AN APPLICATION OF THEORY-DRIVEN EVALUATION
IN EDUCATIONAL MEASUREMENT

DISSERTATION

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

By
Gregg Manley Gascon, M.P.A.

* * * * *

The Ohio State University
2006

Dissertation Committee:                        Approved by
Helen Marks                                          Advisor
John Opfer                                          College of Education
William E. Loadman

* * * * *
ABSTRACT

In the 2004-05 school year, the Canton City School District embarked on a plan to increase student achievement by giving teachers the tools to conduct formative and value-added student assessments in reading and mathematics. These initiatives were the major thrust of continuous improvement plans created at both the district and building levels. The purpose of this study is to report on the findings of the Canton City School District’s theory-driven evaluation of its initial implementation of these assessment systems.

Five hypotheses were tested within two cohorts of students in the eleven elementary school buildings identified by the Ohio Department of Education in school year 2004-05 as at risk or in school improvement. This study found that the relationship between predicted achievement scale scores generated by the state’s value-added assessment system and observed reading Ohio Achievement Test scale scores in the sample groups of students were positive, significantly strong and linear in the 2004-05 and 2005-06 school years; these predicted scores correctly classified proficient readers more than 75 percent of the time across cohorts. The use of value-added predicted scores as level-1 variables in unconditional and random effects ANCOVA hierarchical linear models were associated with varying levels of ICC values dependent upon model and cohort parameters. The study found no statistically significant difference between reading achievement proficiency levels in the year before and year of the implementation of the state’s value-added assessment system.
Dedicated to my loving family in appreciation of their sacrifices and prayers on my behalf
ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my adviser, William E. Loadman, Ph.D. His support, encouragement and guidance during my doctoral education have been invaluable to me. I would also like to thank Helen Marks, Ph.D., and John Opfer, Ph.D., members of my candidacy examination and dissertation committees, for their critical support of this study. I am also grateful to Xiadong Liu, Ph.D., for his outstanding instruction in multivariate statistical approaches in the education sciences. He remains a highly valued professor and friend.

I wish to thank Michele Evans-Gardell, Ph.D., Superintendent Dianne Talarico and Canton Professional Educators Association (CPEA) President Sam Dorto of the Canton City School District for their assistance with the evaluation. I would like to express my sincere gratitude to Dr. Michelle Evans-Gardell for providing documentation of her work in the context of this evaluation, and Mike Nicholson from Battelle for Kids for providing me with background information on the projected and predicted score algorithms of the value-added model. I would also like to thank my colleagues Fritz Fekete, Don Holub, Andy Jewell, Briana McKay, Laura Simonini and Pat Turner at the Ohio Education Association who gave me the time to complete this study.
In the spirit of this study, I would also like to thank all of the teachers who have had a profound impact on the direction of my work (in reverse chronological order): in graduate school, William Loadman, Ph.D., Xiadong Liu, Ph.D., Charles A. Washington, Ph.D.; in college, Loy Littlefield, Ph.D. and Jaak Panksepp, Ph.D.; in secondary school, Jeanette Storms; in primary school, Sam Salvo, Ph.D.

I would also like to thank David Du, R.N., Debra Holm, R.N., and Pastor Steven Ray of China Service Ventures and the students of the Qi Gang Primary School, Tang Geng Primary School and Li Ji Ja Primary School of Henan Province, China, for the motivation to complete this dissertation. The longitudinal school health program evaluation being conducted there holds the potential to help many students achieve their goals for better health and academic achievement; I am indebted to all of them for their kind invitation to assist in the evaluation.
VITA

February 20, 1966 ........................................ Born – Fairview Park, Ohio

1984 – 1988 ........................................... B.A., Psychology
Bowling Green State University

Cleveland State University

1992 – 1997 .......................................... Researcher
Cuyahoga Research Group, Inc.
Parma Heights, Ohio

1997 – present ...................................... Researcher
Ohio Education Association
Columbus, Ohio

FIELDS OF STUDY

Major Field: Education

Area of Emphasis: Quantitative Research, Evaluation and Measurement
TABLE OF CONTENTS

Abstract ........................................................................................................................................ ii

Dedication .................................................................................................................................. iii

Acknowledgments .................................................................................................................... iv

Vita ........................................................................................................................................ vi

List of Tables ........................................................................................................................ xi

List of Figures ........................................................................................................................ xiii

Chapters:

1. Introduction .........................................................................................................................1
   Statement of the problem .................................................................................................4
   Canton City School District’s academic status, 2004-05 ..............................................8
   Assumptions underlying the study .............................................................................12
   Specific research questions .........................................................................................13
   Significance of the study .............................................................................................15
   Delimitations ..................................................................................................................17
   Definitions and operational terms ..............................................................................17
      Theory-driven evaluation .........................................................................................17
      Change model ..........................................................................................................19
      Action model ...........................................................................................................19
      Assessment ...............................................................................................................20
      Tests ..........................................................................................................................20
      Measurement ............................................................................................................20

2. Literature Review .............................................................................................................21
   Elementary and Secondary Act of 1965 ......................................................................21
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation plan development</td>
<td>25</td>
</tr>
<tr>
<td>Decisions</td>
<td>27</td>
</tr>
<tr>
<td>Politics</td>
<td>27</td>
</tr>
<tr>
<td>Multidimensional measurement</td>
<td>28</td>
</tr>
<tr>
<td>Standards</td>
<td>29</td>
</tr>
<tr>
<td>Eclectic flexibility</td>
<td>30</td>
</tr>
<tr>
<td>Education Sciences Reform Act of 2002</td>
<td>32</td>
</tr>
<tr>
<td>Scientifically Based Research Standards</td>
<td>33</td>
</tr>
<tr>
<td>Scientifically Valid Education Evaluation</td>
<td>34</td>
</tr>
<tr>
<td>Scientifically Valid Research</td>
<td>34</td>
</tr>
<tr>
<td>U.S.D.O.E. Notice of Final Priority</td>
<td>35</td>
</tr>
<tr>
<td>Conceptual framework of program theory</td>
<td>36</td>
</tr>
<tr>
<td>Practical evaluation taxonomy</td>
<td>40</td>
</tr>
<tr>
<td>Origin and ramifications</td>
<td>44</td>
</tr>
<tr>
<td>Actionable data</td>
<td>48</td>
</tr>
<tr>
<td>Educational assessment</td>
<td>51</td>
</tr>
<tr>
<td>Summative assessment</td>
<td>55</td>
</tr>
<tr>
<td>Value-added assessment</td>
<td>58</td>
</tr>
<tr>
<td>Formative assessment</td>
<td>62</td>
</tr>
<tr>
<td>3. Procedures</td>
<td>67</td>
</tr>
<tr>
<td>Design</td>
<td>67</td>
</tr>
<tr>
<td>Principles</td>
<td>71</td>
</tr>
<tr>
<td>Guidelines</td>
<td>74</td>
</tr>
<tr>
<td>Derivation of general and specific research hypotheses</td>
<td>77</td>
</tr>
<tr>
<td>Participants</td>
<td>81</td>
</tr>
<tr>
<td>Instruments</td>
<td>81</td>
</tr>
<tr>
<td>Predicted Ohio Achievement Test scores in reading</td>
<td>81</td>
</tr>
<tr>
<td>Observed Ohio Achievement Test scores in reading</td>
<td>82</td>
</tr>
<tr>
<td>Teacher assessment questionnaire</td>
<td>83</td>
</tr>
<tr>
<td>Structured interview protocols</td>
<td>84</td>
</tr>
<tr>
<td>Statistical treatment</td>
<td>84</td>
</tr>
<tr>
<td>Limitations</td>
<td>85</td>
</tr>
<tr>
<td>4 Results</td>
<td>87</td>
</tr>
<tr>
<td>School improvement plans</td>
<td>87</td>
</tr>
<tr>
<td>Program rationale</td>
<td>90</td>
</tr>
<tr>
<td>Initial implementation lessons</td>
<td>94</td>
</tr>
<tr>
<td>Teacher assessment questionnaire</td>
<td>95</td>
</tr>
<tr>
<td>Structured interviews</td>
<td>101</td>
</tr>
<tr>
<td>Value-added assessment</td>
<td>102</td>
</tr>
<tr>
<td>Formative reading assessment</td>
<td>103</td>
</tr>
<tr>
<td>Formative mathematics assessment</td>
<td>104</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Canton City School District academic ratings by building, 2004-05</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Fourth-grade student demographics, 2004-05</td>
<td>108</td>
</tr>
<tr>
<td>3</td>
<td>Fifth-grade student demographics, 2004-05</td>
<td>109</td>
</tr>
<tr>
<td>4</td>
<td>Observed fourth-grade reading scale scores across schools, 2004-05</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>Unconstrained model of fourth-grade student reading achievement, 2004-05</td>
<td>110</td>
</tr>
<tr>
<td>6</td>
<td>Contingency table of predicted and observed proficiency status, fourth grade, 2004-05</td>
<td>113</td>
</tr>
<tr>
<td>7</td>
<td>One-way ANCOVA with random effects model of fourth-grade student reading achievement, 2004-05</td>
<td>116</td>
</tr>
<tr>
<td>8</td>
<td>Observed fifth-grade reading scale scores across schools, 2004-05</td>
<td>119</td>
</tr>
<tr>
<td>9</td>
<td>Unconstrained model of fifth-grade student reading achievement, 2004-05</td>
<td>120</td>
</tr>
<tr>
<td>10</td>
<td>Contingency table of predicted and observed proficiency status, fifth grade, 2004-05</td>
<td>123</td>
</tr>
<tr>
<td>11</td>
<td>One-way ANCOVA with random effects model of fifth-grade student reading achievement, 2004-05</td>
<td>125</td>
</tr>
<tr>
<td>12</td>
<td>Fourth-grade student demographics, 2005-06</td>
<td>128</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>Program rationale</td>
<td>93</td>
</tr>
<tr>
<td>2</td>
<td>Action theory</td>
<td>94</td>
</tr>
<tr>
<td>3</td>
<td>Predicted and observed fourth-grade reading achievement scale scores across schools, 2004-05</td>
<td>112</td>
</tr>
<tr>
<td>4</td>
<td>Fourth-grade student reading achievement scale scores, 2004-05</td>
<td>115</td>
</tr>
<tr>
<td>5</td>
<td>Level-1 residuals in the one-way random effects ANCOVA HLM of fourth-grade reading achievement, 2004-05</td>
<td>118</td>
</tr>
<tr>
<td>6</td>
<td>Predicted and observed fifth-grade reading achievement scale scores across schools, 2004-05</td>
<td>122</td>
</tr>
<tr>
<td>7</td>
<td>Fifth-grade student reading achievement scale scores, 2004-05</td>
<td>124</td>
</tr>
<tr>
<td>8</td>
<td>Level-1 residuals in the one-way random effects ANCOVA HLM of fifth-grade reading achievement, 2004-05</td>
<td>127</td>
</tr>
<tr>
<td>9</td>
<td>Predicted and observed fourth-grade reading achievement scale scores across schools, 2005-06</td>
<td>132</td>
</tr>
<tr>
<td>10</td>
<td>Fourth-grade student reading achievement scale scores, 2005-06</td>
<td>135</td>
</tr>
<tr>
<td>11</td>
<td>Level-1 residuals in the one-way random effects ANCOVA HLM of fourth-grade reading achievement, 2005-06</td>
<td>138</td>
</tr>
<tr>
<td>12</td>
<td>Predicted and observed fifth-grade reading achievement scale scores across schools, 2005-06</td>
<td>141</td>
</tr>
</tbody>
</table>
13 Fifth-grade student reading achievement scale scores, 2005-06 .................................................................144

14 Level-1 residuals in the one-way random effects ANCOVA HLM of fifth-grade reading achievement, 2005-06 .................................................................147

15 Two generations of Wallin, Brembert, Haglund & Holm (1993) citations .................................................................151

16 Percentage of fourth-grade students at or above reading proficiency, 2003-04 through 2005-06 .................................................................155
CHAPTER 1

INTRODUCTION

In the 2004-05 school year, the Canton City School District embarked on a plan to increase student achievement by giving teachers the tools to conduct formative and value-added student assessments in reading and mathematics. These initiatives were the major thrust of continuous improvement plans created at both the district and building levels. The purpose of this study is to report on the findings of the Canton City School District’s theory-driven evaluation of its initial implementation of these assessment systems. These findings were generated through the utilization of the pilot test approach to program evaluation (Chen, 2005), and focused on the eleven elementary school buildings identified by the Ohio Department of Education in school year 2004-05 as at risk or in school improvement. In that context, the evaluation plan called for the generation of several hierarchical linear models (HLM) to construct a framework for the analysis of summative student assessment data in the context of value-added assessment and generation of baseline student achievement data for comparative and predictive purposes. This study thus comprises an important element of the District’s school improvement effort. It also constitutes an empirical demonstration of the pilot testing approach within theory-driven evaluation, and through its use illustrates the utility of theory-driven evaluation in education program planning.
The theory-driven approach to program evaluation is compatible with educational applications but rarely used. Its holistic approach to assessment takes into account the complexity of the educational process, whether one views that process from the point of view of educational administration, instruction, learning or educational settings. Its use of the conceptual framework of program theory provides a lens through which to view the myriad of interactions that mediate education program development, implementation and outcomes. Its emphasis on the importance of both scientific and stakeholder credibility in a program evaluation reflects the applied nature of educational science. Finally, its view of evaluation as necessary to the improvement of programs is congruent with the perspectives of practitioners in the field. While theory-driven evaluation holds great promise for process and outcome program evaluations, its potential in program plan development may be even greater due to its ability to change the shape of the program on the basis of empirical evidence before or during program implementation.

In his taxonomy of program evaluation means and ends, Chen (2005) classifies the pilot testing approach as an evaluation that takes place in the program planning stage to identify problems that may occur during the implementation of a program (p. 48). As such, its primary objectives are to determine the field feasibility of a program and to identify the problems that may arise in its implementation. As a tool of development, the evaluation that emerges in a pilot test must be responsive to the concerns of stakeholders while maintaining scientific credibility.

In order to generate useful information for program improvement, Chen advances four principles that must be adhered to in such an evaluation: Implementation agents and clients must participate in small-scale program trials; the subjects involved in trials must
be typical of the persons whom the program was designed to assist; data gathering methods must be flexible; and results should be used for program development rather than evidence of program effectiveness (Chen, 2005, pp. 120-121).

This study will provide an empirical demonstration of Chen’s conception of pilot testing in program plan development in order to clarify and extend the theoretical framework of the approach; assist in the development of the program theory of a formative assessment system; advance a program evaluation framework for formative and summative assessments within the context of value-added assessment; develop a sample measurement instrument of teacher assessment; develop sample baseline measures of student achievement growth in reading; and test a method for linking formative to value-added and summative assessments within the structure of a hierarchical linear model. It will culminate in a field test of the initial implementation of the program theory by predicting fourth and fifth grade student reading achievement using two multilevel models, an unconstrained and a one-way ANCOVA with random effects model.

In these equations, the intraclass correlation coefficient (ICC), a measurement of the proportion of variance in the dependent variable that is accounted for by level-2 units (Luke, 2004), is predicted to be higher in the latter model. The models developed herein will be used as baseline data for the District to compare student academic achievement results in the 2006-07 school year using predicted and formative assessment scores as level-1 predictors. Generating baseline measures of the main program theory construct under study represents an extension of Chen’s pilot testing approach, one which has important implications for educational program planning.
Statement of the Problem

The programs in question are the elementary school reading and mathematics programs of the Canton City School District in Ohio. Historically, student achievement on summative achievement tests in many of the schools in the District has not met educational achievement benchmarks. Within the context of Ohio’s accountability system, the District and many of its buildings have not met state and federal educational standards. In this context, District administrators have created district and building improvement plans for implementation in the 2005-06 school year.

The pilot testing approach was chosen by the District to guide program development efforts towards the implementation of three educational assessment systems to improve student achievement. The first, a value-added assessment developed by Dr. William Sanders (SAS in Schools) and implemented in Ohio through Battelle for Kids, will be a required component of all Ohio school district assessment programs by the 2006-07 school year. For that reason, the District chose to implement the system one year early in order to identify ways in which the system could be utilized by principals and teachers in each building. The value-added assessment system utilizes summative assessment data in grades 3 – 8 to create value scores per grade per subject per building based upon each student’s previous academic achievement scores. Due to the implementation schedule of state-mandated achievement tests in Ohio, this system will become increasingly reliable each year as more data points become available for each student. This system will be implemented throughout the District in the 2004-05 school year. The second, ETS Pulliam’s formative assessment system, is a program that the District is implementing voluntarily in mathematics in grades K – 12 and in reading in
grades 3 - 12 as a core component of its District improvement plan. This system will be implemented in targeted classrooms in the 2004-05 school year, then implemented district-wide in the following school year. The third, the Dynamic Indicators of Basic Early Literacy Skills Sixth Edition (DIBELS) will be utilized as an assessment of reading progress in grades K – 3, and in grades 4 – 6 for Title 1 and special education students only. This system will also be implemented in targeted classrooms in the 2004-05 school year, then implemented district-wide in the following school year. All three assessments are being employed in the context of District and building continuous improvement plans to orient student achievement upwards to meet the established benchmarks of Ohio’s accountability system for public schools.

The current public education accountability system of Ohio as developed through Senate Bill 1 (2001), the 2001 reauthorization of the Elementary and Secondary Act of 1965 (2002), House Bill 3 (2003) and House Bill 493 (2004) is built upon four measures: state indicators, a performance index, a growth calculation and a measure of adequate yearly progress. These measures reflect the fact that the state has transitioned its statewide testing system towards screening, diagnostic and achievement tests.

The first measure, state indicators, is composed of state achievement and proficiency examinations, attendance and graduation rates. In school year 2004-05, Ohio public schools and districts were assessed on 23 performance indicators; 111 districts were rated Excellent, 297 were Effective, 175 were in Continuous Improvement, 21 were in Academic Watch, and 5 were in Academic Emergency. Thus, nearly 96 percent of Ohio districts earned Excellent, Effective or Continuous Improvement designations, though these ratings may be earned through indicators or through the performance index.
The second measure, the performance index, calculates the cumulative achievement of all tested students in all subjects tested by the state (reading (R), writing (W), mathematics (M), science (S) and social studies (SS)) in grades 3 – 8 and the Ohio Graduation Test (OGT). In school year 2004-05, achievement tests in reading and mathematics were operative in grades 3, 6 and 8. In school year 2005-06, reading and mathematics assessments will be conducted in grades 3-8. By school year 2007-08, science and social studies achievement tests will be added to grades 5 and 8. Student scores on these achievement tests are weighted by performance level: advanced scores yield 1.2 point, accelerated yields 1.1 points, proficient yields 1.0, basic yields 0.6, and a limited score yields 0.3 points. All points earned by a school or district are averaged and multiplied by 100 to generate a school and district index (Ohio Department of Education, 2005).

The third measure, a temporary growth measure, is computed through performance index score comparisons over time. For districts or buildings in academic emergency or watch, other provisions apply. If a school or district increases its performance index by 10 or more points over two years, and its index score increased in each of the previous two years, and the most recent score is at least 3 points higher than the previous year, the school or district can raise its rating up one level, but no higher than continuous improvement.

The fourth measure, adequate yearly progress (AYP), is a federal indicator. It computes the percentage of students and subgroups of students who score at the proficient level and above in reading and mathematics, the percentage of students who participate in that testing, and the percentage of students who attend classes and graduate
during the school year. AYP assessment goals are set at the district, building and subgroup level, while AYP attendance and graduation goals are set at the district and building level. AYP goals are established by the state in consultation with the federal government.

By federal law, AYP goals must be raised continuously from 2001 to 2014 so that all children become proficient in tested subjects in 2014. Buildings and districts can meet AYP requirements in one of three ways. In the first, they can meet or exceed the aforementioned AYP goals. In the second, they can meet or exceed those goals with a two-year average. In the third, they can qualify for a safe harbor if they receive a 10 percent reduction in the percent of non-proficient students over the previous year and meet attendance or graduation rate goals. For the purposes of these and related student achievement calculations, students must meet the Ohio Department of Education’s full academic year criteria by being present in a building by Friday of the first week of October (“count week”) and being continuously enrolled through March 19 (~100 days).

Beyond these goals, it should be noted that districts and buildings are responsible for testing at least 95 percent of their students who are enrolled at the time of testing; moreover, they must test at least 95 percent of students in nine subgroups: Native American, Hispanic, White, Limited English Proficient, African-American, Asian/Pacific Islander, Multi-Racial, Economically Disadvantaged, and Disabled. These nine subgroups and all students must meet the aforementioned AYP goals. The minimum number of students tested per group is 30 for eight subgroups and 45 in the disabled student subgroup.
Together, these four measures—state indicators, performance index, growth calculation and AYP—determine school and district ratings. In school year 2004-05, 55.5 percent of districts and 75.7 percent of buildings met AYP. Failure to meet AYP goals is accompanied by increasingly severe penalties under state and federal law. Under H.B. 66, a building or district can miss AYP and earn Excellent or Effective designation for up to two years—third year missing AYP for more than one student group, designation drops to Continuous Improvement. For students and teachers, the consequences of failing to meet building or district adequate yearly progress (AYP) goals can be devastating: staff can be fired, building governance replaced, and students released to other schools. Thus, administrators, teachers and students are well motivated to reach the high academic goals and standards of Ohio’s accountability system.

_Canton City School District’s Academic Status, 2004-05_

In the 2004-05 school year, the Canton City School District received an ‘Academic Watch’ rating from the Ohio Department of Education. Thus, in that year, the District was in District Improvement, Year 2 status. Its students had taken achievement tests in reading in grades 3-5, 8 and 10, and mathematics achievement tests in grades 3, 6-8 and 10. The District met 3 of the 23 state indicators, achieved a performance index score of 74.2 out of 120, and missed Adequate Yearly Progress for the third consecutive year. Given the consequences associated with such an academic performance, the District revised its district improvement plan in order to orient student achievement upwards towards state and federal education standards. Table 1 outlines the 2004-05 performance of Canton City School Districts’ schools.
<table>
<thead>
<tr>
<th>Building Name</th>
<th>Designation</th>
<th>2006 School Improvement Status</th>
<th>2004-05 Performance Index Score</th>
<th>2005 Overall AYP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen ES</td>
<td>Academic Watch</td>
<td>At Risk</td>
<td>71.1</td>
<td>Not Met</td>
</tr>
<tr>
<td>Barbara Schreiber ES</td>
<td>Academic Watch</td>
<td>At Risk</td>
<td>73.5</td>
<td>Not Met</td>
</tr>
<tr>
<td>Belden ES</td>
<td>Academic Emergency</td>
<td>At Risk</td>
<td>67.4</td>
<td>Not Met</td>
</tr>
<tr>
<td>Belle Stone ES</td>
<td>Academic Watch</td>
<td>At Risk</td>
<td>71.0</td>
<td>Not Met</td>
</tr>
<tr>
<td>Canton Digital Academy</td>
<td>Academic Emergency</td>
<td>At Risk</td>
<td>38.7</td>
<td>Not Met</td>
</tr>
<tr>
<td>Cedar ES</td>
<td>Continuous Improvement</td>
<td>OK</td>
<td>81.7</td>
<td>Met</td>
</tr>
<tr>
<td>Clarendon ES</td>
<td>Continuous Improvement</td>
<td>OK</td>
<td>81.4</td>
<td>Met</td>
</tr>
<tr>
<td>Compton ES</td>
<td>Academic Watch</td>
<td>At Risk</td>
<td>70.8</td>
<td>Not Met</td>
</tr>
<tr>
<td>Dueber ES</td>
<td>Academic Watch</td>
<td>At Risk</td>
<td>71.4</td>
<td>Not Met</td>
</tr>
<tr>
<td>Fairmount ES</td>
<td>Continuous Improvement</td>
<td>OK</td>
<td>72.8</td>
<td>Met</td>
</tr>
<tr>
<td>Gibbs ES</td>
<td>Academic Watch</td>
<td>Improvement Year 4</td>
<td>71.2</td>
<td>Not Met</td>
</tr>
<tr>
<td>Harter ES</td>
<td>Continuous Improvement</td>
<td>At Risk</td>
<td>88.6</td>
<td>Not Met</td>
</tr>
<tr>
<td>Mason ES</td>
<td>Continuous Improvement</td>
<td>OK</td>
<td>87.8</td>
<td>Met</td>
</tr>
<tr>
<td>McGregor ES</td>
<td>Academic Watch</td>
<td>At Risk</td>
<td>74.4</td>
<td>Not Met</td>
</tr>
<tr>
<td>Summit ES</td>
<td>Academic Emergency</td>
<td>At Risk</td>
<td>57.3</td>
<td>Not Met</td>
</tr>
<tr>
<td>Worley ES</td>
<td>Continuous Improvement</td>
<td>OK</td>
<td>82.7</td>
<td>Met</td>
</tr>
<tr>
<td>Youtz ES</td>
<td>Academic Watch</td>
<td>Improvement Year 1</td>
<td>71.4</td>
<td>Not Met</td>
</tr>
<tr>
<td>Portage Montessori MS</td>
<td>Effective</td>
<td>OK</td>
<td>94.9</td>
<td>Met</td>
</tr>
<tr>
<td>Crenshaw MS</td>
<td>Continuous Improvement</td>
<td>Improvement Year 1*</td>
<td>77.8</td>
<td>Met</td>
</tr>
<tr>
<td>Hartford MS</td>
<td>Academic Emergency</td>
<td>Improvement Year 3</td>
<td>68.4</td>
<td>Not Met</td>
</tr>
<tr>
<td>Lehman MS</td>
<td>Academic Watch</td>
<td>Improvement Year 4</td>
<td>75.0</td>
<td>Not Met</td>
</tr>
<tr>
<td>Souers MS</td>
<td>Academic Watch</td>
<td>Improvement Year 3</td>
<td>72.5</td>
<td>Not Met</td>
</tr>
<tr>
<td>McKinley HS</td>
<td>Continuous Improvement</td>
<td>Improvement Year 2</td>
<td>85.2</td>
<td>Not Met</td>
</tr>
<tr>
<td>Timken HS</td>
<td>Academic Watch</td>
<td>Improvement Year 2</td>
<td>75.8</td>
<td>Not Met</td>
</tr>
</tbody>
</table>

* Delay

Table 1: Canton City School District academic ratings by building, 2004-05
Of the 17 elementary schools in the District, 1 was in the fourth year of school improvement, 1 was in the first year of school improvement, 10 were at risk and 5 were O.K. Of the 4 middle schools, 1 was in the fourth year of school improvement, 2 were in the third year of school improvement, and 1 was in the first year of school improvement. Both of the high schools were in their second year of school improvement. Thus, the District worked with each building identified as at risk or in school improvement to increase their academic improvement. To facilitate that process, the District contracted with Learning Points Associates in the summer of 2005 to develop and implement an inquiry-based data retreat on academic assessment that would act as the foundation for the revision of school improvement plans. This comprehensive program was designed to allow school improvement teams to activate assessment knowledge in the context of school improvement. This program was part of a wider effort to create an assessment culture to facilitate discussion and decision-making based on data.

At the start of the 2005-06 school year, the District worked with Learning Points Associates to lead building-based teams through an inquiry process to create the foundation for building improvement plans. During the data retreats, each building analyzed summative student assessment data to create hypotheses or working theories about the possible causes of observed trends and patterns in the data. The teams worked under the assumption that these hypotheses or working theories could be rejected or accepted, but must be done so with additional data. Each team wrote one goal, posted the goal and allowed other data retreat participants to offer feedback. After the August 2005 data retreats, the goals of each team were edited and revised in consultation with the District’s senior leadership team. By the end of September 2005, each school had
submitted at least two goals to the Director of Testing, Evaluation and Research and received feedback in turn.

By October 2005, building improvement teams had begun to write continuous improvement plans following the format supplied by the District at the data retreat. These plans were informed by the grade- and building-level value-added assessment data available from Battelle for Kids. Each school improvement plan was constructed in such a way that each could be blended into one plan following the norm among other districts in Ohio. These plans were then implemented and monitored throughout the year using a variety of educational data gathered in each building. Schools in year 2, 3 and 4 of school improvement status were then visited by the Superintendent and Teachers’ Association President during the public presentation of their plan. Assistance from the District’s central office was offered to all buildings on the basis of building needs, especially in the buildings that were in the later years of school improvement status. Support for interim data analysis for building principals and teachers was also offered throughout the year.

Through all of these efforts- the process of developing the program theory underlying the deployment of formative assessment systems in reading and math, the initial deployment of the value-added assessment system data throughout each school building in the District, and the creation of school and district continuous improvement plans- the District seeks to develop its assessment program in order to increase the slope of the growth trajectory of student and school-level student achievement in reading and math in the context of Ohio’s accountability system. To accomplish this goal, the District adopted the pilot testing approach to program plan development in order to determine the field feasibility of the assessments systems and to identify the problems that may appear
with the full implementation of the formative assessment systems in the 2006-07 school year. For the District, this evaluation will assist in the development of the intervention in the context of the District and school improvement plans. In this context, the following seven research objectives of this study are advanced:

1) To conduct an empirical demonstration of the pilot testing approach to program plan development in the context of theory-driven evaluation in order to clarify and extend the theoretical framework of the approach;

2) To assist in the development of the program theory of a formative assessment system;

3) To provide a program evaluation framework for formative and summative assessments within the context of value-added assessment that assists the District in augmenting student achievement growth trajectories;

4) To develop a measurement instrument of teacher assessment for the District;

5) To develop baseline measures of student achievement growth in reading;

6) To test a method for linking summative and value-added assessments within the structure of a hierarchical linear model.

7) To propose a method for analyzing formative and summative assessments within the context of a value-added assessment system through the construction of several hierarchical linear models.

Assumptions Underlying the Study

In the process of conducting the study, two assumptions were made outside of the methodological assumptions arising under the evaluation and those required under the constraints of the statistical approach advocated herein; both sets of these other assumptions are detailed in chapters 3 and 4. The assumptions made a priori constitute important constraints on the ability of the evaluation to be useful to program stakeholders. To put this study into its proper context requires that these assumptions be identified succinctly. Thus, specifically, the following assumptions were made:
1) That the District’s full implementation of three separate assessment systems in the 2006-07 school year is feasible— that the ‘complex chains of reciprocal action’ that support the necessary components of program implementation are predictable, present and controllable in this setting (Pressman & Wildavsky, 1984).

2) That the data associated with all student achievement records (e.g., demographic, enrollment, assessment) are sufficiently free from error outside of the measurement error associated with psychometric assessments.

These assumptions will have to be re-visited by the District in its expected program evaluation of the full implementation of the three assessment systems in the 2006-07 school year. While this study will influence the first assumption, the set of relationships that control aspects of the second are outside the control of the evaluator. In raising these issues, the study acknowledges the role that tests of the veracity, reliability and validity of such data plays in the responsibilities of the District, test developers, test administrators and the Ohio Department of Education.

Specific Research Questions

Theory-driven evaluations involve crucial tests of program theory to validate their use. Program theories consist of action and change models. While the theoretical base of theory-driven evaluation (Chen, 1990; 2005) has not addressed this issue explicitly, data collection and analysis in such evaluations must be aligned to verify action and change models in order to test the program rationale. Moreover, within each model, the research questions proposed in the evaluation must be connected to data collection points. At each point, the method used to collect the data must be appropriate to the research question posed. In a program’s development phase, this congruency assumption would have evaluators create data collection instruments for critical components of the action model.
in the program rationale. Thus, this study will generate data collection procedures and instruments to answer a series of research questions in the program’s action model, the portion of the program theory upon which the pilot testing approach to evaluation focuses.

In alignment with the research objectives of the study, the principles and guidelines of the pilot test evaluation and the focus of the study, five research hypotheses were advanced in the articulation of baseline student achievement data generated through hierarchical linear models of the value-added and summative reading achievement assessment systems:

1) The relationship between predicted achievement scale scores generated by the state’s value-added assessment system and observed reading Ohio Achievement Test scale scores in the sample groups of students is positive, significantly strong and linear in the 2004-05 and 2005-06 school years.

2) The predicted achievement scale scores generated by the state’s value-added assessment system will correctly classify the sample group of students who score at or above proficient correctly no more than 80 percent of the time in the 2004-05 and 2005-06 school years.

3) When the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scores of the sample groups are placed in an unconditional hierarchical linear model, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderately low (i.e., between 15 to 19 percent) in the 2004-05 and 2005-06 school years.

4) When the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups are placed in an one-way random effects ANCOVA, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderate (i.e., between 20 to 24 percent) in the 2004-05 and 2005-06 school years.
5) When the percentage of sample group students scoring at or above proficiency in the state’s summative reading assessment in school years 2004-05 and 2005-06 are analyzed, a statistically significant gain among cohorts will not be observed.

