learning transpire due to the interaction of the students with regard not only to ability as Vygotsky suggests, but also to age. Because age is the defining factor of multiage structure, the degree of achievement afforded by this one variable when not confounded by extraneous variables needs to be explored.

3. A longitudinal study be conducted to address the trends within quantitative measures of academic achievement of individual students within each organizational structure. The criterion-referenced test using a RIT scale provides this type of developmental approach independent of age or grade level. However, initial comparability of students established by a test of cognitive abilities as well as achievement would be an essential requirement for test analysis.

4. A longitudinal study be conducted to address patterns of performance of cohorts of students within each organizational structure. The same representative cohort of students in school at ages 9, 13, and 17, as an example, could be tracked at elementary, middle school, and high school levels. Group trends from this type of cross-sectional design would provide data for informed program implementation and evaluation across developmental levels.

Although compelling evidence for one organizational structure over another has not resulted from this study, it seems clear from the research that the goals each school set had an impact on its students' academic achievement in literacy. In addition, the professionalism of teachers willing to engage in collaborative planning and assessment-driven instruction made a difference. Principals that realized a school's effectiveness was at stake provided support and direction through such collaborative leadership.

Yet in the long run, perhaps the words of Howard Gardner (1999) are cogent:

The point is that there is no direct tie between a scientific theory and a set of educational moves. Whether one believes in one intelligence or twenty, and whether one thinks early experiences are more important than later ones, or the reverse, one is still free to implement any number of educational approaches. Indeed, in an art like teaching, the proof comes down to whether an approach

works; it matters little whether the theory was correct. And, conversely, even if the theory is both correct and elegant, if it cannot be mobilized for concrete educational consequences, the theory matters not a whit to the educators. (p. 144)

Overall, one of the initial premises of this research was to provide data of a component-building nature to facilitate the necessary collaboration within the profession and between the school and the community for school improvement (Goodlad, 1979).

This study continued the research necessary to meet the fundamental assumption as stated by Lipsitz et. al (1997) in that

collection, analysis, and utilization of data...[is] the heart of professionalism. When schools embrace data-based decisionmaking as a school-improvement tool, they make measurable progress in attaining their objectives. They are able to plan next steps in such critical areas as creating small communities for learning, strengthening the core academic program, and reconnecting schools and communities based upon verified performance. (p. 536)

Once again, and finally, free and equal opportunity of education is fundamental to our democratic society. All children should have the benefit of what is best for learning. While more choices are available in public schools today than ever before, private and homeschools are increasing. Public schools must meet the demands of a diverse community through both fiscal and educational accountability. This study emphasizes the need for this accountability.

References

- Administrative Rules of Montana, Title 10 Education, (Compiled by Mike Cooney, Secretary of State) Vol. 4, Part I (1997). Helena, MT.
- Allington, R.L., & Cunningham, P.M. (1996). Schools that work: Where all children read and write. New York: HarperCollins.
- Anderson, B.L. (1980, July). Writing Assessment Alternative for the '80s. Paper presented at the annual meeting of the American Association of School Administrators, Chicago. (ERIC Document Reproduction Service No. ED 192 359)
- Anderson, R.C. (1992, April). The nongraded elementary school: Lessons from history. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (Eric Document Reproduction Service No. ED 348 161)
- Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, I.A. (1985). Becoming a nation of readers: The report on the commission on reading. Washington, D.C.: National Institute of Education.
- Archer, J. (1999, December 8). Unexplored territory. Education Week, 19, 22-25.
- Arter, J. (1993). Managing your assessment with confidence & style or how to conduct a large-scale writing assessment and support instruction too. Portland, OR: Northwest Regional Educational Laboratory.
- Arter, J., Culham, R., Pollard, J., & Spandel, V. (1994). The impact of training students to be self-assessors of writing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. 370 975)
- Babbie, E. (1998). The practice of social research (8th ed.) Belmont, CA: Wadsworth.
- Barry, B.A. (1997, March). Motivating students to write: Implementing creative theory to overcome the habitual and encourage autotelic flow. Paper presented at the annual meeting of the Conference on College Composition and Communication, Phoenix, AZ.
- Baumann, J., Jones, L., & Seifert-Kessell, N. (1993). Using think alouds to enhance children's comprehension monitoring abilities. The Reading Teacher, 47, 184-195.

- Beggs, D., & Buffie, E. (Eds.). (1967). Nongraded schools in action: Bold new venture. Bloomington: Indiana University Press.
- Begley, S. (1996, February 19). Your child's brain. Newsweek, 55-62.
- Bergen, D. (1995). Parent and student views of multiage classrooms. Childhood Education, 71, 192.
- Berliner, D., & Biddle, B. (1995). The manufactured crisis: Myths, fraud, and the attack on America's public schools. Reading, MA: Addison-Wesley.
- Bickman, L., & Rog, D.J. (1998). Handbook of applied social research methods. Thousand Oaks, CA: Sage.
- Bogdan, R. & Biklen, S. (1992). Qualitative research for education. Boston: Allyn & Bacon.
- Borg, W., Gall, J., & Gall, M. (1993). Applying educational research: A practical guide (3rd ed.) New York: Longman.
- Bracey, G. (1996). The sixth Bracey report on the condition of public education. Phi Delta Kappan, October, 127-137.
- Bracey, G. (1999). Going loopy for looping. Phi Delta Kappan, October, 169.
- Brand, A.G. (1991). Constructing tasks for direct writing assessment: A frontier revisited. (Report No. CS213116). SUNY Brockport. (ERIC Document Reproduction No. ED 340 037)
- Bredenkamp, S. (1997). Developmentally appropriate practice in early childhood programs. (NAEYC Publication No. 234). Washington, DC: National Association for the Education of Young Children. (ERIC Document Reproduction Service No. ED 403 023)
- Brossell, G. (1986, April). Essay test topic development. Paper presented at the Annual Conference on Writing Assessment, Cleveland, OH. (ERIC Document Service No. ED 270 002)
- Brown, B.F. (1970). Position statement of chairman. In Report of a National Seminar: Models of Nongrading Schools. Dayton, OH: The Institute for Development of Educational Activities, an affiliate of the Charles F. Kettering Foundation.

- Brown, K.G., & Martin, A.B. (1987). Student achievement in multigrade and single grade classes. In R. Fogarty (Ed.), The multiage classroom: A collection. (1993, pp. 159-165). Palatine, IL: IRI/Skylight. (ERIC Document Reproduction Service No. ED 369 574)
- Calkins, L. (1986). The art of teaching writing. Portsmouth, NH: Heinemann.
- Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.
- Carbone, R.C. (1961). A comparison of graded and non-graded elementary schools. Elementary School Journal, 62 (2), 82-88.
- Case, R. (1931). The platoon school in America. Stanford, CA: Stanford University Press.
- Chaney, R. (2000, January 9). Here comes the neighborhood. The Missoulian, pp. A1, A5.
- Christensen, L., & Stoup, C. (1991) Introduction to statistics for the social & behavioral sciences (2nd ed.). Pacific Grove, CA: Brooks/Cole.
- Comer, J. P. (1997). Waiting for a miracle: Why schools can't solve our problems-and how we can. New York: Dutton.
- Committee reports on multiage education. (1997, Spring). What's happening in Your Public Schools: Missoula County Public Schools Shareholders' Report, 2. Missoula County Public Schools Shareholders' Report.
- Cooper, C.R., & Odell, L. (1977). Evaluating writing: Describing, measuring, judging. State University of New York at Buffalo: National Council of Teachers of English.
- Creswell, J. (1994). Research design: Qualitative & quantitative approaches. Thousand Oaks, CA: Sage.
- CTB/McGraw-Hill. (1996, July). TerraNova Prepublication Technical Bulletin. (Issue No. 53689). Monterey, CA: Author.
- CTB/McGraw-Hill. (1997a). TerraNova Content Objectives. Monterey, CA: Author.
- CTB/McGraw-Hill. (1997b). TerraNova Spring Norms Book. (Issue No. 53688-P). Monterey, CA: Author.
- CTB/McGraw-Hill. (1997c). TerraNova Technical Bulletin I. Monterey, CA: Author.

