COMMODITIZING EDUCATION:
THE VALIDITY OF TESTING PRACTICES AT AN URBAN ELEMENTARY SCHOOL

Craig Allen Seale
B.A. (English), California State University, Chico, 1991
B.A. (Philosophy), California State University, Chico, 1991

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF ARTS

in

ENGLISH
(Composition)

at

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

SPRING
2010
Student: Craig Allen Seale

I certify that this student has met the requirements for format contained in the
University format manual, and that this thesis is suitable for shelving in the Library
and credit is to be awarded for the thesis.

__________________________, Department Chair     ________________________
Dr. Sheree Meyer              Date

Department of English
Abstract of

COMMODITIZING EDUCATION: THE VALIDITY OF TESTING PRACTICES AT AN URBAN ELEMENTARY SCHOOL

by

Craig Allen Seale

The author writes a case study examining the narrowing effect that high-stakes testing was having on his elementary school. Using a definition of testing validity put forth by Samuel Messick, the author determines that certain aspects of his school’s testing protocol do not measure up to the standard of being valid testing procedures. Staff interviews, district assessment data, and secondary sources from the field of Education compliment the anecdotal evidence which portrays an elementary school increasingly strained by the burden of high-stakes testing.

_______________________, Committee Chair
Dr. Dan Melzer

_______________________
Date
ACKNOWLEDGMENTS

I would not have been able to undertake such an endeavor without the daily support and encouragement from my wife and grade-level partner, Brooke. I also want to thank Dr. Melzer, Dr. Underwood, and Dr. Meyer for their assistance.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION TO THE STUDY</td>
<td>2</td>
</tr>
<tr>
<td>Research Questions and Methods</td>
<td>9</td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td>11</td>
</tr>
<tr>
<td>Organization of the Study</td>
<td>13</td>
</tr>
<tr>
<td>2. LITERATURE REVIEW</td>
<td>15</td>
</tr>
<tr>
<td>Gate-Keeping</td>
<td>15</td>
</tr>
<tr>
<td>The Narrowing Effect</td>
<td>16</td>
</tr>
<tr>
<td>Reliability</td>
<td>22</td>
</tr>
<tr>
<td>Validity</td>
<td>24</td>
</tr>
<tr>
<td>Score Inflation</td>
<td>28</td>
</tr>
<tr>
<td>Loose-Coupling</td>
<td>36</td>
</tr>
<tr>
<td>3. PARKER ELEMENTARY: THE CASE STUDY</td>
<td>41</td>
</tr>
<tr>
<td>Chapter/Section</td>
<td>Page</td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
</tr>
<tr>
<td>Data-Based Student Performance</td>
<td>43</td>
</tr>
<tr>
<td>The Push Toward Proficiency</td>
<td>55</td>
</tr>
<tr>
<td>Performance Rewards &amp; Test Prep</td>
<td>61</td>
</tr>
<tr>
<td>Instructional Hours</td>
<td>65</td>
</tr>
<tr>
<td>Administrative Policies vs. Classroom Realities</td>
<td>68</td>
</tr>
<tr>
<td>Commoditizing Education</td>
<td>73</td>
</tr>
<tr>
<td>4. RECOMMENDATIONS AND CONCLUSIONS</td>
<td>83</td>
</tr>
<tr>
<td>Testing Validity</td>
<td>83</td>
</tr>
<tr>
<td>Restructuring Leadership</td>
<td>94</td>
</tr>
<tr>
<td>Conclusion</td>
<td>104</td>
</tr>
</tbody>
</table>

Appendices

Appendix A: Teacher Survey | 106
Appendix B: Support Staff Survey | 108
Works Cited | 109
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Figure 1. 5th Grade ELA Scores 2008-09</td>
<td>9</td>
</tr>
<tr>
<td>2.</td>
<td>Figure 2. STAR Test Categories</td>
<td>48</td>
</tr>
<tr>
<td>3.</td>
<td>Figure 3. ELA Assessments: Girls</td>
<td>49</td>
</tr>
<tr>
<td>4.</td>
<td>Figure 4. ELA Assessments: Boys</td>
<td>50</td>
</tr>
<tr>
<td>5.</td>
<td>Figure 5. Vocabulary: District Test</td>
<td>51</td>
</tr>
<tr>
<td>6.</td>
<td>Figure 6. Spelling: Prompt</td>
<td>52</td>
</tr>
<tr>
<td>7.</td>
<td>Figure 7. Grammar: Test vs. Prompt</td>
<td>54</td>
</tr>
<tr>
<td>8.</td>
<td>Figure 8. 4x16-foot ELA Data Wall</td>
<td>76</td>
</tr>
<tr>
<td>9.</td>
<td>Figure 9. Data Wall (close-up)</td>
<td>76</td>
</tr>
</tbody>
</table>
“As we develop these tests, we must provide safeguards to ensure that they do not result in unhealthy competition among our schools or geographic areas. Even more important, we must be certain that they do not encourage our teachers to ‘teach to the tests’ rather than the curriculum.”

Governor Edmund G. Brown’s opening remarks to the 1961 California legislature (qtd. in Underwood 62).
Chapter 1

INTRODUCTION TO THE STUDY

Public schools across the country have been trending towards practices which assess and measure student and teacher performance. Of these methods, the most commonly used are standardized tests. At Parker Elementary (a pseudonym)—the school in which I have been teaching for the past eleven years—there has been an increasing emphasis placed upon the use of such assessments. In fact, the extent to which these assessment devices are emphasized has brought me to question whether our reliance upon standardized testing has affected the educational experience of our students in such a way as to become unbalanced and narrowed. My thesis explores the many aspects of how Parker implements assessments, in particular, the impact they have on students, teachers, and instructional resources. My concern is that the academic experience is becoming commoditized into a product bereft of many of the distinguishing qualities which differentiate one child’s educational career from another. Furthermore, as standardized assessments continually narrow the focus of classroom instruction, the overall education for our students becomes increasingly barren and fatigued. Clearly, in a state as massively large and diverse as California, the intention to standardized instruction must be brought into alignment with the regional, cultural and societal factors which comprise our public educational landscape.

How much testing is needed in order to adequately evaluate student and teacher performance? With limited resources—including a finite number of instructional hours
over the course of a school year—what percentage of time can be devoted to testing? As
time is given to testing, how are other educational components valued? Are Science and
Social Studies second-tier disciplines, given that they have little or no representation on
the California Standardized Testing and Reporting exams (STAR test)? How is Physical
Education prioritized being that it is mandated under state law, but does not factor into a
school’s Adequate Yearly Progress (AYP—a derivative from the STAR test data)? Even
if there were instructional minutes and resources left over for the visual and performing
arts, should any time be given to non-academic endeavors, especially for an under-
performing school like mine? These are the questions which echo in the hallways at
Parker Elementary. By examining how my school allocates its instructional resources
towards testing, I will discuss the various considerations which comprise the overall
educational experience provided to our students.

The axis upon which I examine the testing practices at Parker is notion of
“validity” as put forth by Samuel Messick. He contends that a testing practice must first
be scrutinized before it can be considered a valid measurement. This scrutiny includes
looking at the theoretical premises which precede the development of the testing
procedures, along with evidenced gathered afterward to indicate its success. Messick’s
definition of test validity can be used to determine whether Parker’s testing
methodologies are sound. Therefore, the theoretical rationales by which testing materials
are developed, the empirical evidence which supports the testing procedures, and the
decisions which result from the testing-generated data, can not only provide a fair
judgment of a testing protocol, they can also reasonably evaluate the overall academic rigor of a school like mine.

Having conducted a substantial amount of time researching the literature in the fields of assessment, interviewing the staff at my school, examining all Reading and Language Arts testing data accumulated over the 2008-2009 academic year, and written and reflected extensively on this topic for the better part of two years, it seems perhaps a bit disingenuous to suggest uncertainty as to how my school will fare under the scrutiny of the validity argument.\(^1\) I know that the results will be varied, some indications being that we are moving in a direction which favors a valid and evolved use of instructional resources including assessment measures. Other indications will point to a woeful lack of vision and wherewithal needed to keep up with the high-stakes consequences associated with contemporary public schooling. Nonetheless, one thing is clear: looking at the assessment practices at Parker Elementary through the lens of validity will initiate a discussion regarding school accountability that really started in the public forum a long time ago, arguably with the 1983 publication of *A Nation at Risk*; it will also shed light on some possible directions our school and school district might pursue, or at least consider; and it will also make manifest the fact that if little is done to alter the direction of our school’s present course of action, a certain future will be upon us in which the educational experience for all stakeholders—none more important than the children—will move us perilously close to a commoditized educational product.