Outside of these specific research questions, the program evaluation will also seek to catalogue information relative to the action theory concerns of the District by examining issues related to the implementing organization and implementers. This effort is related to the process of gathering baseline information on student achievement and the assessment practices and beliefs of teachers related. This latter effort will be done to extend the theoretical framework of the pilot testing approach by concentrating study resources on the conditions that exist before the intervention occurs. The need for this extension occurred in the design phase of the present study, and is rooted in the tradition of single-subject research in psychology (Kazdin, 1982; McCormick, 1995). The generation of such baseline data is of paramount importance for program stakeholders in the program planning phase. In doing so, these efforts will help the study achieve its goals: to assist in the program development of three assessment systems in the Canton City School District, and to advance the theory and practice of the pilot testing approach to theory-driven evaluation.

Significance of the Study

One consequence of this study is that program stakeholders will receive a plan through which to frame the issues that evolve out of District and school improvement plans. For program stakeholders, the stakes of the program and its development are high: without a plan to improve student achievement dramatically, the schools of the District
and the District itself face severe sanctions. A variety of other factors make this study significant to educators and evaluators outside of the District.

As an empirical test of the utility of the state’s mandated value-added assessment system, this study will shed light on the potential of the system to contribute to higher academic achievement in a sample group of students. Specifically, the early implementation of the state’s value-added assessment system will allow the District to test whether or not the implementation of the system leads to statistically significant increases in the percentage of students at or above proficient levels in the schools targeted for intervention by the District. This in turn will influence how the District may need to use additional assessment systems to augment student performance. The assumption of the stakeholders is that the implementation of the value-added assessment system alone will not cause a statistically significant increase in student achievement.

As an empirical demonstration of the pilot testing approach to program development, it represents a contribution to evaluation theory and practice. By articulating the congruency assumption in theory-driven evaluations, it enhances the practice of evaluation by creating opportunities for synchronicity between theory, data collection, and data analysis methods, all of which help to establish stakeholder and scientific credibility, a key goal of theory-driven evaluation (Chen, 2005). By proposing a theory of actionable data and specifying its consequences for the placement of data collection nodes at critical points in the assumption chains of action and change models, it contributes to the continuing dialogue of theory-driven evaluation improvements. Finally, by surfacing the statistical relationships that exist between summative and value-added assessments systems, it provides the District with an elegant measurement
framework through which to view the academic growth of students within classrooms within schools.

*Delimitations*

In any program evaluation, the delimitations of a study must be set by program stakeholders and the evaluator. In the present study, the stakeholders established the population, sample, and selection criteria. Given the objectives of the evaluation and the stage of program development, the evaluator worked with the District Director of Testing, Evaluation and Research to select an evaluation approach, while the evaluator chose the evaluation strategy. In that context, the statistical approach advocated herein was also under the control of the evaluator.

*Definitions and Operational Terms*

*Theory-driven evaluation*

A theory-driven evaluation is a holistic assessment of a social program or intervention based upon the conceptual framework of program theory that is oriented towards scientific and stakeholder credibility and provides a foundation for program improvement (Chen, 1990; 2004; 2005). Change and action models dominate the landscape of program theory. According to Chen, program theory is “a specification of what must be done to achieve the desirable goals, what other important impacts may be anticipated, and how these goals and impacts would be generated” (Chen, 1990b, p. 43). Historically, program theory has been assembled from four areas: the literature base of empirical theory, the implicit ideas that program stakeholders maintain, program observations, and research (Donaldson, 2001). Thus, to elaborate program theory, one must surface the implicit and explicit assumptions of program stakeholders,
simultaneously describing and prescribing program design, implementation, operations and outcomes (Chen, 2005). Once completed, it is memorialized in a program rationale and subjected to scrutiny.

When a program is in the developmental stages of growth, evaluators can facilitate the development of program planning by clarifying stakeholders’ ideas in a systematic program theory. This is accomplished by specifying the goals of the program, specifying the change model, specifying the action model, fine-tuning and diagramming the program theory (Chen, 2003). Thus, once the goals of the program are identified, the role of the evaluator in the development of a program is to surface the program theory of the intervention or program.

In the theory-driven approach to evaluation, ‘theory’ refers to “a frame of reference that help humans to understand their world…” providing “not only guidelines for analyzing a phenomenon but also a scheme for understanding the significance of research findings” (Chen, 1990, p. 17). By incorporating stakeholder and social science theories into program evaluations, the theory-driven approach stands as an alternative to methods-based evaluation approaches. By incorporating the action and change theories of the program into the evaluation, theory-driven approaches allow evaluators to identify the locus of program failure or success in concrete terms.

By incorporating the theory-driven approach into the program planning process, evaluators can help provide a framework to guide the development of a plan, enumerate the components of the program that will contribute to its success, facilitate communication about the program, and provide a foundation for the evaluation of the program (Chen, 2003). By connecting theory, data collection and analysis with the action
and change models of program theory, evaluators can help program stakeholders connect program theory to program evaluation. This offers a method through which theory and experimentation can check one another, an idea which has been at the forefront of scientific advocacy for some time (Platt, 1964).

**Change model**

Descriptively, the assumptions in a program’s change model indicate the causes behind the behaviors, condition or situation that the program was designed to correct. The causal mechanisms identified could be the result of belief, experience, empirical data or theory. As the basis of the program design, their validity pre-determines the effectiveness of the program. When identified as a set, these causal assumptions form the change model of the program. The change model informs the program rationale.

According to Chen (2005), the change model is comprised of goals and outcomes, determinants and interventions. Briefly, goals are the general aims of the program, while outcomes are their measurable reflection. In Chen’s conception, the determinants are the factors responsible for causing the problem that the program was designed to remedy. While programs may employ multiple determinants, the prioritization therein is essential to the success of the change model. Interventions are the set of efforts and actions undertaken by the program to change a determinant.

**Action model**

Prescriptively, the set of assumptions regarding program components and operations required to activate the change model are referred to as the program’s action model. These assumptions inform day-to-day program processes, and, to the extent that they are valid, contribute to program effectiveness. The action model in Chen’s
framework is composed of intervention and service delivery protocols, implementing organizations, program implementers, associate organizations/community partners, the ecological context of the program and the target population. These elements form the implementation plan of the program, providing the basis of the program plan.

Assessment

Assessment is a broad term meant to encompass “any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects or programs” (AERA, APA, & NCME, 1999, p. 172).

Test

A test is a “device or procedure in which a sample of an individual’s behavior is obtained, evaluated and scored using standardized procedures” (AERA, APA, & NCME, 1999, p. 183).

Measurement

Measurement, strictly speaking, encompasses a process whereby rules are created to assign numbers to represent attributes, traits or behaviors (Reynolds, Livingston & Willson, 2006). Measurement differentiates and describes (Hopkins, 1998).
CHAPTER 2
REVIEW OF THE LITERATURE

To identify and develop the ideas germane to the study, it was necessary to identify, review and analyze a literature base from educational evaluation and assessment. Within the field of educational evaluation, it is necessary to trace the development of theory-driven evaluation from earlier forms of evaluation in order to identify its potential benefits over alternative approaches. Within the field of educational assessment, the major assessment approaches—summative, value-added and formative—were explored in order to identify the relative contribution of each in education and to identify how teachers might use data from each to inform instruction. Through these investigations, the work of theorists and practitioners in both fields will be examined in order to identify the study’s potential contribution to the body of knowledge in theory-driven educational evaluations of academic achievement and carefully limit the problem under investigation in the study.

*Elementary and Secondary Act of 1965*

The origin of modern program evaluation theory and practice in education can arguably be traced back to the Elementary and Secondary Act of 1965 (Worthen, Sanders & Fitzpatrick, 1997). The Act constituted an extension of federal aid to education that was unparalleled in its purpose and reach. With its passage, the nature of educational federalism changed irreversibly from a unilateral to a bilateral relationship (Bailey &
Mosher, 1968). It was built upon five foundations: Title I grants for compensatory education of culturally and economically deprived children; Title II for the improvement of educational resources; Title III for supplementary educational centers and services, Title IV for regional educational laboratories; and Title V to assist state educational agencies (Public Law 89-10). Five-sixths of the Act’s funds flowed through Title I programs.

Operationally, Title I funds were dispensed by the federal government to state educational agencies (SEA) and then to local educational agencies (LEA) on the basis of competitive three-year grants. The overall allocation of funds was set by the federal government, while SEAs set grant parameters for LEAs in consultation with federal guidelines. In return for this limited discretion, SEAs and LEAs were responsible for evaluating the educational effectiveness of efforts funded under the Act. Those programs which could not demonstrate effectiveness would no longer have funding. This tension was reflected in the evaluation imperatives in Titles I and III:

Title I, Section 205 (a) (5) “that effective procedures, including provision for appropriate objective measurements of educational achievement, will be adopted for evaluating at least annually the effectiveness of the programs in meeting the special educational needs of educationally deprived children.”

Title III, Section 304 (a) (6) “provide for making an annual report and such other reports, in such form and containing such information, as the Commissioner may reasonably require to carry out his functions under this title and to determine the extent to which funds provided under this title have been effective in improving the educational opportunities of persons in the area served, and for keeping such records and for affording such access thereto as the Commissioner may find necessary to assure the correctness and verification of such reports.”

This language reflects the tension between two of the Act’s evaluation advocates, Senator Robert Kennedy and the Department of Health, Education and Welfare’s
William Gorham (McLaughlin, 1975). Kennedy viewed the evaluation language as a tool to ensure political accountability; he wanted to ensure that federal funds created local opportunities for poor schoolchildren in each classroom of an LEA. For Gorham, educational achievement data was required to conduct analyses pursuant to the creation of production functions for education.

Senator Kennedy’s views emanated from concerns of minority constituents who viewed the failure of public education in the poor quality of schools attended by poor children, and in the lack of concern displayed by school officials for the students, their parents, and their opinions regarding education. As the junior Senator from New York, Kennedy’s efforts were guided by his desire to give these constituents a voice. For Kennedy, the object of Sections 205 and 304 was to trade greater accountability for additional resources. Kennedy’s notion of accountability was rooted in outcomes reporting, the production of which was designed for communication purposes at three levels: parents of students in poor school districts, the local communities those districts served, and the government agencies tasked with responsibility for education. He and United States Office of Education (USOE) Commissioner Keppel seemed to believe that a public review of educational outcome data would pressure schools to change in alignment with community expectations, and cause stakeholders throughout the education system to realize the great variation in educational needs and practices across the nation.

Gorham, the Assistant Secretary for Program Coordination and later Program Evaluation at DHEW, was responsible in part for implementing the Planning Programming Budgeting System (PPBS) in order to bring rational decision-making to political programs. In this sense, the vague language of Sections 205 and 304 represented
the convergence of the accountability and efficiency purposes of evaluation. The data that resulted from ESEA 1965 was expected to drive cost-benefit analyses that would identify promising educational programs at different investment levels. Thus, while Kennedy’s attention was fixed on the analysis of programs themselves, Gorham’s interest was in the analysis of strategies employed to pursue educational goals.

With hindsight, it is plain that the field of educational evaluation was unprepared for the mandate represented in ESEA 1965. The field had experience with small-scale evaluation projects and research such as that conducted by Ralph W. Tyler (Madaus & Stufflebeam, 1989), but did not have the capacity to undertake large-scale projects. ESEA 1965 represented a change in the educational evaluation demand curve that could not be met with the resources it could bring to bear at the time of the legislation’s passage. In the rush to compliance, local education agency administrators and teachers were pressed into service unprepared for their task (Worthen, Sanders & Fitzpatrick, 1997). Placed into unfamiliar roles, they responded with number-gathering or hypothesis-testing exercises that failed miserably (Popham, 1975). In response to a growing recognition of these and related problems, social scientists from related fields began the work of responding to the challenges posed by ESEA 1965.

Nine months after the passage of ESEA 1965, the Director of the Evaluation Center at the Ohio State University addressed a national evaluation conference to identify the evaluation challenges posed by the Act (Stufflebeam, 1966). In his address, Stufflebeam voiced the opinion that the purpose of evaluation in the Act was to assist local, state and federal administration in making decisions. To accomplish this purpose, he advocated process evaluations for local decision-making efforts, and outcome
evaluations for state and federal officials. He found three challenges to meeting the Act’s objectives: inadequate evaluation experience in education, inadequate numbers of trained evaluators, and a lack of salient evaluation designs. Five years later, in his work on the Phi Delta Kappan Commission on Evaluation, Stufflebeam expanded the list to include the inadequacy of evaluation theory, the inadequate specification of evaluation instruments and the inadequacy of systems with which to operate evaluation programs (Stufflebeam, et al., 1971, as cited in Worthen & Sanders, 1973).

One of the first papers to respond to these challenges and other suggested solutions was by Robert E. Stake, a psychometrician who worked at the Center for Instructional Research and Curriculum Evaluation at the University of Illinois. In making his response, he directly addressed the concerns voiced in two prominent papers in the field written by Lee J. Cronbach (1963) and Michael Scriven (1967).

*Evaluation Plan Development*

Stake’s seminal paper “The Countenance of Educational Evaluation” represents one of the first comprehensive attempts to design an evaluation plan for education stakeholders that could meet the new federal mandates (Stake, 1967). In the paper, Stake presents his view of evaluation as multi-faceted and contingent upon the program’s stage of development. He observes that the school environment faced by evaluators is complex and composed of factors which the federal mandate did not anticipate. The two factors mentioned most frequently by Stake includes the antecedent conditions of teachers and students, and the dynamic classroom transactions that influenced the processes of teaching and learning. He indicates that Michael Scriven’s view of evaluation as judgment (Scriven, 1967) does not take these two factors into account, and builds on Lee
Cronbach’s description of student achievement (Cronbach, 1963) to articulate a model of evaluation capable of responding to the complex interplay of factors within students, within classrooms and within schools.

To begin, Stake argues that evaluators must first be able to identify the rationale of the educational program: the background philosophy and purpose for which it was created. Stake’s model includes a plan for data collection within two broad matrices: descriptions built from intentions, and observations and judgments built from standards and judgments. The two matrices are connected by antecedent conditions, transactions and outcomes. Stake argues that evaluation efforts should detect the congruence between intentions and observations, logical contingencies between intended antecedents, transactions and outcomes and empirical contingencies between observed antecedents, transactions and outcomes. He makes a clear distinction between the evaluations that emphasize the relative merit of programs through comparisons with similar programs and comparisons between a program and a fixed standard. Stake makes a further contribution to the development of evaluation by pointing out that different types of evaluation questions should guide the evaluator’s use of design features within the model’s framework. The matrices which are developed in the process allow for a clear analysis of the relative contributions of description and judgment at each phase of the evaluation (Popham, 1975). This participant-oriented evaluation approach (Worthen, Sanders & Fitzgerald, 1997), like Stake’s concentration on program rationale, established an emphasis within educational evaluation on the importance of stakeholders and their belief systems.
**Decisions**

In 1971, Daniel Stufflebeam and his colleagues presented an evaluation approach that oriented attention towards four types of educational decisions often made by administrators: planning, structuring, implementing and recycling (Stufflebeam et al., 1971). In doing so, he categorized evaluation objectives and methods around context, input, process and product evaluations. To conduct each type of evaluation, Stufflebeam recommended six steps: focusing the evaluation, collecting information, organizing information, analyzing information, reporting and administration of the evaluation (Worthen, Sanders & Fitzgerald, 1997). By clarifying the various roles that evaluation efforts could play in the administration of educational programs, Stufflebeam broadened the definition of evaluation beyond the strict boundaries alluded to in ESEA 1965, drawing attention to the vital role evaluation could play in educational program improvement. By responding directly to the needs of education administrators, Stufflebeam ensured that his model would receive widespread use in the field and reflection in the literature. Both outlets were to shape the further development of evaluation.

**Politics**

More than any other evaluation theorist, Carol Weiss’ increasingly sophisticated recommendations for experimental designs in educational evaluation (1972; 1987) reflect the impact that field work with evaluation had on her views. Weiss began her work by advocating experimental and quasi-experimental designs in educational evaluation in the context of political settings; these settings included differing interpretations of program purposes, changes in the target population and community support, and changing
allegiances of program stakeholders. Throughout the years, other theorists have supported Weiss’ advocacy for experimental designs (Fitz-Gibbon, 1978; Boruch, 1997). She would eventually observe that evaluation findings were most often utilized when the changes they advocated were marginal, or when the changes they advocated were aligned with pre-existing organizational intent (Weiss, 1997). Eventually, frustrated with the lack of direct use of evaluation results by organizations, she sought out new roles for evaluation within policy development, implementation and evaluation (Shadish, Cook & Leviton, 1991).

The lasting contributions to program evaluation theory and practice in education that Weiss makes concerns her observations of the political and social contexts of evaluation, and her emphasis on tailoring evaluation designs to the program milieu. Her concentration on the utilization of evaluation results supports her view of evaluation as a tool for improving social programs through the measurement of program results relative to program goals. By combining evaluation utilization to its contexts, she demonstrated the way in which stakeholder needs and the need for program results could both be met in an evaluation.

*Multidimensional Measurement*

The importance of multiple measurements across a variety of cognitive, affective, organizational and social domains was articulated forcefully by one of education’s premier measurement theorists, Lee J. Cronbach (Cronbach et al., 1980, Cronbach, 1982). One of his earliest papers on the subject emphasized the decisions that evaluation could illuminate: course improvement, decisions about individuals, and administrative regulation (Cronbach, 1963). Cronbach felt that evaluation was best utilized when it
sought to discover how educational courses produced academic gains and what factors influenced their effectiveness. Even though Cronbach’s chief investigative aim was causation, the range of quantitative and qualitative methods he advocated in the employment of evaluation services was wide, including follow-up studies to determine the long-term impact of educational programs (Worthen & Sanders, 1973). Essentially, Cronbach’s emphasis on evaluation in the context of program development pre-dates Scriven’s articulation of the difference between formative and summative evaluation.

In his 1982 book on program evaluation, Cronbach helped to bring specificity to the elements of evaluation design that helped bring clarity to the importance of delineating units of analysis. He properly highlighted the fact that a few stakeholders could dominate evaluation concerns, and articulated flexible evaluation designs to account for a multiplicity of views regarding program purpose. His attention to social, organizational and political factors in the evaluation experience underscores Weiss’ contributions to evaluation theory; his accommodation of these factors takes her observations into account when designing empirical evaluation designs.

Standards

In 1994, the Joint Committee on Standards for Educational Evaluation published its set of standards for the evaluation profession. This effort came among similar efforts by other evaluation organizations, including the Evaluation Research Society and American Evaluation Society. The Joint Committee Standards are arguably the most well-known in the field (Worthen, Sanders & Fitzgerald, 1997). The idea behind the Standards was to embody the characteristics of what evaluations should contain, how
they should be conducted, and how their quality should be determined. Four conceptual
bases form the foundation of the Standards: utility, feasibility, propriety and accuracy.

The utility standards identify the ways in which the needs of stakeholders will be
met in terms of information, the basis for judgment, reporting and utilization. The
feasibility standards identify the extent to which evaluations are practical, politically
viable and cost effective. Propriety standards refer to the legal and ethical conduct of
program evaluations, while the accuracy standards identify the extent to which
evaluations consider the technical merits of plans, operations and conclusions (Joint
Committee, 1994).

The Standards are an important component of the professionalization of program
evaluation, as are its professional associations (American Educational Research
Association Division H, American Evaluation Association, Canadian Evaluation Society,
European Evaluation Society, Australasian Evaluation Society, and Latin American
Evaluation Society) and professional journals (Evaluation and Program Planning,
Evaluation Practice, Evaluation Review, Educational Evaluation and Policy Analysis, and
New Directions for Program Evaluation). While the professional associations and
journals allow practitioners to learn from others and hone their craft and science, the
Standards provide a guide to the quality of practice throughout the field.

Eclectic Flexibility

As a sociologist working in a number of applied social science positions, Peter
Rossi brought concepts and methods from economics, political science, social studies and
public administration to bear on a number of evaluation engagements (Rossi, 2004). This
eclecticism was further enhanced through seven editions of a seminal text book in the
field noted for its egalitarian treatment of theories and methods in evaluation (Rossi, Lipsey & Freeman, 2004).

In this theoretical work, Rossi expounded a view of evaluation methods selection known as the “good-enough” rule: in light of program evaluation constraints with respect to funding, staffing and time, the evaluator should choose the best design for the situation that presents itself (Rossi, Lipsey & Freeman, 2004, p. 238). In other words, evaluators should choose the design that is most appropriate to the parameters of the project, understanding that all choices will force trade-offs.

For Rossi, evaluations are oriented towards administrative decision-making. Like Stake before him, Rossi believes that these decisions are best predicated on the stage of development that a program occupies at the time. In this respect, administrative decisions include those that create, sustain, modify or end social programs (Rossi, 2004). Each evaluation must be carefully tailored to the needs of the evaluation sponsor and key stakeholders. Rossi identifies three aspects of an evaluation design that must be so tailored: the questions the evaluation must answer, the methods and procedures proposed to answer the questions, and the nature of the relationships between the evaluator and key stakeholders. Once these aspects of the evaluation have been addressed, three components of the evaluation plan must be elicited: the purpose of the evaluation, the structure of the program and its relative stage of development, and the resources that are available for accomplishing program evaluation objectives.

The last step in the process of tailoring the program evaluation is to determine the activities that will constitute evaluation operations. Rossi lists five such activities: needs assessment, program theory, implementation, impact and efficiency. By identifying the
steps necessary to tailoring the evaluation to the needs of the evaluation sponsor and the key stakeholders, Rossi demonstrated an effective method through which to increase the utilization of evaluation results.

*Education Sciences Reform Act of 2002*

While the development of modern program evaluation was catapulted forward with the passage of the Elementary and Secondary Act of 1965, its progress was redirected with the passage of the Education Sciences Reform Act of 2002 (ESRA). The ESRA was passed by Congress in order to “provide for improvement of Federal education research, statistics, evaluation, information, dissemination and for other purposes” (P.L. 107-279). The law created an Institute of Education Sciences out of the research and statistics functions of the USDOE, and reoriented the staff and staff efforts through a National Center for Education Research, National Center for Education Statistics, National Center for Education Evaluation and Regional Assistance, and more than twenty comprehensive centers designed to offer educational technical assistance to local education agencies. In the process, the Institute of Education Sciences (IES) has become the research arm of the USDOE. One goal that has been articulated for the IES is for that agency to help make education an evidence-based field. Each Center is tasked with work plans in alignment with that goal.

For the National Center for Education Research, national research and development centers funds research in line with the IES goal of finding out ‘what works’ in instructional science. This funding amounts to over $100 million per year in data collection and evaluation efforts alone. The National Center for Education Statistics continues to collect, analyze and report on studies of the condition and progress of
education which include the National Assessment of Educational Progress (NAEP). The National Center for Education Evaluation and Regional Assistance is responsible for coordinating evaluations of federal programs, sponsoring and reviewing evaluations which concern education. Lastly, new comprehensive centers provide research dissemination services, training, professional development and technical assistance through existing regional educational laboratories and the National Center for Education Evaluation and Regional Assistance. The work of IES and its Centers is guided by the Elementary and Secondary Act of 2001 (P.L. 107-110) and Education Sciences Reform Act of 2002 (P.L. 107-279). Three definitions within the latter control the definition of research standards, program evaluation and research for all USDOE programs, projects, contracts and grants: scientifically based research standards, scientifically valid education evaluation, and scientifically valid research.

*Scientifically Based Research Standards*

“Title 1, Section 102 (18)

(A) The term “scientifically based research standards” means research standards that-

(i) apply rigorous, systematic, and objective methodology to obtain reliable and valid knowledge relevant to education activities and programs; and

(ii) present findings and make claims that are appropriate to and supported by the methods that have been employed.

(B) The term includes, appropriate to the research being conducted-

(i) employing systematic, empirical methods that draw on observation or experiment;

(ii) involving data analyses that are adequate to support the general findings;

(iii) relying on measurements or observational methods that provide reliable data;

(iv) making claims of causal relationships only in random assignment experiments or other designs (to the extent such designs substantially eliminate plausible competing explanations for the obtained results);
(v) ensuring that studies and methods are presented in sufficient detail and clarity to allow for replication or, at a minimum, to offer the opportunity to build systematically on the findings of the research;
(vi) obtaining acceptance by a peer-reviewed journal or approval by a panel of independent experts through a comparably rigorous, objective, and scientific review; and
(vii) using research designs and methods appropriate to the research question posed.”

**Scientifically Valid Education Evaluation**

“Title 1, Section 102 (19)
The term “scientifically valid education evaluation” means an evaluation that-
(A) adheres to the highest possible standards of quality with respect to research design and statistical analysis;
(B) provides an adequate description of the programs evaluated and, to the extent possible, examines the relationship between program implementation and program impacts;
(C) provides an analysis of the results achieved by the program with respect to its projected effects;
(D) employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible; and
(E) may study program implementation through a combination of scientifically valid and reliable methods.”

**Scientifically Valid Research**

“Title 1, Section 102 (20)
The term “scientifically valid research” includes applied research, basic research, and field-initiated research in which the rationale, design, and interpretation are soundly developed in accordance with scientifically based research standards.”

These definitions of research and evaluation are based on the definition of scientifically based research found in the Elementary and Secondary Act of 2001, and were meant to bring the practice of education research and evaluation in line with the aforementioned IES goal. Thus, in one act of Congress, the definition of what constitutes education research standards, program evaluation and research for the purposes of federal funding changed entirely. For education program evaluation theorists, the new definitions
may cause a shift towards methods that are aligned with federal policy, and away from
the unique program parameters specified by program stakeholders. A recent notice in the
Federal Register quickened the pace of change.

*U.S.D.O.E. Notice of Final Priority*

On November 4, 2003, the Department of Education posted a notice of proposed
priority regarding scientifically based evaluation methods in Federal Register 68 (213). In
the notice, the Secretary of Education proposed a priority for funding evaluations that
was based on the following methods (ordered by priority); experimental designs featuring
random assignment; quasi-experimental designs with matched comparison conditions;
regression discontinuity designs; and single-subject designs. The language of the notice
makes it clear that the Department will use this priority to fund impact evaluations only.
That intent was crystallized in the Department’s notice of final priority on January 25,
2005, in Federal Register 70 (15): In spite of receiving comments from almost 300
parties, the Department changed nothing substantially, adding only minor technical
corrections and a section detailing definitions. In essence, ESRA and the Final Notice
indicate that federal funding for educational program evaluations in the future will be
limited to impact evaluations that utilize research designs that have the potential to
demonstrate causality.

The effects that these legislative and regulatory mandates will have on evaluation
theory and practice in education are not yet apparent. While the mandates impact federal
funding specifically, the efforts of the IES to disseminate the reasons behind the policy
shift is causing state and local education agencies to consider program evaluation theories
that can accommodate the federal mandates while meeting the needs of local decision
makers for programs at every stage of program development. Recent publications by the
National Research Council regarding the encouragement of systematic research
initiatives between schools, state education agencies and universities underscores the
importance of identifying a comprehensive framework of research and evaluation that
can meet the needs of all education evaluation stakeholders (Donovan, Wigdor & Snow,
2003; Donovan & Pellegrino, 2004). One system that may be able to accommodate the
demands of program stakeholders in their development, implementation and evaluation
of programs is Huey-Tsyh Chen’s conceptual framework of program theory and practical
evaluation taxonomy (Chen, 1996; 2005).

Conceptual Framework of Program Theory

Chen’s framework is predicated on the assumption that effective program
evaluations must embody three characteristics: future directedness, scientific and
stakeholder credibility, and a holistic approach (Chen, 2005). According to Chen,
education program stakeholders engage in evaluation efforts in order to improve the
quality of programs. Evaluation is used as a tool to augment the efforts of program staff
to reach program goals. Their belief that the design employed by the evaluator must meet
their needs is just as important as the scientific merit of the design in Chen’s framework.
Since the program operates in a social system, its processes and effects must be viewed in
the context in which it operates. For Chen, one way in which to ascertain the presence of
these three characteristics is through the framework of program theory.

According to Chen (1990), program theory is “a specification of what must be
done to achieve the desirable goals, what other important impacts may be anticipated, and
how these goals and impacts would be generated” (p. 43). Program theory requires the
surfacing of the implicit and explicit assumptions of program stakeholders, simultaneously describing and prescribing program design, implementation, operations and outcomes (Chen, 2005).

Descriptively, these assumptions indicate the causes behind the behavior, condition or situation the program was designed to correct. The causal mechanisms identified might be the result of belief, experience, empirical data or theory, but they must be articulated. As the basis of the program design, their validity pre-determines the effectiveness of the program. When identified as a set, these causal assumptions form the change model of the program.

Prescriptively, the set of assumptions regarding program components and operations required to activate the change model are referred to as the program’s action model. These assumptions inform the day-to-day processes of the program, and, to the extent that they are valid, contribute to program effectiveness. Working together, the program’s change and action models form the program theory. By connecting the validity of the change and action model assumptions to the effectiveness of program impacts and processes, Chen extends earlier efforts to differentiate program failure from theory failure (Suchman, 1969).

To clarify how the change and action models work within the program theory framework, the components of each model need to be extracted and examined in a program rationale. According to Chen (2005), the change model is comprised of goals, outcomes, determinants and interventions. Briefly, goals are the general aims of the program, while outcomes are their measurable reflection. Determinants are represented in the factors that are responsible for causing the problem the program was designed to
correct. While programs may employ multiple determinants, the prioritization therein is essential to the success of the change model. Interventions are the set of efforts and actions undertaken by the program to change a determinant.

In a similar manner, the action model in Chen’s framework is composed of intervention and service delivery protocols, implementing organizations, program implementers, associate organizations/community partners, the ecological context and the target population. These elements form the plan to implement program processes.

Briefly, the intervention protocol specifies the nature, content and course of action of the intervention. The service delivery protocol identifies the specific steps required to implement the intervention in the field. Following this delineation, the capacity of the organization tasked with the responsibility of implementing the program must be analyzed in order to determine its ability to deliver the intervention effectively. In much the same way, the staff tasked with the responsibility for implementation must be assessed for their ability to do the job. If either the organization or the implementers are not equipped to deliver the intervention, a plan for capacity-building must be identified and employed. Chen advocates that program designers also assess the micro- and macro-level contextual support necessary to program success: the former concerns the support required for subjects to remain in the treatment program, the latter refers to the support the program requires from the economic, legislative, and community environments.

Finally, three assumptions regarding the target population must be analyzed for their validity. In order to deliver the scarce resources available to the program effectively to the target population, they must be identified through operationally-defined eligibility criteria. Once identified, they must be screened to assess whether or not they are actually
ready to receive service- if they can accept the intervention, and if they are ready to receive the intervention. Once these assumptions have been tested, the program plan should begin to identify the salient relationships between the change and action models in text and graphic forms. This will help program developers understand the sequential order of essential operations and the relationships between model components. While the text can explain the validity of the logic behind the action and change models, a graphic display of the program rationale can be used to demonstrate the interdependencies of the models that will, once in operation, indicate the probability of program success. Once the program theory has been thus explicated internally to staff and key stakeholders and externally to the funding agency, the evaluator must choose the appropriate evaluation approach.

Chen advocates a contingency view to evaluation approach selection as opposed to a universalist view. While the universalist advocates one approach or method for every program and in every situation, the contingency theorist does not assume the superiority of one approach or method over another a priori. Instead, the contingency theorist uses professional judgment to determine the best evaluation approach and method to fit the unique characteristics of the program, the stage of program development, the purpose of the evaluation, and the context in which it will play out. It is from this perspective that Chen developed his ideas regarding program theory (Chen, 1990; Chen & Rossi, 1992), his ideas behind the conceptual framework of program theory (Chen, 2004), and his practical evaluation taxonomy (Chen, 1996; 2005).
Practical Evaluation Taxonomy

Chen’s development of a practical evaluation taxonomy offers a framework through which evaluators can classify phenomena and choose appropriate responses that have antecedents in modern program evaluation history (Scriven, 1967; Stufflebeam, 1971). For Chen, evaluators enter the taxonomy by clarifying the development stage of the program in recognition of the fact that stakeholder needs vary by stage (Chen, 2005). While the stages are arranged sequentially, Chen recognizes the nonlinearity of program development; there is no fixed sequence to program growth and evaluation. With this caveat in mind, the stages, evaluation strategies and evaluation approaches in the taxonomy will be reviewed in order to explicate the dynamic nature of evaluation approach selection in the Chen framework.

In the planning stage, evaluators work with program developers to identify the ways in which evaluation can assist developers in the identification of program theory in the planning and development of the program. This involves the clarification of the change and action models of the program. As stakeholders begin to put the program into effect in the initial implementation, their primary need becomes one of feedback on the implementation of the program. Evaluators in this stage best serve stakeholders by designing sound feedback systems that help stakeholders identify whether or not their action models are operating as designed. As the program implementation matures, additional feedback loops for funding agency review are activated, and stakeholders begin to require assistance with the data infrastructure of the program’s action and change model requirements. If the program is ready for outcome assessment, evaluators may assist in program improvement, outcome monitoring, or outcome assessment. As
stakeholder needs change across program development stages, the strategies and approaches chosen by the evaluator must change to meet program stakeholder needs.