- Cuban, L. (1984). How teachers taught: Constancy and change in American classrooms, 1890-1980. Research on Teaching monograph series. (Eric Document Reproduction Service No. ED 383 498)
- Daily Report Card (1995, November 15). From Coleman, Atlanta Journal/Constitution, 11/3. Education Commission of the States and the National Education Goals Panel [On-line]. Available Rptcrd@gwuvm.gwu.edu and ERIC Document Reproduction Service No. 015 024.
- Davis, R. (1992). The nongraded primary: Making schools fit children. (Stock No. 21-00192). Arlington, VA: American Association of School Administrators.
- Dean, J.P., Eichhorn, R.L., & Dean, L.R. (1969). Establishing field relations. In G.J. McCall & J.L. Simmons (Eds.), Issues in participant observation: A text and reader (pp. 68-70). Reading, MA: Addison-Wesley.
- Diederich, P.B. (1974). Measuring growth in English. Urbana, IL: National Council of Teachers of English monograph. (Eric Document Reproduction Service No. ED 097 702)
- Denzin, N.K. (1970). The research act: A theoretical introduction to sociological methods. Chicago: Aldine.
- de Tocqueville, A. (1945). Democracy in America (Henry Reeve, Trans. edited by P. Bradley). New York: Vintage Books. (Original work published in 1835).
- Dewey, J. (1916). Democracy and education. New York: Macmillan.
- Drucker, P. (1994, November). The age of social transformation. Atlantic Monthly, 53-80.
- Elser, T. (1997). Holistic Developmental Writing Scales. Next Generation Learning Tools. Available through Instructional Media Services, University of Montana.
- Falk-Ross, F. (1997). Developing metacommunicative awareness in children with language difficulties: Challenging the typical pull-out system. Language Arts, 74, 206-216.
- Farr, R. (March, 1992). Assessing achievement. Literacy and Learning (pp. 23-27). Bloomington, IN: Office of Research. (Eric Document Reproduction Service No. ED 343 109)

- Fisher, C. (1997). Multiage classrooms an alternative to gradedness: Review of literature, qualitative study and recommendations for school-site policy. (Doctoral dissertation, University of Delaware, 1997). Dissertation Abstracts International, UMI 9733566.
- Flippo, R.F. (1997). Sensationalism, politics and literacy. Phi Delta Kappan, 79, 301-304.
- Fogarty, R. (Ed.). (1993). The multiage classroom: A collection. Palatine, IL: IRI Skylight. (ERIC Document Reproduction Service No. ED 369 574)
- Forman, E.A., & Cazden, C.B. (1994). Exploring Vygotskian perspectives in education: The cognitive value of peer interaction. In R.B. Ruddell, M.R. Ruddell, & H. Singer (Eds.), Theoretical models and processes of reading (4th ed., pp. 155-178). Newark, DE: International Reading Association.
- Freed, M.N., Hess, R.K., & Ryan, J.M. (1989). The educator's desk reference: A sourcebook of educational information and research. New York: American Council on Education.
- Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. New York: Basic Books.
- Gardner, H. (1991). The unschooled mind: How children think and how schools should teach. New York: Basic Books.
- Gardner, H. (1999). Intelligence reframed. New York: Basic Books.
- Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In The Interpretation of Cultures (pp.193-233). New York: Basic Books.
- Gergen, D. (1996, September 30). A social contract for schools. U.S. News & World Report, 121, 76.
- Gipe, L. (1992). School improvement network directory: 1992 supplement. Portland, OR: Northwest Regional Educational Lab. (ERIC Document Reproduction Service NO. 353 650)
- Glass, G.V., & Stanley, J.C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.
- Glickman, C.D. (1998). Revolutionizing America's schools. San Francisco: Jossey-Bass.

- Goodlad, J.I. (1979). What schools are for. Los Angeles: University of California and Institute for Development of Educational Activities, Inc. Phi Delta Kappa Education Foundation.
- Goodlad, J.I. (1984). A place called school. New York: McGraw-Hill.
- Goodlad, J. I., & Anderson, R. (1987). The nongraded elementary school. (Rev. ed.). New York: Teacher College Press.
- Graves, D. (1983). Writing: Teachers and children at work. Portsmouth, NH: Heinemann.
- Graves, M.F., & Avery, P.G. (1997). Scaffolding students' reading of history. The Social Studies, 88, 134-135.
- Gray, J. (1982). Properties of writing tasks: A study of alternative procedures for holistic writing assessment. Report presented by the Bay Area Writing Project. Berkeley, CA. (ERIC Document Reproduction Service No. ED 230 567)
- Gregory, K. (1991, January). More than a decade's highlight? The holistic scoring consensus and the need for change. Paper presented at the Annual Meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 328 594)
- Guba, E.G., & Lincoln, Y.S. (1981). Effective evaluation. San Francisco: Jossey-Bass.
- Guitérrez, R., & Slavin, R. (1992). Achievement effects of the nongraded elementary school: A retrospective review. Baltimore, MD: Center for Research on Effective Schooling for Disadvantaged Students. (ERIC Document Reproduction Service No. ED 346 960)
- Harste, J.D. (1989). New policy guidelines for reading. Urbana, IL: National Council of Teachers of English.
- Harvard Graduate School of Education. (March 1986). Repeating a grade: Does it help? The Harvard Education Letter 2:2. Boston, MA: Harvard University.
- Hawk, A. W., & Cross, J.L. (1987, April). Scoring writing samples in educational research: Selecting and developing an appropriate procedure for evaluating elementary student writing. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Reproduction Service No. 283 849)

- Heald-Taylor, B. (1996). Three paradigms for literature instruction in grades 3 to 6. The Reading Teacher, 49, 456-466.
- Healy, J.M. (1990). Endangered minds: Why children don't think and what we can do about it. New York: Touchstone.
- Hiebert, E.H. (1994). Becoming literate through authentic tasks: Evidence and adaptations. In R.B. Ruddell, M.R. Ruddell, & H. Singer (Eds.), Theoretical models and processes of reading (4th ed., pp. 391-413). Newark, DE: International Reading Association.
- Holmes, C. T. (1983). The fourth r: Retention. Journal of Research and Development in Education, 17, 1-6.
- Holmes, C. T., & Matthews, K. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. Review of Educational Research, 54(2), 225-236.
- Hopkins, K.D. (1992). CTBS/4 review. In J. Kramer & J. Conoley (Eds.), The Eleventh Mental Measurements Yearbook (pp. 216-218). Lincoln, NE: Buros Institute of Mental Measurement.
- Hopkins, K.D., & Glass, G. V. (1978). Basic statistics for the behavioral sciences. Englewood Cliffs, NJ: Prentice-Hall.
- Howell, D.C. (1997). Statistical methods for psychology (4th ed.) Belmont, CA: Duxbury Press.
- Huitema, B. (1980). The analysis of covariance and alternatives. NY: John Wiley & Sons.
- Indrisano, R., & Chall, J. (1995). Literacy development. Journal of Education, 177, 63-82.
- Jacob, E. (1998). Culture, context, and cognition. In M.D. LeCompte, W.L. Millroy, & J. Preissle (Eds.), The handbook of qualitative research in education (pp. 293-335). San Diego, CA: Academic Press.
- Jahrig, G. (1997, October 6). Climb time. The Missoulian, pp. B1-2
- Jahrig, G. (1998, August 10). Lessons through the ages. The Missoulian, pp. B1-2.
- Jick, T. (1979). Mixing qualitative and quantitative: Triangulation in action. In J. Van Maanen (Ed.), Qualitative Research (pp. 135-48). Beverly Hills, CA: Sage.

- Johnson, E.R., Merrell, K.W., & Stover, L. (1990). The effects of early grade retention on the academic achievement of fourth-grade students. Psychology in the Schools, 27, 333-338.
- Kasten, W., & Clarke, B. (1993). The multi-age classroom: A family of learners. Katonah, NY: Richard C. Owen.
- Katz, L. G. (1992). Nongraded and mixed-age grouping in early childhood programs. Urbana, IL: ERIC Clearinghouse on Elementary and Early Childhood Education. (ERIC Document Reproduction Service No. ED 353 148)
- Katz, L. G. (1996). The benefits of mixed-age grouping. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education. (ERIC Document Reproduction Service No. ED 382 411)
- Katz, L.G., Evangelou, D., & Hartman, J. A. (1990). The case for mixed-age grouping in early education. Washington, DC: National Association for the Education of Young Children. (ERIC Document Reproduction Service No. 326 302)
- Keenan, N. (March 1997). The report to the Montana 1997 state legislature [On-line]. Available: <http://www.metnet.mt.gov/OPI/Supt.html>
- Keliher, A. (1931). A critical study of homogeneous grouping. New York: Teachers College, Columbia University.
- Keppel, G. (1973). Design and analysis: A researcher's handbook. Englewood Cliffs, NJ: Prentice-Hall.
- KERA: The Kentucky Education Reform Act. (1997, December). Phi Delta Kappan, 79, 264-276.
- Lehman, B. & Scharer, P. (1996). Reading alone, talking together: The role of discussion in developing literary awareness. The Reading Teacher, 50, 26-35.
- Lincoln, Y., & Guba, E. (1985). Naturalistic inquiry. Newbury Park, CA: Sage.
- Lipsitz, J., Mizell, M.H., Jackson, A.W., & Austin, L. (1997). Speaking with one voice. Phi Delta Kappan, 78, 533-540.
- Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum.