\(^1\) The terms English and Language Arts (ELA) and Reading and Language Arts are used interchangeably.
That is not to say, however, that this thesis will be merely criticizing the shortcomings of standardized testing; many teachers, including myself, understand the value derived from such accountability measures. The romantic vision that all teachers are equally skilled and motivated must be abandoned. Reliable and equitable monitoring of teacher performance is a prudent practice. In this age, when a child’s welfare is increasingly vulnerable to the untoward influences like single-parenthood, substance abuse and/or gang violence, it is incumbent upon the schools to take the stewardship of their students more earnestly than ever before. Whether or not we agree with all the outcomes from this evolution towards accountability measures (none less significant than the 2001 reauthorization of the Elementary and Secondary Act of 1965, Public Law 107-110, commonly referred to as “No Child Left Behind” or “NCLB”), it must be recognized that implicit within the educational process is the stewardship of our nation’s children. Thomas Sizer says, “NCLB and the state initiatives that preceded it have served the purpose of putting in front of all citizens the stunning weaknesses of our schooling systems, especially as it serves—or fails to serve—its most vulnerable young citizens” (61).

Furthermore, assessments provide the fuel to drive both accountability measures, as well as classroom instruction. Edys S. Quellmalz, author of the article “Designing Writing Assessments: Balancing Fairness, Utility, and Cost” says that “if schools are to promote language competence, educators, students, and parents must have information describing the status and progress of language skills development” (64). Quellmalz and I agree that there is no avoiding the fact that we live in an era of accountability, where
measurable data must be produced in order to satisfy both educational and budgetary demands.

While time will be taken to show how my school effectively utilizes testing data to better inform and direct instructional resources, the essence of my study will question the sheer volume of testing done at Parker, for it is within this consideration of testing volume that the threat of commoditization lurks. As instructional resources are depleted, I believe a compelling case will be made regarding the question of educational balance. I will compare my school with other school districts in order to better see how assessments are prioritized and used. At Parker last year, the total number of Reading and Language Arts assessments for 1st through 6th grades totaled approximately 29,856 tests. To put that into a physical context, if each assessment were a single page of paper (which is not always the case), we delivered over 600 pounds of ELA assessments last year. That total does not include Kindergarten, which is rigorously assessed and scrutinized under a program called Reading First. Nor do these totals include the Mathematics curriculum which assesses approximately 20 times a year per student per grade level.

Locating those assessment quantities within the context of the instructional hours of an academic year helps illuminate the role these assessments play in the overall educational experience of the student. In the Shriver Unified School District—SUSD (a pseudonym), we are contracted to teach 990 instructional hours over the course of a single academic year. Those 990 hours are spread out over 180 school days, averaging 5 ½ hours per day. In “A Student-Centered Approach to Assessment,” Franklin L. Smith and Zollie Stevenson Jr. report that a comparable urban school district in Washington
D.C. administers the exact same number of instructional hours per year as we do. That same school district devotes 21 hours an academic year to test taking, “exclusive of teacher-made classroom tests. This is not excessive when you consider that 969 hours remain for instruction” (Smith 79). In comparison, our school allots 30 hours per year to ELA assessments alone. Adding the approximately 15 hours or so for Math testing each year, the total number of instructional hours set aside for assessments totals 45 hours per year. Already, the amount of testing administered by SUSD more than doubles the amount allocated by the District of Columbia Public Schools (DCPS). Factor in the 10 hours or so of STAR testing each spring, and that testing number totals 55 hours per year. Any other testing—including the state mandated PE testing, Science, Social Studies, etc.—are not included in that 55 hour total; that total includes ELA, Math and STAR testing only. Those 55 testing hours account for nearly 6% of the allotted instructional time in our district, as compared with the DCPS, which devotes 2% of their instructional time to testing.

Naturally, the finite amount of time available for instruction must be rationed carefully. Although obvious, the relationship between academic improvement and the amount of instructional time devoted to a targeted goal cannot be overstated. Instructional hours are the blood within an academic body. As will be specified regarding my school, the manner in which those hours are properly managed is the degree to which a school experiences success. Furthermore, the manner in which those hours are properly managed is the degree to which the educational experience has a rich and meaningful resonance for the students.
Before going any further, it would be remiss for me not to locate myself within the context of this case study. I have taught for eleven years at the 5th grade at Parker Elementary alongside my wife as my grade-level partner. She and I were hired before we had met each other to fill the two vacancies at the 5th grade at Parker. We immediately started working collaboratively as grade-level partners, pooling our ideas, energy and resources towards our two classes. Three years later, our professional esteem had blossomed into love and marriage. I mention this primarily because of its uniqueness; very few couples I know have the opportunity to professionally immerse themselves with their spouse. As may become more apparent in the general tone and my presentation of my study, my passion for Parker Elementary includes not only my devotion to a profession that I find extremely gratifying, but also the opportunity I have on a daily basis to work with someone I admire beyond description.

Therefore, it is my first priority to write a thesis that will honor and represent the practical and centrist mindset of my colleagues. My coworkers are fair-minded and trustworthy individuals who have negotiated the difficult task of balancing curriculum, assessment, limited instructional resources, and precious human assets. While I look forward to telling our story within the forum of academic discourse, I hope my project speaks to the extended Parker family. I aim to write a paper in which my scrutiny is thoroughly grounded within the practical realities of the Parker experience, but which most of all, honors the educational community to which so many of us are devoted.
Research Questions and Methods

Three years ago I questioned whether the amount of testing at Parker was raining huge amounts of negative feedback upon our students. The combination of a Parker’s history of underperformance plus a burdensome testing regimen seemed like a recipe for psychological and emotional distress. It was for this reason that I went through the SUSD testing database and counted the number of 5th grade ELA tests for the 2008-09 academic year. I also coded each test score, placing it one of the five STAR test categories: Advanced, Proficient, Basic, Below Basic, or Far Below Basic (see Figure 1).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced</td>
<td>59</td>
<td>101</td>
<td>64</td>
<td>21</td>
<td>13</td>
<td>103</td>
<td>94</td>
<td>16</td>
<td>106</td>
</tr>
<tr>
<td>Proficient</td>
<td>30</td>
<td>78</td>
<td>56</td>
<td>56</td>
<td>39</td>
<td>43</td>
<td>117</td>
<td>69</td>
<td>79</td>
</tr>
<tr>
<td>Basic</td>
<td>33</td>
<td>54</td>
<td>46</td>
<td>155</td>
<td>141</td>
<td>44</td>
<td>27</td>
<td>122</td>
<td>55</td>
</tr>
<tr>
<td>Below Basic</td>
<td>162</td>
<td>62</td>
<td>100</td>
<td>109</td>
<td>154</td>
<td>73</td>
<td>5</td>
<td>124</td>
<td>58</td>
</tr>
<tr>
<td>Far Below Basic</td>
<td>81</td>
<td>82</td>
<td>113</td>
<td>41</td>
<td>41</td>
<td>89</td>
<td>30</td>
<td>48</td>
<td>85</td>
</tr>
</tbody>
</table>

FIGURE 1. 5TH GRADE ELA SCORES 2008-09

I found that in fact a large number of test scores did reside in the lowest categories; however, there were also a large numbers of scores in the higher categories as well. What I was left with was the conclusion that there were a large number of ELA tests last year and every year. From there, my research focus shifted to an examination of the overall volume of testing and its narrowing effect on instruction.
I did, however, want to include in the discussion a description of how testing data is used at Parker to improve student performance. In order to fully portray the manner in which testing data is being used at Parker it became evident that teacher voices would need to be sought. I wrote a brief, single-page survey (see Appendix 1) and gave it all but the newly-hired teacher. The basis of the questioning was to examine to what extent the teachers could either verify our testing practices met the standard of testing validity. To that end my questions asked about the theoretical premises that went into our testing procedures, along the empirical evidence and actions which resulted from our test data. Twelve out of 18 teachers responded. I wrote a similar survey for support staff and received back four more responses. I tried to make a point of encouraging staff members to be themselves. I reassured them that I was not looking for a chorus of like-minded voices; rather, I wanted to locate myself within the actual composition of Parker stakeholders. While organizing my school-based research on the one hand, I took the direction of my advisors to draw upon the foundational authors within the discussion of assessment reform. The confluence of broad-based staff responses, testing data, and educational research and commentary converged into a pool of resources within which I could reasonably draw conclusions, make inferences and suggest changes for Parker elementary.