In Chen’s taxonomy, evaluation strategies are differentiated from evaluation approaches: the former concerns the general direction of the evaluation, the latter describes the procedures and philosophy that will guide the evaluation. Method selection and data interpretation follow the approach chosen for the evaluation. Chen advocates that evaluators first clarify the strategy the stakeholders want to take and then review approach options. This procedure mirrors his observation that stakeholder credibility precedes scientific credibility (Chen, 1990). Chen identifies four categories of evaluation strategies: merit assessment, development, enlightenment and partnership (Chen, 2005).

Merit assessment strategies determine the merit of programs by examining their processes or outcomes. Two types of strategies comprise this category of strategies: performance assessment and performance monitoring strategies. Performance assessment strategies identify the merit of a program through their implementation or outcomes using robust, evidence-based research designs. Common approaches used within this strategy are fidelity and outcome evaluations. Performance monitoring strategies use indicators to track program progress across implementation and outcome domains throughout a period of time. Common approaches used with such strategies are process and outcome monitoring approaches.

When programs are in the planning or development phase, development strategies help stakeholders to gather formative data quickly in order to improve program prospects. These types of strategies fall into three categories: background information provision, troubleshooting and development facilitation. Background information provision
strategies seek information relative to community needs, target population characteristics and intervention options; typical evaluation approaches used herein include needs assessment and formative research. Troubleshooting strategies help to identify weak components of program theory and generate options to overcome them; formative evaluation, program reviews/development meetings, relevancy testing (change model) and pilot testing (action model) approaches are often utilized within this category of strategies. Development facilitation strategies actually enlist evaluators to work as facilitators, identifying potential problems in program theory as the program is being developed; evaluation approaches that have been used within this category in the past include conceptualization facilitation and concept-mapping approaches.

The enlightenment strategy goes beyond merit strategies by seeking to identify why, how and when programs work as well as if they work or not. This involves the explication of the program’s change and action models. The theory-driven approach is used most often with this strategy.

The last strategy that Chen includes in his taxonomy is the partnership strategy. This is used most often when evaluators are asked to sit on the development team planning and implementing programs. One approach that has been used with this strategy in the past is the bilateral empowerment approach.

In order to facilitate a contingency view of program evaluation, Chen provides a framework based upon program stages and evaluation purposes to guide evaluators and practitioners through his evaluation taxonomy. In working through these steps, Chen encourages evaluators to frame the discussion in terms of future action-directedness to reinforce the view that evaluation is a means to an end, not an end in itself.
Stakeholders rarely come to evaluators with operational definitions and well-articulated program development plans, so the first step in applying the practical taxonomy is for an evaluator to help stakeholders to identify the program stage that they would like to evaluate. If the stakeholders can identify where their program stands within the taxonomy, the evaluator can review with the stakeholder the options an evaluation has in that context. Following this steps, the evaluator helps the stakeholder to identify whether or not the evaluation will be primarily internal, looking at process and implementation operations, or external, investigating assessment or accountability measures. Once the program stage and the audience have been determined, the evaluator and stakeholder can discuss evaluation strategy and approach options in a manner consistent with Rossi’s “good-enough” rule, identifying the trade-offs that each offers relative to the purpose of the evaluation (Rossi, 2004). Once discussed, the evaluator should choose a potential strategy and approach and communicate to the stakeholder the costs and benefits associated with the choice. This choice represents the scientific credibility of the evaluation, while the extent to which the stakeholder agrees and supports the evaluator’s choice is indicative of stakeholder credibility. Without both, the evaluation has a low probability of success.

After completing each step, the evaluator and stakeholder can determine if a single-entry or multiple-entry is appropriate. In the former, the evaluator and stakeholder begin their work at a single program stage. If a multiple-entry evaluation is more appropriate, the evaluator and stakeholders would work to determine the timing and coordination of entries relative to the purpose of the evaluation. In both cases, Chen’s conceptual framework details the requirements at each stage: program planning, initial
and mature implementation and outcome stages. In each stage, though the stakeholder interacts with the evaluator and helps to guide the evaluation, the evaluator is ultimately responsible for ensuring that the evaluation is oriented towards future action, attains scientific and stakeholder credibility, and takes a holistic approach that best fits the unique nature of the program and its context.

**Origin and Ramifications**

Huey-Tsyh Chen’s development of theory-driven evaluation grew out of a search for an approach to evaluation that could overcome the contributions that inadequate methodologies made to that lack of effectiveness so often found in analyses of the effects of evaluations (Chen & Rossi, 1980). Studying at the University of Massachusetts under mentor Peter Rossi, Chen sought for a methodology that could approach the balance between research emphases that statisticians such as William Gosset in the past had sought when pursuing internal and external validity (Chen, 2004b).

For Chen and Rossi, evaluations that gave witness to program failures could not adequately distinguish between a failure of implementation and a failure of the theory behind the intervention. The problem as they saw it was in the articulation of research and program design (Chen & Rossi, 1980). The approach they advocated to overcome this deficiency rested on two propositions: that every social program has some effect on its social system, no matter how small or large, intended or unintended; and that the knowledge that stakeholders and social science possess anticipates some of the effects of every social program.

The multi-goal, theory-driven approach created by Chen and Rossi is based on the argument that social science rarely informs social reforms, and that administrators
typically charge programs with vague, immeasurable goal statements that are based on desirability rather than feasibility. Thus, one of the first reforms the multi-goal, theory-driven approach makes is to have evaluators articulate all the intended and unintended outcomes of a proposed program that are most likely given stakeholder experience and social science knowledge. In this effort, evaluators become active participants in the search for a theoretical model that explains both the social problem under scrutiny and the program that seeks to alleviate its effects.

Chen and Rossi’s approach concerned itself primarily with the effects of programs. In their first paper on the subject, their work centered on social science theory, the prediction of program outcomes, implicit program theories, intervening processes, target population composition, program processes, and the context of programs. Later work by Chen and Rossi (1983, 1987) expanded on these themes. Chen’s seminal work on the subject (1990a) focused on advancing an approach to evaluating the effectiveness of programs through normative, causative and composite evaluations. A work published during the same year explicated program theory in a more compact frame (Chen, 1990b) While the midpoint of the decade found him articulating distinctions between process and outcome evaluations (Chen, 1996), it wasn’t until much later that Chen turned towards questions concerning the role of theory-driven evaluation in program plan development.

While Chen had never focused his attentions on the role of theory-driven evaluation in program development, he advanced the steps required for such an evaluation in an article written primarily for evaluation practitioners (Chen, 2003). Using the theory-driven approach, Chen advanced the idea that evaluators can facilitate the
development of program planning by clarifying stakeholders’ ideas into a systematic program theory when a program is in the early stages of development (Chen, 2003). Given that the goals of the program had been identified, this could be accomplished by specifying the goals of the program, specifying the change and action models, fine-tuning and diagramming the program theory.

To utilize the theory-driven approach in program planning and development, evaluators need to familiarize themselves with issues and strategies such as the existing documentation of the program, the role the evaluator may take in program planning, the participatory modes and formats that the evaluator may take with stakeholders, and the theorizing approaches that the evaluator may take to surface program theory. For Chen, four applications of theory-driven evaluation in program development could be fruitful: planning a new program, clarifying an existing program, facilitating communications about a new or existing program, and providing a foundation for program monitoring and outcome evaluation. A year later, Chen published a book on the subject of practical program evaluation that further explicated the ways in which the theory-driven approach could be useful in a variety of program development stages including program planning, initial implementation, mature implementation, and outcome (Chen, 2005).

Chen’s practical taxonomy for program evaluation means and ends organizes evaluation and strategies and approaches by the stage of program development and the purpose of an evaluation. Chen advances this taxonomy as a way in which to inform and communicate the importance of choosing the right research strategy. One of the most intriguing evaluation approaches advanced by Chen in the taxonomy is the pilot testing approach. This approach is one option for program planning when the purpose of the
evaluation is primarily troubleshooting. The primary objectives of a pilot test approach to program evaluation are to determine the field feasibility of a program and to identify the problems that may arise in its implementation.

As a tool of development, the evaluation that emerges from a pilot testing approach must be responsive to the concerns of stakeholders while maintaining scientific credibility and remaining cognizant of the ecological context of program operations. In order to generate useful information for program improvement, four principles must be met: implementation agents and clients must participate in trials, the samples used for trials must be typical of those clients whom the program was designed to assist, data gathering methods must be flexible, and results should be used for program development rather than evidence of program effectiveness (Chen, 2005, pp. 120-121).

A review of the literature using Academic Search Premier, Social Sciences Abstracts (1983-2000, 2001-2005), Education Abstracts (1985-2005) and the general search tool of the Social Science Citation Index (1980-2005) using the terms ‘theory-driven’, ‘theory’, and ‘evaluation’ alone and in combinations found no instances of a published empirical work with the theory-driven approach to program development, although other types of pilot test evaluations were in evidence (e.g., pilot testing in process evaluations in Troxell et al., 2005). Chen’s taxonomy (2005) cites only one use of the approach in the context of its explication (Wallin, Bremberg, Haglund, & Holm, 1993). A personal communication with the author in 2005 failed to identify any other studies reporting on the use of this approach. This lack of empirical work on the pilot testing approach in theory-driven evaluation is regrettable, though understandable given the myriad evaluation approaches available in the literature (Stufflebeam, 2001).
However, the advantages that this approach, taxonomy and orientation embodies—its orientation towards program improvement in order to solve problems, the goal of achieving scientific and stakeholder credibility, and a holistic approach to assessment—is precisely what program stakeholders need in the program planning stages of program development.

In the practice of program evaluation, evaluators work with program developers and administrators to identify measurable outcomes (Chen & Rossi, 1980). These outcomes are informed by data which may be collected deliberately through appropriate methods and analyzed correctly through techniques that are sensitive to the limitations of the methods employed; no one methodology is superior in this respect (Chen, 1994; 1997). Regardless of the methodology chosen, the role of data and its utility in the program evaluation is of paramount importance. In this light, a theory of actionable data is articulated in order to inform data collection placement for the purpose of program development, and also to posit why formative assessment data may be more useful to teachers and principals than summative or value-added assessment data.

Actionable Data

Actionable data are those that possess high utility for decisions at critical decision points. This type of data can be contrasted with data used to provide contextual information, background information, or information that relates to the state of a person, place or phenomena at another point in time. Actionable data possesses four characteristics: they are reliable, valid, timely and comprehensible. Studies with clinical decision support systems for physicians practicing medicine in health care facilities underscore the tenets of the theory.
Clinical decision support systems deliver patient-specific recommendations to physicians in health care settings that are based upon optimal, evidence-based care. Some of these systems have achieved one or more of the following benefits: the improvement of prescribing practices, the reduction of medication errors, the improvement of preventative care service delivery, and the improvement of adherence to recommended care standards (Kawamoto et al., 2005). Using a combination of univariate analyses, multiple logistic regression analyses and direct experimental evidence, researchers have identified a number of features of clinical decision support systems that support improved clinical practice.

The univariate analyses found five features by examining success rates of clinical decision support systems with and without various features: integration with charting or order entry system, computer-based generation of decision support, the automatic provision of decision support as part of clinician workflow, a request to document reasons for not following system recommendations, and the provision of a recommendation rather than just an assessment. The multiple logistic regression analyses yielded four such features: automatic provision decision support as part of clinician workflow, provision of a recommendation and an assessment, provision of decision support at the time and location of decision making, and computer-based decision support. Lastly, a survey of direct experimental evidence found four features associated with greater effect: provision of decision support at the time and location of decision making, a request to document reasons for not following system recommendations, and the provision of periodic feedback about clinician compliance with system recommendations. Other studies gave support to system features that gave results to both
Clinician and patient. Thus, four features of clinical decision support systems were found by at least two methods to support improved clinical practice: computer-based generation of decision support, automatic provision decision support as part of clinician workflow, a request to document reasons for not following system recommendations, and the provision of a recommendation rather than just an assessment.

Writing fourteen years ago, one contributor to Science saw that any information system attempting to provide data directly to users had to address two factors: quality assurance and format design (Lide, 1981). The author of the article placed the responsibility of the growing need for critical data on the proliferation of the problems that crossed disciplinary lines. It is precisely this environment of complexity that the education sciences are situated.

Combining the fields of educational administration, teaching, learning, settings, and outcomes, each stakeholder in the system- administrators, teachers, students and parents- has a vital need for information in which utility for decision making is high. Decision scientists and philosophers of science have termed this type of information ‘actuarial’, stating that such information is automatic, or specified in advance, and based on empirical research (Dawes, Faust & Meehl, 1989).

It is the underlying premise of the current program evaluation that this type of data is missing from the two types of student assessment systems mandated by the state of Ohio: summative achievement tests and its value-added assessment system. Thus, for the District and schools to be able to produce summative assessment achievement gains larger than those predicted by the value-added system in 2006-07 and beyond, a new system- two standardized, formative assessment systems- will be implemented to produce
actionable data for teachers to guide instruction. For their part, advocates of formative assessment argue that such assessment is the principal means of raising educational standards, and should be central to any national assessment policy (Department of Education and Science, 1988). While all teachers use some type of formative assessment in the process of educating students, the development and implementation of a standardized formative assessment system allow for a more reliable and valid measurement of educational program indicants and variables.

*Educational Assessment*

Educational assessment encompasses the process of gathering information for student, curriculum and educational policy decisions (Nitko, 2001). At its most basic level, assessment is measurement. To be useful, any measurement procedure must be reliable and valid. Reliability refers to the accuracy or precision of a measurement procedure, while validity refers to the degree to which data from such a procedure are relevant to the inferences made from them (Thorndike, 2005). At the student level, assessment data informs decisions relative to placement, progress and achievement. At the program level, assessment data helps to identify areas for curricular improvement and evaluation. At the policy level, assessment data informs educational goal-setting and support decisions.

Canton City School District’s incorporation of a formal system of formative assessment within its larger accountability system is reflective of changing conceptions of curriculum, learning theory, group processes and measurement. While recognizing the necessity of complying with the state mandate for value-added assessment, District leaders recognized that teachers needed additional assessment tools that reflected...
teachers’ understanding of the connections between instruction and assessment. These connections are reflected in changing conceptions of assessment theory and practice in America at the beginning of the twenty-first century.

According to Shepard (2001), the dominant paradigms in education in the twentieth century were scientific measurement, learning theories based on behaviorism and a hereditarian theory of intelligence, and a curriculum developed for social efficiency (p. 1068). Scientific measurement grew out of the work of Edward Thorndike’s efforts to introduce objective measurements of learning and intelligence, and was utilized extensively by psychologists and educators in the early part of the twentieth century. The objective tests which emerged from these perspectives reflected views of learning and curriculum which emphasized rote memorization and recall. At the time, American men such as Goddard and Terman viewed intelligence as a static biological asset that categorized the potential of people on the basis of a score, rejecting the views of Binet, one of the founders of intelligence testing, who viewed IQ as dynamic and malleable (p. 1069). This view of intelligence was echoed in the work of behaviorists such as Skinner who viewed learning as sequential, motivated by external reinforcement, and readily transferable to similar tasks and challenges. The American view of intelligence and learning were readily adopted by a society eager to assimilate a tide of immigrants into its culture and provide workers for a newly industrialized state. In this paradigm, educational assessment was conducted after knowledge had been transferred from teacher to student, and measured the extent to which students had absorbed the discrete blocks of the curriculum presented. In this system, assessment, instruction and learning could be conceived of as separate spheres of activity.
In a review of educational assessment literature, Brookhart (2004) traced how the practice of classroom assessment was influenced by conceptions of the study of individual differences, the study of groups, and the study of measurement. What she found was that changing perspectives on each of these components over the last twenty years created a variety of tensions between the theory and practice of assessment. Specifically, she found that inventories of teacher assessment practices taken at odd intervals over the last quarter of the twentieth century demonstrated the waning dominance of a single approach to educational assessment. In practice, new findings from fields such as psychology, sociology and measurement offered teachers new ways to support student learning that more accurately reflected their changing perceptions of learning, curriculum and instruction. These perceptions began to form a new paradigm of the role of classroom assessment in teaching and learning in the last decade of the twentieth century.

Shepard (2001) finds the primary cause of the shift to be an ascendant belief in all human knowledge as something that is constructed. In the context of education, the context of instruction influences the shape and form of knowledge. Shepard summarizes this understanding with a series of propositions: intellectual abilities are socially and culturally developed (p. 1074), new learning is shaped by prior knowledge and cultural perspectives (p. 1075), and cognitive performance depends upon dispositions and personal identity (p. 1076), for example. This new perspective of learning and intelligence, in alignment with a changing culture and economy, created a new view of the curriculum in which the principles of learning support active movement of all students towards subject matter standards that encourage higher-order thinking and
problem-solving. This new perspective created a need for educational assessments that could assist in student learning and teacher instruction. The new assessments that emerged were deployed in classrooms to support active learning, and emphasized feedback mechanisms, metacognition, and an integration of curricular materials that heretofore were separated. These formative assessments emphasized constructivist views of learning, and oriented teachers towards learning what students needed to know to understand fundamental concepts within a curricular domain.

According to Bloom et al., (1971), summative assessment refers to tests given at the end of a course of instruction to judge the extent of student learning (117). This approach to assessment dominates the field of education and is reflected in each state’s use of large-scale, standardized achievement tests of reading and mathematics in grades three through eight as a result of the 2001 reauthorization of the Elementary and Secondary Act (PL 107-110). A value-added assessment system relies upon the results of summative assessment data to predict individual student performance on the basis of past student scores on one or more summative assessments across one or more years. As such, the value-added interpretation of summative assessment scores both measures and predicts student progress within particular time frames (Ballou, Sanders & Wright, 2004). Formative assessments are administered frequently for the purpose of improving instructional methods and identifying the progress and extent of student learning. While teachers have always used a variety of sources to identify the growth of student learning, standardized test-based formative assessment systems have begun to emerge over the last decade. Regardless of test form, all educational assessments utilized are guided by
standards that articulate test construction, evaluation, and documentation, fairness in testing and testing applications (e.g., AERA, APA, & NCME, 1999).

Summative assessment

In the context of education, the term ‘summative assessment’ refers to a class of systematically administered objective tests given by a teacher, school, district or state education agency in order to document learning and compare student results with a given standard or the performance of similar students. Generally administered, summative assessments are utilized to determine the extent of student knowledge at a point in time and a measurement of the individual’s development of a given ability (Anastasi & Urbina, 1997). In education, summative assessments help a teacher to decide how much students have learned, what grades to give, and if student instruction should be modified (McMillan, 2001). Typically, summative assessments are delivered as large scale, high stakes assessments that are tied to tracking, promotion or graduation sanctions under district or state accountability plans (Heubert & Hauser, 1999). As the stakes increase, greater emphasis is placed on the technical quality of the assessments (AERA, APA, & NCME, 1999).

Measurement, strictly speaking, encompasses a process where by rules are created to assign numbers to represent attributes, traits or behaviors (Reynolds, Livingston & Willson, 2006). Measurement differentiates and describes (Hopkins, 1998). When measurement is applied in a summative assessment setting, it performs these functions against a given standard or relative to others who are taking the same assessment. In the case of the former, criterion-referenced tests seek to identify student progress towards a given performance or learning standard. In the latter, norm-referenced tests describe
student performance against the performance of an appropriate reference group (Hopkins, 1998).

Precise measurement demands a number of technical qualities in its pursuit of objectivity. The first is validity. According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), validity “refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). Three types of validity are widely assessed in educational testing: content, criterion-related and concurrent.

Content validity seeks to identify the extent to which an assessment samples the content and cognitive abilities required by a course of instruction. Criterion-related validity refers to the extent to which performance on an assessment is able to predict performance on some related criterion. Concurrent validity, on the other hand, indicates the extent to which performance on one given assessment relates to performance on another, usually well-established, assessment (Hopkins, 1998). Additionally, the use of summative assessment results for high-stakes decisions (Heuser & Hauser, 1999) creates the need to explain an additional form of validity: consequential validity. In 1971, *Griggs v. Duke Power Company* put forth the proposition that the validity of a test was tied to its use and interpretation in the process of selection processes. This legal opinion led to the theoretical articulation of several related types of validity, including consequential validity (Messick, 1980; 1989).

The second technical quality exhibited by summative assessments to assure objectivity is reliability. Briefly, the reliability of a measure indicates the consistency with which results are generated through repeated measurements. Reliability does not
guarantee that what is being measured is what is desired to be measured, it refers simply to repeatability. Thus, a measure could be reliable without being valid, but a measure cannot have validity without being reliable. For a summative assessment to be utilized correctly, then, it must provide evidence that it meets the twin standards of validity and reliability.

Summative assessments figure prominently in district, state and national accountability systems. Accountability systems based upon these assessments across successive test administrations have been utilized to determine the progress of student learning across the nation (Goertz and Duffy, 2001). In a large-scale assessment, such scores may be summed across classrooms per grade and averaged to yield a building score on a given assessment. In so doing, the building scores become a relative school performance measure compared horizontally across schools. This is the essence of the Elementary and Secondary Act (ESEA) of 2001, a federal education law which requires state and local education agencies to measure relative school performance through successive test score comparisons.

The data yielded by such summative assessments may be utilized by teachers to inform instruction. If the summative assessment is given at the conclusion of a section within a course, data from the assessment may be used in order to inform the instruction of subsequent sections within that course. If the summative assessment is given at the conclusion of a course, data from the assessment may be passed on to the student’s teacher in the following grade. Of course, results from such assessments may also be used to guide an analysis of a curriculum, teaching method, or sequence of instruction (Mertler, 2002).
Nitko, citing Hieronymous (1976), delineates the appropriate use of summative assessment results by teachers inside and outside the classroom more generally. Inside the classroom, test results can be used to describe the academic development level of students, describe qualitative strengths and weaknesses of students and groups of students, measure student performance against predetermined standards, and provide students and parents with feedback regarding student learning. Outside the classroom, test results can be used by teachers to alter curriculum decisions, educational programming or instructional practices (Nitko, 2004, p. 374). In all cases, however, data from summative assessments lead to inferences regarding student performance, not growth. That measure of academic development is informed through value-added and formative approaches to assessment and educational evaluation.

*Value-added assessment*

Value-added assessment refers to a number of statistical approaches that analyze standardized student test data in order to identify the growth in student test scores over time (Tekwe, Carter, & Ma, 2004). The most cited approach in the literature has been the education value-added assessment approach authored by Dr. William Sanders (Sanders and Horn, 1994). In this approach, students are tracked within classrooms and schools longitudinally in order to generate benchmarks of annual student growth within grades and subjects for comparative and predictive purposes.

By examining an individual student’s own academic progress against a normative sample of that student’s own summative assessment scores across three to five years, value-added assessment systems can predict a child’s academic progress in the future, creating the opportunity for relative comparisons within and between students in order to
assess the progress of the student against a given benchmark. If successive tests are calibrated on a common scale, baseline scores from the prior year’s assessment can be compared to scores from the current year’s score in a mean prior score approach. Alternatively, if successive tests are calibrated on different scales, baselines can be calculated from predicted scores.

The term ‘value-added’ refers to the fact that academic scores outside of the student’s historical point in the norm distribution relative to the system’s prediction for the child lead to judgments regarding teacher, classroom or school effects. In this sense, teachers, classrooms or schools ‘add value’ if the student performs above the predicted level of student performance on a standardized subject matter test at the end of a school year. Teachers whose students score higher than predicted are assumed to have caused the observed increase in student scores. Thus, while such assessment systems track student achievement scores, inferences are typically made at the teacher level.

In Sanders’ value-added assessment system, students act as their own statistical control. That is, by using each student’s own prior standardized test scores in prediction calculations, variables that influence academic progress such as parental education and income and the child’s gender, race and ethnicity are included in all statistical calculations at the student level. In essence, the use of a student’s prior academic achievement accounts for background student variables tracked by a district as well as those that are omitted by standards student assessment systems (Ballou, Sanders & Wright, 2004). According to the authors of one of the largest evaluations of value-added models for teacher accountability, “VAM [value-added models] might actually provided less biased and more precise estimates of teacher effects” (McCaffrey, Lockwood, Koretz
& Hamilton, 2003). This is due to the fact that in such assessment systems, growth calculations are conducted within rather than across students. Thus, a value-added assessment system, though based on summative test data, is seen to represent the academic growth of students rather than simple snapshots of student performance.

In America, value-added assessment systems have been established in education primarily for accountability purposes, though their implementation has led to important research findings regarding the relationship between socioeconomic status and academic growth, effects associated with students entering new buildings, the distribution of academic growth in a variety of education settings and teacher effectiveness (Sanders, 1998). Value-added assessment systems that utilize Sanders’ methodology are currently in place in schools in Colorado, North Carolina, Ohio, Pennsylvania and Tennessee. Even more schools across the nation are interested in adding value-added growth measures to their state and federal education accountability systems. According to the April 26, 2006 edition of the *Education Digest*, the list of those that have submitted proposals to the U.S. Department of Education to participate in a pilot program on using value-added measures to satisfy federal education accountability standards are four times that number.

In Ohio, Battelle for Kids began sponsoring a program to provide schools with value-added analysis in 2002 through Project S.O.A.R. (Schools’ Online Assessment Reports). Initially, 42 school districts joined in the initiative. In 2003, a value-added assessment system became part of the state’s accountability system with the passage of H.B. 3 in 2003. Districts choosing to implement the value-added assessment system early were invited to join Project S.O.A.R. through Battelle for Kids. In 2006, the Ohio Department of Education began a fourth grade pilot program for all school districts,
providing value-added analyses for math and reading in grades four through eight on the basis of the Project S.O.A.R. Both Battelle and the Ohio Department of Education rely on the statistical software package EVAAS™ to generate value-added analyses.

While states and districts may use the data from education value-added assessment systems primarily for accountability purposes, teachers can use the data to improve the academic performance of their students. In Project S.O.A.R., the value-added reports that are provided to schools indicate student achievement and progress across classrooms and within a grade. In school-level reports, teachers can identify student performance in a subject through static and growth measures to identify areas of achievement (high and low) and progress (high, average and low). Online school-level reports identify the mean student score and percentile of student achievement, the mean predicted score and percentile of student achievement, the school effect (progress) and its attendant standard error. Thus, these reports can be used by school building improvement teams to inform decisions regarding educational program planning and professional development for school improvement. In student-level reports, teachers can also view static and growth measures of student academic performance. Online student-level reports detail student academic achievement by year according to the child’s performance relative to district, school or student percentiles. Reports run at the beginning of the year indicate the student’s projected score on the district’s summative assessment in that subject. While these data may be used for planning and instructional purposes, the goal is to increase student achievement.

Districts, schools and teachers implementing the value-added assessment system may find many ways to use the data to inform instruction. By the 2007-08 school year,
value-added measures will be added to district and school building report cards in grades four through eight in mathematics and reading based on testing conducted in the spring of 2007. Thereafter, the informal ways in which value-added assessment have been used by districts to augment student performance may be codified in ‘best-practice’ workshops and meetings within and between school districts.

*Formative assessment*

Formative assessment refers to the process of assessment that includes all those formal and informal activities used by teachers to support learning. This includes activities such as teacher-constructed tests, homework, projects, papers, observations and conversations that occur in classrooms that are performed to gain information about learning (Linn & Gronlund, 2000). In contrast to summative assessment, formative assessment occurs as an individual is working on a concept, project or course of study. The phrase ‘assessment for learning’ is often used in this context. Formative assessments help teachers identify whether or not students understand the material, which students need extra help and how to facilitate student learning (McMillan, 2001).

As an educational intervention, formative assessment has been used to support teacher decision-making in the classroom for decades, though its use in successful teaching goes back far further (Wiliam, 1996). It relies primarily upon evidence about desired and actual levels of performance on a construct and suggested actions to close the gap. It supports a variety of instructional decisions that teachers use to guide their instruction of individual students. It can be standardized, as in the case of an assessment system characterized by rules, procedures and standards, or unstandardized, as in the case of a series of teacher observations, interviews with students, or similar qualitative data.
collection techniques. Regardless, the characteristic that defines formative assessment is its generation of actionable data relative to classroom decisions regarding student learning.

One meta-analysis of 21 formative assessment studies of P-12 students found that, when used as a feedback tool to inform the interactive process of teaching and learning, formative assessment contributed to significant learning gains (Fuchs & Fuchs, 1986). Empirical work conducted since then on different subjects, in different settings and using different methods (e.g., Bergan et al., 1991; Martinez & Martinez, 1992; White & Fredericksen, 1998) have replicated and extended these summary findings.

Black (2001) concisely summarizes the features of formative assessment that contribute to its success:

- “All such work involves new ways to enhance feedback from students to their teacher that require new modes of pedagogy so involving significant changes in classroom practice.

- Underlying the various approaches are assumptions about what makes for effective learning— in particular that pupils have to be actively involved.

- For assessment to function formatively, the results have to be used to adjust teaching and learning.

- Several of these studies show that improved formative assessment helps the (so-called) low attainers more than the rest, and so reduces the spread of attainment while also raising it overall.

- The way in which assessment can affect the motivation and self-esteem of pupils, and the benefits of engaging pupils in self-assessment, both deserve careful attention (Butler, 1987, 1988)” (p. 11).

The ways in which formative assessment impacts adjustments to teaching and learning are myriad, but center on teacher and student roles and responsibilities (Black & Harrison, 2001; Wiliam, Lee, Harrison & Black, 2004). To be useful as an intervention,
though, the nature of the formative assessment must be clearly delineated. A description of the two formative assessment instruments to be utilized by Canton City School District teachers will suffice to make the point.

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) was developed by Dr. Roland Good and Dr. Ruth Kaminski of the University of Oregon. DIBELS is a set of standardized early literacy measures that are administered individually as a student is learning how to read. As a formative assessment, they identify the student’s progress along constructs such as initial sound fluency, phoneme segmentation fluency, nonsense word fluency, letter naming fluency. These short-cycle assessments can be given by teachers at established benchmarks (i.e., beginning, middle and end of term) or within established benchmarks to determine the growth and character of student learning. The assessment of reading skills is linked so that teachers can focus on assisting students with learning basic phonemic awareness skills before advancing on to phonics and fluency. The assessments require very little time to administer, and have demonstrated reliability and validity in a variety of settings (see, for example, Hintze, Ryan & Stoner, 2003).

While more than 1.7 million students have been assessed on the DIBELS system nationally, the Canton City School District has collaborated with Wireless Generation™ to administer the DIBELS to students through a Palm Pilot in order to standardize data collection and make results available immediately with teachers (Evans-Gardell, 2006). The District intends to utilize DIBELS as an assessment of reading progress in grades K – 3 and in grades 4 – 6 for Title 1 and special education students only.

ETS Pulliam is also working with the Canton City School District to deliver a standardized formative assessment system in reading in grades three through twelve and
in mathematics, grades three through twelve. While the reading assessments will be aligned with the efforts to continue assessing the fluency measures assessed by DIBELS in the early grades, the mathematics portion of the formative assessment will assess the District’s mathematics curriculum, Everyday Mathematics, in alignment with state and national mathematics standards. This math curriculum, developed by the University of Chicago School Mathematics Project, is a standards-based, paced program of instruction that ETS Pulliam will align with in order to assess student progress through annual mathematics learning objectives. In both reading and mathematics, the District will work with ETS Pulliam’s Focus on Standards and online IDMS system to assess students on a regular basis. In this system, curriculum specialists and reading and mathematics teachers can work together with district evaluation and research staff to design short-cycle, focused assessments that reflect short-term lesson goals or progress towards a long-term course goal. While the content validity of the instruments is transparent through a comparison with state standards and the testing blueprints, internal consistency reliability is gauged through the hand-held device that administers the assessment. Due to the fact that the assessments are based on national and state standards, student work can be examined against the standards to identify areas of strengths or weaknesses by strand (e.g., geometry and spatial sense).

Both the DIBELS and ETS Pulliam formative assessment systems offer teachers a quick, efficient and standardized way in which to formatively assess student academic progress in the classroom in a valid and reliable manner. While the District understands that such systems will be utilized differently between and within schools, it recognizes that, in order for teachers and schools to meet the goals established in the District and
school improvement plans, teachers need such systems in order to effect change in classrooms. In order for any formative assessment to be effective however, the data that the assessment yields must be put into action in the classroom. This understanding lies at the heart of the change model of the program rationale that is the subject of District and school improvement efforts.
CHAPTER 3

PROCEDURES

The procedures employed in the present study flowed from the design principles and guidelines advanced within Huey-Tsyh Chen’s conceptual framework of program theory and taxonomy for program evaluation means and ends (Chen, 2005). These design principles and guidelines establish the steps to be taken when conducting a theory-driven evaluation of a program in the social sciences, and are built in alignment with the implicit and explicit assumptions of a program’s rationale. In conducting a theory-driven evaluation, then, it is necessary to articulate a program theory before connecting the tasks in the work plan to the principles and guidelines of Chen’s framework and taxonomy. In the present study, the design principles and guidelines will be explained presently, while the program theory will be explicated in the context of study results.