- Ludwick, J. (1998, June 29). Montana's high school graduation rate 5th in U.S., college rate above average. The Missoulian, p. A1.
- Lyman, H.B. (1971). Test scores and what they mean. (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Mallows, C.L. (1983). Data description. In G.E.P. Box, T. Leonard, & C. Wu (Eds.), Scientific inference, data analysis, and robustness (pp.135-152). New York: Academic Press.
- Mantzicopoulos, P. Y. (1997). Do certain groups of children profit from early retention: A follow-up study of kindergartners with attention problems. Psychology in the Schools, 34, 15-27.
- Maryland School Performance Assessment Program. (1993, May). Maryland Assessment Consortium Summer Workshop. Baltimore, MD: Author.
- Mason, D.A., & Burns, R.B. (1996). Simply no worse, and simply no better may simply be wrong: A critique of Veenman's conclusion about multigrade classes. Review of Educational Research, 66, 307-332.
- Mason, D.A., & Stimson, J. (1996). Combination and nongraded classes: Definitions and frequency in twelve states. The Elementary School Journal, 96, 441-450.
- McCarthy, S. (1994). Authors, text, and talk: The internalization of dialogue from social interaction during writing. Reading Research Quarterly, 29, 201-231.
- McLean, J.E. (1992, November). The utility, reliability, and validity of holistic scoring for writing assessment samples. Symposium presented at the annual meeting of the Mid-South Educational Research Assn., Knoxville: TN (ERIC Document Reproduction Service No. 353 327)
- McLoughlin, W. (1970). Continuous pupil progress in the nongraded school: Hope or hoax? The Elementary School Journal, November, 90-97.
- Merriam, S.B. (1998). Qualitative research and case study applications in education. San Francisco: Jossey-Bass.
- Miles, M., & Huberman, A. (1994). Qualitative data analysis. Thousand Oaks, CA: Sage.
- Miller, B. (1990). A review of the quantitative research on multigrade instruction. Journal of Research in Rural Education, 7(1), 1-8.

- Miller, B. (1991). A review of the qualitative research on multigrade instruction. Journal of Research in Rural Education, 7(2), 3-12.
- Miller, R. (1967). The nongraded school: Analysis and study. New York: Harper & Row.
- Missoula County Public Schools. (1996a). Missoula achievement level testing: Teacher's guide to the MALT. Document available through Missoula County Public Schools, 215 South Sixth West, Missoula, MT.
- Missoula County Public Schools. (1996b). Title I elementary schoolwide program plan description. Documents for both schools available through Montana Title I, Office of Public Instruction, Helena, MT.
- Missoula County Public Schools. (1997). Communication arts curriculum. Document available through Missoula County Public Schools, Administration Building, 215 South Sixth West, Missoula, MT.
- Missoula County Public Schools. (1998a). Elementary/middle/high school K-12 longitudinal CTBS analysis 1991-1998. Prepared by Christopher, J. and McKean, B. Document available through Missoula County Public Schools, Administration Building, 215 South Sixth West, Missoula, MT.
- Missoula County Public Schools. (1998b). MCPS Missoula achievement level tests administration guide. Document available through Missoula County Public Schools, Administration Building, 215 South Sixth West, Missoula, MT.
- Missoula County Public Schools. (1998-99). "Charting a course" strategic plan 1995-2000 (1998-99). Document available through Missoula County Public Schools, Administration Building, 215 South Sixth West, Missoula, MT.
- Mohr, L.B. (1990). Understanding significance testing. Newbury Park: Sage.
- Muse, I., Smith, R., & Barker, B. (1987). The one-teacher school in the 1980s. Las Cruces, NM: ERIC Clearinghouse on Rural Education and Small Schools. (ERIC Document Reproduction Service No. ED 287 646)
- Myers, M. (1985). The teacher-researcher: How to study writing in the classroom. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills and the National Council of Teachers of English.

- National Council of Teachers of English and the International Reading Association. (1996). Standards for the English Language Arts. Urbana, IL: National Council of Teachers of English. Newark, DE: International Reading Association. (ERIC Document Reproduction Service No. ED 389 003)
- Northwest Evaluation Association. (1996). NWEA Achievement Level Testing Technical Manual. Portland, OR: Author.
- Northwest Evaluation Association. (1999, August). NWEA Achievement Level Test Norms. Portland, OR: Author.
- Nye, B.A. (1995). Are multiage/nongraded programs providing students with a quality education? Nashville: TN: Center of Excellence for Research in Basic Skills. (ERIC Document Reproduction Service No. ED 384 998)
- Onset of puberty in girls beginning at earlier age. (1997, April 8) . The Missoulian, p. A3. Citing M. Herman-Giddens, University of North Carolina at Chapel Hill, Pediatrics, April, 1997.
- Otto, H. (1969). Nongradedness: An elementary school evaluation (Bureau of Laboratory Schools Monograph No. 2) Austin: University of Texas, Department of Education.
- Parker, F. (1993, July). Turning points: Books and reports that reflected and shaped U.S. education, 1749-1990. (Eric Document Reproduction Service No. ED 369 695)
- Patton, M.Q. (1980). Qualitative evaluation methods. Beverly Hills: Sage.
- Pavan, B. (1992a). The benefits of nongraded schools. Educational Leadership, 50, 22-5.
- Pavan, B. (1992b, April). The waxing and waning of nongradedness. Paper presented at annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 346 608)
- Pawluk, S. T. (1992). A comparison of the academic achievement of students in multigrade elementary classrooms and students in self-contained single-grade classrooms (Doctoral dissertation, Montana State University, 1992). Dissertation Abstracts International, 53, AAC 9312278.
- Pearson, R. (1996). Homeschooling: What educators should know. Information Analyses (070). (Educational Document Reproduction Service No. 402 235)
- Popham, W.J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

- Pratt, D. (1986). On the merits of multiage classrooms. In R. Fogarty (Ed.), The multiage classroom: A collection. (pp. 48-60). Palatine, IL: IRI Skylight. (ERIC Document Reproduction Service No. ED 369 574)
- Proett, J., & Gill, K. (1986). The writing process in action: A handbook for teachers. Urbana, IL: National Council of Teachers of English.
- Pulliam, J.D., & Van Patten, J. (1995). History of education in America (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Ralph, J., Keller, D., & Crouse, J. (1994). How effective are American schools? Phi Delta Kappan, 76, 2, p. 144 (7).
- Ray, B. (1996). Home education research facts. Montana Homeschool Reference Guide, 2, 1-20. Seeley Lake, MT: Hane.
- Reichardt, C. S., & Cook, T.D. (1979). Beyond qualitative versus quantitative methods. In C.S. Reichardt & T.D. Cook (Eds.), Qualitative and quantitative methods in evaluation research (pp. 7-32). Beverly Hills, CA: Sage.
- Riis, J.A. (1971). How the other half lives: Studies among the tenements of New York. New York: Dover Press (Unabridged republication of 1901 edition of Riis' original 1890 work).
- Robinson, S. (1997). Building knowledge for a nation of learners: A framework for education research-1997. Office of Educational Research and Improvement Report[Online]. Available: [<http://www.ed.gov/offices/OERI/RshPriority/plan/chap2c.html>]. Robinson's endnote for this statistic: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing in the United States: A Statistical Profile, 1993-94, NCEs 96-124 (Washington, D.C.: U.S. Government Printing Office, 1996) 24, Table 3.
- Roderick, M. (1995). Grade retention and school dropout: Policy debate and research questions. Phi Delta Kappa Research Bulletin, 15, 1-6. (Eric Document Reproduction Service No. 397 213)
- Rollins, S. (1968). Developing nongraded schools. Itasca, IL: F.E. Peacock.
- Shannon, T. (1994). The changing local school board. Phi Delta Kappan, 75, 387-390.
- Shaughnessy, M.F. (1993). Vygotsky's zone of proximal development: Implications for gifted education. Report Type 055. (ERIC Document Reproduction Service No. 358 620)

- Shepard, L. A., & Smith, M. L. (1990). Synthesis of research on grade retention. Educational Leadership, *47*, 84-88.
- Shepherd, G. D., & Ragan, W. B. (1982). Modern elementary curriculum. Fort Worth, TX: Holt, Rinehart & Winston.
- Slavin, R.E. (1983). Component building: A strategy for research-based instructional improvement (Report No. 337). Baltimore, MD: Johns Hopkins University, Center for Social Organization of Schools. (ERIC Document Reproduction Service No. 231 794)
- Slavin, R.E. (1986). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. Review of Educational Research, *57*, 293-336.
- Slavin, R.E., Karweit, N.L., & Wasik, B.A. (1993). Preventing early school failure: What works? Educational Leadership, *50*, 10-18.
- Slavin, R.E., Madden, N.A., Dolan, L.J., Wasik, B.A., Ross, S., & Smith, L. (1994). Whenever and wherever we choose: Replication of success for all. Phi Delta Kappan, *75*, 639-47.
- Smith, F. (1983). Essays into literacy. Portsmouth, NH: Heinemann.
- Smith, K. (1993). Attitudes toward multiple aged classrooms of third, fourth, fifth, and sixth grade students. (ERIC Document Reproduction Service ED 3361 088)
- Spandel, V., & Culham, R. (1993). New directions in writing assessment. Portland, OR: Northwest Regional Educational Laboratory.
- Stanley, J.C., & Hopkins, K.D. (1972). Educational and psychological measurement and evaluation. Englewood Cliffs, NJ: Prentice-Hall.
- Stone, S.J. (1997). Foundations for successful multiage classrooms. Altoona, WI: National Education Institute.
- Tanner, C. K., & Decotis, J. (1995). The effects of continuous-progress nongraded primary school programs on student performance and attitudes toward learning. Journal of Research and Development in Education, *28*, 135-143.
- Tanner, C. K., & Galis, S.A. (1997). Student retention: Why is there a gap between the majority of research findings and school practice? Psychology in the Schools, *34*, 107-14.