Limitations of the Study

Attempting to portray the assessment procedures for an entire school is a bit misleading, as differences exist between grade levels. The primary grades have worked
under the scrutiny of the Reading First grant, which has requirements unique to them. The primary grades also have different expectations in terms of the district generated tests, as well as the state-mandated tests. Furthermore, some grades have different types of state-mandated tests: fourth grade has its own Writing exam; fifth grade has a Science assessment. Therefore, any attempt to represent the overall assessment practices of my school should be considered more accurately as a confederation of practices among each grade level which carry with it some generalizable patterns.

The procedure by which I received staff feedback was limited by various time constraints, both by me and my coworkers. The manner in which I solicited staff input was through the use of a brief survey. The aim being to ground my study within a chorus of voices from Parker, I feel that the survey accomplished its purpose. However, a more detailed representation could have been accomplished through the use of a more lengthy survey, particularly one that would have been better tailored to the idiosyncrasies of each particular grade level. Furthermore, some type of follow-up discussion—perhaps in the form of a roundtable forum—would have added to the breadth of responses.

A significant restructuring occurred at the district level during the two academic years which my study covers. The changes at Shriver Unified School District have been massive, and have resulted in significant changes in all phases of operations. In particular to my study, the district’s assessment database changed during the course of my writing, although the total volume of testing between the 08-09 academic years and this year has remained very similar.
One additional limitation exists in regards to the generalizability of my case study. While Parker elementary certainly resembles other urban elementary schools—and may very well mirror similar practices and perspectives found on other campuses—it would be ill-conceived to base conclusions beyond my own school. If I have learned anything over the course of this case study, the notion that one-size-fits-all has been thoroughly exposed in my opinion as a phrase which provides virtually no real-world application in terms of elementary school life. The same holds true for my elementary school case study conclusion: Parker elementary conclusions do not necessarily fit all other elementary school scenarios.

Organization of the Study

The arrangement of chapters will allow for a systematic analysis of a case study that is otherwise extremely fluid. The second chapter will review a significant portion of the current literature regarding this field of interest. As a Composition thesis, much of the research will point back to the issue of literacy; however, the majority of research has been compiled from the field of Education. The axis between Composition and Education will pivot upon the discussion of “validity,” primarily the contributions from Samuel Messick.
The third chapter will examine the particular details surrounding assessment procedures at Parker Elementary. In some ways, one could say that everything that occurs at a school site pertains to assessment. My study will certainly include many far-ranging topics which directly relate to assessment: some anecdotal, some data-based, all grounded in the story of Parker, which in many ways mirrors the experiences found at other urban elementary schools.

The final chapter will extend some possible remedies to address the current dilemma facing assessment protocol at Parker Elementary. None of the recommendations will satisfy all those involved in the educational process at Parker, be they students and family members, teachers, support staff, administrative and district personnel, board members or political leaders who engage in educational matters. However, it is my belief that many of the considerations made in this final chapter will appeal to the majority of the stakeholders, which is ultimately the purpose of this project.
“Thus, the implementing regulations for NCLB imply that at the end of the twelve-year period specified by the statute, mildly retarded students in the United States should outscore roughly a third of the students in two of the highest-scoring countries ever to participate in international assessments of mathematics (Japan and Korea)” (307).

Measuring Up by Dr. Daniel Koretz, PhD.

Harvard University Graduate School of Education
Chapter 2
LITERATURE REVIEW

My case study utilizes several secondary sources from the field of Education, many of which oppose the increasing use of standardized tests. Education being a vocation which subsumes a high degree of child advocacy, there tends to be a strongly liberal mindset inherent in the literature, particularly in the discussion of assessment reform. Most of these sources align with the data from my case study, providing a solid foundation of voices united against the continued proliferation of standardized testing measures. A much smaller number of critical voices will be heard in support of the function assessments play in the general accountability of school performance—smaller in number not because the argument is any less valid, however. The overwhelming professional sentiment is that standardized testing is inherently problematic, and therefore the most critical responses will speak to that end.

Gate-Keeping

Many respected voices within the debate regarding assessment reform raise very strong concerns about assessments in general. Brian Huot remarks that testing is merely a “means for gate-keeping and upholding standards” (61). Echoing the sentiments of Michel Foucault, Huot says that “the concept of the examination is closely related to acts of punishment and hegemony toward those in society who hold positions of vulnerability” (174). Terry Underwood says in the opening pages of his book that he has a “deeply rooted conviction that standardized multiple-choice tests of literacy are irremediably flawed as measuring instruments and have unwanted, detrimental
consequences in that they diminish students’ opportunities to learn and therefore their life chances” (Underwood 16). Although his remarks are directed towards assessments of literacy in particular, his broad critique of testing could reasonably be applied to all forms of educational testing in my opinion. It didn’t appear through the course of the teacher responses to my survey that many teachers questioned the reliability of the testing measures. However one teacher did mention as an aside, “…that is assuming the data is valid.” A healthy dose of suspicion may eventually lead to an effective questioning and evaluation of our assessment protocol.

Huot summarizes a common view towards testing which many, including me, share, that testing is “something done because of a deficit in student training or teacher responsibility” (1). Unfortunately for the public school teachers who must operate under such burdensome legal mandates, testing is a reality with which we have to accept and comply. Overindulging in anti-testing sentiment is a luxury few public school teachers can afford.

The Narrowing Effect

The primary issue found in nearly all of my research was some suggestion of how standardized testing was having a “narrowing effect” upon the overall educational experience at public schools. “Teachers emphasized that instruction with the goal of getting students through these tests distorted the teaching and learning process” (Darling-Hammond, 84). These distortions occur within every fiber of the educational fabric: curriculum is aligned to testable items and state standards; instruction is altered to accommodate testing schedules and testing emphases; components of classroom
experiences which were commonplace during my generation like Social Studies, Science, PE, Art, Music, and Drama are being either scaled-back or discontinued entirely (FairTest.org, “NCLB” 1); school-wide routines like visits to the computer room or library, as well as assemblies or field trips, are far more difficult to schedule; explicit pressure to perform well on tests is exerted from the upper echelons of the state and district administrators on down to teachers, students and families; and variations in instruction or methods specifically tailored to suit the diversity found within a class—as well as across the classrooms in California—has been met with increasing resistance as the goal of standardization takes further prominence. These are the manifestations of the narrowing effect assessment and accountability measures are having of school systems.

In their article “The Impact of High-Stakes Assessments on Our Schools” Tanya M. Suarez and Nancy C. Gottovi remarked, “the focus of classroom instruction is narrowing, with the tests determining the curriculum” (83). References to the narrowing effect echo numerous times throughout these various articles on assessment reform (Darling-Hammond; Madaus; Au; Koretz; Padulla; Abrams; Morton; and Wei); it is the epicenter of the controversy and is referred to either directly or indirectly with great frequency and fervor. Measuring Up by Daniel Koretz, and The Right to Learn: A Blueprint for Creating Schools that Work by Linda Darling-Hammond both provided a comprehensive look at high-stakes testing. Addressing literacy and writing in particular were (Re)Articulating Writing Assessment by Brian Huot and The Portfolio Project: A Study of Assessment, Instruction, and Middle School Reform by Terry Underwood.
Finally, I used *Savage Inequalities* by Jonathan Kozol as a narrative model for describing the events at my school.

Darling-Hammond states that “60 percent of the teachers reported that the emphasis on testing had affected their own teaching” (87). That percentage is significantly lower than the responses I received from my own staff survey in which 100% of the respondents who explicitly mentioned how testing affected their own teaching. Nearly all the teachers’ responses were comparable to this statement: “The tests are used to determine what the students are understanding and what they are struggling with in each subject area.” Darling-Hammond mentions the other side of the controversy when she cites George Madaus as having said that “national standards and tests would necessarily narrow curriculum” (225).