Design

In Chen’s framework, pilot testing is to be conducted in accordance with four design principles. In conducting an empirical demonstration of this evaluation approach, these principles will guide the research design of the study. The relevant design principles here are four: Pilot testing requires program implementers and recipients to participate in trials; small but nevertheless typical samples are required for valid inferences about program operations to be drawn; the methods of gathering data must be flexible; and pilot
test findings should only be used for program development purposes (Chen, 2005, 120-121).

The guidelines to applying the taxonomy of program evaluation means and ends within the context of the theory-driven approach (Chen, 2005) require the evaluator to work with program stakeholders to identify the development stage of the program under study, select an evaluation strategy that matches stakeholders’ purposes, choose an evaluation approach and research methods that maximize the timeliness, rigor, and thoroughness of the evaluation at the lowest possible cost, and communicate to stakeholders facts about the chosen evaluation strategy, approach and research methods (pp. 58-61).

The research design of this study was further shaped by the need to develop the program theory of the intervention, including the program rationale and plan. Briefly, the program rationale presents stakeholders with an outline of how a particular intervention will solve a particular problem that the program addresses. It explains the goal of the program, the target population, the intervention that will occur, and the reasons why it will work. The program plan in turn describes how the program will address the particular program; it is the implementation plan of the program. Together, the rationale and plan form the infrastructure of the program, its program theory.

In this study, the program under scrutiny is the District’s implementation of a value-added assessment and initial implementation of two formative assessment systems in the 2005-06 school year. The first component of that program, the state’s value-added assessment system, was implemented in all buildings during the 2004-05 school year. The second and third components of that program, the DIBELS and ETS Pulliam
formative assessment systems in reading and mathematics, were initially implemented in targeted classrooms. To determine the field feasibility of the assessments systems and to identify the problems that may appear with the full implementation of the formative assessment systems in the 2006-07 school year, the District and evaluator chose the pilot testing approach to program plan development. For the District, this evaluation will assist in the development of the intervention in the context of the District and school improvement plans. In this context, the following seven research objectives of this study are advanced:

1) To conduct an empirical demonstration of the pilot testing approach to program plan development in the context of theory-driven evaluation in order to clarify and extend the theoretical framework of the approach;

2) To assist in the development of the program theory of a formative assessment system;

3) To provide a program evaluation framework for formative and summative assessments within the context of value-added assessment that assists the District in augmenting student achievement growth trajectories;

4) To develop a measurement instrument of teacher assessment for the District;

5) To develop baseline measures of student achievement growth in reading;

6) To test a method for linking summative and value-added assessments within the structure of a hierarchical linear model.

7) To propose a method for analyzing formative and summative assessments within the context of a value-added assessment system through the construction of several hierarchical linear models.

In this context, the overall research design reflects the steps required in the design principles and guidelines advanced within Huey-Tsyh Chen’s conceptual framework of program theory for practitioners and taxonomy for program evaluation means and ends (2005), the need to improve the educational program of the District and its schools, the
need to establish scientific and stakeholder credibility, and the desire to take a holistic approach to program assessment.

When employing the approach, a deficit was quickly discovered. What was missing from the approach was an understanding of the need to produce baseline data on the outcome variable to generate points of comparison for subjects involved in the intervention. In the present study, this was achieved by analyzing the relationships between predicted and observed achievement scale scores of fourth and fifth grade students in reading to compare this relationship during the 2004-05 and 2005-06 school years. These grades and this assessment were chosen across the targeted schools due to the scheduled adoption of achievement tests in Ohio by the State Board of Education. In other words, there wasn’t enough data points to run baseline figures for mathematics achievement or for any other grade in the elementary schools. By creating baselines in reading achievement in fourth and fifth grade, the relationship between predicted and actual scale scores could serve as a point of reference, a baseline of student performance to provide subjects with their own control data before the intervention began.

Given the flexible nature of data collection procedures utilized in theory-driven evaluation (Chen & Rossi, 1992; Chen, 1997), and the need to employ a holistic approach to program assessment (Chen, 2005), qualitative and quantitative data collection methods were used within a mixed-methods framework (Chen, 1997; Cresswell, 2003; Tashakkori & Teddlie, 2003). Each principle and guideline is discussed below in the context of its application in the design of the study.
Principles

The first principle of pilot testing is that actual program implementers and recipients must participate in the trials that inform the evaluation (Chen, 2005, p. 120). In this context, the perspectives of each program stakeholder and target population member must be allowed to inform the evaluation. In the current study, this principle was applied through the active involvement of school improvement teams, District assessment specialists, teachers and students in the process of examining the program’s initial implementation. Here, data collection efforts operated at three levels.

At each school, the Director of Testing, Evaluation and Research worked with school improvement teams to develop school-wide plans to improve student achievement. These plans were constructed during and after the District’s two-day inquiry-based data retreat during August 2005. After each had completed its initial school goals at the retreat, they worked with the Director to identify their goal, the data sources, their specific strategies to achieve their goals, and their evaluation plans. Each team developed at least one such goal (Appendix A). School improvement plans developed by school teams at the District’s elementary school data retreat were gathered upon their completion, sent to the District Director of Testing, Evaluation and Research, and sent to the office of the evaluator where they were coded (Ryan & Bernard, 2000) by the evaluator to identify themes across schools concerning assessment and instruction that could impact program implementation.

Within three grades of three schools, teachers were asked to complete a questionnaire in order to discover potential problems in implementation in developing baseline data concerning classroom instruction, subject-matter training and professional
development, assessment practices and beliefs (Appendix B). The items in this questionnaire were taken directly or developed from the National Center for Education Statistics’ Teacher Questionnaire, Grade 4, of the International Association for the Evaluation of Education Achievement’s Trends in International Mathematics and Science Study. These interview data were content analyzed (Silverman, 2000) by the evaluator in order to inform the field feasibility of a program and to help in the identification of the problems that may arise in the implementation of the new assessment systems in the 2006-07 school year.

Within selected classrooms, teachers were granted access to the two formative assessments and the value-added assessment system. Here, selected teachers underwent structured interviews (Fontana & Frey, 2000) with the Director of Testing, Evaluation and Research in order to identify the challenges and opportunities that they saw in each assessment, implementation issues, and the adjustments to instruction (if any) that they made as a result of using the various assessments (Appendix C). These interview data were content analyzed (Silverman, 2000) by the evaluator in order to identify the challenges and opportunities in each new assessment system for the 2006-07 school year. In a limited number of cases, student growth in reading and mathematics was analyzed formatively by the District to ensure fidelity to the intervention.

The second principle calls for the use of small but nevertheless typical samples of program implementers and recipients (Chen, 2005, p.21). Thus, in the present study, strategic subsets of program implementers and recipients within the targeted schools were created to test crucial components of the program theory and to guide the pilot testing approach to the evaluation.
Samples of schools, classrooms and students were developed by the District’s Director of Testing, Evaluation and Research in line with the need for District improvement as identified by the District’s senior leadership (i.e., superintendent, assistant superintendent, and directors) and the need for school improvement as identified by the school improvement teams. Here, samples were chosen on the basis of expected yield; that is, the District chose the samples that it felt could yield the greatest amount of information. As such, the samples chosen met the requirements of Chen’s model and the needs of program stakeholders.

The third principle calls for flexible data gathering methods (Chen, 2005, p. 121). To employ this principle, the evaluator worked with program stakeholders to design data collection instruments that could inform crucial assumptions of the action model. While some instruments were developed in advance of the program evaluation for related projects, others were developed by the District as it implemented the pilot testing approach to the evaluation. Moreover, some of the assessment instruments were created by the state and others through private sector vendors using the subject matter standards of Ohio. The information gleaned from all these sources reflected a variety of paradigms resolved through a dialectic approach (Greene & Caracelli, 2003).

The fourth principle mandates that evaluation results be used only for program development purposes rather than as evidence for the effectiveness of the program itself. Thus, the research design of the study will reflect methods appropriate to such questions. When utilizing methods traditionally associated with tests of a program’s effectiveness (e.g., hierarchical linear modeling), the evaluator will clearly communicate that any preliminary findings are to be used for program development purposes only. In this
context, it will culminate in a field test of the initial implementation of the program theory by predicting fourth and fifth grade student reading achievement using two general hierarchical linear models. In these equations, the intraclass correlation coefficient (ICC), a measurement of the proportion of variance in the dependent variable that is accounted for by level-2 units (Luke, 2004), is predicted to be higher in a one-way random effects ANCOVA than in an unconditional model. The models developed herein will be used to generate baseline data for the District to compare student academic achievement results in the 2006-07 school year using predicted and formative assessment scores as level-1 predictors, and a teacher assessment variable as a level-2 predictor. Generating baseline measures of the main program theory construct under study represents an extension of Chen’s pilot testing approach, one which has important implications for educational program planning.

**Guidelines**

Four general guidelines of Chen’s theory-driven approach (Chen, 2005) were utilized to guide the research design of the present study (pp. 58-61). The requirements of each guideline were interwoven with the aforementioned designed principles, as Chen requires that program evaluation designs be directed towards the future, oriented towards stakeholder and scientific credibility, and holistic (p. 7-9).

The first guideline requires the evaluator to work with program stakeholders to identify the development stage of the program under study. Given the stage-based timing of the proposed program (i.e., the value-added assessment system in the 2004-05 school year, then the formative assessment systems in the 2005-06 school year), program stakeholders and the evaluator determined during initial conversations in winter 2005 that
the program was in the planning stage. At this stage of program growth and development, stakeholders concentrate their resources on developing the basis for a program that is to be implemented in the future. Here, the primary focus of program evaluation is to assist program stakeholders to identify the strategies and approaches that will inform design and development decisions (Chen, 2005, pp. 49-50).

The second general guideline advanced by Chen calls for program stakeholders and the evaluator to select an evaluation strategy that matches the purposes of stakeholders. During the first conversations with program stakeholders, it became clear that the preferred strategy of the District was a development-oriented evaluation to assist and inform initial implementation. In the context of Chen’s program evaluation taxonomy of program evaluation means and ends, in such evaluations program stakeholders and program evaluators pursue one of two strategies: the troubleshooting or the development partnership strategies. After reviewing what each provided, program stakeholders and the evaluator opted for the former.

Chen’s third guideline in applying the taxonomy is to choose an evaluation approach and research methods that maximize the timeliness, rigor, and thoroughness of the evaluation at the lowest possible cost. Given the choice of a troubleshooting strategy, Chen’s program evaluation taxonomy (2005) identifies three possible approaches: relevancy testing, pilot testing and the commentary or advisory approach. Again, after reviewing what each provided, program stakeholders and the evaluator chose the pilot testing approach. Chen’s third general guideline also requires the evaluator and stakeholders to choose an evaluation approach and research methods that maximize the timeliness, rigor, and thoroughness of the evaluation at the lowest possible cost, and
communicate to stakeholders facts about the chosen evaluation strategy, approach and research methods. Program stakeholders and the evaluator reviewed the research methods available to the study in light of the questions to be asked and the needs of the District during the pilot testing evaluation. The consensus opinion was to orient research methods towards those with the highest information yields relative to the action model questions. Eventually, what emerged to analyze was a mix of District and school improvement plans, teacher questionnaires, teacher interviews, predicted and actual student achievement scores.

The fourth general guideline for Chen is to communicate to stakeholders the facts about the chosen evaluation strategy and approach as well as research methods. In this study, communication efforts were undertaken between the program stakeholders and the evaluator about the extent of the intervention and its evaluation.

During the beginning of the 2005-06 school year, the Director of Testing, Evaluation and Research shared the value-added assessment system concept with several groups, including the District’s senior leadership, principals, curriculum specialists and union leadership. In most groups, the conversations centered on the professional development opportunities of such a system, though questions of the confidentiality of student records occasionally appeared in the context of sharing student achievement results within a building for planning purposes. During these initial meetings, participants discussed the goal of increasing the knowledge and awareness of all staff relative to value-added assessment. This was to be accomplished by working with school principals and teachers, respectively.
Value-added assessment data of Canton City School District students was due to be delivered in August 2005 in time for teacher reflection and analysis by the organization that is authorized in Ohio to distribute and market SAS for Schools value-added assessment system, Battelle for Kids. However, after a number of delays, the data were finally delivered to the District in October 2005. This caused a number of shifts in the timelines to communicate the intended intervention and its evaluation. Thus, assessment meetings with each building were delayed until November. Conversations with Battelle for Kids since that time indicate that data due to be delivered in September 2006 will be delivered on time.

**Derivation of General and Specific Research Hypotheses**

In alignment with the research objectives of the study, the principles and guidelines of the pilot test evaluation and the focus of the study, five research hypotheses are advanced in the articulation of baseline student achievement data generated through hierarchical linear models of the value-added and summative reading achievement assessment systems:

1) The relationship between predicted achievement scale scores generated by the state’s value-added assessment system and observed reading Ohio Achievement Test scale scores in the sample groups of students is positive, significantly strong and linear in the 2004-05 and 2005-06 school years.

2) The predicted achievement scale scores generated by the state’s value-added assessment system will correctly classify students who score at or above proficient correctly no more than 80 percent of the time in the baseline year in the 2004-05 and 2005-06 school years.

3) When the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scores of the sample groups are placed in an unconditional hierarchical linear model, the proportion of variance in
the dependent variable that is accounted for by level-2 units will be moderately low (i.e., between 15 to 19 percent) in the 2004-05 and 2005-06 school years.

4) When the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups are placed in an one-way random effects ANCOVA, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderate (i.e., between 20 to 24 percent) in the 2004-05 and 2005-06 school years.

5) When the percentage of sample group students scoring at or above proficiency in the state’s summative reading assessment in school years 2004-05 and 2005-06 are analyzed, a statistically significant gain among cohorts will not be observed.

To test the first hypothesis, the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups will be analyzed in SPSS 14.0 for Windows in order to generate a Pearson correlation and coefficient of determination to describe the relationships between the variables in each sample. A scatterplot of the relationships will also be used to depict the relationships graphically. Here,

\[ H_0 = \rho \leq 0 \]
\[ H_1 = \rho > 0 \]

The hypothesis test will be computing using a \( t \) statistic with significance established at the .05 level.

To test the second hypothesis, contingency tables of predicted and observed proficiency scores will be constructed and analyzed. To test the third hypothesis, the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample
groups will be uploaded into HLM-6 and analyzed through an unconstrained model to estimate between-groups effects with an intraclass correlation coefficient (ICC). This null model is presented below in Equations 1 (hierarchical) and 2 (mixed-effects).

\[ Y_{ij} = \beta_{0j} + r_{ij} \]  
\[ \beta_{0j} = \gamma_{00} + u_{0j} \]  

\[ Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \]  

Here, \( Y_{ij} \) represents the reading achievement score for a particular student in a particular school, \( \gamma_{00} \) represents the grand mean across all students in all schools, \( u_{0j} \) represents the variability between schools, and \( r_{ij} \) represents the variability within schools. The deviance of the unconstrained model will be compared to that under the one-way random effects ANCOVA to determine the relative degree of model fit under both (Luke, 2004).

To test the fourth hypothesis, the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups will be uploaded into HLM-6 and analyzed through a one-way random effects ANCOVA to estimate between-groups effects with an intraclass correlation coefficient (ICC). This model is presented below in Equations 3 (hierarchical) and 4 (mixed-effects).
\[ Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \]  \hspace{1cm} (3)
\[ \beta_{0j} = \gamma_{00} + u_{0j} \]
\[ \beta_{1j} = \gamma_{10} \]

\[ Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + r_{ij} \]  \hspace{1cm} (4)

In Equation 3, the variable \( Y_{ij} \) represents the scores of individual I in school J on the Ohio Achievement test in reading. The variable \( \beta_{0j} \) represents the unadjusted mean score of every \( J \)th school on the outcome (\( Y \)) variable when all predictor variables (\( X \)) are held at zero; the level-1 coefficient is random. In this hierarchical linear model, there is an intercept for each classroom included in the analysis, and it is centered on the group mean. This was done in order to contextualize individual scores within school. The within-classroom error term is represented by the variable \( u_{ij} \). This term is assumed to be normally distributed within each school, while its variance is assumed to be relatively homogenous across schools. The \( \beta_{1j} \) variable represents the slope of the achievement score on predicted scores for the \( J \)th school; here the level-1 coefficient is fixed.

Given that predictors were utilized in only the first level of the one-way random effects ANCOVA model to generate baseline data, two assumptions will be tested: homogeneity of variance and normality (Raudenbush & Bryk, 2002).

To test the fifth hypothesis, the percentage of sample group students scoring at or above proficiency in the state’s summative reading assessment in school years 2004-05...
and 2005-06 will be compared through a non-parametric test for two independent samples. The asymptotic significance of the hypothesis test will be set at the .05 level.

**Participants**

The participants include students in grades 4 and 5 from the eleven elementary schools in the District judged by the state and federal governments to be ‘at-risk’ or in ‘school improvement’ status during the 2004-05 school year. To test the fifth hypothesis, the same grade level will be tested in the same schools during the 2005-06 school year. Demographic analyses of all participants are presented in chapter four.

**Instruments**

*Predicted Ohio Achievement Test scale scores in reading*

De-identified student assessment data was prepared by the District in preparation for the present study. The predicted scale scores are a mathematical prediction of student performance based upon multiple observations of actual student achievement results generated by SAS at School through Battelle for Kids in line with Dr. William Sanders’ value-added assessment methodology (Sanders & Horn, 1994). These data consist of the predicted student achievement scores on the March 2005 and 2006 administrations of the Ohio Achievement Test in reading for fourth and fifth grade students in eleven Canton City School District elementary schools identified by the Ohio Department of Education as at risk or in school improvement during the 2004-05 school year. These data were downloaded from the Battelle for Kids website by the Director of Testing, Evaluation and Research in June and September 2006.

In this system, predicted scores are calculated as a residual during the analysis. Predictions are estimated from at least three prior assessments and current student test
data. Missing predictor points occur when SAS in School was not able to link current
demographic data for a particular student to that student’s prior test record. In this
context, students of high mobility bear a higher probability of a missing prediction score
than those who are stable in one system. Moreover, fewer missing scores are expected to
occur as the state mandates the wider use of achievement testing in reading and
mathematics in grades 3 – 8 in 2006-07 and beyond.

*Observed Ohio Achievement Test scale scores in reading*

De-identified student assessment data were prepared by the District in preparation
for the present study. These data consist of the observed student achievement scores on
the March 2005 and 2006 administrations of the Ohio Achievement Test in reading for
fourth and fifth grade students in the eleven Canton City School District elementary
schools identified by the Ohio Department of Education as at risk or in school
improvement during the 2004-05 school year. Statistical summaries, item analyses and
test blueprints for the fourth and fifth grade Ohio Achievement Test in reading can be
found in Appendix D.

The Ohio Grade Four and Five Reading Achievement Test Blueprints developed
by the Ohio Department of Education’s Office of Assessment were developed in 2003.
The Blueprint lays out the structure of both assessments in the context of Ohio’s reading
standards. In both tests, forty-nine raw points are distributed through reading process,
reading applications and vocabulary standards through literary text, informational text in
short, medium and long passages. The Department reports scores annually in terms of
raw and scale scores.
The reliability of the March 2005 fourth grade Ohio Achievement Test in reading was .88. Cut score points (scale) for the test were as follows: basic, 384; proficient, 400; accelerated, 435; and advanced, 467. Similarly, the reliability of the March 2005 fifth grade Ohio Achievement Test in reading was .87. Cut score points (scale) for the test were as follows: basic, 384; proficient, 400; accelerated, 441; and advanced, 459. The reliability of the March 2006 fourth grade Ohio Achievement Test in reading was .89. Cut score points (scale) for the test were as follows: basic, 384; proficient, 400; accelerated, 435; and advanced, 467. Similarly, the reliability of the March 2006 fifth grade Ohio Achievement Test in reading was .90. Cut score points (scale) for the test were as follows: basic, 384; proficient, 400; accelerated, 441; and advanced, 459. The Ohio Department of Education utilizes common item equating to produce comparable assessments over time. After each administration, raw test items are calibrated and equated so that proficiency standards are maintained at a scaled score of 400 (Ohio Department of Education, 2005b). The content validity of the assessments is rooted in their derivation from the Ohio Department of Education’s English Language Arts and Mathematics Academic Content Standards (adopted by the State Board of Education in December 2001).

**Teacher assessment questionnaire**

The instruments utilized herein include a 22-question teacher questionnaire concerning assessment and a series of three interview protocols for teachers providing feedback on the new assessment systems (Appendix E). The questions in the teacher questionnaire was taken entirely or adapted from the Trends in International Mathematics Study (TIMSS) Teacher Questionnaire, Grade 4, for 2003. They were submitted to three
groups in an initial test of the instrument and teacher feedback process. To protect their confidentiality, they are referred to as School 1 fifth grade teachers \((n = 3)\), School 2 third grade teachers \((n = 2)\), and School 3 fifth grade teachers \((n = 2)\). Their responses will be used by the District to improve the instrument and distribute it to teaching staff next school year.

**Structured interview protocols**

The structured interview protocols consist of nine questions regarding the challenges and opportunities posed by the assessment system in question. Three protocols were developed by the District to tap into teacher assessment opinions regarding the value-added and formative assessment systems. In the protocols, the interviewer seeks to identify the adjustments made to their instruction as a result of the data yielded by each assessment, and asks them to reflect on what value they and other teachers see in the system. Interviews were conducted with teachers from four elementary schools in an initial test of the instrument and teacher feedback process. Teacher responses will be used by the District to improve the instrument and utilize it with teaching staff next school year.

**Statistical Treatment**

Student achievement data were analyzed in the context of a hierarchical linear model (HLM). Such models are utilized by researchers interested in studying observations which are nested within various possible units of analysis. In the social sciences, much of the research is conducted in such hierarchical data structures (Raudenbush & Byrk, 2002). HLM models are appropriate to educational research questions that involve sampled observations of effects within students, classrooms and
schools (Wang, 1999). Such multilevel models are often required in education research for theoretical and statistical reasons (Goldstein, 2003).

Theoretically, multilevel models are required for observations that are nested in some type of hierarchy (Luke, 2004). In the present case, students, classrooms or schools could be potential units of analyses. By embedding observations of student achievement within classrooms or within classrooms within schools, however, errors of inference are not as likely to occur.

Statistically, the use of a classical general linear model requires that a number of assumptions have been met; namely, independent observations, interval scale of the dependent variable, homogeneity of variance, normality, and the absence of outliers/extreme values. With HLM, however, these assumptions need not be met when modeling multilevel phenomena with maximum likelihood estimation techniques (Luke, 2004).

In the present study, the District required baseline student performance data to be estimated for both theoretical and statistical reasons. Theoretically, the change in student achievement hypothesized by the intervention is to be brought about by changes in building and teacher practices. Statistically, many of the assumptions of classical general linear modeling would be violated in any analysis of score variation between students within and between classrooms of any particular school.

Limitations

The chief limitation of the study is related to the timing of the program evaluation. Had each of the associate organizations been prepared to implement their assessment systems in a single school in the beginning of the school year, training and
initial implementation issues could have been more fully explored. As a pilot testing
evaluation, such issues go to the heart of the utility of the evaluation. To overcome the
consequences of this limitation, the evaluator will work with the District to draft training
evaluation instruments for implementation in the 2006-07 school year.

One other troublesome area of the study is the number of missing predictor scores
in the fourth and fifth grade datasets in the 2004-05 and 2005-06 school years. This is a
direct consequence of two factors: the state’s value-added assessment system and the
nature of urban education research. First, value-added assessment requires multiple
assessment data points per student in order to make reliable predictions and estimates.
Due to the state’s implementation schedule for its achievement tests, this system will
yield fewer and fewer missing data points overall each year. Second, urban education
research is often hampered by the high degree of mobility of students within and between
districts. Districts that are in suburban or rural settings, therefore, will possess fewer
missing predictor scores by fiat. The problem of missing predictor scores is also another
reason behind the District’s proposed implementation of standardized formative
assessment systems. By its very nature, formative assessment systems generate student
assessment data on a regular basis prior to a summative assessment. This fact yields the
conclusions that level-1 estimates in a hierarchical linear model of student achievement
should be fairly robust if the predictors correlate strongly with the outcome variable.
CHAPTER 4

RESULTS AND DISCUSSION

In the 2004-05 school year, the Canton City School District received an ‘Academic Watch’ rating from the Ohio Department of Education. Its students had taken achievement tests in reading in grades 3-5, 8 and 10, and mathematics achievement tests in grades 3, 6-8 and 10. The District met 3 of the 23 state indicators, achieved a performance index score of 74.2 out of 120, and missed Adequate Yearly Progress for the third consecutive year. At the time of this writing, the District is in District Improvement, Year 2 status. Given its level of performance and the consequences it faces, the District revised District and building improvement plan early in the 2005-06 school year. To facilitate that process, the District contracted with Learning Points Associates in the summer of 2005 to develop and implement an inquiry-based workshop on academic assessment that would act as the foundation for the revision of school improvement plans.

School Improvement Plans

In the context of the study, the eleven Canton City School District elementary school buildings identified by the Ohio Department of Education in school year 2004-05 as at risk or in school improvement were asked to participate in a data retreat sponsored for all elementary schools by the District. This comprehensive program focused on student learning, the school improvement cycle, and the essentials of data use: developing a leadership team, collecting data, analyzing patterns and generating hypotheses,
developing goals and strategies to reach them, defining evaluation criteria and making the
government of the plan developed by the school (Learning Point Associates, 2004). The
goal of the retreat was to allow school improvement teams to activate assessment
knowledge in the context of school improvement. While a summary of those plans is
available in Appendix A of this study, a content analysis (Ryan & Bernard, 2000) was
performed by the evaluator on the text of all schoolwide plans in support of the
evaluation’s aim: to determine the field feasibility of the program and to identify the
problems that may arise in its implementation (Chen, 2005).

After each team had completed its initial school goals at the retreat, they worked
with the Director of Testing, Evaluation and Research to complete their schoolwide
planning forms. Though the form contained seven prompts for each subject, three are
germane to this study: the identification of the improvement goal(s), the action steps
intended to achieve the goal(s), and their evaluation criteria. According to the guidelines
for developing goal statements, teams were urged to state their goal(s) clearly, make them
measurable, and base them upon patterns observed directly from that data (e.g., previous
summative achievement test results) presented to them in the course of the data retreat.
Teams were then directed to reflect upon research and experience to guide their choice of
strategies designed to reach those goals. Here, trainers urged teams to generate a number
of different strategies based upon hypotheses generated from thoughtful reflection on the
data presented to them during the training. Finally, teams were asked to delineate specific
evaluation strategies to determine the extent to which their work had been successful
(Learning Point Associates, 2004). Each prompt was to be answered by teams through
consensus and recorded upon a central schoolwide planning form. Upon its completion,
each team’s plan was photocopied and sent to the Director’s office where it was coded (Ryan & Bernard, 2000) by the evaluator in order to identify themes across schools concerning assessment and instruction that could impact program implementation.

The first prompt for teams required a goal statement to direct their schoolwide plan for reading improvement. In an analysis of the text from this first prompt in reading, the overarching theme of each team was the acquisition of AYP targets in achievement across grades 3 – 5. Hereafter, however, variability increased as teams chose various avenues to reach that goal. Four themes emerged within the second prompt in reading, action steps. Here, formative assessment, targeted interventions, teaching reading strategies, and reorienting clusters of classroom practices were the predominant themes. The third main prompt in reading, evaluation criteria, yielded one clear theme of assessment information. Most of the teams stated how they were going to use the results of formative assessments to periodically assess student progress towards the goal of student achievement. A second theme to emerge therein was the use of Ohio Achievement Tests in reading in 2006 to evaluate the success of team plans.

In mathematics, the same three prompts were offered for team reflection and planning. Again, the first prompt for teams required a goal statement to direct their schoolwide plan for reading improvement. Here the dominant theme was the acquisition of AYP targets in grade 3 mathematics, though a second, weaker theme emerged targeting a given proficiency level separate from such targets. The second prompt required teams to identify several action steps to guide school staff efforts toward the achievement of the intended goal(s). Two themes dominated all activities here: the use of formative assessment to guide instruction, and the use of focused activities to strengthen
student comprehension in various content areas. Greater thematic variability was observed within the third prompt, the proposed evaluation criteria for identifying the extent to which teams were successful in working with students to achieve their desired goal(s). Here, several themes emerged: while formative and summative assessment results were utilized most strongly, teacher observation and reflection also figured strongly in teams’ plans. Across teams, the main difference observed in the coding and analysis process was specificity and the logical connections between goals, strategies and evaluation criteria. Across subject matter, reading demonstrated the greatest variability within strategy, especially in relation to the choice of curricular programming. While each team maintained a different timeline for implementing its various activities, all responded to each prompt (e.g., improvement goal, data source, action step description, evaluation, resources needed, persons responsible and timeline) of the schoolwide plan form.

Program Rationale

To support these plans, the evaluator worked with the Canton City School District’s Director of Testing, Research and Evaluation to develop a program rationale for the District’s implementation of the state’s value-added assessment system and its initial implementation of two formative assessment systems in reading and mathematics. To assess reading progress, the District began to implement the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Sixth Edition assessment in grades K – 3 and in grades 4 – 6 for Title 1 and special education students only. To assess reading and mathematics achievement, the District began to utilize ETS Pulliam’s formative assessment system in reading in grades 3 – 12 and in mathematics in grades K – 12.
Following the data retreat, District stakeholders and the evaluator began by developing the program rationale and plan for the intervention, and to discuss data collection procedures and instruments to answer the questions posed under the action model of the program theory.

All theory-driven evaluations involve crucial tests of program theory to validate their use. Program theory consists of change and action theories; the change model is referred to as the program rationale, while the action model constitutes the program plan. In a sense, the program plan fulfills the prescriptions of the program rationale. Together, the rationale and plan form the infrastructure of the program (Chen, 2005, p. 73). In the context of a pilot testing evaluation, however, Chen recommends that only implementation issues- the assumptions of the action model- are tested (Chen, 2005, p. 121). However, in conducting an empirical demonstration of the approach, program stakeholders and the evaluator found that it was necessary to establish baseline data on the crucial components of the study. While Chen does not advocate such in his model, it is imperative to provide stakeholders and the evaluator with some point of comparison so that changes that accompany the implementation of the intervention can be compared with pre-intervention performance. While the intervention itself is tied back to the program rationale in Chen’s model, one flaw is that the pilot testing design does not identify the need to establish pre-intervention performance levels of the subject of interest.

In order to provide utility to program planners, a program rationale must clearly state the problem that is to be resolved, the target population, the intervention employed to affect the assumed causes of the problem, and the goals of the program. Meetings with
program stakeholders in the winter and spring of 2005 generated a program rationale which was used in the development of the pilot testing evaluation (Figure 1). The rationale is represented graphically in Figure 1. According to program stakeholders, the primary problem that the program was designed to correct was the low student achievement of District students on statewide standardized examinations. Given the size of the District and the extent of the problem in the elementary schools, the District chose to target students in grades 3 – 5 during the pilot test. The intervention consisted of the introduction of the state’s value-added and two formative assessment systems, while the determinant, the factors that are responsible for causing the problem the program was designed to create, was three-fold: a lack of inquiry-based school improvement training, the lack of a standardized, reliable, and valid formative assessment system, and assessment training. In this context, the goal of the program is to increase student achievement within a cycle of continuous improvement.

The specific research questions of the study came from the program rationale and its use in the program’s action model (Figure 2). Figure 2 represents the action model of the program theory as developed by the Director of Testing, Evaluation and Research and the evaluator. This action model is the subject of inquiry in the pilot testing evaluation.
In Figure 2, the implementing organization is the District and its schools, while the responsibility for implementing the assessment systems belongs to the Office of Testing, Evaluation and Research, the schools and classroom teachers. The associate organizations are those that are partnering with the District to put the assessment systems into place. Similarly, the ecological context of the program enumerates those giving internal micro- and macro-level support to the program. The intervention protocol specifies the content and activities of the intervention, while the service delivery protocol
indicates the steps that implementers took to deliver the intervention. The target population details those who are to receive the intervention.

Figure 2: Action model

**Initial Implementation Lessons**

Data were gathered from the teacher assessment questionnaires and targeted structured interviews to further inform the program rationale of the District’s theory-driven evaluation of its initial implementation of three assessment systems. In the case of both data collection instruments, a purposive sample of schools and teachers was developed by the District’s Director of Testing, Evaluation and Research in line with the need for District improvement as identified by the District’s senior leadership (i.e.,
superintendent, assistant superintendent, and directors) and the need for school improvement as identified by the building-level continuous improvement teams.