- Tanner, D. (1993). A nation 'truly' at risk. Phi Delta Kappan, 75.4, 288-298.
- Tewksbury, J. (1967). Nongrading in the elementary school. Columbus, OH: Merrill.
- Tierney, W.G. (1992). Utilizing ethnographic interviews to enhance academic decision-making. In D. Fetterman (Ed.). Qualitative approaches in institutional research. San Francisco: Jossey-Bass.
- Van Horn, R. (1999). Inner-city schools: A multiple-variable discussion. Phi Delta Kappa, 81, 291-297.
- Veenman, S. (1995). Cognitive and noncognitive effects of multigrade and multi-age classes: A best-evidence synthesis. Review of Educational Research, 65, 319-81.
- Veenman, S. (1996). Effects of multigrade and multi-age classes reconsidered. Review of Educational Research, 66, 323-400.
- Venezky, R.L. (1995). What is literacy? Selected definitions and essays from the literacy dictionary: Vocabulary of reading and writing. (R. Hodges, Ed). Newark, DE: International Reading Association.
- Viadero, D. (1996). Mixed blessings. Education Week, 15, 31-33.
- Vygotsky, L. (1978). Mind in society: The development of higher psychological processes. (M. Cole, V. John-Steiner, S. Scribner & E. Souberman Trans., Eds.). Cambridge, MA: Harvard University Press.
- Vygotsky, L. (1962). Thought and language (A. Kozulin, Trans). Cambridge, MA: MIT Press (Original work published in 1934)
- Walters, D. M., & Borgers, S. B. (1995). Student retention: Is it effective? The School Counselor, 42, 300-310.
- Webb, M. (Ed.). (1993). My folks and the one-room schoolhouse. Topeka, KS: Capper.
- Wertsch, J.V. (1985). Vygotsky and the social formation of mind. Cambridge, MA: Harvard University Press.
- Wildt, A.R., & Ahtola, O. (1978). Analysis of covariance. Beverly Hills, CA: Sage.
- Wiles, J. (1976). Planning guidelines for middle school education. Dubuque, IA: Kendall/Hunt.

- Wood, D.J., Bruner, J.S., & Ross, G. (1976). The role of tutoring in problem-solving. Journal of Child Psychology and Psychiatry, 17, 89-100.
- Yin, R.K. (1984). Case study research: Design and methods. Beverly Hills, CA: Sage.
- Young, A., & Boyle, R. (1994, April). Grade-level status effects in multiage groupwork: The lady bountiful syndrome. Paper presented at the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 379 078)
- Zemelman, S., & Daniels, H.(1988). A community of writers. Portsmouth, NH: Heinemann.
- Zemelman, S., Daniels, H., & Hyde, A. (1993). Best practice: New standards for teaching and learning in America's schools. Portsmouth, NH: Heinemann.

Appendix A

Selected Summary of Frequently-Cited or Recent Research
of Cognitive/Affective Results Within Organizational Structures

Author(s)	Date	Sample	Data Collection -	Conclusion
Bender	1996	MA	CTBS -	No S.S. differences
Advisor: B.Pavan		Grades 4-5	Coopersmith Survey -	No differences Parental differences: Single grade felt more creative
Bledsoe	1994	MA Gr. 1-3	Pre- and post achievement test-	No S.S. differences
		46 students		
Brown/Martin	1987	Reading MG	Attitude Scale - GPA and	No differences No S.S. differences
		8 Canadian schools Grades 1-5	Achievement test scores - Report cards -	No S.S. Teachers favored SG
Byrnes	1994	MA	Interviews with students/parents-	Olders felt unchallenged parents agreed
Carbone	1961	Ages 6-8 NG	Achievement -	Unit-aged, graded scored significantly higher
			Mental health -	4 out 5 no difference Graded scored higher in social

(appendix continues)

Author(s)	Date	Sample	Data Collection -	Conclusion
Coon-Carty	1998	MA Gr. 1 & 4 193 students	Pre/Post achievement tests- Self-concept	No S.S. Fourth grade males in MA significantly lower control over performance
Gutiérrez and Slavin	1992	NG	Multiple Meta-analysis	Mixed results: positive for NG as group, but not for individualized instruction
Huffman	1995	MA 120 third grade students from 4 schools	Self-report -	No main effects
Lison	1997	MA only One class	Interaction Protocol	Interaction on age and sex; helping behaviors differ
McLoughlin	1967	NG 8 school districts		Differences favored SG
Milburn	1981	MA 6-11 years Canada 350 students	4 tests	No. S.S. however younger performed better on vocab
Miller, B.	1990/91	MG/rural 21 studies K-6	Achievement -	No S.S., but academically favors MG
Montgomery	1995	NG/159 gr.2-4 4 parochial schools	Affective - Self-esteem Index	S.S. difference favors NG
Muse, Smith, & Barker	1987	Rural, one-teacher 204 students/3 states	Multiple measures from each all different	No S.S.

(appendix continues)

Author(s)	Date	Sample	Data Collection -	Conclusion
Nye	1995	NG 1500 K-4 7 schools	Quantitative -	NG performs as well as or better, but no S.S.
Pavan	1973/ 1992	Study continues NG, MG, UG 64 studies From '68-90	Standardized Achievement -	Favors NG Not S.S.> Boys, African-Americans, & under-achievers do better
			Affective -	By simple count favors NG
Pawluk	1992	MG 288 Grade 5-8 Private, parochial	CTBS 4 subjects -	No S.S in any subject
Pratt	1986	30 studies	Varied -	MA; no consistent effect
			Affective -	Benign effect
Smith, K.	1993	MA 45 grade 3-5 4 classrooms	Multiage Attitude Survey -	S.S differences in correlations between grade and negative attitude
Tanner/ Decotis	1995	NG 4 schools 343 K-1	Attitude -	No S.S. Differences
			Kindergarten Assessment -	No S.S.
			Report cards -	Favored NG

(appendix continues)

Author(s)	Date	Sample	Data Collection -	Conclusion
Veenman	1995	MG MA	34 studies - 11 studies -	No S.S. Few differences
Young/ Boyle	1994	MA. Grades 3-5	Interviews with 11 pairs -	Olders saw youngers as incapable

Note. MA means multiage, MG multigrade, NG nongraded, SG single grade, SU single unit. Multiple indicates a variety of tests were administered to different students with no one test used as comparison among groups of students. SS indicates statistically significant.

Appendix B

List of Terms Used to Define Organizational Structures and Class Grouping

As Compiled from Literature Review*

blended	multiage
combination	multigrade
continuous progress	multiple-aged
double year	multiunit
family grouping	nongraded
graded	single grade
heterogenous	single unit
homogeneous	split
horizontal grouping	traditional
horizontal streaming	ungraded
individual guided	unit graded
mixed age	unit level
mixed group	vertical grouping
	vertical streaming

*The use of some terms interchangeably or without clear definitions of classroom structure created confusion in the body of literature about organizational structure. This list may not be exhaustive.

Appendix C

Initial Interview Guide: Feature Analysis of Instructional Programs, Policies and/or Practices from Conceptual Framework Compiled from Literature Review*

I. Physical setting

Places for small groups to work

Wide selection of whole, original books, and materials of high quality

Student-centered and experiential

Social environment

Cross-age experiences, including parents, community

II. Programs with:

Developmentally appropriate materials

Individual differences accommodated through varied materials and assignments

Differentiated learning

Collaborative and cooperative learning

Heterogeneous, flexible grouping

Teacher as facilitator/mentor

Opportunities for students to share literacy in a reflective manner

Process goals, especially in writing

Six-trait writing curriculum

Writing across the curriculum via inservice; and all grade and age levels

(appendix continues)

Initial Interview Guide continues

III. Assessment and Evaluation

Individual narratives/Anecdotal records

Diagnostic/Formative/Continuous

Multiple measures of assessment to include:

Authentic/Performance assessments: checklists, portfolios, rubrics...

as well as standardized tests

*The list is not exhaustive. It is intended to be a beginning for questions regarding school practice. The items listed were among those frequently mentioned in literature regarding best practice and/or multiage and nongraded classrooms, and represent this researcher's subjective selection prior to data collection.