Considering the remarkable diversity within our state, any process which moves instruction towards a more standardized “one-size-fits-all” approach seems counter-intuitive. In Helen Wei’s 2002 report to the American Educational Research Association, she says that:

> existing research indicates, mandated standardized testing policies constrain teachers’ use of time, often leading to teaching to the test and increased instructional hours spent on test preparation—leaving less time for culturally relevant instruction to occur. (11)

FairTest.org speaks very bluntly about “NCLB’s test-and-punish approach to school reform relying on limited, one-size-fits-all tools that reduce education to little more than test prep” (“NCLB” 1). The narrowing effect is very real and central within the literature of assessment reform. Brian Huot calls this byproduct of testing a “negative
force because so many current assessment practices do not even attempt to address teaching and learning, yet they nonetheless narrow or guide instruction” (150). As I’ve discovered over the course of my research—something that I have intuitively known for the duration of my teaching career—the one-size-fits-all approach certainly doesn’t work for Parker, and may not work anywhere. One teacher remarked that, “It is unfortunate, but there are so many factors affecting student performance that have nothing to do with their abilities (class size, home life, distractions, etc.) But we can’t meet all of their individual needs.” Clearly, individual needs cannot be met with a single solution. That kind of wisdom (or lack thereof)—in which the only tool in one’s toolbox is hammer—produces the kind of warped response which treats every problem as a nail. A student who comes to class without jacket can’t be fixed that way.

Wayne Au says that as a result of standardized testing “curricular content is narrowed to tested subjects, subject area knowledge is fragmented into test-related pieces, and teachers increase the use of teacher-centered pedagogies” (“High-Stakes” 258). Reducing curriculum into fragments of “test-related pieces” will be discussed in greater detail in Chapter 3 as it pertains to my school site. However, I will mention that Parker adopted a curricular component last year called “Test Prep” which was explicitly geared towards delivering test-related content and testwiseness skills. Parker Elementary’s response to the pressures of high-stakes testing was precisely how FairTest.org described it:

Teachers and administrators feel enormous pressure to ensure that test scores go up. Schools narrow and change the curriculum to match the test. Teachers teach
only what is covered on the test. Methods of teaching conform to the multiple-choice format of the test. (“Testing Damage” 1)

In a more measured statement Daniel Koretz says “we have studies showing an inappropriate reallocation of resources away from aspects of the curriculum that are important but aren't tested” (“No Child” 4). Or, as another teacher mentioned, “I used the test prep period to teach test-wise awareness and skill. The time period was very limited.” Resources—particularly instructional minutes—are short. The pressure from high-stakes testing reduces teaching in many instances to a staging area for testwiseness.

Not surprisingly, in the 19 states in which state-mandated testing is considered “high-stakes” (outcomes are connected to significant rewards or sanctions), the narrowing effect is occurring most acutely. In terms of instruction in Fine Arts, Physical Education, Foreign Languages and Industrial/Vocational Education, states like California are reporting “the largest decreases in instructional time” (Pedulla 52). The states are under the most pressure are responding accordingly by stripping down instruction to a bare core aligned to high-stakes testing. However, whether the narrowing effect is directly linked to assessment outcomes, on a more fundamental level, it is working against the better judgment of educators across the country.

Perhaps most disconcerting was the substantial proportion of teachers in both types of testing programs (76% of high-stakes teachers and 63% of low-stakes teachers) who reported that their state testing program has lead them to teach in ways that contradict their own notions of sound educational practice. (Abrams 23)
And as mentioned earlier, this pressure is passed onto the students, especially in light of the fact that many teachers are experiencing some degree of discord in regards to their own teaching practices:

educators’ anxiety about school performance is passed on to students who increasingly experience pressure and sanctions as teaching is increasingly focused on test-related drills, which are becoming narrower as the tests become less-authentic and less performance oriented. (Darling-Hammond, 243)

To what extent one resists or accepts the criticism of testing practices, there is an inherent problem in the construction of standardized testing items in which rote facts are decontextualized so as to make them suitable for multiple-choice responses. Darling-Hammond cites national and international reports which indicate that the “large majority of U.S. schools emphasize rote learning with heavy doses of lecture, drill-and-practice, memorization, and multiple-choice and short-answer testing” (9). She goes on to further bolster this claim with the support of the 1981 National Assessment of Educational Progress (NAEP), the National Research Council, and the National Councils of Teachers of English and Mathematics that “low student performance on problem-solving and critical-thinking tasks [may be attributed] to schools’ overemphasis on multiple-choice tests of basic skills” (58). The reason for this testing of basic skills is because it is difficult to assess students on higher-order functions. Generating multiple-choice questions and answers requires the use of lower-level cognitive items “at the expense of reasoned and disciplined thought [thus] it is not surprising that students fail to develop more comprehensive thinking and analytic skills” (Darling-Hammond 56). Underwood says that “the multiple-choice format may well serve…if the aim is to test recall of what
Sheridan Blau refers to as ‘brute signifiers’” (42). Oftentimes I think that my job has been reduced to an assembly-line process of transmitting decontextualized facts.

Reliability

How and why did the educational movement trend so stridently towards these standardized assessment practices? It is not unreasonable to be drawn to the allure of standardized measurements, considering they so strongly suggest a uniform fairness.

Arguing against Grant Wiggins’ call for the use of “authentic assessments,” James Terwilliger in his article “Semantics, Psychometrics, and Assessment Reform: A Close Look at ‘Authentic’ Assessments,” makes an ethical claim that cannot be easily dismissed. He says that, “validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force” (26). Edys S. Quellmalz echoes this same sentiment about fairness when she seeks “to insure that criteria are sufficiently precise to permit standard, replicable application” (66). However, as Daniel Koretz point out, this attraction to the perceived glow of testing uniformity may be nothing more than “a patina of scientific credibility” (*Measuring* 8). The appearance of fairness in the form of standardization may in fact be completely undermined by the numerous problems that occur as a result.

The word “reliability” has a specific function in the world of assessments and psychometrics. It is a synonym for consistency when applied to testing, and points directly to the targeted goal in standardized tests. I, however, use the term in a more colloquial manner, instead likening it to way an individual might “rely” on a friend. It
stands to reason that we stakeholders in education rely on the data we are given, trusting it to be an accurate measurement. So as I use the word, consider that I am suggesting a broader meaning than the assessment term; I am looking to assessments—as we all do—as something I should be able to trust.

Unfortunately, trustworthiness is not always an accurate claim when testing and scoring mechanisms are not functioning properly. Kathleen Rhoades and George Madaus, in their comprehensive article “Errors in Standardized Tests: A Systemic Problem,” document numerous scoring errors within the mechanics of standardized tests, affecting “well over 1.5 million students and 4,000 school” (2). They list so many examples of standardized testing errors that I thought of including only a few of the more recent ones from my own state:

- In Fresno county during the 2001 academic year, “[curriculum publisher] Harcourt admitted to using the wrong set of norms in calculating some California student scores and school averages—an error that inflated scores in 22 schools. In six of the schools, $750,000 had been erroneously paid out in bonus money to both staff and school. In the other 16, staff were informed that they were no longer eligible for bonuses or that their bonuses would be reduced. (52)
- Similarly, inaccurate scoring resulted in erroneously awarding bonus money in San Diego County as a result of the STAR test in 2000. (51)
- Harcourt Brace erroneously classified newly English-proficient students as being “Limited English Proficient” (LEP), which inflated the aggregate scores for LEP students. Early press reports cited the large gains as evidence that California’s Proposition 227 (which decreased the numbers of LEP students in bilingual education) worked. (38)

Rhoades and Madaus said that despite public perception, media representation and political momentum pushing our public school schools towards a system of standardized testing protocols, their study provided 25 years worth of “testimony to counter the
implausible demands of educational policy makers for a single, error-free, accurate, and valid test used with large groups of children” (28). It is not a fatal blow to the use of contemporary assessments method to cut through the illusion of reliability. In fact, if assessment measures are to something within which we can truly rely and trust, we are duty bound to examine and critique the extent to which a measurement is trustworthy. That is the nature of testing validity.

Validity

Perhaps the most important secondary sources for my project are the contributions from two of the preeminent scholars who have written on the subject of validity, Samuel Messick and Lee Cronbach; Messick’s definition of validity will form the basis of my evaluation of our assessment procedures at Parker Elementary, as seen in my fourth chapter. In his article “Meaning and Values in Test Validation: The Science and Ethics of Assessment,” Messick writes that “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (5). A dense and foundational quotation, I want to unpack two key concepts: 1) an assessment protocol should have some identifiable rationale as its basis. That rationale should be ground in theoretical assumptions, as well as empirical evidence acquired from classroom experiences; 2) the results of the testing should inform some reasonable inferences about student and teacher performance, along with actions which would manifest itself in future student learning (Those “actions” resulting from testing data should help drive future instruction, a concept I will discuss in more detail in
Chapter 3’s “Response to Intervention.”). Our principal succinctly captured the
magnitude of validity—particularly the issue of inferences and actions based upon testing
data—when he responded by saying, “This is a huge question.” In order for a
measurement to attain the reasonable standard of validity, it must, however, satisfy both
conditions. Much of my staff surveying involved questions seeking to discover to what
extent these two conditions are in existence at Parker.