**Teacher assessment questionnaire**

Within three grades of three schools, teachers were asked to complete a cross-sectional questionnaire in order to discover potential problems in implementation in developing baseline data concerning classroom instruction, subject-matter training and professional development, assessment practices and beliefs (Appendix B). The items in this questionnaire were taken directly or developed from the National Center for Education Statistics’ Teacher Questionnaire, Grade 4, of the International Association for the Evaluation of Education Achievement’s Trends in International Mathematics and Science Study. These questionnaire data were content analyzed (Silverman, 2000) by the evaluator in order to inform the field feasibility of a program and to help in the identification of the problems that may arise in the implementation of the new assessment systems in the 2006-07 school year.

The questionnaire was divided up into four sections concerning the demographic background of teachers, mathematics instruction, reading instruction, and background information on teachers’ perspective of the school. The middle two sections are germane to this study. In the first prompt within the mathematics instruction section, teachers were asked to reflect on their training and experience in mathematics content and instruction in order to assess how ready they felt to teach each strand of the Ohio Department of Education’s Mathematics Academic Content Standards (adopted by the State Board of Education in December 2001): number, number sense and operations; measurement;
geometry and spatial sense; patterns, functions and algebra; and data analysis and probability.

Teachers in the first building did not report feeling ready to teach two content standards in the curriculum: patterns, functions and algebra and data analysis and probability. When asked to identify their professional development experiences in mathematics over the past two years, teachers exhibited variability in their training. When asked to identify the percentage of time that students spent on various activities within the classroom, teachers reported that the majority of class time was spent on working problems, while the least was in reviewing homework. Similarly, uniformity was seen on the frequency of testing (e.g., about once a week), while the items typically used were true-false and multiple choice formats. When asked to identify from a closed-ended list all the decisions that teachers could make from formative assessment data, teachers did not identify all the decisions that could be made from such data. However, when asked to identify from a closed-ended list all the decisions that teachers could make from summative mathematics assessment data, teachers demonstrated a solid understanding of the uses of such data in regard to classroom decision making.

In the first prompt within the reading instruction section, teachers were asked to reflect on their training and experience in reading content and instruction in order to assess how ready they felt to teach each strand of the Ohio Department of Education’s English Language Arts Academic Content Standards (adopted by the State Board of Education in December 2001): phonemic awareness, word recognition and fluency; acquisition of vocabulary; reading processes; reading applications concerning informational, technical and persuasive text; and reading applications concerning literary
text. Here, teachers exhibited a uniformity of responses regarding their teaching readiness as somewhat ready. Interestingly, uniformity was also observed among teachers when asked to report their professional development experiences during the past two years. Teachers here reported very few reading experiences. When asked to identify the percentage of time that students spent on various activities within the classroom, teachers reported that the majority of class time was spent on small group reading instruction. Some similarities were observed in the amount of time spent in the classroom on comprehension activities and read aloud/think aloud activities. Uniformity was similarly observed when teachers were asked to report the frequency of classroom examinations (e.g., about once a week) and the item formats used for reading examinations (e.g., multiple choice). Lastly, when asked to identify from a closed-ended list all the decisions that teachers could make from summative reading assessment data, teachers demonstrated a solid understanding of the uses of such data in regard to classroom decision making.

The second building that was sampled for teacher assessment information demonstrated that in some buildings, team teaching practices must be taken into account when asking elementary school teachers about their instruction and assessment practices. Due to this phenomenon, data collection was limited, but valuable in the mathematics portion of the questionnaire. In this building, the relationship between teaching readiness per content area and professional development experience was seen again. This finding is not as strong as in the other building, though, due to the fact that mathematics instruction is delegated to one teacher in the grade, while reading instruction is a shared responsibility of all.
In the first prompt within the reading instruction section, teachers were asked to reflect on their training and experience in reading content and instruction in order to assess how ready they felt to teach each strand of the state’s reading standards. Here, uniformity was observed across teachers and content areas. Great variability, however, was seen in professional development experiences, demonstrating that readiness is not directly tied to such experiences. Similarly, variability was demonstrated among teachers when asked to indicate the percentage of time spent on various activities in a typical week of reading lessons. This variability may be a result of differentiated instruction, building practice, or some other untested variable. This building replicated the pattern of testing in reading that the first building demonstrated: testing conducted about once a week. Here, however, a mixture of item formats were observed including short answer/extended response and essays. Similarly, when teachers were asked to identify from a closed-ended list all the decisions that teachers could make from formative assessment data, teachers did not identify all the decisions that could be made from such data. Finally, as has been seen throughout buildings and subject matter, when asked to identify from a closed-ended list all the decisions that teachers could make from summative reading assessment data, teachers demonstrated a solid understanding of the uses of such data in regard to classroom decision making.

Teachers in the third building sampled demonstrated a strong sense of readiness in teaching mathematics, and uniformity was observed across professional development experiences which were quite strong. The distribution of activities reported by teachers in a typical week of mathematics lessons demonstrated the same patterns of instruction found elsewhere (e.g., 40 percent of time spent on working problem), though assessment
in this building was uniformly conducted about every two weeks. In these classroom assessments, uniformity was demonstrated on the item formats typically used in such tests (e.g., short answer/extended response). In a pattern consistent across buildings and subject matter, teachers did not identify all the decisions that could be made from formative mathematics data. However, teachers in this building demonstrated variability when asked to indicate all of the decisions that could be supported from summative assessments administered to students relative to mathematics.

When teachers in the third building responded to the reading instruction assessment, teachers did not report a strong level of readiness across reading standard content areas, and reported very few professional development experiences related to reading instruction and assessment. Here, however, it is important to coordinate instructional readiness and professional development experiences by grade level, in that certain portions of the reading standards (e.g., phonemic awareness) are stressed in the lower grades of the elementary while others (e.g., reading applications concerning informational, technical and persuasive text) are more appropriate to the higher grades. This is an issue both for future versions of the teacher assessment questionnaire and teachers’ individual professional development plans. When asked to identify the percentage of time that students spent on various activities within the classroom, teachers reported that the majority of class time was spent on small group reading instruction and comprehension activities, while similarities were observed between teachers on the percentage of time spent on read aloud/think aloud and story event sequencing activities. As was the case in mathematics instruction in this building, teachers reported administering reading assessments about every two weeks using short answer/extended
response item formats. Interestingly, when asked to identify from a closed-ended list all the decisions that teachers could make from formative reading assessment data, teachers did not identify all the types of classroom decisions that could be made from such data. Interestingly, the same finding occurred when teachers were asked to identify from a closed-ended list all the decisions that teachers could make from summative reading assessment data. When these responses were reviewed against the teachers’ professional development experiences, it was noted that the teachers did not report any professional development experiences over the past two years in reading assessment.

A content analysis of the teachers’ answers across schools led to a number of insights to inform the District of several issues regarding the field feasibility and implementation of the value-added and two formative assessment systems in the 2006-07 school year. Two conclusions were drawn as a result of the purposive sampling of teacher assessment practices. First, the questionnaire yields valuable data in the context of school improvement within mathematics and reading instruction and assessment. The trends and issues that this instrument uncovered have provided valuable insights into this theory-driven evaluation. Second, the instrument, while valuable in and of itself, cannot substitute for observations of classroom teacher practices. While it can identify opportunities for alignment and integration between teacher training and experience, perceptions of content knowledge and teacher practices, the myriad complexities of teaching demand that a variety of tools be used in any study of instruction and learning. Within the data yielded by the instrument, several issues arose that inform this evaluation.
Within mathematics and reading instruction, four main issues arose that inform the evaluation. The first issue is the alignment of professional development to the content standards being taught in mathematics in each grade. Within schools and between grades, the relationship between instructional readiness and professional development was observed to be a fairly strong one, indicating that the alignment between the two can be tested within teacher within building and correlated with various teacher perception and practice variables. The second issue is the integration of instructional practices and testing. While teachers reported periodic testing of students, the questionnaire could not get at the larger question of the integration of lessons taught, lessons and homework performed, and the systematic demonstration of skills that were learned. Third, assessment queries in both subjects demonstrated that teachers are currently utilizing formative assessment on a weekly or biweekly basis. Finally, the variability demonstrated among teachers on the types of decisions that formative assessment, and, to some extent, summative assessment indicates that teacher education, training and professional development on these topics may not be uniform.

*Structured interviews*

Within selected classrooms, teachers were granted access to one of two formative assessments and the value-added assessment system. Here, selected teachers utilizing both underwent structured interviews (Fontana & Frey, 2000) with the Director of Testing, Evaluation and Research in order to identify the challenges and opportunities that they saw in each assessment, implementation issues, and the adjustments to instruction (if any) that they made as a result of using the various assessments (Appendix C). These interview data were content analyzed (Silverman, 2000) by the evaluator in
order to identify the challenges and opportunities in each new assessment system for the
2006-07 school year.

Value-added assessment

The first structured interview protocol concerned value-added assessment. Teachers in two buildings were given access to the system in one grade in order to test the connections between the associate (e.g., Battelle for Kids) and implementing (District and schools) organizations. In the protocol, teachers were asked to comment on the challenges they experienced in working with the data, the opportunities that they saw in using the data, and the adjustments that they made in their instruction as a result of using the data. In both buildings sampled, principals accessed value-added data for the teachers, providing them with grade-level data regarding student achievement in reading and mathematics.

In the first building sampled, teachers saw the opportunities for the use of value-added data in the classroom and across teachers in the grade, but found that their lack of direct access impeded their use of the data in the initial implementation of the system. Teachers reported that they found the information on student growth to be quite useful, and indicated that they used growth data for intervention purposes. All teachers asked to obtain professional development on using the value-added system in order to maximize its use in the classroom and in order to act as a resource to others in the building, and reported that they wanted to know more about the value-added assessment system.

In the second building sampled, teachers reported that the principal’s generation and explanation of value-added reports helped the teachers to surmount challenges to working with the student achievement data. In this building, teachers saw a number of
opportunities for working with the data. In two cases, teachers used value-added assessment data at the student level to work with students individually to understand the progress that they were expected to make during the year, and motivate them to achieve higher results. Once the school year was nearing its end, the teacher met again with the student to discuss how they performed on their summative assessments relative to the results that they were projected to make. Teachers in this school reported using the data at grade-level meetings to group students for intervention by their projected scores in advance of the summative assessments; they also sought to replicate this process in the following year. In this school, no teacher felt as if they could serve as a resource to their colleagues in the building. All preferred to have the principal act as the chief resource for value-added assessment data and reports in the school.

With the value-added assessment system, it is clear that past building practices and relationships mediated the success of the initial implementation of the system. In each school, the principal played a central role in generating data, explaining results and in identifying the parameters of the system’s use in the context of instruction. To the degree that teachers are granted direct and indirect access to the data and are allowed to utilize it in the classroom, information regarding the challenges they face and the opportunities they perceive will be impacted and reported upon in the context of the structured interview. This variability must be captured by the instrument.

**Formative reading assessment**

The second structured interview protocol concerned the DIBELS formative assessment system in reading. Selected teachers in four buildings were given access to the system in one grade in order to test the connections between the associate (e.g.,
University of Oregon Center on Teaching and Learning) and implementing (District and schools) organizations. In the protocol, teachers were asked to comment on the challenges they experienced in working with the data, the opportunities that they saw in using the data, and the adjustments that they made in their instruction as a result of using the data.

In response to the first prompt concerning implementation challenges, four out of the eight teachers interviewed reported technology concerns. Two of the four reported utilization concerns, while two identified specific problems with the operation of the Palm Pilot. Two other teachers expressed their concern over finding the time to give the assessment, while two others were satisfied with the system and its operation. Across all teachers, time and manpower issues were central concerns in a district-wide scale up.

When asked to reflect on the opportunities that the assessment held for the District, teachers were united in their commentary regarding the utility of the system. Teachers who used the system felt as though it could offer diagnostic support for instruction, that the data offered immediate feedback regarding students’ learning gaps, and that the system would help support teachers’ work in the classroom. Overall, teachers reported positive experiences with the administration of the system and the interpretation of the results.

*Formative mathematics assessment*

The third structured interview protocol concerned the ETS Pulliam formative assessment in mathematics. Teachers in one building were given access to the system in one grade in order to test the connections between the associate (e.g., ETS Pulliam) and implementing (District and schools) organizations. In the protocol, teachers were asked to
comment on the challenges they experienced in working with the data, the opportunities that they saw in using the data, and the adjustments that they made in their instruction as a result of using the data. In both buildings sampled, principals accessed ETS Pulliam data for the teachers, providing them with grade-level data regarding student achievement in mathematics.

Teachers here felt that the technology provided a number of challenges in the initial implementation of the system. For some students, the tests took longer to complete than expected. Some of the students experienced access problems, and some problems with the technology were observed. Similar to the findings observed in the value-added assessment system, the principal controlled access to the data and bore the responsibility for generating reports. All teachers reported that the students perceived the test length as a challenge. Teachers did not report any changes to their instruction as a result of the data, but indicated that there might be such adjustments if full implementation was accompanied by shorter tests and integration with current teacher instructional practices. All the teachers did not feel as if they could act as building resources on this assessment system.

The results of teacher interviews on the ETS Pulliam formative assessment system in mathematics were not encouraging. The benefits associated with this system at the elementary school level were not emphasized by participants of the structured interview, while the challenges to implementation and utilization stood at the forefront of teacher perceptions. Many of these challenges were located in the system’s technology. Others related to existing building practices and relationships. Clearly, these issues will mediate the success of the implementation of this assessment system in 2006-07, and are,
to some extent, embedded in the student achievement patterns observed in the year before
the initial implementation of the value-added and two formative assessment systems.

**Baseline student achievement in reading**

In alignment with the research objectives of the study, the principles and
guidelines of the pilot test evaluation and the focus of the study, five research hypotheses
are advanced in the articulation of baseline student achievement data generated through
hierarchical linear models of the value-added and summative reading achievement
assessment systems:

1) The relationship between predicted achievement scale scores generated
by the state’s value-added assessment system and observed reading
Ohio Achievement Test scale scores in the sample groups of students is
positive, significantly strong and linear in the 2004-05 and 2005-06
school years.

2) The predicted achievement scale scores generated by the state’s value-
added assessment system will correctly classify students who score at or
above proficient correctly no more than 80 percent of the time in the
baseline year in the 2004-05 and 2005-06 school years.

3) When the predicted achievement scale scores generated by the state’s
value-added assessment system and the observed reading Ohio
Achievement Test scores of the sample groups are placed in an
unconditional hierarchical linear model, the proportion of variance in
the dependent variable that is accounted for by level-2 units will be
moderately low (i.e., between 15 to 19 percent) in the 2004-05 and
2005-06 school years.

4) When the predicted achievement scale scores generated by the state’s
value-added assessment system and the observed reading Ohio
Achievement Test scale scores of the sample groups are placed in an
one-way random effects ANCOVA, the proportion of variance in the
dependent variable that is accounted for by level-2 units will be
moderate (i.e., between 20 to 24 percent) in the 2004-05 and 2005-06
school years.

5) When the percentage of sample group students scoring at or above
proficiency in the state’s summative reading assessment in school years
2004-05 and 2005-06 are analyzed, a statistically significant gain among cohorts will not be observed.

Each hypothesis will be tested according to the plan set forth in chapter 3.

School Year 2004-05

Before the initial research analyses commenced, data were screened to examine the existence and potential impact of missing data and multivariate outliers (Mertler & Vannatta, 2002). The fourth grade reading achievement dataset included 532 cases, 94 of which (17.6 percent) were associated with missing predicted scores. The fifth grade reading achievement dataset included 551 cases, 57 of which (10.3 percent) were associated with missing predicted scores. Conversations with the provider of the predicted score data, Battelle for Kids, indicated that the value-added assessment system does not issue a predicted scale reading achievement score for students with an insufficient number of achievement test scores from which to generate a predicted score; in most cases, this occurs because the child has no such scores on record due to grade level or mobility. In the current study, the level of missing data stems from both causes. Clearly, this problem confronts schools within the district differently, and poses a concern for the statistical analyses and evaluation observations which follow.

Missing cases were examined by school in each dataset. In the fourth grade, missing predicted scores by school ranged from 2.1 percent (Youtz) to 42.1 percent (Schreiber). In the fifth grade, missing predicted scores by school ranged from 4 percent (Belle Stone) to 21.1 percent (Schreiber). A review of Tables 4 and 8 indicate that there appears to be no correlation between missing predictors and the number of students per school or the average reading achievement within the school. In line with Battelle for
Kids’ assertions, there are fewer missing predictor scores overall in the fifth grade than in the fourth grade. Multivariate outliers (predicted and observed scale scores) were examined for data entry errors, subject to population misfits, and individual differences (Tabachnick & Fidell, 1996). No such outlier was removed from further analyses.

**Descriptive Data**

The participants include students in grades 4-5 from the eleven elementary schools of the District judged by the state and federal governments to be ‘at-risk’ or in ‘school improvement’ status during the 2004-05 school year. Demographic analyses of the students in the grade four and five samples are presented in Tables 2 and 3.

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Gender</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>American Indian</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Black/African American</td>
<td>96</td>
<td>113</td>
<td>209</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>121</td>
<td>134</td>
<td>255</td>
<td></td>
</tr>
<tr>
<td>Multi-racial</td>
<td>35</td>
<td>27</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>254</td>
<td>278</td>
<td>532</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Fourth-grade student demographics, 2004-05
### Table 3: Fifth-grade student demographics, 2004-05

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Black/African American</td>
<td>104</td>
<td>122</td>
<td>226</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>White</td>
<td>120</td>
<td>135</td>
<td>255</td>
</tr>
<tr>
<td>Multi-racial</td>
<td>25</td>
<td>37</td>
<td>62</td>
</tr>
<tr>
<td>Total</td>
<td>251</td>
<td>300</td>
<td>551</td>
</tr>
</tbody>
</table>

**Fourth-grade student reading achievement**

An unconditional model was run in HLM-6 with restricted maximum likelihood estimation as a null model wherein fourth grade student reading achievement scale scores was the outcome variable. Using the reading achievement datasets in the eleven schools of interest, the model was run to generate the appropriate fixed effects estimates. Table 4 indicates the variability between schools in fourth-grade scale scores of reading achievement and identifies the percent of fourth grade students who scored at or above the proficient level in the March 2005 administration of the Ohio Achievement Test in reading; the percent of all fourth-grade students in the targeted schools scoring at or above the proficient level is 55.5 percent (295/532).
<table>
<thead>
<tr>
<th>School</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Predicted Scores</th>
<th>Percent at and Above Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>48</td>
<td>397.4</td>
<td>23.2</td>
<td>14.6%</td>
<td>50.0</td>
</tr>
<tr>
<td>Belden</td>
<td>45</td>
<td>403.6</td>
<td>29.3</td>
<td>31.1%</td>
<td>62.8</td>
</tr>
<tr>
<td>Belle Stone</td>
<td>68</td>
<td>400.2</td>
<td>25.1</td>
<td>14.7%</td>
<td>54.1</td>
</tr>
<tr>
<td>Compton</td>
<td>31</td>
<td>396.5</td>
<td>31.5</td>
<td>9.7%</td>
<td>48.3</td>
</tr>
<tr>
<td>Dueber</td>
<td>35</td>
<td>393.4</td>
<td>30.2</td>
<td>20.0%</td>
<td>44.4</td>
</tr>
<tr>
<td>Gibbs</td>
<td>44</td>
<td>405.6</td>
<td>28.7</td>
<td>25.0%</td>
<td>51.4</td>
</tr>
<tr>
<td>Harter</td>
<td>67</td>
<td>430.3</td>
<td>27.0</td>
<td>16.4%</td>
<td>78.4</td>
</tr>
<tr>
<td>McGregor</td>
<td>51</td>
<td>402.1</td>
<td>27.4</td>
<td>7.8%</td>
<td>57.4</td>
</tr>
<tr>
<td>Schreiber</td>
<td>38</td>
<td>405.2</td>
<td>24.0</td>
<td>42.1%</td>
<td>60.0</td>
</tr>
<tr>
<td>Summit</td>
<td>48</td>
<td>388.5</td>
<td>29.9</td>
<td>9.6%</td>
<td>33.3</td>
</tr>
<tr>
<td>Youtz</td>
<td>57</td>
<td>402.4</td>
<td>22.8</td>
<td>2.1%</td>
<td>56.1</td>
</tr>
<tr>
<td>Sum</td>
<td>532</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>402.3</td>
<td>27.2</td>
<td>17.7%</td>
<td>54.2</td>
</tr>
</tbody>
</table>

Table 4: Observed fourth-grade reading scale scores across schools, 2004-05

In Table 4, great variability is observed in proficiency levels when the unit of analysis is at the school level. Table 5 shows the estimates for an unconstrained hierarchical linear model of fourth grade student reading achievement across schools.

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average School Mean</td>
<td>γ₀₀</td>
<td>402.42</td>
<td>3.30</td>
<td>121.94</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean</td>
<td>μᵢⱼ</td>
<td>10.20</td>
<td>10</td>
<td>93.18</td>
<td>0.000</td>
</tr>
<tr>
<td>Student Residual</td>
<td>rᵢⱼ</td>
<td>26.97</td>
<td>727.18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Unconstrained model of fourth-grade student reading achievement, 2004-05
The estimate for the grand-mean reading achievement is 402.42 with a standard error of 3.30; the cut scale score point for the proficient level of this assessment is \( \geq 400 \). The estimated variability in the school means is 10.20. According to the value of the chi-square statistic, there appears to be significant variation among schools in their reading achievement scale scores \( (\chi^2(10) = 93.18, p = .000) \), a finding that was foreshadowed in Table 4. The deviance of the model represents one measure of the lack of fit between the model and the data (Luke, 2004); in the unconstrained model, deviance is equal to 5030.66. While this value may not be informative in and of itself, this value can be compared to other models that utilize the same data in order to compare models to one another.

Calculating an intraclass correlation coefficient (ICC) allows for a measurement of the proportion of variance in the dependent variable that is accounted for by level-2 units (Luke, 2004). Other authors refer to this as the cluster effect (Raudenbush & Bryk, 2002). In this model, schools accounted for 13 percent of the variance in the student reading achievement \( (\frac{104.09}{727.18 + 104.09}) \). This is a moderately low value for an ICC, and lower than that which was hypothesized (i.e., between 15 to 19 percent). Given the nature of the intervention, a full hierarchical model was pursued with the prediction that the ICC value would increase relative to the unconstrained model due to school effects.

Following this first analysis, a scatterplot of the predicted and actual scale scores in fourth grade reading achievement by school identifier (IRN) was generated in SPSS 14.0 for Windows.
Figure 3: Predicted and observed fourth-grade reading achievement scale scores across schools, 2004-05

Figure 3 demonstrates the strong, linear, positive correlation between the predictor, predicted reading achievement scale score, and the outcome, reading achievement scale score across schools ($r = .79, p < .01$), providing support in this dataset for the first hypothesis of the study. Here, the coefficient of determination indicates that 62 percent of the variability in fourth-grade student reading achievement scale scores can be predicted from the relationship with the state’s value-added prediction scores. In the figure, reference lines on the x- and y-axis have been imposed on the scatterplot at the cut scale score for the proficient level. In Figure 3, quadrant 1 (top right) contains the values of
students who were predicted and observed at or above the proficient level, while quadrant 3 (bottom left) contains the values of those that were predicted and observed to score below the proficient level. A review of the false negative scores in quadrant 2 (top left) reveals the paucity of points in this category. The greatest opportunity for student achievement and growth, then, lies in the false positive points in quadrant 4 (bottom right). Here, the school IRN markers delineate the spread of students by school across categories with Harter (IRN 015123) dominating quadrant 1.

In a related analysis, a contingency table was constructed in order to empirically test the second hypothesis of the study, that the predicted achievement scale scores generated by the state’s value-added assessment system would correctly classify students who score at or above proficient correctly no more than 80 percent of the time in the baseline year. Table 6 demonstrates that this hypothesis is supported by the observations in the fourth grade dataset.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Below</td>
<td>Proficient</td>
<td>Proficient</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proficient</td>
<td>or Higher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below Proficient</td>
<td>123</td>
<td>29</td>
<td>152</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient or Higher</td>
<td>74</td>
<td>212</td>
<td>286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>197</td>
<td>241</td>
<td>438</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Contingency table of predicted and observed proficiency status, fourth grade, 2004-05

By examining Table 6, it can be demonstrated that the predictive proficiency model utilized in the state’s value-added assessment system correctly classified 74
percent of students (212/286) as proficient, with 26 percent of students (74/286) registered as false positives in this dataset. Alternatively, the system classified 81 percent (123/152) of students as not proficient, with 19 percent (29/152) registered as false negatives in this dataset. It should be noted that the number of students with predicted scores generated from the value-added assessment system (438) is far less than the total number of students who took the fourth-grade reading achievement test (532) for reasons noted above.

In all, fifty-five percent (295/532) of fourth-grade students did not achieve the state standard for reading at or above the proficient level. If the intervention to be implemented (i.e., value-added and standardized formative assessments systems) is successful, the percent of students failing to achieve this standard should significantly decrease. To be successful, the District will have to work with each of its elementary schools to help 75 percent of its students to reach the state standard for reading at or above the proficient level. A review of Table 6 indicates that the growth necessary to achieve this goal at each of the target schools ranges from 41.7 (75.0 – 33.3) at Summit to 12.2 (75.0 – 62.8) at Belden; 78.4 percent of the students in one of the targeted schools (Harter) scored above the state reading standard in the fourth grade.

An analysis of the variability within and between schools in observed reading achievement scale scores across schools is evident in Figure 4. Here, the theoretical justification for a nested model of student achievement become evident, as patterns of student achievement within and between schools demonstrate the variability below district reading assessment scale score averages that a formative assessment system seeks to minimize. In Figure 4, the variability seen in Table 6 is expanded to include students
scoring at the 25th and 75th percentiles (the bottom and top of the boxes, respectively), the
median (the horizontal line within the boxes), outliers (circles) and extreme values (stars).
Here, the variability in observed scale scores is striking; the difference in the level and
range of reading achievement between Harter Elementary and Summit is made plain.

Figure 4: Fourth-grade student reading achievement scale scores, 2004-05

A one-way random effects ANCOVA was run in HLM-6 with restricted
maximum likelihood estimation with student scores in level-1 and schools in level-2; the
outcome variable was the scale score of the fourth grade reading achievement
assessment. Diagnostically, it is interesting to note that only four iterations were required for convergence of the EM algorithm, indicating that the prediction scores were highly informative (Raudenbush & Byrk, 2002). Table 7 displays the estimates for the one-way random effects ANCOVA model of student reading achievement.

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\gamma_{00}$</td>
<td>402.32</td>
<td>3.62</td>
<td>111.06</td>
<td>10</td>
</tr>
<tr>
<td>Slope</td>
<td>$\gamma_{10}$</td>
<td>0.98</td>
<td>0.04</td>
<td>24.5</td>
<td>436</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean</td>
<td>$\mu_{0j}$</td>
<td>11.73</td>
<td>137.61</td>
<td>10</td>
<td>271.32</td>
</tr>
<tr>
<td>Student Residual</td>
<td>$r_{ij}$</td>
<td>15.69</td>
<td>246.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: One-way ANCOVA with random effects HLM of fourth-grade student reading achievement, 2004-05

The estimate for the expected value of reading achievement is 402.32 with a standard error of 3.62; again, the cut scale score point for the proficient level of this assessment is $\geq 400$. This indicates that students in schools where the predicted reading score is at the classroom mean, students are expected to have a scale reading achievement score of 402.32. There appears to be a significant positive relationship between scale reading achievement scores and predicted reading achievement scale scores, ($t = 24.50, p = .000$). We can expect, on average, to see approximately a .98-point increase in scale reading achievement score as the mean predicted reading achievement scale score increases by one point.
According to Table 7, the estimated variability among school means is 11.73. There is significant variation among schools in their observed scale score reading achievement ($\chi^2(10) = 271.32, p = .000$). Schools accounted for about 39 percent of the variance in the student reading achievement (137.61/[137.61+246.08]) when predicted scale scores are added as a level-1 predictor, a significant finding. Moreover, the deviance of the model is equal to 3686.79, indicating that this model with its lower deviance value represents a better fit with the data. This finding provides support in this dataset for the study’s fourth hypothesis.

Given that predictors were utilized in only the first level of the one-way random effects ANCOVA model to generate baseline data, two assumptions were tested: homogeneity of variance and normality (Raudenbush & Bryk, 2002). A test of the assumption of homogeneity of variance was run; common variance within each of the $J$ level-2 units was found ($\chi^2 (10) = 7.32, p > .500$). If the variance had been found to depend systematically as a function of the predictors at level-1, an inference of bias among level-2 elements would have been justified.

To examine the tenability of the assumption of a normal distribution of level-1 errors in the one-way random effects ANCOVA model, a residual analysis was conducted to identify violations which could negatively impact the estimated standard errors for fixed effects estimates (Raudenbush et al., 2004) and biased the estimation of level-2 effects (Raudenbush & Byrk, 2002). The Q-Q plot of level-1 residuals in Figure 5 indicates that no such violation was found. This being done, the analysis turned to fifth grade reading achievement in the 2004-05 school year.
Figure 5: Level-1 residuals in the one-way random effects ANCOVA HLM of fourth-grade reading achievement, 2004-05

Fifth-grade student reading achievement

An unconditional model was run in HLM-6 with restricted maximum likelihood estimation as a null model wherein fifth grade student reading achievement scale scores was the outcome variable. Using the reading achievement datasets in the eleven schools of interest, the model was run to generate the appropriate fixed effects estimates. Table 8 indicates the great variability observed between schools in fifth-grade scale scores of reading achievement and identifies the percent of fourth grade students who scored at and above the proficient level in the Ohio Achievement Test in reading; the percent of all
fifth-grade students in the targeted schools scoring at or above the proficient level is 52.4 percent (265/506).

<table>
<thead>
<tr>
<th>School</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Predicted Scores</th>
<th>Percent at and Above Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>49</td>
<td>400.8</td>
<td>26.7</td>
<td>16.3%</td>
<td>61.4</td>
</tr>
<tr>
<td>Belden</td>
<td>58</td>
<td>398.6</td>
<td>22.0</td>
<td>5.2%</td>
<td>54.5</td>
</tr>
<tr>
<td>Belle Stone</td>
<td>50</td>
<td>391.6</td>
<td>27.2</td>
<td>4.0%</td>
<td>54.8</td>
</tr>
<tr>
<td>Compton</td>
<td>45</td>
<td>393.1</td>
<td>23.6</td>
<td>8.9%</td>
<td>40.5</td>
</tr>
<tr>
<td>Dueber</td>
<td>41</td>
<td>396.6</td>
<td>31.9</td>
<td>12.2%</td>
<td>57.1</td>
</tr>
<tr>
<td>Gibbs</td>
<td>48</td>
<td>394.6</td>
<td>26.8</td>
<td>10.4%</td>
<td>42.9</td>
</tr>
<tr>
<td>Harter</td>
<td>74</td>
<td>426.7</td>
<td>25.2</td>
<td>13.5%</td>
<td>83.9</td>
</tr>
<tr>
<td>McGregor</td>
<td>57</td>
<td>391.1</td>
<td>26.5</td>
<td>10.5%</td>
<td>46.3</td>
</tr>
<tr>
<td>Schreiber</td>
<td>38</td>
<td>403.1</td>
<td>30.6</td>
<td>21.1%</td>
<td>63.3</td>
</tr>
<tr>
<td>Summit</td>
<td>50</td>
<td>393.3</td>
<td>29.6</td>
<td>8.0%</td>
<td>45.5</td>
</tr>
<tr>
<td>Youtz</td>
<td>41</td>
<td>394.9</td>
<td>24.5</td>
<td>4.9%</td>
<td>48.6</td>
</tr>
</tbody>
</table>

**Sum**     | 551|
**Mean**    | 398.6| 26.8| 10.5%| 54.4|

Table 8: Observed fifth-grade reading achievement scale scores across schools, 2004-05

Table 9 shows the estimates for an unconstrained hierarchical linear model of fifth-grade student reading achievement across schools. The estimate for the grand-mean math achievement is 398.68 with a standard error of 3.14. The estimated variability in the classroom means is 9.70. There appears to be significant variation among classrooms in their reading achievement scale scores ($\chi^2(10) = 97.12, p = .000$). As was the case in the fourth grade model, the deviance in the unconstrained model is quite high, 5198.93. This value will be used in comparison with an ANCOVA with random effects to test the relative fit of the models.
Calculating an intraclass correlation coefficient (ICC) allows for a measurement of the proportion of variance in the dependent variable that is accounted for by level-2 units (Luke, 2004). In this model, schools accounted for 12 percent of the variance in the student reading achievement (94.05/[710.93 + 94.05]). Again, this dataset does not provide support for the study’s third hypothesis. Given the nature of the study, however, a full hierarchical model was pursued with the prediction that the ICC value would increase relative to the unconstrained model due to school effects.