Appendix D

Research Timeline with Design Schemata

	<u>Data Collection Methods</u>	
<u>1998</u>	<u>Interviews</u>	<u>Test</u>
September: Entry from principals	Principals/ teachers	
October	Teachers	MALT
November: End of first quarter	Superintendent	
December	Principals	
<u>1999</u>		
January:		Writing sample
February		
March		
April: End of third quarter		TerraNova
May		MALT
June	Teacher survey	Writing sample
	Superintendent	Collect scores
June 12: Begin formal data analysis		
Writing assessment by trained raters (3-half days)		
July and August	Principal follow-ups	

(appendix continues)

Research Timeline with Design Schemata continues

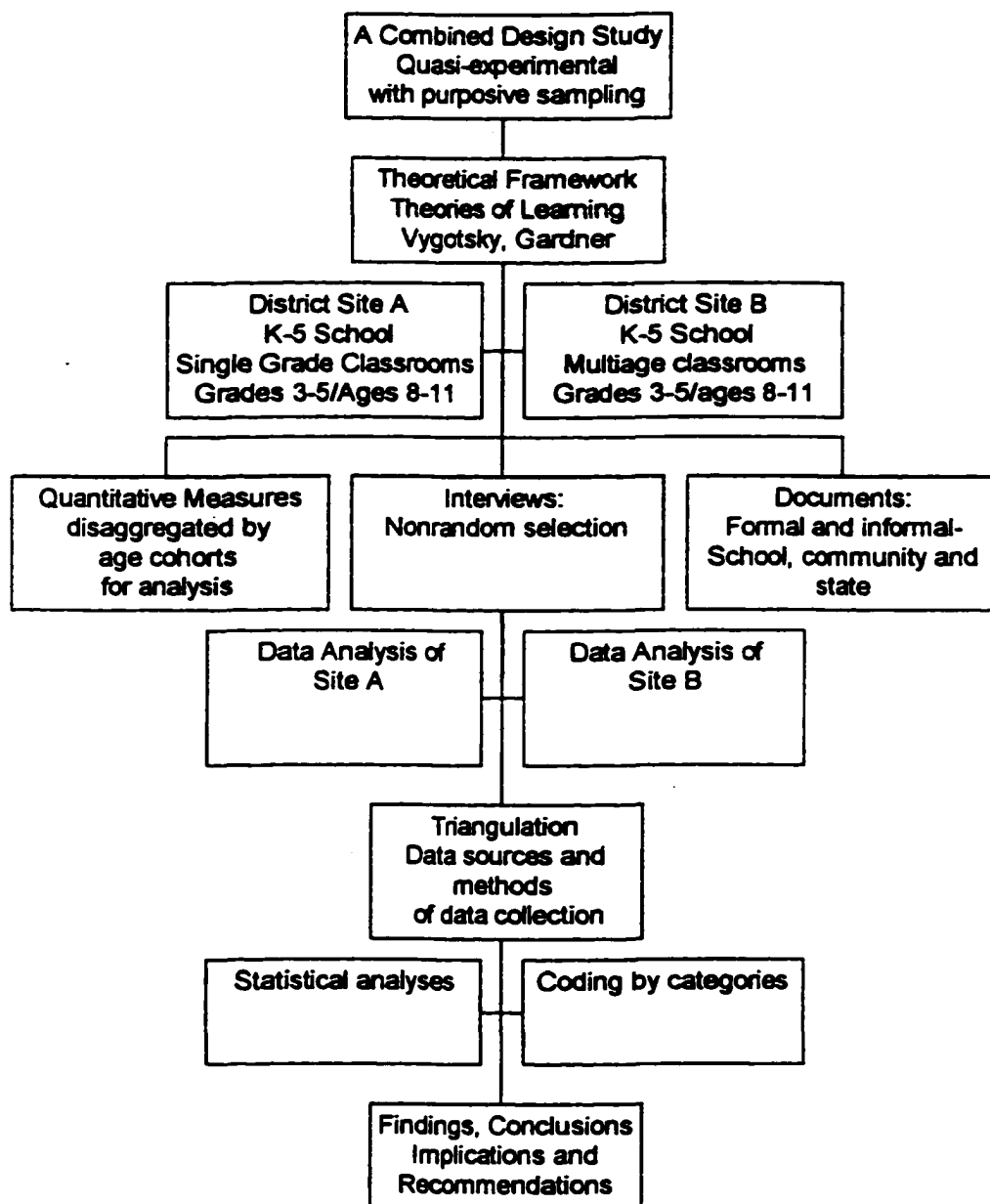


Figure 2. Schemata for Research Process for Timeline-

Appendix E

TerraNova Characteristics in Reading Test Levels 10-21/22

Demonstrate understanding of literal meanings of passage through:

identifying stated information

indicating sequence of events

define grade-level vocabulary

Analyze text by:

drawing conclusions

inferring relationships such as cause and effect

identify themes and story elements

Evaluate and extend meaning by:

making predictions

distinguishing between fact and opinion

judging author's purpose, point of view

Identify reading strategies by:

summarizing content

comparing information across texts

using graphics and text structure

Note. Compilation of main concepts in CTB/McGraw-Hill (1997). TerraNova Content Objectives, p. 34.

Appendix F

TerraNova Characteristics in Language Test Levels 11-21/22

Demonstrate understanding and knowledge of:

Sentence structure

Complete and effective sentences

Subject and verb agreement

Punctuation and capitalization

Combining sentences for clarity

Writing strategies

Information sources

Outlines

Topic and concluding sentences

Connective and transitional words and phrases

Supporting statements

Sequencing ideas

Relevant information for expository prose

Editing skills

Capitalization and punctuation

Parts of speech in existing text

Note. Compilation of main concepts presented in CTB/McGraw-Hill. (1997). TerraNova

Content Objectives, p. 34.

Appendix G

District Curriculum Goals/Objectives and Subgoals Used as Blueprint in Construction of MALT in Reading

1. Word Meaning
 - a. Understand words/sentences in context
 - b. Interpret multiple meanings
 - c. Recognize synonyms, antonyms, homonyms
 - d. Recognize component structure (prefixes, suffixes, word origins)
2. Literal Comprehension
 - a. Classify facts
 - b. Interpret directions
 - c. Recall/identify main idea
 - d. Recall details
 - e. Sequence details
3. Interpretive Comprehension
 - a. Recognize cause and effect relationships
 - b. Draw inferences
 - c. Predict events
 - d. Summarize/synthesize
4. Critical analysis
 - a. Understand and recognize bias, assumptions, stereotypes

(appendix continues)

4. Critical analysis continued
 - b. Evaluate conclusions and resolutions
 - c. Identify fact/opinion
 - d. Determine merit, accuracy, persuasive qualities
 - e. Evaluate validity
 - f. Evaluate quality of work/information/ideas
 - g. Comparative works/information
 - h. Apply and transfer knowledge
-

Note. From Missoula Achievement Level Test: Teacher's Guide to the Malt, (1996, p. 7),
Missoula, MT: Missoula County Public Schools.

Appendix H

District Curriculum Goals/Objectives and Subgoals

Used as a Blueprint in Construction of MALT in Language

1. Composing/Writing Process

- a. Prewriting skills
- b. Drafting and revising skills
- c. Editing and proofreading processes

2. Composition Structure

- a. Appropriate format
- b. Sentence forms appropriate to practice
- c. Develop paragraphs
- d. Composition forms

3. Basic Grammar/Usage

- a. Basic sentence patterns
- b. Phrases
- c. Clauses
- d. Noun forms
- e. Distinguish verb tenses
- f. Irregular verb forms
- g. Subject-verb agreement
- h. Adjective forms

(appendix continues)

3. Basic Grammar/Usage continued

- i. Adverbs
- j. Pronoun forms
- k. Pronoun antecedent agreement
- l. Negative forms

4. Conventions

- a. Appropriate end punctuation
- b. Commas
- c. Apostrophes
- d. Enclosing punctuation
- e. Underlining for titles
- f. Beginning capitalization
- g. Capitalize proper nouns and adjectives
- h. Capitalize pronoun I

Note. From Missoula Achievement Level Test: Teacher's Guide to the Malt, (1996, p. 8),

Missoula, MT: Missoula County Public Schools.

Appendix I

Criteria for Good Writing Prompts and Pre and Post/Directions for Students

An effective writing prompt has these characteristics:

1. Contains clear instructions;
2. Consists of carefully chosen words (i.e., explain, tell a story about, convince, etc.) If you want to elicit a certain mode, be sure the directional words will encourage writing in that mode, as distinguished from other modes;
3. Allows assessment of writing—not knowledge, not reading. Avoid prompts that would give some students an advantage because of their knowledge base. Rather select prompts that allow students to write from their personal experiences;
4. Is focused;
5. Is brief, but not cryptic;
6. Allows for mental elbow room;
7. Is free from bias (gender, race, culture, socio-economic background);
8. Respects students' privacy; does not encourage writing that could easily become too personal;
9. Has no "built-in" answer (can't be answered YES or NO);
10. Avoids inflammatory issues;
11. Is interesting (select something you'd enjoy writing about);
12. Is appropriate for the grade level(s) being assessed;
13. Allows for the best writing by both the most capable and least capable writers;

(appendix continues)

Prompts/Directions continue

14. Avoids built-in positives and negative (e.g., “Write an essay on what makes life wonderful”)

Note. From Managing Your Assessment with Confidence & Style (1993) developed by Dr. Judy Arter, Evaluation and Assessment, Northwest Regional Educational Laboratory, Portland, OR.