Cronbach’s contribution to validity seems much more humanistic than Messick’s,
although it is perhaps for that reason that it is more fluid and therefore more problematic
to apply. Besides calling Messick and Cronbach “the two most influential validity
theorists of the last few decades,” (157) Brian Huot says “for Cronbach validity ‘must
link concepts, evidence, social and personal consequences, and values’” (53). Looking
for discrete measurements of “social and personal consequences, and values,” would be
extremely difficult for me. While I certainly appreciate the ontology of Cronbach’s view,
I would not know how to evaluate and measure social and personal consequences and
values. Although Cronbach’s definition is more in keeping with the human qualities
inherent in the classroom experience, I found Messick’s definition easier to explicate and
apply to Parker’s testing practices. However, of particular use is Cronbach’s compact
proclamation that “validation is never finished” (5). Cronbach’s understanding of the
fluidity of a validation process perfectly suits the constant flux of any school system,
Parker Elementary being no exception.

Certainly, close contact with teachers and students regarding the nature of testing
procedures and outcomes—along with their residual impacts on curriculum and
instruction—would be the foundation of how validity is determined. Huot asserts that “if decisions based on an assessment must promote teaching and learning, as current validity theory dictates, then we must be accountable to those people who are most expert about teaching and learning—students and teachers” (Huot 179). As I will discuss in the recommendations in Chapter 4, one considerable shortcoming of our current assessment system is the lack of input on a local level, particularly involving teachers and students. Darling-Hammond defines the validity of an assessment practice by asking “whether stakes that undermine these beneficial influences [in student learning] are warranted and defensible” (243). All endeavors come at a cost. The question becomes whether the costs are justified in light of the benefits to student learning.

One approach to examining school practices through the lens of validity is what Underwood refers to as “Backwards Planning…looking at desired learning outcomes and then…com[ing] up with actions with a demonstrable potential to support student growth toward those outcomes” (7). Harvard professor Richard Elmore uses a similar term—“Backwards Mapping”—in order to determine the validity of a particular school practice (a “unit”). He asks, “What is the ability of this unit to affect the behavior that is the target of the policy? And what resources does this unit require in order to have that effect?” (“Mapping” 604). In any event, the validation process is one in which predetermined criteria is used as a reference by which certain procedures are reflected upon and evaluated. Perhaps the most humanistic of all the various explanations of validity is Lucas’s 1988 article, “Toward Ecological Evaluation,” which seeks “to ensure that assessment systems do not damage what they seek to measure” (qtd. in Underwood
This reminds me of a witticism told to me from a long-time educator and mentor of mine, Arthur Clanin, who used to say to his students with a smile: “I am only hurting you for your own good.” It is hard to imagine that I deliberately comply with a testing regimen which causes stress and hardship for my students. Nevertheless, I could only hope for a testing program which measured up to Lucas’s standard of validity.

“This is the question of validity, which is the single most important criterion for evaluating achievement testing... The question to ask is how well supported the conclusion is” (Koretz, Measuring 31). How well supported are our testing practices and the outcomes they produce in student achievement? This might be the most important question we can ask of ourselves at Parker Elementary. What empirical evidence is used either in support for or against our assessment methods? What theoretical rationales form the basis of our testing practices? What types of inferences are formed from the testing data? And what actions generated as a result of our tests? These will be the questions that I will pursue in chapters 3 and 4, seeking to find whether or not they are being considered at my school. Not only will this discussion of validity provide a reference point from which I can address more subjective terms like “balance”, “commoditization”, and “the narrowing-effect,” it will also provide basis for an entire line of interview questions.

Score Inflation

Perhaps the most corroding element undermining the reliability of standardized testing is the very real possibility that the data produced from these methods of testing are not the accurate measurements we assume them to be. “The obvious question...is
whether scores on the high-stakes test have become an inflated and therefore misleading indicator of what students actually have learned” (Koretz, Measuring 11). In my experience, there are two types of movement towards addressing testable items which result in score inflation. The first type has been a slow building inertia aligning curriculum and teaching practice with the state standards; this has gained strength over the years. There is also the short-term surge which occurs right before an exam, when a teacher seeks to address the particular items found on an upcoming test; the most dramatic example of this “surge” occurs seasonally before the STAR testing in spring. However, both types of teaching-to-the-test—coupled with the explicit Test Prep curriculum—indicate a strong likelihood that pressure is being applied to student outcomes which may be distorting the actual domain which we are wanting to observe and measure. “Scores on the tests used for accountability have become inflated, badly overstating real gains in student performance. Some of the reported gains are entirely illusory, and others are real but grossly exaggerated.” (Koretz, Measuring 235). For the first time, I am distinguishing in my mind the difference between student proficiency and proficient scores. I have always considered them to be one and the same. However, practices like Test Prep are producing proficient scores without a concrete link to deep student understanding or real mastery (proficiency).

It seems perfectly understandable why this type of score inflation occurs. This naturally occurring phenomenon is explained by noted social scientist Donald T. Campbell. In his article “Assessing the Impact of Planned Social Change,” Campbell
says that “the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor” (qtd. in Koretz, *Measuring* 58). Clearly, the “social indicator” being used for “social decision-making” is the data generated from standardized test scores. Therefore, the natural outcome is for those indicators to naturally corrode under the weathering pressure of high-stakes rewards and/or sanctions.

The responses to these high-stakes pressures are interventions and curriculum devices specifically designed to achieve the highest possible student responses on tests, which may not align with higher-order understand or long-term retention. “Test-prep” sounds legitimate enough, as it connotes the “rigor” so desperately sought after by district administrators. However, what have students learned from curriculum which is so narrowly focused and facts decontextualized? A teacher interview by Jonathan Kozol answered this question by saying that children “have learned that education is a brittle, abstract ritual to ready them for an examination’” (144). Another teacher interviewed said, “preparation for the state exams controls curriculum….Literature gets lost. The driving notion here is that skills learned in isolation are more useful than skills learned in context” (Kozol 161). And without a doubt, I too am constantly driving that realization into my students’ consciousness on a daily basis.

Coming back to the illusion of reliability in standardized testing, Darling-Hammond reminds us that test-based curricula or test-prep materials have the “trappings of intellectual rigor” but “miss the point” in terms of teaching higher-order skills (60).
Clune and White call these kinds of test-direct materials “interim assessments,” saying that they could “be useful for shaping instruction, help[ing] students prepare for the state assessments, and provid[ing] useful information for school improvement plans” (8). While true in moderation and with a close eye on the validity criteria, interim assessments could provide teachers with guidance for future instruction. We refer to these interim exams as “formative tests.” One teacher at Parker said, “Formatives can help guide forward teaching or skills not mastered by groups of students or particular students.” Used sparingly in a low-stakes manner, the formative tests can guide future instruction. However, it does continue to consume instructional time, which must be an overarching concern. Left unexamined, these measures can produce overinflated assessment results and distorted educational goals as test proficiency supplants higher-order thinking and understanding.

In a topic which I will discuss in more detail as to how it affects my school, the concern is that these testing pressures are eroding away the fabric and texture of the educational experience. Abrams and Madhaus refer to the “substantial allocation of instructional time for specific test preparation” as “a source of concern” (27). At Parker Elementary, where explicit test preparation has consumed as much as 30 minutes a day, that amount of time would typically be considered a substantial allocation of instructional time and resources. To be completed 4 or 5 times a week, test prep accounts for up to 150 minutes per week. With instructional minutes fixed at 1,585 minutes each week that means that the Test Prep period alone consumes nearly 10% of the instructional time allocated over an entire week (9.5%).
Instructional time is everything. Every minute accounted for, and must be treated with the same fidelity and stewardship with which the students and curriculum are treated. Connie A. Bridge, Margaret Compton-Hall, and Susan Chambers Cantrell, in their article, “Classroom Writing Practices Revisited: The Effects of Statewide Reform on Writing,” draw the connection between instructional hours and improved academic indicators. Addressing the 1995 Kentucky Education Reform Act, they showed that test scores and writing proficiency improved when instructional time increased (158). The more time that is devoted to any targeted domain of knowledge the more learning results. Therefore, the removal of instructional minutes equals a direct and immediate lessening of learning opportunities.