Following this first analysis, a scatterplot of the predicted and actual scale scores in fifth grade reading achievement by school identifier (IRN) was generated in SPSS 14.0 for Windows. Figure 6 demonstrates the strong, linear, positive correlation across schools between the predictor, predicted reading achievement scale score, and the outcome, reading achievement scale score ($r = .81$, $p < .01$). This finding provides support in this dataset for the study’s first hypothesis. Here, the coefficient of determination indicates that 66 percent of the variability in fifth-grade student reading achievement scale scores can be predicted from the relationship with the state’s value-added prediction scores. In
the figure, reference lines on the x- and y-axis have been imposed on the scatterplot at the cut scale score for the proficient level.

In Figure 6, quadrant 1 (top right) contains the values of students who were predicted and observed at or above the proficient level, while quadrant 3 (bottom left) contains the values of those that were predicted and observed to score below the proficient level. A review of the false negative scores in quadrant 2 (top left) reveals the paucity of points in this category. The greatest opportunity for student achievement and growth, then, lies in the false positive points in quadrant 4 (bottom right). In quadrant 4, these opportunities are distributed throughout the targeted schools in the intervention.
A contingency table analysis demonstrates that the predicted achievement scale scores generated by the state’s value-added assessment system correctly classified students who score at the proficient level or higher correctly less than 80 percent of the time in the baseline year (Table 10). By examining Table 10, it can be demonstrated that the predictive proficiency model utilized in the state’s value-added assessment system correctly classified 76 percent of students (247/325) as proficient, with 24 percent of students (78/325) registered as false positives in this dataset. Alternatively, the system classified 91 percent (153/169) of students as not proficient, with 9 percent (16/169)
registered as false negatives in this dataset. Thus, these findings provide support for the study’s second hypothesis.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed Below Proficient</th>
<th>Proficient or Higher</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Proficient</td>
<td>153</td>
<td>16</td>
<td>169</td>
</tr>
<tr>
<td>Proficient or Higher</td>
<td>78</td>
<td>247</td>
<td>325</td>
</tr>
<tr>
<td>Total</td>
<td>231</td>
<td>263</td>
<td>494</td>
</tr>
</tbody>
</table>

Table 10: Contingency table of predicted and observed proficiency status, fifth grade

If the intervention to be implemented (i.e., value-added and standardized formative assessment systems) is successful, the percent of students failing to achieve this standard should significantly decrease. To be successful, the District will have to work with each of its elementary schools to help 75 percent of its students to reach the state standard for reading at or above the proficient level. A review of Table 8 indicates that the growth necessary to achieve this goal at each of the target schools ranges from 34.5 (75.0 - 40.5) at Compton to 11.7 (75.0 - 63.3) at Schreiber; 83.9 percent of the students in one of the targeted schools (Harter) scored above the state reading standard in the fifth grade.

As was the case in the fourth grade, a visual analysis of the variability within and between schools in observed reading achievement scale scores across schools is evident in the fifth-grade student scores in Figure 7. Here again, the theoretical justification for a nested model of student achievement become evident, as patterns of student achievement
within and between schools demonstrate the variability below district reading assessment scale score averages that a formative assessment system seeks to minimize.

In Figure 7, the variability seen in Table 8 is expanded to include students scoring at the 25th and 75th percentiles (the bottom and top of the boxes, respectively), the median (the horizontal line within the boxes) and outliers (circles). In the fifth grade, the consistently average performance level of students around the proficient level across schools is clearly observable.

Figure 7: Fifth-grade student reading achievement scale scores, 2004-05
A one-way random effects ANCOVA was run in HLM-6 with restricted maximum likelihood estimation with predicted student scale scores in level-1 and schools in level-2; the outcome variable was the observed scale score of the fifth grade reading achievement assessment. Only three iterations were required for convergence of the EM algorithm, indicating that the prediction scores were highly informative (Raudenbush & Byrk, 2002).

Table 11 shows the estimates for the one-way ANCOVA with random effects model of student reading achievement. The estimate for the expected value of reading achievement is 398.05 with a standard error of 3.29. This indicates that students in schools where the predicted reading score is at the classroom mean, students are expected to have a scale reading achievement score of 398.05. There appears to be a significant positive relationship between observed scale reading achievement scores and predicted reading achievement scale scores, ($t = 26.69, p = .000$). We can expect, on average, to see approximately a 1.03-point increase in scale reading achievement score as the mean predicted reading achievement scale score increases by one point.

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\gamma_{00}$</td>
<td>398.05</td>
<td>3.29</td>
<td>120.87</td>
<td>10</td>
</tr>
<tr>
<td>Slope</td>
<td>$\gamma_{10}$</td>
<td>1.03</td>
<td>0.04</td>
<td>26.69</td>
<td>492</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean</td>
<td>$\mu_{ij}$</td>
<td>10.66</td>
<td>113.63</td>
<td>10</td>
<td>283.59</td>
</tr>
<tr>
<td>Student Residual</td>
<td>$r_{ij}$</td>
<td>15.7</td>
<td>246.34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: One-way ANCOVA with random effects HLM of fifth-grade student reading achievement, 2004-05
The estimated variability among school means is 10.66. There is significant variation among schools in their observed scale score reading achievement ($\chi^2(10) = 283.59, p = .000$). Schools accounted for about 32 percent of the variance in the student reading achievement ($113.63/[246.64 + 113.63]$) when predicted scale scores are added as a level-1 predictor, a significant finding. Moreover, the deviance of the model is equal to 4154.33, indicating that this model with its lower deviance value represents a better fit with the data. These findings provide support for the study’s fourth hypothesis, though the ICC is larger than originally predicted.

Given that predictors were utilized in only the first level of the one-way random effects ANCOVA model to generate baseline data, two assumptions were tested: homogeneity of variance and normality (Raudenbush & Bryk, 2002). A test of the assumption of homogeneity of variance was run; common variance within each of the $J$ level-2 units was found ($\chi^2(10) = 10.90, p = .365$). To examine the tenability of the assumption of a normal distribution of level-1 errors in the one-way random effects ANCOVA model, a residual analysis was conducted to identify violations which could negatively impact the estimated standard errors for fixed effects estimates (Raudenbush et al., 2004). The Q-Q plot of level-1 residuals in Figure 8 indicates that no significant violation was found, though a little wobble is in evidence at both tails of the distribution.
Figure 8: Level-1 residuals in the one-way random effects ANCOVA HLM of fifth-grade reading achievement, 2004-05

*School Year 2005-06*

As was the case in the previous dataset, data were screened to examine the existence and potential impact of missing data and multivariate outliers (Mertler & Vannatta, 2002). In the 2005-06 school year, the fourth grade reading achievement dataset included 490 cases, 123 of which (25.1 percent) were associated with missing predicted scores; the fifth grade reading achievement dataset included 506 cases, 76 of which (15 percent) were associated with missing predicted scores.

Missing cases were then examined by school in each dataset. Fourth grade missing predicted scores by school ranged from 13.9 percent (Compton) to 39.0 percent (Schreiber), while fifth grade missing predicted scores by school ranged from 2 percent
(McGregor) to 33.3 percent (Compton). A review of Tables 14 and 18 indicate that there appears to be no correlation between missing predictors and the number of students per school or the average reading achievement within the school. In line with Battelle for Kids’ assertions, there are fewer missing predictor scores overall in the fifth grade than in the fourth grade. Multivariate outliers (predicted and observed scale scores) were examined for data entry errors, subject to population misfits, and individual differences (Tabachnick & Fidell, 1996). No outlier was removed from further analyses.

**Descriptive Data**

Demographic analysis of 2005-06 students in grades four and five are presented in Tables 12 and 13. The demographic pattern observed in the 2004-05 cohort is replicated in the present cohort.

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Gender</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Total</td>
</tr>
<tr>
<td>American Indian</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Black/African American</td>
<td>88</td>
<td>117</td>
<td>205</td>
</tr>
<tr>
<td>White</td>
<td>107</td>
<td>114</td>
<td>221</td>
</tr>
<tr>
<td>Multi-racial</td>
<td>31</td>
<td>32</td>
<td>63</td>
</tr>
<tr>
<td>Total</td>
<td>226</td>
<td>264</td>
<td>490</td>
</tr>
</tbody>
</table>

Table 12: Fourth-grade student demographics, 2005-06
<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Gender</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Total</td>
</tr>
<tr>
<td>American Indian</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Black/African American</td>
<td>101</td>
<td>106</td>
<td>207</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>White</td>
<td>111</td>
<td>120</td>
<td>231</td>
</tr>
<tr>
<td>Multi-racial</td>
<td>31</td>
<td>30</td>
<td>61</td>
</tr>
<tr>
<td>Total</td>
<td>244</td>
<td>262</td>
<td>506</td>
</tr>
</tbody>
</table>

Table 13: Fifth-grade student demographics, 2005-06

*Fourth-grade student reading achievement*

An unconditional model was run in HLM-6 with restricted maximum likelihood estimation as a null model wherein fourth grade student reading achievement scale scores was the outcome variable. Using the reading achievement datasets in the eleven schools of interest, the model was run to generate the appropriate fixed effects estimates. Table 14 indicates the variability between schools in fourth-grade scale scores of reading achievement and identifies the percent of fourth grade students who scored at or above the proficient level in the March 2006 administration of the Ohio Achievement Test in reading. Across the targeted schools, the percent of fourth-grade students scoring at or above the proficient level is 58.6 percent (287/490).
### Table 14: Observed fourth-grade reading scale scores across schools, 2005-06

<table>
<thead>
<tr>
<th>School</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Missing Predicted Scores</th>
<th>Percent at and Above Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>54</td>
<td>401.8</td>
<td>25.6</td>
<td>18.5%</td>
<td>57.4%</td>
</tr>
<tr>
<td>Belden</td>
<td>54</td>
<td>393.8</td>
<td>25.7</td>
<td>27.8%</td>
<td>42.6%</td>
</tr>
<tr>
<td>Belle Stone</td>
<td>45</td>
<td>407.3</td>
<td>34.7</td>
<td>20.0%</td>
<td>60.0%</td>
</tr>
<tr>
<td>Compton</td>
<td>36</td>
<td>411.0</td>
<td>27.8</td>
<td>13.9%</td>
<td>75.0%</td>
</tr>
<tr>
<td>Dueber</td>
<td>46</td>
<td>415.3</td>
<td>30.2</td>
<td>30.4%</td>
<td>76.1%</td>
</tr>
<tr>
<td>Gibbs</td>
<td>44</td>
<td>396.9</td>
<td>34.7</td>
<td>27.3%</td>
<td>47.7%</td>
</tr>
<tr>
<td>Harter</td>
<td>68</td>
<td>417.9</td>
<td>37.4</td>
<td>16.2%</td>
<td>64.7%</td>
</tr>
<tr>
<td>McGregor</td>
<td>36</td>
<td>413.9</td>
<td>24.2</td>
<td>36.1%</td>
<td>66.7%</td>
</tr>
<tr>
<td>Schreiber</td>
<td>41</td>
<td>404.7</td>
<td>23.1</td>
<td>39.0%</td>
<td>58.5%</td>
</tr>
<tr>
<td>Summit</td>
<td>39</td>
<td>398.6</td>
<td>31.2</td>
<td>25.6%</td>
<td>46.2%</td>
</tr>
<tr>
<td>Youtz</td>
<td>27</td>
<td>408.7</td>
<td>36.1</td>
<td>29.6%</td>
<td>48.1%</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>490</td>
<td></td>
<td></td>
<td></td>
<td>58.5%</td>
</tr>
</tbody>
</table>

Table 14: Observed fourth-grade reading scale scores across schools, 2005-06

Table 15 shows the estimates for the unconstrained hierarchical linear modeling model of fourth grade student reading achievement across schools.

### Table 15: Unconstrained model of fourth-grade student reading achievement, 2005-06

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average School Mean</td>
<td>$\gamma_{00}$</td>
<td>406.33</td>
<td>2.50</td>
<td>162.54</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean</td>
<td>$\mu_{ij}$</td>
<td>6.82</td>
<td>46.50</td>
<td>10</td>
<td>33.43</td>
</tr>
<tr>
<td>Student Residual</td>
<td>$r_{ij}$</td>
<td>30.90</td>
<td>954.54</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 15: Unconstrained model of fourth-grade student reading achievement, 2005-06
The estimate for the grand-mean math achievement is 406.33 with a standard error of 2.50; the cut scale score point for the proficient level of this assessment is $\geq 400$. The estimated variability in the school means is 6.82. According to the value of the chi-square statistic, there appears to be significant variation among schools in their reading achievement scale scores ($\chi^2(10) = 33.43, p = .000$). The deviance of the model represents one measure of the lack of fit between the model and the data (Luke, 2004); in the unconstrained model, deviance is equal to 4758.61. While this value may not be informative in and of itself, this value can be compared to other models that utilize the same data to test the relative fit in order to compare models to one another.

Calculating an intraclass correlation coefficient (ICC) allows for a measurement of the proportion of variance in the dependent variable that is accounted for by level-2 units (Luke, 2004). In this model, schools accounted for 4.6 percent of the variance in the student reading achievement ($46.50/\left[954.54 + 46.50\right]$). Again, this finding does not provide support for the study’s third hypothesis in this dataset. The ICC obtained is a low value for an ICC, suggesting that a multilevel model incorporating second-level variables may not be warranted. Given the nature of the study, however, a full hierarchical model was pursued with the prediction that the ICC value would increase relative to the unconstrained model due to school effects.

Following this first analysis, a scatterplot of the predicted and actual scale scores in fourth grade reading achievement by school identifier (IRN) was generated (Figure 9). Figure 9 demonstrates the strong, linear, positive correlation between the predictor, predicted reading achievement scale score, and the outcome, reading achievement scale
score across schools \((r = .81, p < .01)\); support for the study’s first hypothesis is thus provided for in this dataset. Here, the coefficient of determination indicates that 66 percent of the variability in fourth-grade student reading achievement scale scores can be predicted from the relationship with the state’s value-added prediction scores.

Figure 9: Predicted and observed fourth-grade reading achievement scale scores across schools, 2005-06

In Figure 9, reference lines on the x- and y-axis have been imposed on the scatterplot at the cut scale score for the proficient level. In Figure 9, quadrant 1 (top right) contains the values of students who were predicted and observed at or above the
proficient level, while quadrant 3 (bottom left) contains the values of those that were predicted and observed to score below the proficient level. A review of the false negative scores in quadrant 2 (top left) reveals the paucity of points in this category. The greatest opportunity for student achievement and growth, then, lies in the false positive points in quadrant 4 (bottom right).

In a related analysis, a contingency table was constructed in order to empirically test the second hypothesis of the study, that the predicted achievement scale scores generated by the state’s value-added assessment system would correctly classify students who score at or above proficient correctly no more than 80 percent of the time in the baseline year. Table 16 demonstrates that this hypothesis is not supported by the observations in the fourth grade dataset.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proficient or Higher</td>
<td>Below</td>
</tr>
<tr>
<td>Predicted</td>
<td>Below Proficient</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Proficient or Higher</td>
<td>185</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>209</td>
</tr>
</tbody>
</table>

Table 16: Contingency table of predicted and observed proficiency status, fourth grade, 2005-06

By examining Table 16, it can be demonstrated that the predictive proficiency model utilized in the state’s value-added assessment system correctly classified 82 percent of students (185/226) as proficient, with 18 percent of students (41/226) registered as false positives in this dataset. Alternatively, the system correctly classified
83 percent (117/141) of students as not proficient, with 17 percent (24/141) registered as false negatives in this dataset. The number of students with a predicted score (367) was far less than the number of students who took the fourth grade reading achievement tests (490), however, for reasons pointed out above.

Overall, forty-one percent (203/490) of fourth-grade students did not achieve the state standard for reading at or above the proficient level. If the full intervention (i.e., value-added and standardized formative assessments systems) to be implemented is successful, the percent of students failing to achieve this standard should significantly decrease. To be successful, the District will have to work with each of its elementary schools to help 75 percent of its students to reach the state standard for reading at or above the proficient level. A review of Table 10 indicates that the growth necessary to achieve this goal at each of the target schools ranges from 32 (75 – 43) at Belden to 8 (75 – 67) at McGregor; two schools, Compton and Dueber, scored at or above the state reading standard in the fourth grade.

An analysis of the variability within and between schools in observed reading achievement scale scores across schools is evident in Figure 10. Here, the theoretical justification for a nested model of student achievement become evident, as patterns of student achievement within and between schools demonstrate the variability below district reading assessment scale score averages that a formative assessment system seeks to minimize.
In Figure 10, the variability seen in Table 14 is expanded to include students scoring at the 25th and 75th percentiles (the bottom and top of the boxes, respectively), the median (the horizontal line within the boxes), outliers (circles) and extreme values (stars). Here, the variability in observed scale scores is striking; the difference in the level and range of reading achievement between Harter Elementary and Summit is made plain.

A one-way random effects ANCOVA was run in HLM-6 with restricted maximum likelihood estimation with student scores in level-1 and schools in level-2; the outcome variable was the scale score of the fourth grade reading achievement.
assessment. In this dataset, only three iterations were required for convergence of the EM algorithm, indicating that the prediction scores were highly informative (Raudenbush & Byrk, 2002). Table 17 shows the estimates for the one-way random effects ANCOVA model of student reading achievement.

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Score</td>
<td>$\gamma_{00}$</td>
<td>404.16</td>
<td>3.30</td>
<td>122.51</td>
<td>10</td>
</tr>
<tr>
<td>Predicted Score</td>
<td>$\gamma_{01}$</td>
<td>1.12</td>
<td>0.05</td>
<td>24.46</td>
<td>365</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>Variance</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean</td>
<td>$\mu_{ij}$</td>
<td>SD</td>
<td>10</td>
<td>110.25</td>
<td>10</td>
</tr>
<tr>
<td>Student Residual</td>
<td>$r_{ij}$</td>
<td>Variance</td>
<td>17.10</td>
<td>292.51</td>
<td></td>
</tr>
</tbody>
</table>

Table 17: One-way ANCOVA with random effects HLM of fourth-grade student reading achievement, 2005-06

The estimate for the expected value of reading achievement is 404.16 with a standard error of 3.30; again, the cut scale score point for the proficient level of this assessment is $\geq 400$. This indicates that students in schools where the predicted reading score is at the classroom mean, students are expected to have a scale reading achievement score of 404.16. There appears to be a significant positive relationship between scale reading achievement scores and predicted reading achievement scale scores, $(t = 24.46, p = .000)$. We can expect, on average, to see approximately a 1.12-point increase in scale reading achievement score as the mean predicted reading achievement scale score increases by one point.
The estimated variability among school means is 10.50. There is significant variation among schools in their observed scale score reading achievement ($\chi^2(10) = 146.29, p = .000$). Schools accounted for about 27 percent of the variance in the student reading achievement ($110.25/[292.51+110.25]$) when predicted scale scores are added as a level-1 predictor, a significant finding. This finding provides support in this dataset for the study’s fourth hypothesis. Moreover, the deviance of the model is equal to 3150.05, indicating that this model with its lower deviance value represents a better fit with the data.

Given that predictors were utilized in only the first level of the one-way random effects ANCOVA model to generate baseline data, two assumptions were tested: homogeneity of variance and normality (Raudenbush & Bryk, 2002). The software program utilized, HLM6, assumes homogeneity of residual variance at level-1. A test of the assumption of homogeneity of variance was run; common variance within each of the $J$ level-2 units was found ($\chi^2(10) = 12.32, p > .263$).

To examine the tenability of the assumption of a normal distribution of level-1 errors in the one-way random effects ANCOVA model, a residual analysis was conducted to identify violations which could negatively impact the estimated standard errors for fixed effects estimates (Raudenbush et al., 2004). The Q-Q plot of level-1 residuals in Figure 11 indicates that no such violation was found.
Figure 11: Level-1 residuals in the one-way random effects ANCOVA HLM of fourth-grade reading achievement, 2005-06

Fifth-grade student reading achievement

Using the 2005-06 dataset, an unconditional model was run in HLM-6 with restricted maximum likelihood estimation as a null model wherein fifth grade student reading achievement scale scores was the outcome variable. Using the reading achievement datasets in the eleven schools of interest, the model was run to generate the appropriate fixed effects estimates. Table 18 indicates the variability between schools in fifth-grade scale scores of reading achievement and identifies the percent of fifth grade students who scored at and above the proficient level in the Ohio Achievement Test in reading; the percent of fifth-grade students scoring at or above the proficient level is 52.4 percent (265/506).

138
<table>
<thead>
<tr>
<th>School</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Missing Predicted Scores</th>
<th>Percent at and Above Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>46</td>
<td>389.6</td>
<td>26.3</td>
<td>15.2%</td>
<td>32.6%</td>
</tr>
<tr>
<td>Belden</td>
<td>41</td>
<td>396.2</td>
<td>27.9</td>
<td>19.5%</td>
<td>43.9%</td>
</tr>
<tr>
<td>Belle Stone</td>
<td>52</td>
<td>398.9</td>
<td>28.4</td>
<td>11.5%</td>
<td>48.1%</td>
</tr>
<tr>
<td>Compton</td>
<td>33</td>
<td>391.8</td>
<td>29.6</td>
<td>33.3%</td>
<td>30.3%</td>
</tr>
<tr>
<td>Dueber</td>
<td>37</td>
<td>393.6</td>
<td>31.2</td>
<td>24.3%</td>
<td>48.6%</td>
</tr>
<tr>
<td>Gibbs</td>
<td>35</td>
<td>407.8</td>
<td>29.6</td>
<td>5.7%</td>
<td>54.3%</td>
</tr>
<tr>
<td>Harter</td>
<td>60</td>
<td>438.6</td>
<td>32.4</td>
<td>11.7%</td>
<td>83.3%</td>
</tr>
<tr>
<td>McGregor</td>
<td>51</td>
<td>410.8</td>
<td>33.8</td>
<td>2.0%</td>
<td>68.6%</td>
</tr>
<tr>
<td>Schreiber</td>
<td>41</td>
<td>394.0</td>
<td>28.8</td>
<td>17.1%</td>
<td>41.5%</td>
</tr>
<tr>
<td>Summit</td>
<td>48</td>
<td>391.1</td>
<td>35.2</td>
<td>25.0%</td>
<td>43.8%</td>
</tr>
<tr>
<td>Youtz</td>
<td>62</td>
<td>406.9</td>
<td>30.5</td>
<td>9.7%</td>
<td>59.7%</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>506</td>
<td></td>
<td></td>
<td>15.0%</td>
<td>50.4%</td>
</tr>
</tbody>
</table>

Table 18: Observed fifth-grade reading achievement scale scores across schools, 2005-06

Table 19 shows the estimates for the unconstrained hierarchical linear modeling model of fifth grade student reading achievement across schools. The estimate for the grand-mean math achievement is 401.9 with a standard error of 4.36. The estimated variability in the classroom means is 13.71. There appears to be significant variation among classrooms in their reading achievement scale scores ($\chi^2(10) = 117.84, p = .000$). The deviance of the model represents one measure of the lack of fit between the model and the data (Luke, 2004); in the unconstrained model, deviance is equal to 4915.07.
<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average School Mean</td>
<td>$\gamma_{00}$</td>
<td>401.90</td>
<td>4.36</td>
<td>92.20</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean</td>
<td>$\mu_{ij}$</td>
<td>13.71</td>
<td>187.99</td>
<td>10</td>
<td>117.84</td>
</tr>
<tr>
<td>Student Residual</td>
<td>$r_{ij}$</td>
<td>30.58</td>
<td>934.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 19: Unconstrained model of fifth-grade student reading achievement, 2005-06

Calculating an intraclass correlation coefficient (ICC) allows for a measurement of the proportion of variance in the dependent variable that is accounted for by level-2 units (Luke, 2004). In this model, schools accounted for 16.7 percent of the variance in the student reading achievement ($187.99 / [934.8 + 187.99]$). This finding does not provide support for the study’s fourth hypothesis.

Following this first analysis, a scatterplot of the predicted and actual scale scores in fifth grade reading achievement by school identifier (IRN) was generated. Figure 12 demonstrates the strong, linear, positive correlation between the predictor, predicted reading achievement scale score, and the outcome, reading achievement scale score across schools ($r = .84, p < .01$). Here too, support for the study’s first hypothesis is found. In this dataset, the coefficient of determination indicates that 71 percent of the variability in fifth-grade student reading achievement scale scores can be predicted from the relationship with the state’s value-added prediction scores, a very strong association.

In the figure, reference lines on the x- and y-axis have been imposed on the scatterplot at the cut scale score for the proficient level. In Figure 12, quadrant 1 (top right) contains the values of students who were predicted and observed at or above the
proficient level, while quadrant 3 (bottom left) contains the values of those that were predicted and observed to score below the proficient level. A review of the false negative scores in quadrant 2 (top left) reveals the paucity of points in this category. The greatest opportunity for student achievement and growth, then, lies in the false positive points in quadrant 4 (bottom right).

Figure 12: Predicted and observed fifth-grade reading achievement scale scores across schools, 2005-06
In a related analysis, a contingency table was constructed in order to empirically test the second hypothesis of the study, that the predicted achievement scale scores generated by the state’s value-added assessment system would correctly classify students who score at or above proficient correctly no more than 80 percent of the time in the baseline year. Table 20 demonstrates that this hypothesis is supported by the observations in the fifth grade dataset.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
<th>Below</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proficient or Higher</td>
<td>Below Proficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted</td>
<td>Below Proficient</td>
<td>31</td>
<td>172</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Proficient or Higher</td>
<td>180</td>
<td>47</td>
<td>227</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>211</td>
<td>219</td>
<td>430</td>
</tr>
</tbody>
</table>

Table 20: Contingency table of predicted and observed proficiency status, fifth grade, 2005-06

By examining Table 20, it can be demonstrated that the predictive proficiency model utilized in the state’s value-added assessment system correctly classified 79 percent of students (180/227) as proficient, with 21 percent of students (47/227) registered as false positives in this dataset. Alternatively, the system correctly classified 85 percent (172/203) of students as not proficient, with 15 percent (31/203) registered as false negatives in this dataset. Again, if the intervention (i.e., value-added and standardized formative assessment systems) to be implemented is successful, the percent of students failing to achieve this standard should significantly decrease. To be successful, the District will have to work with each of its elementary schools to help 75
percent of its students to reach the state standard for reading at or above the proficient level. A review of Table 18 indicates that the growth necessary to achieve this goal at each of the target schools ranges from 45 (75 - 30) at Compton to 6 (75 - 69) at McGregor; one school (Harter) scored above the state reading standard in the fifth grade.

As was the case in the fourth grade, a visual analysis of the variability within and between schools in observed reading achievement scale scores across schools is evident in the fifth-grade student scores in Figure 13. Here again, the theoretical justification for a nested model of student achievement become evident, as patterns of student achievement within and between schools demonstrate the variability below district reading assessment scale score averages that a formative assessment system seeks to minimize.
In Figure 13, the variability seen in Table 18 is expanded to include students scoring at the 25th and 75th percentiles (the bottom and top of the boxes, respectively), the median (the horizontal line within the boxes) and outliers (circles). In the fifth grade, the consistently average performance level of students around the proficient level across schools is clearly observable, with Harter being the sole exception.

A one-way random effects ANCOVA was run in HLM-6 with restricted maximum likelihood estimation with predicted student scale scores in level-1 and schools in level-2; the outcome variable was the observed scale score of the fifth grade reading
achievement assessment. Four iterations were required for convergence of the EM algorithm, indicating that the prediction scores were highly informative (Raudenbush & Byrk, 2002).

Table 21 shows the estimates for the one-way ANCOVA with random effects model of student reading achievement. The estimate for the expected value of reading achievement is 397.15 with a standard error of 5.08. This indicates that students in schools where the predicted reading score is at the classroom mean, students are expected to have a scale reading achievement score of 397.15. There appears to be a significant positive relationship between observed scale reading achievement scores and predicted reading achievement scale scores, \((t = 26.3, \ p = .000)\). We can expect, on average, to see approximately a 1.00-point increase in scale reading achievement score as the mean predicted reading achievement scale score increases by one point.

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-ratio</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Score (\gamma_{00})</td>
<td>397.15</td>
<td>5.08</td>
<td>78.21</td>
<td>10</td>
<td>0.000</td>
</tr>
<tr>
<td>Predicted Score (\gamma_{01})</td>
<td>1.00</td>
<td>0.04</td>
<td>26.3</td>
<td>428</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>SD</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Mean (\mu_{ij})</td>
<td>16.61</td>
<td>275.74</td>
<td>10</td>
<td>435.26</td>
<td>0.000</td>
</tr>
<tr>
<td>Student Residual (r_{ij})</td>
<td>16.96</td>
<td>287.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 21: One-way ANCOVA with random effects HLM of fifth-grade student reading achievement, 2005-06

The estimated variability among school means is 16.61. There is significant variation among schools in their observed scale score reading achievement.
(χ²(10) = 435.26, p = .000). Schools accounted for about 49 percent of the variance in the student reading achievement (275.74/[275.74 + 287.80]) when predicted scale scores are added as a level-1 predictor, a significant finding. This is a value that was larger than that which was predicted in the study’s fourth hypothesis. Moreover, the deviance of the model is equal to 3692.45, indicating that this model with its lower deviance value represents a better fit with the data.

Given that predictors were utilized in only the first level of the one-way random effects ANCOVA model to generate baseline data, two assumptions were tested: homogeneity of variance and normality (Raudenbush & Bryk, 2002). A test of the assumption of homogeneity of variance was run; common variance within each of the J level-2 units was found (χ² (10) = 6.29, p = .500). To examine the tenability of the assumption of a normal distribution of level-1 errors in the one-way random effects ANCOVA model, a residual analysis was conducted to identify violations which could negatively impact the estimated standard errors for fixed effects estimates (Raudenbush et al., 2004). The Q-Q plot of level-1 residuals in Figure 14 indicates that no significant violation was found.
Cohort Comparisons

In the context of the theory-driven evaluation employed in this study, five hypotheses were advanced and tested in two cohorts. The first hypothesis advanced the proposition that the relationship between predicted achievement scale scores generated by the state’s value-added assessment system and observed reading Ohio Achievement Test scale scores in the sample groups of students is positive, significantly strong and linear in the 2004-05 and 2005-06 school years. A review of Table 22 finds support for this hypothesis across cohorts and grades.
Table 22: Association between level-1 predictor and outcome variable

The second hypothesis was that the predicted achievement scale scores generated by the state’s value-added assessment system will correctly classify students who score at or above proficient correctly no more than 80 percent of the time in the baseline year in the 2004-05 and 2005-06 school years. Table 23 demonstrates that this hypothesis found support in one out of two cohorts and three out of four grades.

<table>
<thead>
<tr>
<th>Quadrant 1</th>
<th>2004-05</th>
<th>2005-06</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade 4</td>
<td>Grade 5</td>
</tr>
<tr>
<td>r</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>r^2</td>
<td>0.67</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 23: Correct classification of students by level-1 predictor

The third hypothesis advanced was that when the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scores of the sample groups are placed in an unconditional hierarchical linear model, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderately low (i.e., between 15 to 19 percent) in the 2004-05 and 2005-06 school years. Table 24 provides evidence that this hypothesis
did not find support across either cohort. In examining ICC values across grades, three values were observed below the estimate and one within the hypothesized parameters.

<table>
<thead>
<tr>
<th></th>
<th>2004-05</th>
<th>2005-06</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade 4</td>
<td>Grade 5</td>
</tr>
<tr>
<td>Deviance</td>
<td>5030.66</td>
<td>5198.93</td>
</tr>
<tr>
<td>ICC</td>
<td>0.13</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 24: Goodness of fit estimates for the unconditional model

The fourth hypothesis of the study advanced the notion that when the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups are placed in an one-way random effects ANCOVA, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderate (i.e., between 20 to 24 percent) in the 2004-05 and 2005-06 school years. According to Table 25, no cohort or grade generated support for this hypothesis; each accounted for a greater proportion of the variance than originally estimated.

<table>
<thead>
<tr>
<th></th>
<th>2004-05</th>
<th>2005-06</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade 4</td>
<td>Grade 5</td>
</tr>
<tr>
<td>Deviance</td>
<td>3686.79</td>
<td>4154.33</td>
</tr>
<tr>
<td>ICC</td>
<td>0.39</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 25: Goodness of fit estimates for the one-way random effects ANCOVA model
The fifth hypothesis of the study held that when the percentage of sample group students scoring at or above proficiency in the state’s summative reading assessment in school years 2004-05 and 2005-06 are analyzed, a statistically significant gain among cohorts will not be observed. Table 26 provides evidence that this hypothesis received support; a non-parametric test for two independent samples confirmed this fact at the .05 alpha level ($t(1) = 12.71, p = .205$).