(appendix continues)

Prompts/Directions continue

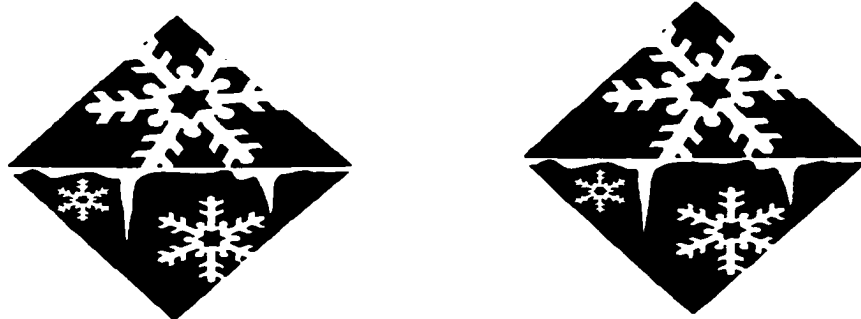
DIRECTIONS FOR THE STUDENT

This morning you will have one half hour to write on the topic:

A Perfect Winter Day in Missoula, Montana

The writing you will do is not a test. It will not be graded. Your audience will be four writing teachers from other schools and your teacher who want to know what kids can do when they write all by themselves on an assigned topic. You may write in cursive or print, whichever is easier for you. You may write on both sides of each page in the blue booklet. If you need more paper, use classroom paper.

Your role is to be yourself. Write from your point of view. Your format is to describe in detail a winter day so that the readers can see in their minds what you are writing about. Tell as much as you like about the whole day. It can be a school day, or a weekend day, but it must be wintertime. This can be a day that has happened to you, or one that you think could really happen to you during winter in Missoula, Montana.



(appendix continues)

Prompts/Directions continues

DIRECTIONS FOR THE STUDENT

This morning you will have one half hour to write on the topic:

A Perfect Spring Day in Missoula, Montana

The writing you will do is not a test. It will not be graded. Your audience will be four writing teachers from other schools and your teacher who want to know what kids can do when they write all by themselves on an assigned topic. You may write in cursive or print, whichever is easier for you. You may write on both sides of each page in the blue booklet. If you need more paper, use classroom paper.

Your role is to be yourself. Write from your point of view. Your format is to describe in detail a spring day so that the readers can see in their minds what you are writing about. Tell as much as you like about the whole day. It can be a school day, or a weekend day, but it must be springtime. This can be a day that has happened to you, or one that you think could really happen to you during spring in Missoula, Montana.



Appendix J

Questionnaire of Writing Assessment for Classroom Teachers

Please feel free to comment beyond a yes or no response. Use the margins or the back. All comments will be anonymous and confidential.

1. Do you feel the writing topic was suitable for your students?

_____ Yes _____ No

2. Did the student direction give them enough guidance?

_____ Yes _____ No

3. Did the directions give you enough information?

_____ Yes _____ No

4. Do you feel from observing the class that enough time was given for this prompt?

_____ Yes _____ No

5. Has your class had experience writing on topics similar to these two prompts?

_____ Yes _____ No

6. Have you had workshops on the teaching of writing according to the six-trait writing analysis?

_____ Yes _____ No

7. Do you use the six-trait writing language in your writing instruction?

_____ Yes _____ No

Appendix K

Demographic Questionnaire for Writing Assessment Raters

Completed at Conclusion of Scoring

The results of this questionnaire will be used for general statements about the demographics of this group in this assessment. All responses will be anonymous and confidential.

1. What is your level of education? B.A./B.S. _____ M.A./M.S. _____ EdD _____ Other _____
2. If teaching, what grade level do you now teach? _____
3. How many years have you been teaching at this level? _____
4. How many total years have you taught? _____ What grades? _____
5. Are you employed by Missoula County Public Schools? _____
6. If you have taught writing using any proscribed model, would you name/describe it?
7. If you have ever participated in a study, project or training in writing assessment, would you name/describe it?
8. Were your working conditions (space, light, food, collegiality, temperature, other) adequate during the scoring of student papers?

Please be candid if any condition detracted from your work _____

9. Were you a willing participant in this study? _____
10. Would you share some brief impressions of this experience? I would like to include some of your comments in the description of this process. These will be anonymous.

* Please let Leslie Ferrell know if you would like a copy of the final results and/or a letter for your professional file to include a note of gratitude for your participation.

Appendix L
Rater Invitation

Dear Colleague,

As per our discussion, I am inviting you to be a rater in a research study to assess writing of elementary students. This assessment is one component of the research for my dissertation. The writing assessment will be conducted on June 29, June 30, and July 1 beginning at 1:30 p.m. at _____. The estimated amount of time will be no more than three hours each for the three consecutive days. Your time will be compensated at \$20.00 per hour. Snacks will be provided.

As a naive rater in this study you will be trained to use holistic scoring. I contacted you specifically because I know that you have at least seven years of full-time teaching experience, have participated in previous writing assessments, and/or completed the Montana Writing Project.

I will be very appreciative of your participation on the rating team. Please call me at _____ as soon as possible if your plans have changed. If they have not, I look forward to seeing you on June 29.

Sincerely,

Leslie Ferrell

Appendix M

Sample by Grade and Ethnic Group

Group	3	4	5	6	7	8	Avg.	Nat'l ¹
American Indian/ Alaskan Native	1.7	1.7	1.7	1.3	1.6	1.6	1.6	1.0
Asian or Pacific Islander	8.6	9.1	8.4	8.6	9.1	9.9	8.9	3.5
Black, not Hispanic	10.0	9.8	10.1	8.6	9.5	9.3	9.6	16.5
Hispanic	9.0	8.6	8.0	7.7	7.3	7.7	8.0	12.3
White, not Hispanic	70.7	70.7	71.8	73.7	72.6	71.5	71.8	66.7

¹Source: U.S. Department of Education, Office of Educational Research and Improvement, Digest of Educational Statistics, 1994, p. 60. Data indicate enrollment in public elementary and secondary schools for the fall of 1992.

Note. From the Northwest Evaluation Association Level Test Norms, 1996 (p. 4), Portland, OR: Northwest Evaluation Association. Reprinted with permission.

Appendix N

Grade Level Means (standard deviations) for NWEA 1998-99 Norms

Grade	Reading			Language						
	Fall	<u>SD</u>	Spring	<u>SD</u>	Average Growth	Fall	<u>SD</u>	Spring	<u>SD</u>	Average Growth
3	186.10	(17.03)	196.14	(16.68)	9.8	188.61	(15.24)	196.69	(15.38)	8.9
4	196.38	(16.44)	203.26	(16.23)	6.5	198.78	(15.19)	204.32	(14.69)	5.7
5	203.83	(16.10)	210.20	(15.95)	5.4	205.11	(14.95)	210.71	(14.23)	4.8

Note. From Northwest Evaluation Association Level Test Norms, 1999, p. 11. Portland, OR: Northwest Evaluation Association. Adapted with permission from NWEA.

Appendix O

Marginal Reliabilities for the NWEA Achievement Level Tests
for Grades 3-5 for Reading and Language-1995

Grade	Reading	Language
3	.932	.939
4	.931	.940
5	.925	.931

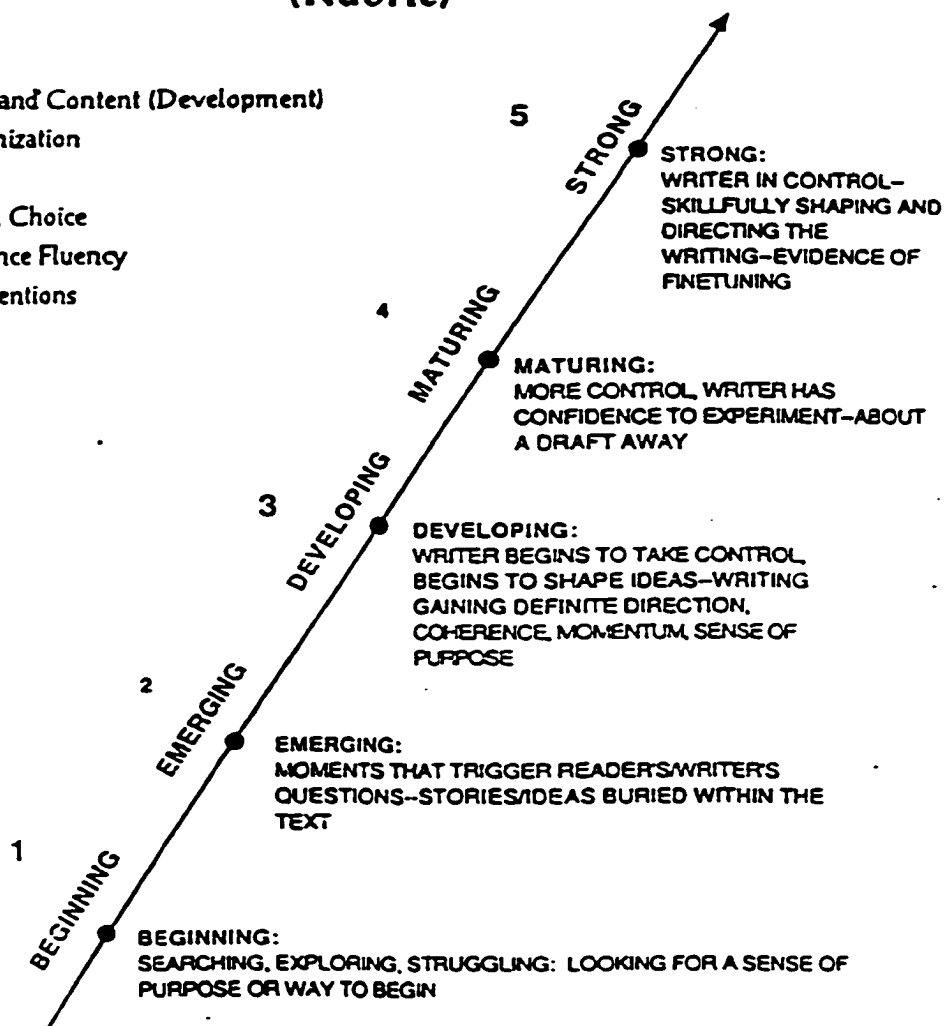
Note. From the Achievement Level Testing Technical Manual, 1996 (p.11), Portland, OR: Northwest Evaluation Association. Adapted with permission from NWEA.