Darling-Hammond echoed Campbell’s law regarding the corruption of measures, applying it to school settings by stating that “evidence from many studies demonstrates that when high-stakes decisions are attached to scores, test can be expected to exert a strong influence on ‘what is taught, how it is taught, what pupils study, how they study, and what they learn’” (58). It would be unreasonable to expect for a school that is within the accountability system could react any other way. Furthermore, it would be impossible to measure to what extent teachers like me place added emphasis upon certain curricular elements, pushing other aspects into the background, all in the name of testing proficiency.

Besides the creation of a new student aptitude—testwiseness—a simple and regrettable reduction in student learning occurs. Summed up in the acronym WYTIWYG, Underwood cleverly states that the result of high states testing = “what you
test is what you get” (36). Studies like Pedulla’s in 2003 have shown that while an increase occurs in the teaching of tested subjects, a reciprocal decrease occurs in the teaching of non-tested subjects. Wayne Au cites Renter’s 2006 article when he says that “one nationwide survey found that 71% of the districts reported cutting at least one subject to increase time spent on reading and math as a direct response to the high-stakes testing mandated under NCLB” (“Social Justice” 45,6). However, when it is all said and done, “inferences about improvement are undermined by certain types of test preparation” (Koretz, Measuring 34). That’s the real problem: what we perceive as improvement may just be a distortion resulting from added test emphasis. “School improvement” or “Adequate Yearly Progress” become phrases of little substance when they are stressed and fatigued under the pressures of test preparation.

Koretz drew a parallel between the affect of public pressure being exerted upon schools as a result of high-stakes testing (particularly NCLB), and the medical profession after mortality statistics were publicized: “79 percent [of doctors surveyed] admitted that ‘the knowledge that mortality statistics would be made public’ had affected their own decisions about whether to perform surgery” (Measuring 239). Not only is this a reasonable response, when kept within a moderate balance of accountability, public expectation, and support, this type of pressure is well-intended and healthy. Doctors and teachers alike—being stewards of our citizenry’s well-being—should be held to level of public scrutiny. As Schoenfeld says, the public “needs to know” how and upon what basis our students and teachers are being evaluated (200). Validity measures would not
exist without an imperative to be held accountable. New York State’s Commissioner of Education in the 1990’s, Thomas Sobol, said that the assessment movement promotes clarity of purpose, quality of work, equity of expectations, and consistency of approach across schools and school districts. The bad news is that for many other students in other places, the [assessment] movement is narrowing the curriculum, imposing a stifling uniformity. (469)

This study shows time and again, that this is about a matter of degree. Clearly, well-intentioned public pressure has considerably overshot the goal of moderation and balance with regards to assessment measures if, as Clune says, we are “beating the dog harder” (qtd. in Darling-Hammond 239). Darling-Hammond says that is an inadequate strategy for meaningful reform. FairTest.org summarized that “after six years, there is overwhelming evidence that the deeply flawed ‘No Child Left Behind’ law is doing more harm than good in our nation’s public schools” (“NCLB” 1). It has taken some time, but many like me are seeing NCLB as an inherently flawed operation. One teacher at Parker said that, “because of NCLB our teaching is such a narrow thing...the only thing schools care about are the tests. It’s kind of sad.”

Theodore Sizer said that NCLB has “served the purpose of putting in front of all citizens the stunning weaknesses of our schooling systems, especially as it serves—or fails to serve—its most vulnerable young citizens” (61). That is an admirable goal, consistent with the public school system’s inherent advocacy for children. The next step however, is to inform the public of the reality of high-stakes testing and the costs incurred at the expense of classroom learning. Marge Scherer said that “as assessment-literate professionals, we are in a much better position to educate those in the general public,
media, and policymaking positions who blindly accept the validity of any test for any purpose” (6). It may not occur to the general public that standardized tests—while providing reliability in a psychometric sense—may not be providing our families with the educational experience to which they rely upon us who work in the public school system. In order to educate the public about how assessments are functioning within the school mechanism, Schoenfeld says it is “essential to publicly specify what is being tested” (199). A reasonable first step would be to inform stakeholders who are not directly involved in testing what the costs and benefits are to current assessment practices.

Norbert Elliot and Paul Zelhart say that the two reasons for a writing assessment are “to ensure the institution's public accountability and to provide information to learners on their progress” (4). Clearly, those two goals exist not only in writing assessments, but within all assessments as well.

The benefits of contemporary testing practices however, can be severely offset by the cumulative costs of narrowing instruction. According to Lisa Guisbond and Monty Neill, the cost-benefit analysis of state-mandated testing procedures fail to satisfy the validity criteria for judging such a system. Quoting Harvard’s Richard Elmore, Guisbond says that the system for measuring a school’s AYP is a “completely arbitrary mathematical function grounded in no defensible knowledge or theory of school improvement” (13). It fails to be measured by a comprehensible system of grading; and it fails to be grounded in a defensible theory of education. By Messick’s definition, state-mandated testing would fail the standard of validity. Frederic Mosher, Susan H. Fuhrman, and David K. Cohen go on to support this point in their article “The Research
That Policy Needs” that large-scale assessment measurements are not adhering to basic thresholds of validity:

However, none of the measures used in state assessments (or even in NAEP) has any direct empirical validation for such an interpretation of current proficiency levels. At best, the measures are based on judges—curriculum specialists, experienced teachers, and sometimes parents or employers—eye-balling items in assessments and choosing those that fit their expectations of what a proficient student should be able to do (and a less proficient student could not). Those judgments are never tested against other observations of the student or against more complex assessments of a student’s effectiveness in real-life settings. (21)

The time for viewing assessment protocols through the validity criteria is upon us. Significant documentation indicates that a well-intentioned system of accountability has run afoul of its intended goal, and is currently undermining the very learning process which it intended to initially measure. For perhaps no greater reason than the fallaciousness of the one-size-fits-all approach to education, standardized practices in assessment and instruction either nullify the accountability gains of testing, or worse yet, wreak further damage upon a beleaguered educational system. “Consider the NCLB requirement that nearly all students, regardless of their level of performance, be assessed using the same tests. Whatever the political virtues of this requirement, it is bad measurement” (Koretz, Measuring 328).

Loose-Coupling

This concluding section of this literature review centers on the historical insights and forward-thinking contained in Richard Elmore’s 2000 article “Building a New Structure for School Leadership,” an article that I as an educator found to be lucidly
The article opens with an ominous warning which accurately captures the desperation felt by many of us in education:

Further, if schools, school systems, and their leaders respond to standards-based reforms the way they have responded to other attempts at broad scale reform of public education over the past century, they will fail massively and visibly, with an attendant loss of public confidence and serious consequences for public education. (“Leadership” 1)

Locating ourselves in the present dilemma facing public schools, there is widespread agreement that what we are doing now is not working; further evidence suggests that what we are doing is neither sustainable, nor anything short of detrimental for the future. Before looking too much ahead at where this failing vector leads, it is important to understand the historical mechanisms at work in this educational discussion.

According to Elmore, the basic structure of the public education system is built around a system he calls “loose-coupling” (“Leadership” 6). This loose-coupling includes many individual structures of learning (classrooms) existing in a patchwork with other similar classrooms. Coupled together, these classes are in essence isolated from one another, however. What occurs within these structures is generally kept within the walls of the classroom, except in rare incidences when like-minded teachers find the opportunity for collaboration. The teaching practices, too, are kept in a state of relative obscurity, tending towards the understanding that teaching is an art-form, and is therefore idiosyncratic (if not borderline mystical): good teachers teach well; not much more is expected as far as an explanation, beyond that. Without question, this disposition towards teaching as an idiosyncratic art-form is widely accepted today by me and others. These isolated centers of learning—existing almost as
a network of cells—developed further as administration functioned primarily to help keep these centers intact. Elmore continues this explanation by saying that since this esoteric teaching practice…

cannot be clearly translated into reproducible behaviors, it requires a high degree of individual judgment, and it is not susceptible to reliable external evaluation. Therefore, the loose-coupling argument continues, the administrative superstructure of the organization—principals, board members, and administrators—exists to “buffer” the weak technical core of teaching from outside inspection, interference, or disruption. (“Leadership” 6)

Anecdotal illustrations abound of principals beloved by teachers who made it their primary mission of clearing a path for the teachers to teach without disruptions: they deal with non-productive student behaviors, parental concerns, neighborhood issues, matters pertaining to facilities and maintenance; when events are staged they are prominently situated to provide leadership and support; they humbly defer to the expertise that is teaching, leaving that job to the teachers; when it comes to supplies, resources or any manner of advocacy on behalf of the teachers, they promptly take the lead role. They form the ultimate diplomatic enclosure around the classroom community, so that nothing can spoil the splendor and ecology within the class. This model of principal embodies the role of “support,” not “expertise” in terms of teaching. Principalship of this nature supports teaching like scaffolding; it is not inherent in any way within the structure of teaching itself. The kinds of schools constructed upon this management model of loose-coupling possess administrators who are detached from instructional or curricular expertise. These types of administrators—who Elmore says are the historical and current
norm—“were inclined to defer all judgments about what students could and should learn, and all decisions about to whom the school is collectively accountable for what, to individual teachers operating in isolation from each other” (“Leadership” 17). These management models are ill-suited for the future, as will be discussed further in Chapter 4, “Restructuring Leadership.”