<table>
<thead>
<tr>
<th></th>
<th>2004-05</th>
<th>2005-06</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent at or above Proficient</td>
<td>55.5</td>
<td>52.4</td>
</tr>
</tbody>
</table>

Table 26: Proficiency observed across schools

Discussion

This study utilized pilot testing, an approach within the practical taxonomy of program evaluation means and ends (Chen, 2005), to provide information to the Canton City School District regarding its program rationale, its field feasibility, and to discover problems that might arise during the scheduled implementation of three separate student assessment systems in the 2006-07 school year. While this approach was articulated by Chen in the context of theory-driven evaluation first in 2005, a review of the literature found only one example of its use in an empirical study (Wallin, Brembert, Haglund & Holm, 1993), an examination of a proposed instructional sequence in the link between nutrition and cancer (Chen, 2005, p. 124). This study has led to a number of different
applications in health education and research (Figure 15), but no additional theoretical work in the pilot testing approach to evaluation.

![Figure 15: Two generations of Wallin, Bremberg, Haglund & Holm (1993) citations](image)

By utilizing this approach in a theory-driven evaluation of a program in educational measurement, this study offers an opportunity to examine the utility of this approach in a different context. One of the advantages that the approach affords evaluators is its clear delineation of principles that must be employed in its utilization. Each principle advanced by Chen in his explanation of this approach (Chen, 2005) found support in this demonstration.
The guidelines to applying the taxonomy of program evaluation means and ends within the context of the theory-driven approach (Chen, 2005) require program implementers and recipients to participate in trials; small but nevertheless typical samples are required for valid inferences about program operations to be drawn; the methods of gathering data must be flexible; and pilot test findings should only be used for program development purposes (Chen, 2005, 120-121).

In the present empirical demonstration, these principles were employed through the design and implementation of the evaluation approach. In that context, each was deployed in a variety of settings using school administrators, teachers and students in order to achieve the objectives of the study. The program stakeholders and evaluator found that these principles acted as helpful guides in designing instruments and evaluating data. However, upon designing the data collection instruments and procedures to inform the program rationale, members of the dissertation committee, the program stakeholders and the evaluator quickly saw the need to propose and add a fifth principle: pilot testing requires baseline data regarding anticipated program outcomes.

In behavior analysis, baseline data are often gathered before a program intervention in order to determine the rate of target behaviors before a given intervention is applied. Once the intervention has begun, assessments are utilized to identify changes in target behaviors (Kazdin, 1982). Based on the research design employed, the intervention then be taken away, changed or compared to other target behaviors or interventions in order to infer correlation or causation. What is crucial in the identification of pre-program target behaviors is that they be assessed in carefully controlled settings so that they differ from the intervention phase of the program in only
one way: the procedures accompanying the intervention are absent (McCormick, 1995, p.5). While this is not always possible in education research (Shavelson & Towne, 2001), every effort was made to achieve these conditions in the present study.

The present study utilized this fifth principle in its research design with respect to student achievement in reading in grades 4 and 5 of the schools of interest. In keeping with the other principles of pilot testing, this study did not require all teachers to submit to assessment questionnaires or structured interviews, but concentrated on the gathering of baseline data on the outcome of interest: student achievement standards and growth. Given the implementation of Ohio’s new achievement tests in reading and mathematics, fourth- and fifth-grade reading was the only baselines that could be established for the study for which value-added predicted scale scores were available.

In the field of educational measurement, the pilot testing approach to evaluation has merit, especially when the addition of the fifth principle, this study’s contribution to the improvement of Chen’s conception of the evaluation approach, is considered. As a tool for program development, it can be used in initial analyses of the field feasibility of proposed assessment systems, instructional approaches, learning theories, and linkages between curricula and other assessment systems. In an effort to more fully understand its utility, it could be employed with other subjects in a variety of settings using different research methods to better understand the utility of the approach in educational measurement.

In the present study, this evaluation approach was utilized in part to conduct an empirical demonstration of its use in program plan development in the context of theory-driven evaluation. In that context, it yielded five achievements: it assisted in the
development of the program theory of a formative assessment system; it provided a program evaluation framework for formative and summative assessments within the context of value-added assessment that assists the District in augmenting student achievement growth trajectories; it assisted in the development of a teacher assessment questionnaire; it guided the development of sample baseline measures of student achievement growth in reading; and it created the context to propose a method for linking formative, summative and value-added assessments within the structure of a hierarchical linear model. While not only achieving the research objectives of this study, it also guided the development of information to guide the development of data collection approaches and instruments to provide data relative to the action and change models in a mature program implementation evaluation. Thus, this evaluation approach delivered everything that was articulated relative to it in Chen’s practical taxonomy of program evaluation means and ends (Chen, 2005, 119-125).

For the District, the use of this evaluation approach will assist in District and school improvement plans aimed at increasing student achievement standards and growth. In reading, historical student achievement results demonstrate the challenge that lies before District schools. Tables 12 demonstrates the extent of this challenge across schools and also indicates the variability associated with year to year comparisons of proficiency required by the federal government under AYP calculations (Kane & Staiger, 2002). In Figure 16, in the 2003-04 school year, the range of fourth-grade students at or above reading proficiency is 34 percent (83 – 49); in the 2004-05 school year, the range is 45 percent (78 – 33); in the 2005-06 school year, the range is 33 percent (76 – 43). An
Ohio Achievement Test for reading was not administered to fifth-grade students in the 2003-04 school year.

Figure 16: Percentage of fourth-grade students at or above reading proficiency, 2003-04 through 2005-06

AYP targets are set annually by the Ohio Department of Education. While these growth targets can be calculated for each school, it must be remembered that each class represents a new group of students, each with their own academic profile. This basic fact of such analyses is what, in part, creates the need for as much summative, value-added
and formative assessment as possible in order to diagnose student learning difficulties, plan instructional activities, place students into learning sequences and monitor student progress. In this context, the need for a predictive model of student achievement is transparent.

The prediction algorithm used inside of the state’s value-added system worked well: school administrators and teachers can use it and the system’s projected scores to organize resources to drive slope changes, but questions about it remain. In both cohorts across grades, the value-added assessment system prediction classifications rose above 80 percent only once. The assessment literature reviewed in chapter 2 indicates avenues for research.

Summative assessments are utilized to determine the extent of knowledge in an individual at a point in time and a measurement of the individual’s development of a given ability (Anastasi & Urbina, 1997). Formative assessments help teachers identify whether or not students understand the material, which students need extra help and how to facilitate student learning (McMillan, 2001). In keeping with the literature base, the District made the decision to implement two standardized, formative assessment systems in order to increase the amount of information to teachers in order to facilitate classroom and instructional decision making. At issue for the District in the initial implementation of these systems was how the schools might utilize the system and the data that each provides. In the pilot testing approach to the evaluation of this intervention, a number of issues were uncovered that have implications for the District. These issues are best addressed by the answering four questions addressed by the pilot testing approach.
Pilot testing evaluation questions

To assist the Canton City School District’s theory-driven evaluation of its initial implementation a value-added and two formative assessment systems within the context of its District and building continuous improvement plans, empirical evidence gathered from teacher assessment questionnaires, structured interviews and student achievement were utilized to provide research evidence in support of the pilot test. Its use in this study has implications for the District and for the evaluation approach. To organize the evidence behind the former, four key questions appropriate to the pilot testing approach will be advanced (Chen, 2004, p. 122):

1) Can the intervention be implemented in the field as intended?

2) Can the implementers anticipate encountering certain problems delivering the intervention?

3) Will clients be receptive to the intervention or resist it?

4) Do any of the program’s organizational procedures impede the implementation process?

Implementation issues

The first question asks if the intervention can be implemented in the field as intended. The implementation plan is summarized in the action model articulated in the fourth chapter. The ecological context of that model, the District and school building improvement teams, provide the support for the assessment systems. Their work in the data retreat led to the development of schoolwide plans for reading and mathematics improvement for the 2005-06 school year. Conversations with the District indicate that this retreat will be replicated in this and succeeding school years, expanding beyond elementary to include middle and high schools as well.
A content analysis (Ryan & Bernard, 2000) was conducted on all schoolwide plans to generate themes and to identify the extent to which all plans demonstrated fidelity to the guidelines of the Learning Points Associates’ data retreat. In the process of coding and analysis, some variability was observed in the plans within the three main outcomes of the training: developing goals, designing specific strategies and defining evaluation criteria. While goal statements tended to orient toward AYP targets in reading and mathematics across the grade span, the strategies and evaluation criteria varied depending upon subject matter. To some extent this is dependent upon curriculum: while the District has implemented the Everyday Mathematics program across all schools, there is no comparable unified reading curriculum. This presents a problem for the activation of the DIBELS and ETS Pulliam reading assessments, as it would in any education system in which instruction, curriculum and assessments are not in alignment. In a similar manner, the linkage between goals, strategies and evaluation criteria was not transparent within many plans. While some variability in mathematics strategies was seen, reading strategies depended upon a variety of reading programs, software and instructional approaches. In order to identify the relative success of the implementation of each system, these disparate contexts will have to be taken into account.

Information from the teacher assessment questionnaire yielded a number of observations regarding classroom instruction, subject-matter training and professional development, assessment practices and beliefs. Data from the questionnaire suggested that the link between instructional readiness and professional development opportunities be examined closely in the context of school improvement planning, and that the intervention and service delivery protocols include formative assessment training so that
all teachers will recognize the many classroom decisions that can be supported by data from such assessments. Evidence gathered from the teacher assessment questionnaires also suggests that elementary school teachers may practice assessment on a schedule conducive to the systematic use of a standardized formative assessment instrument.

Information from the content analysis of the structured interviews of selected teachers also informs the question. In value-added assessment, the analysis suggests that teachers used the system’s reports in ways that were not anticipated by program stakeholders. It was also clear that teachers will hold different preferences for access to the data that will have to be resolved at the school level for school improvement planning purposes as well as administration. Data suggests that, if the associate organization can deliver relevant information before the start of the school year, teachers will respond positively to the implementation of the system. The initial implementation of the ETS Pulliam assessment system in mathematics, on the other hand, met with a number of problems that were related to implementation such as technology, use and reporting. In this context, the implementing organizations and implementers may find it beneficial to review the proposed implementation plans of the associate organization to ensure that teacher and student concerns are adequately resolved.

Obstacles

The second question asks if the implementers can anticipate encountering certain problems in delivering the intervention. An analysis of the schoolwide plans for improvement in reading and mathematics demonstrate that knowledge of the three assessment systems is quite weak. While this observation is to be expected based upon the timeline of the retreat and the initial implementation schedule, it represents an
education and communication challenge to the implementing organizations and implementers. For implementation to be successful, value-added assessment and formative assessment training will need to occur in advance of or in the context of the 2006-07 data retreats. Such training will not only serve to introduce school improvement teams to the new tools, but must demonstrate how, when and why data from these systems should be used to plan instructional activities, diagnose learning difficulties, place students into learning sequences and monitor student progress. This implication is echoed in the findings of the teacher assessment questionnaires.

Many of the teachers who were sampled in the teacher assessment questionnaire reported various levels of instructional readiness across content areas in reading and mathematics. When implementing the value-added and formative assessment systems, teachers will be called upon to interpret student achievement results in these content areas. This may pose a problem for implementation in that teachers’ use of assessment data is dependent upon their content knowledge. Thus, professional development in these areas may need to be mapped against reported teacher readiness perceptions and included in teachers’ individual professional development plans. This gap analysis will have to occur early in the school year in order to effectively evaluate teachers’ use of the assessment system and the data each yields.

Evidence from the interview protocols provides tentative support for these recommendations. Interviews regarding the value-added assessment systems indicated that teachers vary in their need to access data directly. Interviews regarding the ETS Pulliam assessment in mathematics, however, indicated that a number of different
problems in technology, test length and reporting can be anticipated in 2006-07. These findings, while tentative, should be used in developing the program.

School receptivity

The third question asks if the intervention clients will be receptive to the intervention or resist it. Evidence from the schoolwide plans for school improvement in reading and mathematics indicates that many teams have made and will make instructional programming decisions based upon the need to improve student achievement. In that context, implementers will have to clearly communicate the linkages between the assessment systems and the goals for student achievement, the strategies employed to achieve those goals and the evaluation criteria. If students can see the connections between these components of the school and classroom’s instructional program, they may respond positively to the intervention.

Evidence from the teacher assessment questionnaire indicates that a small percentage of time is now spent on assessment activities in a typical week of reading or mathematics instruction. This indicates that for the clients to be receptive to the program, the formative assessments will need to be easy to administer, reliable and content valid and tied directly to the lessons being taught. Information gleaned from the structured interviews support this implication. Students taking assessments through the ETS Pulliam formative assessment system in mathematics reported a number of problems with the system, including test length and administration. Briefly, students felt that the tests were too long, and they wanted the ability to review their answers. While users of the DIBELS system were not queried in the context of this study, conversations between the
implementers and users indicate that the system was well received by the limited number of students who were tested on the system.

Organizational procedures

The fourth and final question posed in the context of the pilot test approach asks if any of the program’s organizational procedures may impede the implementation process. Here, the evidence from the schoolwide improvement planning process would indicate the necessity of a common reading curriculum across elementary schools, the need for formative assessment education, training and practice, and the need to support the knowledge gleaned from the data retreat. While most of these issues will need to be addressed by the implementing organizations, support could be coordinated between implementers and schools at weekly grade-level staff meetings, as evidence from the plans indicates the utility of the approach to educational planning.

The questionnaire data suggests that schools may find it beneficial to work with teachers on assessing instructional readiness by subject content area and coordinating content area professional development with the results of such assessments. The implications of the need for formative assessment training could be repeated here as well. Across classrooms per grade, schools may also benefit from discussing the nature of the formative assessment instruments- as analyses of student performance on a given lesson, or as duplicates of the summative achievement test.

Evidence from the structured interviews suggests that the training program devised by the implementer for value-added assessment resulted in principals who were perceived as well-trained in the system. This in turn suggests the need for the implementers to conduct trainings in formative assessment generally and each system
specifically in each school implementing the programs. The challenges that were met by teachers and students in the ETS Pulliam formative assessment system in mathematics will need to be addressed before it can be implemented.
CHAPTER 5

SUMMARY, CONCLUSIONS AND IMPLICATIONS

This chapter contains three major sections, a summary of the study, conclusions, and implications of the study and suggestions for further research. In the first section, the study’s problem statement, procedures and specific research hypotheses are summarized. The second section articulates the major findings of the study, presenting the hypotheses advanced and evidence of their evaluation. In the third section, the implications of the study’s conclusions are explored and further research suggested.

Summary of the Study

This study sought to assist the Canton City School District’s theory-driven evaluation of its initial implementation of a value-added and two formative assessment systems within the context of its District and school continuous improvement plans by employing the pilot test approach in the eleven District elementary school buildings identified by the Ohio Department of Education in school year 2004-05 as at risk or in school improvement. It utilized a mixed-methods approach in integrating qualitative and quantitative data analyses to inform the field feasibility of a program and to help in the identification of the problems that may arise in the implementation of the new assessment systems. In pursuit of these goals, the study focused on the generation of several hierarchical linear models (HLM) to construct a framework to analyze formative and
summative student assessment data in the context of value-added assessment and to generate baseline student achievement data for comparative and predictive purposes.

Statement of the Problem

In the 2004-05 school year, the Canton City School District received an ‘Academic Watch’ rating from the Ohio Department of Education. The District met 3 of the 23 state indicators, achieved a performance index score of 75.9 out of 120, and was in year 2 of district improvement. In responding to the problem, the District offered a data retreat for all schools, revised school improvement plans in alignment with student results on the state’s summative assessment measures in reading and mathematics, implemented the state’s value-added assessment system through Battelle for Kids one year before mandatory implementation, and partially implement two standardized formative assessment systems in reading and mathematics. A pilot test approach to program evaluation (Chen, 2005) was chosen to indicate the field feasibility of the school improvement efforts and to help in the identification of the problems that might arise in the implementation of the new assessment systems in the context of District and school improvement efforts.

After implementing continuous improvement plans and the value-added assessment system across the District, the Canton City School District received rating of ‘Continuous Improvement’ from the Ohio Department of Education in the 2005-06 school year. The District met 6 of the 25 state indicators, achieved a performance index score of 83 out of 120, and missed Adequate Yearly Progress for the fourth consecutive year.
Statement of the Procedures

Several qualitative and quantitative databases were developed for the program evaluation. These included data from school improvement plans, teacher assessment questionnaires, structured interviews of key teachers, and two summative assessment databases stored in SPSS 14.0 for Windows and HLM-6. In each case, data were gathered in line with the research design principles of theory-driven evaluation (Chen, 2004; 2005) and the expectations of the District and evaluator.

School improvement plans developed by school teams at the District’s elementary school data retreat were gathered upon their completion, sent to the District Director of Testing, Evaluation and Research, and sent to the office of the evaluator where they were coded (Ryan & Bernard, 2000) by the evaluator to identify themes across schools concerning assessment and instruction that could impact program implementation. The evaluator then worked with the Director to develop a program rationale for the District’s initial implementation of the state’s value-added assessment system and two standardized formative assessment systems in reading and mathematics.

Following value-added assessment training by the Director at each elementary school, ETS Pulliam conducted a training session for its formative assessment system in selected buildings. A reading specialist was invited into the District to train selected staff in the use of the DIBELS formative assessment system. All three systems were available to limited numbers of staff in an initial implementation of each system. Afterwards, a small sample of teachers were asked to complete a cross-sectional questionnaire in order to discover potential problems in implementation in developing baseline data concerning classroom instruction, subject-matter training and professional development, assessment
practices and beliefs by completing a teacher assessment questionnaire. Following the questionnaire, those teachers were asked to participate in structured interviews with the Director of Testing, Evaluation and Research in order to identify the challenges and opportunities that they saw in each assessment, implementation issues, and the adjustments to instruction (if any) that they made as a result of using the various assessments.

At the conclusion of the 2004-05 school year, Battelle for Kids generated its predicted scale scores for the Ohio Achievement Tests in reading and mathematics, and these scores were then linked to the observed results of students and stored in de-identified fourth and fifth grade datasets. In the autumn of the 2005-06 school year, Battelle for Kids and the Canton City School District repeated these procedures. These datasets were then uploaded to SPSS 14.0 for Windows and HLM-6 in order to generate descriptive and inferential statistics.

Specific Research Hypotheses

In alignment with the research objectives of the study, the principles and guidelines of the pilot test evaluation and the focus of the study, five research hypotheses were advanced in the articulation of baseline student achievement data generated through hierarchical linear models of the value-added and summative reading achievement assessment systems:

1) The relationship between predicted achievement scale scores generated by the state’s value-added assessment system and observed reading Ohio Achievement Test scale scores in the sample groups of students is positive, significantly strong and linear.

2) The predicted achievement scale scores generated by the state’s value-added assessment system will correctly classify students who score at or
above proficient correctly no more than 80 percent of the time in the baseline year.

3) When the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scores of the sample groups are placed in an unconditional hierarchical linear model, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderately low (i.e., between 15 to 19 percent).

4) When the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups are placed in an one-way random effects ANCOVA, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderate (i.e., between 20 to 24 percent).

5) When the percentage of sample group students scoring at or above proficiency in the state’s summative reading assessment in school years 2004-05 and 2005-06 are analyzed, a statistically significant gain among cohorts will not be observed.

Conclusions

Each of the study’s five specific research hypotheses were advanced in support of the study’s purpose, which was to assist the Canton City School District’s theory-driven evaluation of its initial implementation a value-added and two formative assessment systems within the context of its District and building continuous improvement plans by employing the pilot test approach in eleven elementary school buildings identified by the Ohio Department of Education in school year 2004-05 as at risk or in school improvement. In that context, each of the five specific research hypotheses will be addressed presently, while a review of the empirical evidence gathered in support of the general program evaluation will be presented in the implications section of this chapter.

The first hypothesis advanced the proposition that the relationship between predicted achievement scale scores generated by the state’s value-added assessment
system and observed reading Ohio Achievement Test scale scores in the sample groups of students is positive, significantly strong and linear in the 2004-05 and 2005-06 school years. The analysis presented in chapter 4 found support for this hypothesis across cohorts and grades.

The second hypothesis was that the predicted achievement scale scores generated by the state’s value-added assessment system will correctly classify students who score at or above proficient correctly no more than 80 percent of the time in the baseline year in the 2004-05 and 2005-06 school years. This hypothesis found support in one out of two cohorts and three out of four grades.

The third hypothesis advanced was that when the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scores of the sample groups are placed in an unconditional hierarchical linear model, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderately low (i.e., between 15 to 19 percent) in the 2004-05 and 2005-06 school years. This hypothesis did not find support among cohorts or grades. In examining ICC values across grades, three values were observed below and one within the hypothesized parameters.

The fourth hypothesis of the study advanced the notion that when the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the sample groups are placed in an one-way random effects ANCOVA, the proportion of variance in the dependent variable that is accounted for by level-2 units will be moderate (i.e., between 20 to 24
percent) in the 2004-05 and 2005-06 school years. No cohort or grade provided support for this hypothesis; in fact, the proportion was quite higher than hypothesized.

The fifth hypothesis of the study held that when the percentage of sample group students scoring at or above proficiency in the state’s summative reading assessment in school years 2004-05 and 2005-06 are analyzed, a statistically significant gain among cohorts will not be observed. This hypothesis found support in the data.

*Implications for Program Evaluation*

The employment of the pilot testing approach in this study has led to the identification of how this approach can be useful in educational program planning. It can inform program stakeholders of the feasibility of planned interventions in two main ways. In the first, this approach can help to determine the relative probability of success of a program, given its program rationale. Each element in the action model can be tested after the school has researched the basis of its change model. In the second, this approach can help to identify the practicality of each component of the intervention. If the fifth principle proposed in this study is utilized, that pilot testing requires baseline data regarding anticipated program outcomes, program stakeholders will be able to develop sound goals for planning purposes. This addition to Chen’s conception of the pilot testing approach (Chen, 2005) represents an important advancement in the theory and practice of this approach to evaluation. The use of baseline data allows program stakeholders to depict current levels of the outcome variable and establishes data points for prediction and comparison purposes. For evaluators, the generation of baseline data establishes the context for program rationale parameters. For both, it focuses the concerns of the
evaluation on the outcome variable of interest, providing a foundation for program planners to determine both performance and growth levels.

The utilization of this approach in educational program planning can also help to identify problems that may arise in program implementation. Pilot testing can help to determine the components of the action model that may need to be altered in order to achieve success, and can help identify resources needed to increase the odds of success. While these opportunities for educational program planning can be seen with the pilot testing evaluation approach, significant challenges remain.

The pilot testing approach in particular and the theory-driven evaluation model in general require the constant presence of evaluators and data collection instruments. These are not evaluations that can be done remotely or periodically. They require ‘boots on the ground’. Moreover, they depend upon the experience of staff to develop the program and its rationale and organize the literature behind the rationale and all of its components. These evaluations also require a great deal of time to plan data collection instruments that can test action model components. Time is not often a resource that program stakeholders have, and this can present a significant barrier to the deployment of this approach. Lastly, these evaluations require access to evaluators who are familiar with a variety of research traditions and techniques common to qualitative and quantitative research and evaluation. In sum, though these evaluations require much of program stakeholders and evaluators, the theory-driven approach in general and the pilot testing approach in particular can help to improve educational programs designed to reach the goals that local education agencies set in order to improve student achievement.
Implications for Academic Improvement

While the value-added assessment system proved to yield fairly accurate predictions at the student level, the number of missing predicted scores proved to be quite troublesome. In the 2004-05 dataset, fewer missing scores were found in fifth grade as opposed to the fourth grade. The variability between schools was quite high in both grades, however, and the level of missing data was observed to be unrelated to the number of students in each school or their academic achievement. While the same pattern was observed in the 2005-06 dataset, missing predictor levels were higher overall than in the previous school year. While the methodology employed by SAS for Schools accounts for missing data within its predictive algorithm, the problem of missing scores is felt most acutely at the classroom level. Thus, the utility of the value-added assessment system for teachers is limited to the extent that the system can generate information for each classroom student.

During the 2004-05 and 2005-06 school years, the District was able to implement two-thirds of its assessment intervention. If continuous improvement planning and value-added assessment had been enough to increase student achievement, overall reading achievement would have increased 2004-05 to 2005-06. While this did not occur, two changes are interesting in relationship to the fifth hypothesis of this study.

If the slope of the relationship between predicted and observed reading achievement scale scores are compared between the 2004-05 and 2005-06 school years in both cohorts, the relationship between these variables can be examined. On the one hand, if the relationship between the independent and dependent variables grows stronger, a case could be made for the improved predictive power of the independent variable.
However, if the intervention is successful, the relationship between these variables should weaken. While this is one of the reasons why an absolute measure of academic achievement gains was chosen as a tested hypothesis for this evaluation for state and federal accountability purposes, an examination of the slopes between cohorts is instructive.

Using descriptive and regression data from the 2004-05 and 2005-06 data sets for reading achievement in the targeted schools, a test for differences between slopes ($H_0: b_1 = b_2$) can be conducted. In the fourth grade, a significant difference between the 2004-05 and 2005-06 slopes is detected ($t(803) = -2.13, p < .05$); in the fifth grade, a significant difference is not observed ($t(922) = .369, p > .05$). If the scatterplots and correlations between the independent and dependent variables are examined in Figures 3 and 9 for fourth grade and 6 and 12 for fifth grade, no practical difference is detected.

Again, two interpretations can be advanced in this context. One conclusion, however, is clear. For the purposes of academic achievement, absolute differences in academic achievement is the preferred measure of success for interventions, while predicted scores may prove to be most useful as a screening measure utilized at the classroom level.

A second interesting finding concerning the fifth hypothesis is the absolute difference in academic achievement in two schools, Compton and Dueber, between school years 2003-04, 2004-05 and 2005-06. By looking at change during this time frame, the random nature of academic achievement is observed across schools. For example, no school demonstrates the monotonic growth that state and federal education laws require, but interesting changes are seen in several school districts in the 2004-05 to 2005-06 score subset. While Allen, Belle Stone, McGregor and Summit Elementary
Schools are associated with increases in the percentage of students scoring at or above proficiency, the percentage growth in Compton and Dueber Elementary Schools is substantial. Clearly, the District needs to work with principals in these buildings to identify changes that occurred during those years in terms of staffing, students, curriculum, instruction, assessment and administrative support. While another year of data will provide more evidence for analysis, the timelines for improvement are growing short for the District.

Suggested Further Research

At the conclusion of the 2006-07 school year, the implementing organization and implementers plan to work with the evaluator to carry out a series of statistical tests designed to compare the strength of the association between value-added assessment predictions and formative assessment scores and summative assessment scale scores in reading and mathematics. This will be done in the context of District and school improvement planning efforts, data retreats, and the other recommendations described in detail in the discussions section of chapter 4 and the conclusions section of this chapter. To facilitate that research, a number of different hierarchical linear models (HLM) are advanced to clarify the future directions of research in evaluating these formative assessment systems and value-added assessment system in the context of District and school improvement efforts.

The first model will be tested on a dataset drawn from fifth and sixth grade students in the 11 targeted schools. Here, the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the baseline groups will be uploaded into HLM-6 and
analyzed through an unconstrained model to estimate between-groups effects with an intraclass correlation coefficient (ICC). This null model is presented below in Equations 5 (hierarchical) and 6 (mixed-effects).

\begin{align*}
Y_{ij} &= \beta_{0j} + r_{ij} \\
\beta_{0j} &= \gamma_{00} + u_{0j} \\
Y_{ij} &= \gamma_{00} + u_{0j} + r_{ij}
\end{align*}

In equation 5, $Y_{ij}$ represents the reading achievement score for a particular student in a particular school, $\gamma_{00}$ represents the grand mean across all students in all schools, $u_{0j}$ represents the variability between schools, and $r_{ij}$ represents the variability within schools.

In the second and third models, the predicted achievement scale scores generated by the state’s value-added assessment system and the observed reading Ohio Achievement Test scale scores of the same students will be uploaded into HLM-6 and analyzed through a one-way random effects ANCOVA to estimate between-groups effects with an intraclass correlation coefficient (ICC) against the null model. The second model will be run using the state’s value-added prediction scores as a level one predictor, while the third model will be run using a score extrapolated from the slope of student formative assessment scores. Both models take the general form of Equations 7 (hierarchical) and 8 (mixed-effects).
In Equation 7, the variable $Y_{ij}$ represents the scores of individual $i$ in school $J$ on the Ohio Achievement test in reading. The variable $\beta_{0j}$ represents the unadjusted mean score of every $J$th school on the outcome ($Y$) variable when all predictor variables ($X$) are held at zero; the level-1 coefficient is random. Here two predictors will be assessed: predicted scale scores from the value-added assessment system, and a scale score extrapolated from the slope of student formative assessment scores. In this hierarchical linear model, there is an intercept for each classroom included in the analysis, and it is centered on the group mean. The within-classroom error term is represented by the variable $u_{ij}$. This term is assumed to be normally distributed within each school, while its variance is assumed to be relatively homogenous across schools. The $\beta_{1j}$ variable represents the slope of the achievement score on predicted scores for the $J$th school; here the level-1 coefficient is fixed.

In each model, evidence from the schoolwide plans for school improvement in reading and mathematics and teacher assessment questions will be gathered in order to provide context for the results of the analysis. Composite scores from the latter
instrument may also be developed and utilized as a second level predictor in another hierarchical linear model. In doing so, the implementing organization will seek to discover the formative assessment systems that can help teachers utilize the data from the value-added, formative and summative assessment systems to improve student achievement relative to state and federal goals and District expectations. If this is done, a fourth and fifth model may be proposed.

In these models, the predicted achievement scale scores generated by the state’s value-added assessment system, the formative assessment scores of students and the observed reading Ohio Achievement Test scale scores of the fifth and sixth grade students in the same eleven schools will be uploaded into HLM-6 and analyzed through an intercepts-and- slopes-as-outcomes model designed to estimate between-groups effects with an intraclass correlation coefficient (ICC) against both one-way random-effects ANCOVA models.

In the fourth model, the state’s value-added prediction scores will be used as a level one predictor while in a fifth model, a score extrapolated from the slope of student formative assessments will be used as a level-1 predictor. In both cases, a composite score from the teacher assessment questionnaire will be used as a level 2 predictor. Both models will take the general form of Equations 9 (hierarchical) and 10 (mixed-effects).

\[
Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (9)
\]
\[
\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}
\]
\[
\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}
\]
While the interpretation of the variables in equation 9 are similar to those detailed in equation 7, two improvements to the model are noted: the effect of adding the second level predictor \( W_j \) to account for varying teacher assessment practices between schools, and the modeling of the effects of a cross-level interaction component \( \gamma_{11} W_j X_{ij} \). With these additions, the model becomes capable of estimating the effects of teacher assessment beliefs and practices as well as the interaction between the assessment variable and the value-added predictor in the case of model 4 and the formative assessment predictor in the case of model 5. These may help to generate more questions around the variability between schools in reading achievement.

When the District and its schools implement these systems along with the value-added system in the 2006-07 school year, four results are predicted. First, it is predicted that the percentage of students at or above the proficiency level in reading will be higher than the baseline in 2004-05 and in 2005-06. Second, the percent of false positives should decrease whether one uses the predicted scores from the value-added assessment system or formative assessment scores. Third, the formative assessment scores should make a better predictor in the HLM models employed. Fourth, a hierarchical linear model using formative assessment scores as a level-1 predictor and a composite teacher assessment score as a level-2 predictor should explain more variance in the dependent variable than the models utilized in the baseline or in a model that utilizes the value-added prediction score as a level-1 predictor.
BIBLIOGRAPHY


APPENDIX A

SCHOOL IMPROVEMENT PLANS
School Improvement Plans

In the context of the study, the eleven Canton City School District elementary school buildings identified by the Ohio Department of Education in school year 2004-05 as at risk or in school improvement were asked to participate in a data retreat sponsored by the District. The following is a summary of those plan designs presented in the context of their baseline summative academic achievement in reading and mathematics. All achievement test results are from the Ohio Department of Education (Ohio Department of Education, 2006).

Allen Elementary School

In the 2004-05 school year, 50 percent of fourth grade and 61.4 percent of fifth grade students in Allen Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Similarly, 45.8 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet AYP in reading, but did meet it in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading.

According to its schoolwide plan for reading, the school improvement team from Allen Elementary established goals to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading to 68.3 percent in fourth grade and 68.3 percent for fifth grade in the 2005-06 school year. To achieve these goals, the team chose to implement the following strategies: utilize small group and ability/instructional level reading instruction; use working assessment portfolios to illustrate grade level indicator progress; use data from a variety of assessment including
DIBELS to develop instructional strategies; visit and collaborate similar schools with better reading achievement scores to identify best practices; strength professional development; develop and implement integrated lessons (e.g., K/1, 2/3, 3/4 and 5/6); and meet with parents to orient and train them in utilizing reading strategies at home.