Appendix P

Six-Trait Writing Assessment Rubric

ANALYTICAL TRAIT SCORING GUIDE (Rubric)

- Ideas and Content (Development)
- Organization
- Voice
- Word Choice
- Sentence Fluency
- Conventions



Developed by Vicki Spandel and Ruth Culham of the Northwest Regional Educational Laboratory, June, 1993. This scoring guide is an updated version of the one that appears in Spandel and Stiggins, *Creating Writers*. Addison-Wesley: 1990. The original guide was developed by teachers from the Beaverton, Oregon School District in 1984. The Laboratory gratefully acknowledges the contributions of the more than 10,000 teachers and students whose shared insights and comments are reflected in this revision.

Appendix Q

Writing Assessment Scoring Criteria Fluency and Conventions Scales

Fluency Scale Criteria Review

- Emerging Literacy Phase
- 1 Scribble writing
Real letters copies with no letter/sound correlation evident
Real letters randomly typed if done on computer
Student may write name
Student can often "read" this "kid-writing." but adults cannot
- 2 Unrelated words copied or memorized
Family names. i.e. Mom, Dad, etc.
No story or story line present
Possible new words developmentally spelled
Limited letter/sound correlations may be evident
Generalized knowledge about words. i.e. fun, sun, or fat, cat, bat, etc.
Content unrelated
- 3 Single original sentences
Story beginnings, but no development
"Fat, cat, sat" stories
Plug in new nouns or verbs to a consistent sentence pattern
Recognition and repetition of pattern: pattern stories
- Developing Fluency Phase
- 4 Several distinct related sentences
Same story line apparent and could include a sequence of events
Factual recall of events with no reflection or embellishment with details
Chronological listing often begins: "On Sunday...Last night...etc.
Highly literal, author seems to write all he/she can write, indicating limits of fluency
- 5 Simple narratives or stories
Sequence of events may be presented as a story
Story is embellished with some details or personal reflection
Writing is mostly complete with beginning, middle and end apparent
Pieces may end abruptly with "the end"
Author seems to write all he/she can write, indicating limits to fluency
- 6 Simple narratives or stories flooded with superfluous detail
Story line present, but not always easily followed
Irrelevant embellishment: no item or episode appears more important than another
Increased fluency
quantity may be evident but quality may be low
Tends to ramble and become boring to the reader
- Conscious Control Phase
- 7 Details or reflection selected to further story line
Story holds together and has more developed beginning, middle and end
Author exhibits conscious control over the writing process
More concise, less rambling
Aware of audience
End brings more closure than "the end"
- 8 Increasing clarity and conciseness in the piece
Author may use a style, voice or form to enhance the story but may not follow through
Increasing levels of conscious control: audience awareness
Risk taking with style, voice, or tone may be evident
- 9 Author has a clear purpose and fulfills it
Voice and tone more evident and easily manipulated for effect
Style established and style changes based on audience, form or purpose
Conscious control and audience awareness are consistent
Risktaking evident and often successful

(appendix continues)

Conventions Scale Criteria Review

Fluency Level One

I = Indiscriminate conventions evident

- can not be perceived by the reader

E = Emergent conventions evident which may include:

- directionality, spacing between "words"
- pictures with scribble writing
- list of known letters if handwritten

Fluency Level Two

I = Indiscriminate conventions evident

- can not be perceived by the reader

E = Emergent conventions evident which may include:

- directionality, spacing between "words"
- picture with scribble or phonetically written caption

L = Low conventions evident, interfering with readability, which may include:

- extensive use of temporary (phonetic) spelling
- numerous punctuation/capitalization errors, or lack of consistent application
- numerous fragments and (less commonly) run-on sentences
- numerous usage errors

Fluency Level Three

E = Emergent conventions evident which may include:

- directionality, spacing between "words"
- picture with caption phonetically written

L = Low conventions evident, interfering with readability, which may include:

- extensive use of temporary (phonetic) spelling
- numerous punctuation/capitalization errors, or lack of consistent application
- numerous fragments and (less commonly) run-on sentences
- numerous usage errors

M = Middle conventions evident which may include:

- frequent use of temporary (phonetic) spelling for unfamiliar words
- frequent punctuation/capitalization errors
- frequent fragments and/or run-on sentences
- frequent usage errors

H = High conventions evident which may include:

- occasional use of temporary (phonetic) spelling for unfamiliar words
- occasional punctuation/capitalization errors
- occasional fragments and/or run-on sentences
- occasional usage errors

(appendix continues)

Conventions criteria continues

Fluency Levels Four, Five, and Six

L = Low conventions evident, interfering with readability which may include:

- extensive use of temporary (phonetic) spelling
- numerous punctuation/capitalization errors, or lack of consistent application
- numerous fragments and (less common) run-on sentences
- numerous usage errors

M = Middle conventions evident with may include:

- frequent use of temporary (phonetic) spelling for unfamiliar words
- frequent punctuation/capitalization errors
- frequent fragments and/or run-on errors
- frequent usage errors

H = High conventions evident which may include:

- occasional use of temporary (phonetic) spelling for unfamiliar words
- occasional punctuation/capitalization errors
- occasional fragments and/or run-on sentences
- occasional usage errors

Fluency Levels Seven, Eight, and Nine

L = Low conventions evident, interfering with readability which may include:

- extensive use of temporary (phonetic) spelling
- numerous punctuation/capitalization errors, or lack of consistent application
- numerous fragments and (less commonly) run-on sentences
- numerous usage errors

M = Middle conventions evident with may include:

- occasional use of temporary (phonetic) spelling for unfamiliar words
- occasional punctuation/capitalization errors
- occasional fragments and/or run-on errors
- occasional usage errors

H = High conventions evident which may include:

- rare use of temporary (phonetic) spelling for unfamiliar words
- rare punctuation/capitalization errors
- rare fragments and/or run-on errors
- rare usage errors

Note. From *Holistic Developmental Writing Scales* (1997). Next Generation Learning Tools. Missoula: MT. Available through Instructional Media Services, University of Montana. Reprinted by permission of Dr. Tammy Elser, author.

Appendix R

Principal Interview Protocol

As we discussed earlier, this study explores literacy development of upper elementary age children within different organizational structures. I'd like to know your perspective. I have some general questions to begin and please feel free to elaborate as you wish. I will transcribe these notes and then return to share them with you and see if you feel I have correctly interpreted your ideas. Are there any questions you would like to ask me first?

I. Background/Perspective

1. Tell me about your teaching experiences and resulting philosophy of education.
2. Were you able to choose this assignment?
3. What is the extent of your special training? Workshops?
4. Would you delineate the teaching experience levels and professional development of your 3-5 teachers?

II. Instructional Practices

1. Do you recommend specific practices in the classroom?
2. What are the most common practices in the upper elementary classrooms you've observed?
3. Does the district curriculum prescribe certain practices or approaches?

(appendix continues)

Principal Interview Protocol continues

III. Literacy Development

1. Would you like to describe your school's goals and objectives in this area?
2. Do you have curriculum practices specific to your school needs?
3. Would you describe your assessment and evaluation methods?
4. Are there any stories of your school organization that you would feel pertinent to this research? History of development? Parental requests? District mandates? Other?

IV. Teacher Collaboration

1. How do you feel the teachers felt about conducting a writing assessment outside of regular school requirements? Do you feel they adhered to the instructions?
2. What is your perspective regarding the teaching of writing within your building and/or specific classrooms?
3. How do you feel about the collection, analysis and utilization of data for school program evaluation?