Accepting the premise of loose-coupling as the prevalent condition currently, questions arise: Should this system continue? Could it continue in the newly evolving system of accountability? I agree with Elmore on the matter:

The only thing that could be worse than opening up the instructional core of public schooling to external scrutiny and debate might be not doing so, and watching the public purposes of public education drift away into matters of individual taste and preference. (“Leadership” 11)

With the accountability measures in motion—bolstered by public opinion, political gain, budgetary pressures, and a fair amount of common sense—a future in which these testing practices disappear is unforeseeable. The modification of these practices at my school is quite possible and is clearly my aspiration. Nonetheless, a future in which educational professionals of all levels fail to evolve in accordance with this accountability movement would be disastrous. For one thing, the state of public education is in incredible flux. University of California professor Dr. Jeannie Oakes projects that the state will need between 250,000 and 300,000 new teachers in the next decade (58). The structure of public education in California is going to change; whether that change occurs thoughtfully is another matter entirely. My mentor Arthur Clanin used to say “Change is inevitable. Growth is optional.” If we see the future that Elmore envisions, not changing
will bring about a public educational system most closely resembling the crumbling health-care system “that specializes in clients no one else wants to serve (“Leadership” 11).
“There are no science labs, no art or music teachers. There is no playground. There are no swings. There is no jungle gym. Soap, paper towels and toilet paper are in short supply. There are two working bathrooms for some 700 children. These Children ‘cry out for something more…They do not get it’” (63).

_Savage Inequalities: Children in America’s Schools_

by Jonathan Kozol
Chapter 3

PARKER ELEMENTARY: THE CASE STUDY

When the sink in the boys’ bathroom fell off the wall last month, it revealed the fact that multiple masonry screws had been driven into the same concrete hole. As I looked at the workmanship, I thought to myself that the idea of cramming multiple screws into the same hole couldn’t possibly have seemed like a long-term solution at the time. Although a poorly conceived fix, it did hold for a few years since the contractors who “modernized” our campus in 2001 affixed that same sink to the wall with nails (it only took a few weeks of young boys leaning on the sink to pull it off the wall back then).

Are there connections to be made to Jonathan Kozol’s accounts of disrepair at schools in Chicago, New Jersey, and East St. Louis? Are there larger issues at play in this sink anecdote than just concerns over workmanship? Is there a parallel between driving multiple screws into the same hole—a hole intended for one screw—and the machinations of standardized testing? Sometimes I think I am alone in seeing a relationship between all these issues. However, as I listened to the staff members around my campus, and researched the secondary sources on the matter, I realized there are some people who, like me, view these events as manifestations of a common problem.

My school has many of the same characteristics Kozol describes in his narrative. The upper-grade playground has no play-structure or wall for wall-ball. The music program—which had been fending off cuts in resources for many years—was finally abolished. Library and computer lab resources have been left more for the individual teacher to negotiate, as regular staffing has been pushed down the list of priorities.
Similar to the industrial residues affecting the East St. Louis schools in Kozol’s book, our campus is nestled within vacant lots owned by a regional utility company. Occasionally, students will have to stay inside the classrooms as gaseous fumes drift across our playgrounds from the adjacent lot. And although I haven’t heard other staff members mention this, I am struck by my observation that of the three lots owned by this company, two have the typical plumes of waist-high grasses and weeds. Yet, the one lot from which these fumes seem to emanate is barren and sterile; nothing seems to grow there no matter what time of year it is.

The neighborhood within which my school is situated is home to the highest rate of reporting to Child Protective Services in the county. It also has the highest rate of adult and juvenile probationers in the county. There are over 300 convicted and registered sex offenders within one mile of our campus (which would probably not include the occupants of the homeless tents camped along the outside perimeter of the aforementioned vacant lots.). Virtually all students at Parker elementary qualify for free or reduced breakfast and lunch. Homelessness affects approximately 5½% of the students at Parker. Narratives like Kozol’s aren’t really that unusual in light of the reality at Parker. When he says that “urban schools were, by and large, extraordinarily unhappy places,” I get that (5). Many of the staff gets that too (although we wouldn’t or couldn’t stay there for as long as we have if we weren’t able to turn around some of the grimness into moments of positivity and learning).

Education should be a source of nurturance for the spirit as well as a means of reaching understanding, though it can be, and too often is, conducted in a way that deadens and demoralizes. Tedium, coercive
schooling creates frustrations that must emerge sooner or later in self-deprecation, despair, or violence against others. (Darling-Hammond, 31)

The differences are enormous between the schooling I received as a child versus the schooling my students experience. When I was a school kid I remember a boy named Wallace (a pseudonym) who was unkempt and a child of divorce (a heavy stigma in the early 70’s); he had an older brother who was hit and killed by a car. It was one of the darkest memories of my school days. During my eleven years at Parker, family members have shot and killed other family members. Naturally occurring illness and death have left children without any parents to care for them; incarceration has also left students of mine to be cared for by more-distant relatives or neighbors. Stories of children being physically and sexually abused and/or abusive, kids being thrown out of cars and left on the side of the highway, fifth graders with probation officers or gang affiliations, these are not the stories of fiction or even far-away places. These are our stories, stories from my classroom. This is the jumping-off-place for my case study, because any discussion of the procedures and practices at Parker elementary can only reasonably be considered within the context of the neighborhood we serve.

Data-Based Student Performance

The point of testing data is to produce a plan for improving student learning. In this section I will show how Parker used the 2008-09 ELA test scores to better understand trends in student testing, along with our inferences and actions called “Response to Intervention (RtI).” One support staff member remarked, “The use of data is very important. [Another neighboring school] has made impressive gains as a result of
using data and putting the appropriate interventions in place.” A school’s RtI is its ability to modify instruction to fit individual student needs in an informed and proactive manner. An assessment program which provides a reliable portrait of student and teacher performance is necessary in determining how to make such instructional modifications. As an upper-grade teacher said, “We decide where to focus our interventions and what the needs our students have and try to address them the best we can, that is assuming the data is valid.” Another veteran teacher asked candidly, “Also I probably should know this, but what is RtI??” suggesting, perhaps the lack of full implementation of RtI at Parker. In this age of diverse classroom populations, it is imperative for teachers to provide instruction that meets the individual needs of the students as best as possible. If for no other reason, diverse school populations face a greater obstacle in achieving testing goals merely as a consequence of reporting their various subgroups. “The more groups a school has that must be reported separately, the more likely it is that the school will fail solely because of sampling error” (Koretz, Measuring 169). This is another reason why Parker is more vulnerable to high-stakes testing. It also continues to undermine the one-size-fits-all approach to curriculum and measurement. Furthermore, it points to the fact that diverse school populations like Parker’s have an increased need for RtI.

During the 2008-09 academic year, a large percentage of Parker students had academic shortcomings (learning disabilities or academic abilities far below grade level), English language deficiencies, or learning styles which required some modification to the delivery of instruction. Parker had 443 students of which 306 were Latinos (69%) and 59 Asian (13%), for a combined total of 82%. Of the entire student population, 54% were
designated as “English Language Learners” (EL). Add to that total, the fact that approximately 5% of the student population had a documented learning disability. Furthermore, the number of students who were not “proficient” hovered around 60% of the total number of students. For a school as diverse as Parker, the need for variations in interventions and instruction couldn’t be overstated. In addition to the academic challenges, having multiple at-risk subgroups presents a greater challenge in terms of satisfying AYP goals. According to the 2007 Forum on Educational Accountability:

> Because the needs of special student populations and diverse learners were not considered adequately when the law was drafted, schools that serve significant numbers of these students are in the difficult situation of having to demonstrate and receive credit for performance levels that meet the law’s proficiency requirements in the absence of assessment systems appropriate for these learners. Consequently, schools with large numbers of students in several of these subgroups have a much greater chance of failing AYP requirements. (“Assessment” 22)

Therefore, if RtI was intended to design effective interventions in a timely manner in response to students’ various needs, our RtI would be critically important, considering our acute academic needs and the legal obstacles in order to satisfy NCLB requirements.