To evaluate progress towards these goals, the team decided to compare results on formative assessments from existing District formative assessments on a regular basis, evaluate the results of the Ohio Achievement Test in reading more fully to determine instructional strategies for different levels of student proficiency, and to conduct evaluations after meetings between teachers. The team did not establish goals for mathematics test improvements in its plan.

*Belden Elementary School*

In the 2004-05 school year, 62.8 percent of fourth grade and 54.5 percent of fifth grade students in Belden Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Only 37.2 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet reading AYP in reading or mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading and mathematics.

According to its schoolwide plan for reading, the school improvement team from Allen Elementary established goals to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading to 68.3 percent in fourth grade and 68.3 percent for fifth grade in the 2005-06 school year. At the data retreat, the team identified the main weakness of Belden students in reading to be the reading process. To
overcome this obstacle to higher student achievement in reading, the team resolved to implement a three-step process for improvement. In the first step, each class was to receive a daily focused instruction addressing concepts of print, comprehension strategies and self-monitoring strategies, the three indicators of reading process. In the second step, planning would begin to incorporate reading process activities of the previous week before plans for the following week would be generated. In the third step, after-school tutoring would be made available to third, fourth and fifth grade students needing additional instruction on reading process.

To evaluate the success of these proposed action steps, the team chose to evaluate scores on the Ohio Achievement Tests in reading, observe the daily focused instruction on reading process indicators, present their work to each other in collaborative planning sessions, and review student work on an existing District formative assessment. The team also created a school improvement plan for mathematics achievement in mathematics.

The improvement goal for mathematics established by the team was that 45 percent of third grade students would score at or above proficient on the Ohio Achievement Test in the 2005-06 school year. To reach this goal, the team analyzed summative test data to identify students’ primary weakness in mathematics as the strand in data analysis and probability. To improve students’ understanding of these concepts, teachers agreed to focus at least two times per week on implementing additional activities. In a strategy choice similar to the team from Allen, teachers decided that mathematics activities from each week would be utilized in collaborative planning time to generate plans for the following week.
To evaluate the success of these proposed action steps, the team decided to evaluate scores on the Ohio Achievement Test for math, observe focused activities, present work in collaborative planning sessions, and assess student progress quarterly.

_Belle Stone Elementary School_

In the 2004-05 school year, 54.1 percent of fourth grade and 54.8 percent of fifth grade students in Belle Stone Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Similarly, 49.2 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school met AYP in reading, but did not meet it in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading and mathematics.

According to its schoolwide plan for reading, the school improvement team from Belle Stone Elementary et our to assist 70 percent of students at each grade level in making significant gains with reading informational text. To achieve this goal, the team chose to implement seven integrated strategies in all content areas of instruction: monitoring comprehension; using graphic organizers; answering oral and written questions; generating oral and written questions; recognizing story structure; summarizing; and acquiring more vocabulary. To identify progress towards their goal, the team chose to utilize the Developmental Reading Assessment (DRA) in K – 2 and the Basic Reading Inventory (BRI) in grades 3 – 5 to test students quarterly. The team agreed to use assessment results to guide instruction. In a similar manner, the team established goals for mathematics test improvements in its plan.
The team’s schoolwide plan for mathematics improvement concentrated on two concepts in mathematics: applying program solving and decision-making techniques, and communicating mathematical ideas. To achieve these goals, the team emphasized the need for students to demonstrate the effective use of each of these skills in practicing problem-solving exercises. To evaluate this strategy, the team chose to work with students to assess one class assignment per month using a rubric designed by the mathematics curriculum. The rubrics will be completed and turned in to the teacher to guide student instruction and facilitate student conferences.

_Compton Elementary School_

In the 2004-05 school year, 48.3 percent of fourth grade and 40.5 percent of fifth grade students in Compton Elementary School scored at or above proficient on the Ohio Achievement Test in reading. In contrast to these reading test results, 55.2 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet AYP in reading, but met it in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading and mathematics.

According to its schoolwide plan for reading, the school improvement team from Compton Elementary established goals to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading by 14 percent (to 75 percent) in third grade, 28 percent (to 76.3 percent) in fourth grade and 28 percent (to 68.5 percent) in fifth grade in the 2005-06 school year. To achieve these goals, the team chose to analyze results on state achievement practice tests to establish baseline needs in
four strands in the reading standards: phonemic awareness, acquisition of vocabulary, informational text and literary text. Once student needs in these areas were identified, the team decided that teachers would identify enrichment and remediation groups in order to align instruction to meet student needs. In addition, the team would utilize 180-minute blocks of time to focus on reading and mathematics in classrooms, while Title teachers would get an additional hour per grade for remediation purposes. To evaluate the progress of these steps, the team planned on assessing students quarterly using existing District formative assessments and designing intervention plans for students below goal. Similarly, the team established goals to improve the mathematics test scores of students.

For mathematics, team goals were set to increase third grade mathematics scores on the Ohio Achievement Test by 20 percent to 61.5 percent scoring at or above proficient. The action steps established by the team called for formative assessments in math for grades 3 – 6 using the state practice achievement tests for math to establish baselines and areas of strength and weaknesses across students across strands (e.g., number sense, measurement, geometry, patterns and functions, and data analysis). Once student needs in these areas were identified, the team decided that teachers would identify enrichment and remediation groups in order to align instruction to meet student needs. Similar to the evaluation components of the team’s reading plan, the team planned to utilize 180-minute blocks of time to focus on reading and mathematics in classrooms, while Title teachers would get an additional hour per grade for remediation purposes. To evaluate progress on these action steps, the team planned on assessing students quarterly using existing District formative assessments and designing intervention plans for students below goal.
Dueber Elementary School

In the 2004-05 school year, 44.4 percent of fourth grade and 57.1 percent of fifth grade students in Dueber Elementary School scored at or above proficient on the Ohio Achievement Test in reading. In mathematics, only 36.1 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school met AYP in reading, but did not meet it in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to meet AYP in reading in grades 3 – 5 and mathematics in grade 3.

According to its schoolwide plan for reading, the school improvement team from Dueber Elementary established goals to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading to 71.2 percent in third grade, 68.3 percent in fourth grade and 68.3 percent in fifth grade in the 2005-06 school year. To achieve these goals, the team chose to focus on practicing short answer and extended response questions, learning and applying vocabulary, and increasing professional development opportunities for staff in best practices in vocabulary instruction. To evaluate progress on these action steps, the team planned on utilizing the results of the Ohio Achievement Test in reading. In a similar manner, the team established goals to improve the mathematics test scores of students.

In mathematics, the team goal was simply to make third grade AYP. To reach this goal, the team decided to concentrate on increasing student achievement in data analysis and probability. A number of activities using graphs were proposed. To evaluate progress on this action steps, the team voted to evaluate their progress through the Ohio Achievement Test in mathematics in 2006.
In the 2004-05 school year, 51.4 percent of fourth grade and 42.9 percent of fifth grade students in Gibbs Elementary School scored at or above proficient on the Ohio Achievement Test in reading. In mathematics, 60 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet AYP in reading, but met it in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading and mathematics.

According to its schoolwide plan for reading, the school improvement team from Gibbs Elementary established goals to meet AYP targets in grades 3 – 5 by increasing the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading to 71.2 percent in third grade, 68.3 percent in fourth grade and 68.3 percent in fifth grade in the 2005-06 school year. To achieve these goals, the team focused on the implementation of four action steps. In the first, reading teachers would implement the four blocks language arts model, emphasizing vocabulary by using one new vocabulary word per week in morning announcements, vocabulary lessons, writing extensions, and class word walls. In the second step, reading teachers would implement the four blocks language arts model and emphasize non-fiction reading. In the third, teachers would implement an existing reading intervention program for students identified as ‘at risk’. In the fourth, the teachers would increase reading learning time using an existing reading curriculum package. An evaluation of these action steps was planned using a variety of methods including an analysis of instructional time, student achievement results on
existing formative reading assessments, use of word walls by teachers, and a count of student referrals to increase reading instruction time.

The team also developed goals to increase mathematics achievement in the third grade. According to the team’s schoolwide plan for mathematics in 2005-06, the goal of the school was to increase third grade student scores on the Ohio Achievement Test in mathematics to 60.6 percent scoring at or above proficient. To achieve this goal, the team decided on three action steps: implementing the District’s mathematics curriculum (i.e., Everyday Mathematics) in the classroom; implementing a mathematics software program (i.e., Riverdeep) to support mathematics instruction; and incorporating cooperative learning structures into classroom lessons. Staff meetings would then be used to review and model positive results, while collaborative planning forms would be used to reflect teacher progress on implementation.

To evaluate progress on these schoolwide goals, the team agreed to collaborate on planning sheets and lesson plans, emphasizing that all teachers would use the cooperative structures following the leadership team calendar. The plan also called for all students to complete lessons weekly.

Harter Elementary School

In the 2004-05 school year, 78.4 percent of fourth grade and 83.9 percent of fifth grade students in Harter Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Similarly, 73 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet AYP in reading, but met it in mathematics. In response, school improvement team
members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading and mathematics.

According to its schoolwide plan for reading, the school improvement team from Harter Elementary established goals to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading to 75 percent or more in grades 3 – 5. To achieve these goals, the team chose to create action steps that focused on the acquisition of vocabulary, reading process and informational text strands of the state reading standards. To evaluate progress towards these steps, the team identified several existing formative and summative assessment measures utilized by the District. The team’s approach to mathematics improvement is remarkably similar.

The goal of Harter’s schoolwide plan for mathematics in 2005-06 is to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in mathematics to 75 percent or more in grades 3 – 5. To achieve these goals, the team chose to create action steps that focused on the number and number sense, measurement, patterns and data analysis strands of the state mathematics standards. To evaluate progress towards these steps, the team identified several existing formative and summative assessment measures utilized by the District within its Everyday Mathematics curriculum.

McGregor Elementary School

In the 2004-05 school year, 57.4 percent of fourth grade and 46.3 percent of fifth grade students at McGregor Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Similarly 42.6 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test. The school did not meet AYP in either
reading or mathematics. Due to this unparalleled situation, McGregor’s school
improvement team generated a school-wide plan for the 2005-06 school year designed to
increase the test scores of all students in reading and mathematics based upon test results
gathered from previous summative assessments.

According to its schoolwide plan for reading, the school improvement team from
Gibbs Elementary established goals to meet or exceed AYP targets in grade 3 (71.2
percent), grade 4 (68.3 percent), and grade 5 (68.3 percent). Additionally, the team
decided to decrease the percent below proficiency by 10 percent in the acquisition of
vocabulary strand at all grade levels on the Ohio Achievement Test in reading. To
achieve these goals, the team focused on the implementation of several steps involving
changes at the student and teacher level.

To strengthen curriculum, the team identified a number of connected curricular
programs and formative assessments that could be utilized to assist struggling readers. In
addition, the team targeted tutoring interventions for selected students, bookmobile visits
and readers’ theater programs and word-of-the-day programs to support student learning.
For staff, professional development and studies of best practices were recommended for
implementation. To evaluate the progress of the school relative to these action steps, the
team selected existing District formative assessments to schedule on regular bases.
The team also developed goals to increase mathematics achievement in the third grade.

According to the team’s schoolwide plan for mathematics in 2005-06, the goal of
the school was to increase student scores in grades 3 – 5 on the Ohio Achievement Test
in mathematics to 60.6 percent scoring at or above proficient. In addition, the team set a
goal of decreasing the percent of students performing below proficiency on the data
analysis and probability strand of the state mathematics standards by 10 percent. To achieve this goal, the team planned to implement several steps aimed at changing teaching and learning practices in the school.

To assess student achievement in mathematics formally, the team opted to utilize an existing formative assessment system utilized by the District in all grade levels of the elementary school In addition, Title 1 teachers and volunteers would work one on one with students identified for additional instructional assistance. Flash cards and mathematics night programs were scheduled to be implemented at home, while a schoolwide ‘mad math minute’ program was planned for implementation across grades. In addition, a ‘problem of the day’ would be read each day during the morning announcements, while students would be encouraged to review their answers online on the school’s website. A math marathon was scheduled for January. The importance of uninterrupted math classes was noted as a final action step in this sequence. To evaluate progress on these action steps, the team identified a number of existing formative assessments utilized by the District that are embedded in current instructional programs.

_Schreiber Elementary School_

In the 2004-05 school year, 60 percent of fourth grade and 63.3 percent of fifth grade students at Schreiber Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Similarly, 58.1 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet AYP in reading, but met it in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to increase the test scores of students in reading and mathematics.
According to its schoolwide plan for reading, the school improvement team from Schreiber Elementary established goals to increase the percentage of African-American students scoring at or above proficient on the Ohio Achievement Test in reading in grades 3 – 5 to 50 percent or more. To achieve these goals, the team sought to implement focused Title 1 intervention in those grades using software in alignment with state reading standards. To evaluate progress towards these action steps, the team identified an existing formative reading assessment used by the District and DIBELS. Similarly, the goal of Schreiber’s schoolwide plan for mathematics in 2005-06 is to increase the percentage of African-American students scoring at or above proficient on the Ohio Achievement Test in mathematics in grades 3 – 5 to 50 percent or more. To achieve these goals, the team sought to implement focused Title 1 intervention in those grades using software in alignment with state mathematics standards. To evaluate progress towards these action steps, the team identified an existing formative assessment used by the District to use in short-cycle and pre/post assessments.

Summit Elementary School

In the 2004-05 school year, 33.3 percent of fourth grade and 45.5 percent of fifth grade students in Summit Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Only 23.8 percent of fourth grade students scored at or above proficient on the Ohio Achievement Test in mathematics. The school did not meet AYP in reading or mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to increase the test scores of all students in reading and mathematics.
According to its schoolwide plan for reading, the school improvement team from Summit Elementary established goals to meet AYP in third, fourth and fifth grade reading by increasing the percentage of students scoring at or above proficiency to 71.2 percent, 68.3 percent and 68.3 percent, respectively. To achieve these goals, the team focused on four objectives: administering weekly formative reading assessments that mirrored the Ohio Achievement Test in reading; identifying the strengths and weaknesses of students through a comparative item analysis; aligning reading lessons with the identified weaknesses of individual students; and continually employing a student learning cycle with student assessments, curriculum lessons and instruction. To evaluate progress towards these objectives, the team decided to monitor and chart students’ weekly assessments. The team also developed goals to increase mathematics achievement in the third grade.

According to the team’s schoolwide plan for mathematics in 2005-06, the goal of the school was to increase third grade student scores on the Ohio Achievement Test in mathematics to 60.6 percent scoring at or above proficient. To achieve this goal, the team sought to implement action steps that integrated in-class assignments, during and after-school tutoring programs, existing formative assessments utilized by the District, and needs-based interventions with individual students. Planned action step evaluations included formative assessment analysis and teacher observations.

Youtz Elementary School

In the 2004-05 school year, 56.1 percent of fourth grade and 48.6 percent of fifth grade students in Youtz Elementary School scored at or above proficient on the Ohio Achievement Test in reading. Only 38.6 percent of fourth grade students scored at or
above proficient on the Ohio Achievement Test in mathematics. The school met AYP in reading, but did not meet in mathematics. In response, school improvement team members generated a school-wide plan for the 2005-06 school year designed to, among other things, increase the test scores of all students in reading and mathematics.

According to its schoolwide plan for reading, the school improvement team from Youtz Elementary established goals to increase the percentage of students scoring at or above proficient on the Ohio Achievement Test in reading to 71.2 percent in third grade, and 68.3 percent in both fourth and fifth grade in the 2005-06 school year. To achieve these goals, the team settled on implementing the four blocks language arts model emphasizing the acquisition of vocabulary and non-fiction reading, and utilizing an existing intervention and formative assessment system used by the District in reading. To evaluate the school’s success at meeting these objectives, the team agreed to determine the percentage of teachers posting word walls and analyzing the percentage of instructional time devoted to non-fiction reading and writing.

The improvement goal for mathematics established by the team was that 60.6 percent of third grade students would score at or above proficient on the Ohio Achievement Test in the 2005-06 school year. To reach this goal, the team recommended the implementation of the existing mathematics curriculum and the use of existing mathematics software. In addition, the team planned to incorporate cooperative learning structures into classroom lessons. To evaluate the school’s progress towards these objectives, the team noted several criteria to track, including staff attendance at professional development lessons in mathematics curriculum and software learning sessions.
APPENDIX B

TEACHER ASSESSMENT QUESTIONNAIRE
CANTON CITY SCHOOL DISTRICT TEACHER QUESTIONNAIRE

Identification Label

Teacher ID: 
Building: 

Your school has agreed to participate in a pilot test of Canton City School District’s evaluation of its new assessment system. As part of this study, students will be taking new assessments next year as part of the District’s continuous improvement plan and Ohio’s accountability system. In order to plan strategically for these new assessments, the District is seeking baseline information. This questionnaire is addressed to elementary school teachers who teach reading and mathematics, and seeks information about teachers’ academic and professional background, instructional practices, and attitudes towards teaching reading and mathematics.

It is important that you answer each question carefully so that the information you provide reflects your situation as accurately as possible. Your responses on this questionnaire will remain confidential. Answering this questionnaire should require no more than twenty minutes. To make it as easy as possible for you to respond, most of the questions can be answered by checking or filling in the appropriate circle with a black or blue pen or a pencil. Once you have completed it, please return it to the representative from the Ohio State University. Thank you very much for the time and effort you have put into responding to this questionnaire.

START HERE

Background Information

1. How old are you?
   
   Under 25 O  40 - 49 O
   25 - 29 O  50 - 59 O
   30 - 39 O  60 or older O

2. Are you male or female?
   
   Female O
   Male O

3. By the end of this school year, how many years will you have been teaching altogether?

   Number of years you have taught

4. What is the highest level of formal education you have completed?

   BA O  MA O  Ph.D. O
   BA + 30 O  MA + 30 O

   1
Mathematics Instruction

5. Considering your training and experience in mathematics content and instruction, how ready do you feel you are to teach this topic in your classroom on the following scale: (3 = ready; 2 = somewhat ready; 1 = not ready)?

Write number in space provided

a. Number, number sense and operations
b. Measurement
c. Geometry and spatial sense
d. Patterns, functions and algebra
e. Data analysis and probability

6. In the past two years, have you participated in professional development in any of the following?

Fill in circles as appropriate

a. Mathematics content
b. Mathematics pedagogy/instruction
c. Mathematics curriculum
d. Integrating information technology into mathematics
e. Improving students’ critical thinking/problem-solving skills
f. Mathematics assessment

7. How many minutes per week do you teach mathematics to your class?

Write in the number of minutes per week

8. In a typical week of mathematics lessons for your students, what percentage of time do students spend on each of the following activities?

Write the percent
The total should add to 100%

a. Reviewing homework
b. Listening to lecture-style presentations
c. Working problems with your guidance
d. Working problems on their own
e. Listening to you clarify content and/or procedures
f. Taking tests or quizzes
g. Participating in classroom management not related to the lesson’s purpose (e.g. keeping order)
h. Other student activities

Total 100 %
9. How often do you give classroom mathematics tests or examinations?

   Fill in all that apply
   a. About once a week
   b. About every two weeks
   c. About once a month
   d. A few times a year

10. What item formats do you typically use in your mathematics tests or examinations?

    Fill in one circle only
    a. Short-answer/extended response
    b. True-false
    c. Multiple-choice
    d. Matching
    e. Essays

11. To answer the following question, please consider the following definition: *A formative test or examination is given to students as they are learning a concept in order to judge the quality of their progress in learning a concept. Teachers use such tests in order to improve their teaching and help guide students’ learning. What types of decisions about students do you make from the formative assessments that you administer in your classroom relative to mathematics?*

    Fill in all that apply
    a. Planning instructional activities
    b. Placing students into learning sequences
    c. Monitoring students’ progress
    d. Diagnosing learning difficulties

Are there any other decisions that formative mathematics assessments inform? If so, please write them below in the space provided.

__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________

3
12. To answer the following question, please consider the following definition: A summative test or examination is given to students at the conclusion of a lesson (e.g., Ohio Achievement Test). Teachers use such tests in order to identify the quality of student achievement. What types of decisions about students do you make from the summative assessments that you administer in your classroom relative to mathematics?

*Fill in all that apply*

- a. Feedback to students and parents about achievement
- b. Feedback to teacher about effectiveness
- c. Assigning grades to students

Are there any other decisions that summative mathematics assessments inform? If so, please write them below in the space provided.
Reading Instruction

13. Considering your training and experience in reading content and instruction, how ready do you feel you are to teach this topic in your classroom on the following scale:
   (3 = ready; 2 = somewhat ready; 1 = not ready)?

   Write number in space provided
   a. Phonemic awareness, word recognition & fluency
   b. Acquisition of vocabulary
   c. Reading processes
   d. Reading applications: Informational, technical & persuasive text
   e. Reading applications: Literary text

14. In the past two years, have you participated in professional development in any of the following?

   Fill in circles as appropriate
   a. Reading content
   b. Reading pedagogy/instruction
   c. Reading curriculum
   d. Integrating information technology into reading
   e. Reading assessment

15. How many minutes per week do you teach reading to your class?

   Write in the number of minutes per week

16. In a typical week of reading lessons for your students, what percentage of time do students spend on each of the following activities?

   Write in the percent
   The total should add to 100%

   a. Small group reading instruction
   b. Read aloud (think aloud) activities
   c. Sequencing of story events
   d. Comprehension activities (e.g., generating questions)
   e. Vocabulary instruction
   f. Taking tests or quizzes
   g. Participating in classroom management not related to the lesson's purpose (e.g., keeping order)
   h. Other student activities

   Total

   100 %
17. How often do you give classroom reading tests or examinations?

Fill in one circle only

a. About once a week
b. About every two weeks
c. About once a month
d. A few times a year

18. What item formats do you typically use in your reading tests or examinations?

Fill in one circle only

a. Short-answer/extended response
b. True-false
c. Multiple-choice
d. Matching
e. Essays

19. To answer the following question, please consider the following definition: A formative test or examination is given to students as they are learning a concept in order to judge the quality of their progress in learning a concept. Teachers use such tests in order to improve their teaching and help guide students’ learning. What types of decisions about students do you make from the formative assessments that you administer in your classroom relative to reading?

Fill in all that apply

a. Planning instructional activities
b. Placing students into learning sequences
c. Monitoring students’ progress
d. Diagnosing learning difficulties

Are there any other decisions that formative reading assessments inform? If so, please write them below in the space provided.

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

6
20. To answer the following question, please consider the following definition: A summative test or examination is given to students at the conclusion of a lesson (e.g., Ohio Achievement Test). Teachers use such tests in order to identify the quality of student achievement. What types of decisions about students do you make from the summative assessments that you administer in your classroom relative to reading?

*Fill in all that apply*

a. Feedback to students and parents about achievement  ○
b. Feedback to teacher about effectiveness  ○
c. Assigning grades to students  ○

Are there any other decisions that summative reading assessments inform? If so, please write them below in the space provided.
School Background

21. How would you characterize each of the following within your school on the following scale? (very high = 5; high = 4; medium = 3; low = 2; very low = 1)

Write number in space provided

a. Teachers’ job satisfaction
b. Teachers’ understanding of the school’s curricular goals
c. Teachers’ degree of success in implementing the school’s curriculum
d. Teachers’ expectations for student achievement
e. Parental support for student achievement
f. Students’ desire to do well in school

22. In an average month, how many days do you have the following types of interactions with other teachers?

Write number in space provided

a. Discussion about how to teach a particular concept
b. Working on preparing instructional materials
c. Visits to another teacher’s classroom to observe their teaching
d. Informal observations of my classroom by another teacher

END HERE

Thank you very much for completing this questionnaire. Please return this document to the representative from The Ohio State University.
APPENDIX C

STRUCTURED INTERVIEW PROTOCOLS
Canton City School District Structured Teacher Interview Protocol
on Value-Added Assessment

1) What challenges, if any, have you had in working with the value-added student achievement data delivered through Battelle for Kids?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

2) What opportunities do you see for this assessment in our District?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

3) Have you had any challenges accessing the data? (If so, please describe them)
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

4) What response have you gotten from other classroom teachers regarding the value-added student achievement data delivered through Battelle for Kids?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

5) What adjustments, if any, have you been able to make to instruction as a result of this data? If not, what additional support would you need?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
6) When the District uses assessment from this system in all schools beginning in 2006-2007, what challenges would you see or what advice would you give?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

7) What opportunities would you see for the District with this assessment data?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

8) Do you feel that you could serve as a resource to others in your building? If not, what additional support would you need?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

9) Is there anything else that you would like to share with me regarding your experiences with the value-added assessment system delivered through Battelle for Kids?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

Teacher ID: _________________
1) What challenges, if any, have you had this year implementing the ETS – Pulliam formative mathematics assessment system?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

2) What opportunities do you see for this assessment in our District?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

3) Have you had any challenges accessing the data? (If so, please describe them)

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

4) What response have you gotten from your students regarding the ETS – Pulliam formative assessment system of mathematics achievement?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

5) What response have you gotten from classroom teachers regarding the ETS – Pulliam formative mathematics assessment system?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

6) What adjustments, if any, have you been able to make to instruction as a result of this data? If not, what additional support would you need?
7) If the District were to scale up and assess all K-3rd graders with this assessment, what challenges would you see or what advice would you give?

8) What opportunities would you see for the District with a K-3 scale-up?

9) Do you feel that you could serve as a resource to others in your building? If not, what additional support would you need?

10) Is there anything else that you would like to share with me regarding your experiences with the ETS – Pulliam formative assessment system of mathematics achievement?
Canton City School District Structured Teacher Interview Protocol
Regarding the DIBELS Reading Assessment

1) What challenges, if any, have you had this year implementing the DIBELS?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

2) What opportunities do you see for this assessment in our District?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

3) Have you had any challenges accessing the data? (If so, please describe them)
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

4) What response have you gotten from your students regarding the DIBELS?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

5) What response have you gotten from other classroom teachers regarding the DIBELS?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

6) What adjustments, if any, have you been able to make to instruction as a result of this data? If not, what additional support would you need?
7) If the District were to scale up and assess all K-3rd graders with this assessment, what challenges would you see or what advice would you give?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

8) What opportunities would you see for the District with a K-3 scale-up?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

9) Do you feel that you could serve as a resource to others in your building? If not, what additional support would you need?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

10) Is there anything else that you would like to share with me regarding your DIBELS experience?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
APPENDIX D

OHIO ACHIEVEMENT TEST BLUEPRINTS
<table>
<thead>
<tr>
<th>Selection / # of Items</th>
<th>Length/Type</th>
<th>Standard</th>
<th>Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Multiple Choice (1 point)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Short Answer (2 points)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Extended Response (4 points)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Points</td>
</tr>
<tr>
<td>1</td>
<td>Long (501 - 700)</td>
<td>(5-7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literary Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Process</td>
<td>1-2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Reading Applications</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>2-3</td>
<td>2-3</td>
</tr>
<tr>
<td>2</td>
<td>Long (501 - 700)</td>
<td>(5-7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Informational Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Process</td>
<td>1-2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Reading Applications</td>
<td>2-4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>2-3</td>
<td>2-3</td>
</tr>
<tr>
<td>3</td>
<td>Medium (301 - 500)</td>
<td>(5-7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literary Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Process</td>
<td>1-4</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>Reading Applications</td>
<td>1-3</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>1-2</td>
<td>1-2</td>
</tr>
<tr>
<td>4</td>
<td>Medium (301 - 500)</td>
<td>(5-7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Informational Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Process</td>
<td>1-4</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>Reading Applications</td>
<td>1-3</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>1-2</td>
<td>1-2</td>
</tr>
<tr>
<td>5</td>
<td>Short (up to 300)</td>
<td>(2-5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Informational OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literary Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Process</td>
<td>1-2</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>Reading Applications</td>
<td>1-2</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>0-1</td>
<td>0-1</td>
</tr>
<tr>
<td>Number of Scored Items</td>
<td>29</td>
<td>4 or 6</td>
<td>2 or 3</td>
</tr>
<tr>
<td>Total Numbers</td>
<td>36 or 37 items</td>
<td>49</td>
<td>points</td>
</tr>
</tbody>
</table>

Source: Ohio Department of Education
Office of Assessment
October 16, 2003
Item Distribution

Items are distributed among three item formats on each form of the test according to the following guidelines:

- Each of the four standards is assessed by multiple-choice items and at least one constructed-response item on every form of the test.
- Each operational form will typically include 6 Field Test items.

Number of Points by Standard

Each of the standards will have a minimum number of points on each test form. The chart below shows this distribution.

<table>
<thead>
<tr>
<th>Points By Standard</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Process</td>
<td>14-17</td>
</tr>
<tr>
<td>Reading Applications – Informational, Technical and Persuasive Text</td>
<td>11-13</td>
</tr>
<tr>
<td>Reading Applications – Literary Text</td>
<td>11-13</td>
</tr>
<tr>
<td>Acquisition of Vocabulary</td>
<td>8-9</td>
</tr>
</tbody>
</table>

Reporting of Results

Results will be reported using a scaled score for overall achievement as well as raw score points received for each of the following standards:

- Reading Process: Concepts of Print, Comprehension Strategies and Self-Monitoring Strategies
- Reading Applications: Informational, Technical and Persuasive Text
- Reading Applications: Literary Text
- Acquisition of Vocabulary

Source: Ohio Department of Education
Office of Assessment
October 16, 2009
## OHIO Grade 5 Reading Achievement Test Blueprint

<table>
<thead>
<tr>
<th>Selection/Length/Type</th>
<th>Item Type</th>
<th>Standard</th>
<th>Multiple Choice (1 point)</th>
<th>Short Answer (2 points)</th>
<th>Extended Response (4 points)</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong> Long (601 - 900)</td>
<td>Literary Text</td>
<td>Reading Process</td>
<td>1-2</td>
<td>1</td>
<td>6-8</td>
<td>3-4</td>
</tr>
<tr>
<td>(7-9) Reading Applications</td>
<td></td>
<td>2-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>2-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2</strong> Long (601 - 900)</td>
<td>Informational Text</td>
<td>Reading Process</td>
<td>1-2</td>
<td>1</td>
<td>6-8</td>
<td>3-4</td>
</tr>
<tr>
<td>(7-9) Reading Applications</td>
<td></td>
<td>2-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>2-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3</strong> Medium (301 - 600)</td>
<td>Literary Text</td>
<td>Reading Process</td>
<td>1-4</td>
<td>0-1</td>
<td>6-8</td>
<td>3-5</td>
</tr>
<tr>
<td>(6-8) Reading Applications</td>
<td></td>
<td>1-3</td>
<td>0-1</td>
<td>0-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4</strong> Medium (301 - 600)</td>
<td>Informational Text</td>
<td>Reading Process</td>
<td>1-4</td>
<td>0-1</td>
<td>6-8</td>
<td>3-5</td>
</tr>
<tr>
<td>(6-8) Reading Applications</td>
<td></td>
<td>1-3</td>
<td>0-1</td>
<td>0-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5</strong> Short (up to 300)</td>
<td>Informational OR Literary Text</td>
<td>Reading Process</td>
<td>1-2</td>
<td>0-1</td>
<td>6-8</td>
<td>3-5</td>
</tr>
<tr>
<td>(4-5) Reading Applications</td>
<td></td>
<td>1-2</td>
<td>0-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of Scored Items: 29 4 or 6 2 or 3 --

Total Number of Items: 36 or 37 items 49 points

Source: Ohio Department of Education
Office of Assessment
October 16, 2003
Item Distribution

Items are distributed among three item formats on each form of the test according to the following guidelines:

- Each of the four standards is assessed by multiple-choice items and at least one constructed-response item on every form of the test.
- Each operational form will typically include 6 Field Test Items.

Number of Points by Standard

Each of the standards will have a minimum number of points on each test form. The chart below shows this distribution.

<table>
<thead>
<tr>
<th>Points By Standard</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Process</td>
<td>14-17</td>
</tr>
<tr>
<td>Reading Applications – Informational, Technical and Persuasive Text</td>
<td>11-13</td>
</tr>
<tr>
<td>Reading Applications – Literary Text</td>
<td>11-13</td>
</tr>
<tr>
<td>Acquisition of Vocabulary</td>
<td>8-9</td>
</tr>
</tbody>
</table>

Reporting of Results

Results will be reported using a scaled score for overall achievement as well as raw score points received for each of the following standards:

- Reading Process: Concepts of Print, Comprehension Strategies and Self-Monitoring Strategies
- Reading Applications: Informational, Technical and Persuasive Text
- Reading Applications: Literary Text
- Acquisition of Vocabulary

Source: Ohio Department of Education
Office of Assessment
October 16, 2008