VI. School Demographics

1. What data do you consider most important for this research to consider? Enrollment, transiency rate, SES, events occurring affecting school atmosphere, diverse ethnic and cultural populations, ...

Appendix S

Achievement Level Test Parent Report

Name: _____ School: _____
 ID: _____ Teacher: _____
 Grade: 5 Term: _____

Subject	RIT	%ile	Very Low	Low	Low Average	Average	High Average	High	Very High
Language Usage	212	63	■						
Mathematics	208	55	■						
Reading	207	48	■						

EXPLANATION OF THE TEST SCORES

RIT Score This score is a measure of the student's skill level in the subjects tested. Typically the RIT score ranges from 160 for students in beginning 3rd grade to 260 for the most advanced 8th graders. The RIT score should show growth from year to year.

Percentile %ile This score indicates a student's standing compared to other students in the nation. For example, a percentile score of 50 indicates that 50% of the students in the same grade scored at this level or lower. The percentile score remains the same from year to year if the student maintains the same growth rate.

Performance on this goal was:

Goal Performance

Low	Avg	High

Goals Tested in Language Usage

1. COMPOSING-WRITING PROCESS
2. COMPOSITION STRUCTURE
3. BASIC GRAMMAR USAGE
4. CONVENTIONS

Low	Avg	High

Goals Tested in Mathematics

1. PROBLEM SOLVING-REASONING-CONNECTIONS
2. NUMBER SENSE AND NUMERATION
3. COMPUTATION AND ESTIMATION
4. PROBABILITY AND STATISTICS
5. ALGEBRA
6. GEOMETRY, SPATIAL SENSE, MEASUREMENT

Low	Avg	High

Goals Tested in Reading

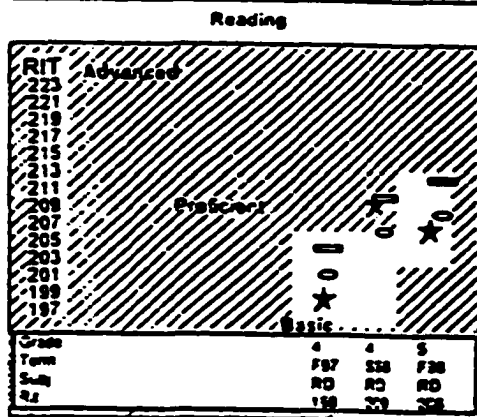
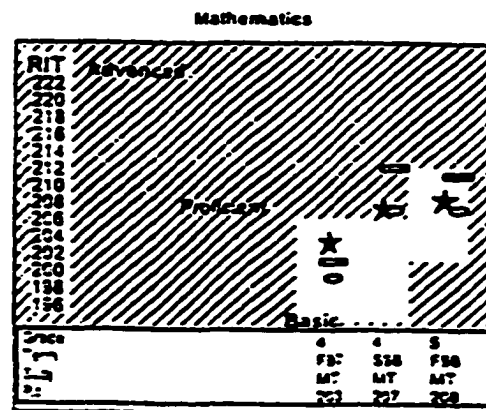
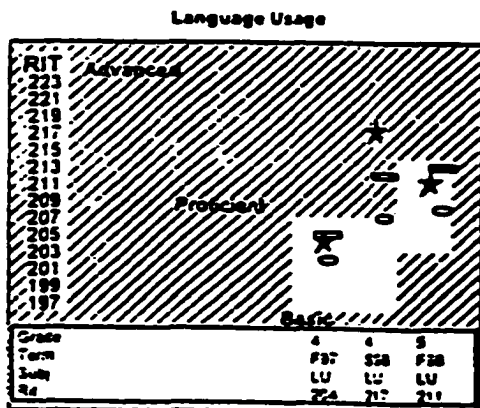
1. WORD MEANING
2. LITERAL COMPREHENSION
3. INTERPRETIVE COMPREHENSION
4. CRITICAL ANALYSIS

Note. From NWEA Achievement Level Test Manual (1996), Portland, OR: Northwest Evaluation Association. Adapted by permission of Northwest Evaluation Association.

Appendix T

Achievement Level Test Longitudinal Report by Student

Name: _____ School: _____
 Grade: _____ ID: _____ Teacher: _____



DEFINITIONS USED IN THE GRAPHS

Basic - Students in the basic range are below the standard and may require more time than other students to attain the next proficient level.

Proficient - Students in the proficient range should, with normal growth, meet standards at all levels.

Advanced - Students in the advanced range may be able to meet proficiency standards at more advanced levels.

Legend: ★ Student Score ▬ District Average ○ Norm Group Average

Note. From NWEA Achievement Level Test Manual (1996), Portland, OR: Northwest Evaluation Association. Adapted by permission of Northwest Evaluation Association.

Appendix U

Summary of Programs, Training, Practices Within Each School Including Schoolwide
Plans to Meet Population Needs (under Section 4, A-E, 5 and 7 of federal plan)

<u>Program or Practice</u>	<u>Control</u>	<u>Experimental</u>
Reading Recovery continuance	X	X
Silent Sustained Reading schoolwide	X	X
Title I/classroom/Spec Ed scheduled planning time	X	X
Increased in-class time for Title I teachers	X	X
Uninterrupted language arts time block for primary		X
Progress/Assessment		
For each student in school (blue folders)	X	
For incoming students	X	X
CCC Successmaker		
all staff trained	X	X
program in each classroom	X	
program in lab		X
Title I staff CCC	X	X
Writing to Read computer lab/Gr. 1	X	
Writing computer lab Gr. 2-5	X	
Mini writing lab in kindergarten classrooms	X	

(appendix continues)

<u>Program or Practice</u>	<u>Control</u>	<u>Experimental</u>
Literacy Resource Library/leveled reading resources and other literacy supports for teachers	X	
Family Resource Center in school	X	X
Bilingual tutoring	X	X
Extended day classes/Hmong	X	
Summer School		X
Summer literacy extension for Title I students		X
		(served 18)
Extended Day Kindergarten	X	
Retention and Dropout Prevention/Native American	X	
Evening tutorial/Native American (Available to all district students)	X	X
School Nurse Outreach program	X	X
Methods to determine if needs are met:	X	X
Primary Reading intervention	X	X
Pre and post running records/K- 2	X	X
Expand to intermediate		X
Observation Survey Data/Kindergarten		X
Based on Reading Recovery	X	X

(appendix continues)

<u>Program or Practice</u>	<u>Control</u>	<u>Experimental</u>
District writing assessment/Gr. 4	X	X
Whole staff assessment of K-5 writing using Holistic Development Scale	X	
Pre and post CCC reports	X	X
Malt Tests, pre and post/ G.3/4/5	X	X
CTBS/ Gr. 3-5	X	X
Block scheduling		X
Student Blue folders passed to next teachers each year contains pre- and post CCC reports, teacher observations, writing samples, work samples, running records, kindergarten observation survey, MALT and CTBS scores, and other teacher-selected pieces	X	
Student portfolio contains running records, Essential Word lists, and three writing pieces		X
Checklists for student progress All students	X	X
Intervention teams	X	X

(appendix continues)

<u>Program or Practice</u>	<u>Control</u>	<u>Experimental</u>
Volunteer program (Community)	X	X
Parent Volunteer program	X	X
Business Partnerships	X	X
Flagship Project	X	
Even Start Program	X	
Summer Feeding Program	X	
Summer Flagship Program	X	
Transition meetings for Special Education and other students in the Spring	X	X
Inclusion of Special Education students in classroom (Resource and extended Resources)	X	X
Professional Development Activities: Consultants/Conferences		
Jerry McVay, MCPS Title I administrator	X	X
Dr. Tammy Elser, Title I Distinguished Educator	X	X
Dr. Andrews, multiage consultant		X
At Risk Conference	X	
Cherry Valley Elementary School Library	X	X

(appendix continues)

<u>Program or Practice</u>	<u>Control</u>	<u>Experimental</u>
Cognitive Coaching training	X	
Computer Curriculum Corporation (CCC)	X	X
Conflict Resolution training		X
Dimensions of Learning	X (16 teachers)	X
Diversity Training	X	
Effective Schools Conference		X (3)
IRA annual convention	X	X
Literacy Learning	X	X
Northwest Regional Lab contact		X
NWAAHPERD conference	X	
Ohio Reading Recovery Conference		X (1)
Options for Curriculum Delivery	X	
Project Adapt	X	X(2)
Reading Recovery	X	X
Running records training		
K-2 teachers		X
Expanding to intermediate		X
K-2 teachers (Phase 1)	X	
Entire staff (Phase 2)	X	

(appendix continues)

<u>Program or Practice</u>	<u>Control</u>	<u>Experimental</u>
Society for Developmental Education		X
SPSS Student Data collection training	X	
Title I conferences	X	X
Wright Group	X	

Note. Data taken from OPI schoolwide plan applications (1996), and supplemented with principal interviews and subsequent member checks. Omissions may occur due to later training by teachers during the year of research. In addition, inservice training mandatory for all schools was not included in the summary.