Within the RtI vernacular is a coding system by which the student population is subdivided into three categories: Tier 1, Tier 2, and Tier 3. Tier 1 (T1) students are those who score “Advanced” or “Proficient” on the STAR test. These students require the least amount of additional intervention, relying almost entirely on the good teaching practices of the core curriculum. These good teaching practices “…might help account for their test scores…a high level of student engagement, clear explanations from teachers before students undertook tasks, a level of enthusiastic activity when it was
appropriate, and spirited discussion among the students” (Koretz, *Measuring 6*). T1 students are referred to as being “proficient,” meaning that they demonstrate a high enough level of mastery that they are located in a category of overall academic success.

Tier 2 (T2) students are those who score “Basic” on the STAR test. In some instances, schools will include those students who fall into the “Below Basic” category as being T2 as well. For the sake of simplicity, I am referring to T2 as only including the Basic student scores. T2 students represent an extremely important group as they are often targeted for improvement into T1 status; they are commonly referred to as “bucket” kids, further conveying this notion that they are to be transported to a higher achievement level. They require more sophisticated differentiation, including but not limited to levelized, homogenous, small-group instruction within the core curriculum. They also receive the lion share of the before or after school interventions as they pose the greatest potential gain in AYP.

Tier 3 (T3) students are those that score the lowest on the STAR test, falling into the “Below Basic” or “Far Below Basic” categories. The interventions necessary for these T3 students are the most intensive in terms of labor and resources, and may include instruction outside of the core curriculum, including possible modified classroom placement and/or Individual Educational Plans (IEPs).

Looking at STAR testing data is certainly useful. However, given that the STAR test occurs once over the course of two weeks in the springtime, a more common assessment practice to examine would be the district assessments which occur much more frequently. The first remarks that needs to be made regarding the data I collected from
the district’s 2008-2009 English and Language Arts assessments (ELA) is that these graphic representations are in the form of histograms, which show the frequency of an event. The core issue for my thesis—as I look at the causes and conditions of the narrowing educational experience at Parker elementary—is the volume of testing. By better understanding how much testing occurred, I can better illustrate why these other negative by-products manifested themselves. Therefore, my histograms represent the frequency with which certain test scores fell within one of five different categories specified by the STAR test: Advanced (100-90%), Proficient (89-80%), Basic (79-70%), Below Basic (69-45%), and Far Below Basic (44-0%). These five categories have been adopted into the vernacular at Parker Elementary as they are the categories used on the STAR test. (see Figure 2).

![STAR Test Categories](image)

**FIGURE 2. STAR TEST CATEGORIES**

Remember, however, that the test scores I am presenting are not STAR test scores. I examined the district-generated ELA assessments which adopted the five STAR test categories in order to better align our testing rigor and reporting in accordance with the STAR test.
Considering the “ELA: Girls” histogram below (Figure 3), the T1 students are those black and blue bars (Advanced and Proficient), the T2 are the green (Basic), and the T3 are the orange and red (Below Basic and Far Below Basic):

![Histogram showing ELA assessments for girls with T1, T2, and T3 categories labeled.]

Comparing the English-Only 5th grade girls (EO) to those who were learning English as a second language (EL), the first item to note is that there were significantly more test scores in the EL category (1,282 EL tests, to 491 EO tests). That demonstrates the fact that nearly three out of every four female 5th graders were English language learners (72%) at Parker Elementary. However, comparing the percentages of those girls who fell into T1, both groups represented themselves fairly equally: there were 45% of the EO girls in T1, and 44% of the EL girls in T1; in other words, more than 40% of the fifth grade girls scored advanced or proficient, whether they were EOs or ELs. Similarly, there were 31% of EO girls in T3, with 35% of the EL girls falling into that lowest T3 category. Clearly, the data indicates that over the course of the entire academic year, in
English and Language Arts, EO and EL girls performed in a virtually indistinguishable manner. The same, however, cannot be said of the boys.

![Bar Chart](https://example.com/bar_chart.png)

**FIGURE 4. ELA ASSESSMENTS: BOYS**

Considering the number of EOs (337) versus ELs (1207) assessments for the boys, it is again clear that the vast majority of the male students were English language learners, nearly four out of five students (78%). The numbers of boys versus girls represented in the total number of test scores however, were fairly balanced, with girls being the slight majority (54%, compared to 47% for the boys). More interesting however, are the number of boys who fell into the T1 and T3 categories. EO boys had 40% of their test scores reach the T1 level, while only 23% of the EL boys reached that mark. Conversely, the EO boys had 38% representation in T3, while the EL boys had a staggering 59% of their test scores fall into the lowest T3 designation.

Some observations can be made after comparing how these test scores fit into the RtI tiers. The first is that the girls generally performed better on the ELA assessments as
both the EO and EL girl-groups found themselves more heavily represented in the T1 category. The EO boys had a comparable number of students in the T1 category as the girls; however, the EL boys were far below the pace of placing themselves in that top category. Similarly, their over-representation in the lowest T3 category was nearly twice that of the girls and the EO boys. Clearly, any discussion of an at-risk sub-group would have to begin with the EL boys. Their overall numbers were large, and their significant underperformance was a cause for great concern. This observation—based upon the testing data generated from our district assessments—presents a compelling argument for the need for performance indicators. Without a data stream to demonstrate student performance, delivery of instruction is left to be relatively uniformed and increasingly idiosyncratic.

A further look into the disparity between the boys and girls can be seen by examining different sub-skills within the ELA assessments. Having already established that the EL boys are the most vulnerable group to fall into the T3 category, the next comparison is between all groups of boys versus all groups of girls.
The “Vocabulary: District Test” (Figure 5) was a computer-scored, multiple-choice assessment. The T1 group for girls was an impressive 58% of the total number of test scores recorded. The T1 group for the boys was much smaller, including only 35% of the total number of recorded test scores. Continuing this pattern, the girls had fewer T3 scores (28%) than the boys (50%).

The “Spelling: Prompt” (Figure 6) score was one of three scores derived from the students writing prompt. It was therefore, manually scored according to a district mandated rubric. Again, the girls posted a larger percentage of their scores in the T1 range than the boys, 30% compared to 14%. Not surprisingly, the boys over-represented the T3 category 58%, as opposed to 34% for the girls. The point I am trying to make, however is not to question necessarily the differences between the performances of different sub-group; rather, it is important that these differences can be observed as a result of testing data, and from there a plan of action can be generated.
FIGURE 6. SPELLING : PROMPT

While both graphs showed a similar pattern of girls out performing the boys, the two graphs differed insofar as the “Vocabulary: District Test” indicated a discriminating affect, while the “Spelling: Prompt” scored show a much more traditional bell curve. Remembering that the Vocabulary test was multiple-choice, the movement of large groups into either the T1 or T3 categories appears to indicate that the test items effectively discriminated between students who did and those who did not possess a certain amount of proficiency within the domain of Vocabulary. Conversely, the Spelling: Prompt generated scores from a manually-scored rubric. Those scores may be indicating one of two possible trends: 1) the writing test did not produce this discriminating effect, as it involved a demonstration of applied, higher-order skills; or 2) the use of the rubric had some gravitating effect, drawing the scores into a more centralized range. Either or both of these possibilities are worth further exploration as they produced an entirely different testing indicator. To lump both of these sub-skills into the same cumulative ELA grade without appreciating how fundamentally different
these assessments measure performance is to miss an inherently unique assessment result, an aspect to consider when defining our school’s testing validity. Again, it is not my intention to draw conclusions based upon the testing data, but rather to point out that conclusions are possible as a result of having testing data, and that these are the real considerations which occur at Parker.

FIGURE 7. GRAMMAR: TEST VS. PROMPT

Consider again these differences in the above graph “Grammar: Test vs. Prompt” (Figure 7). These two histograms show the result of two different Grammar tests. The “District Test” results came from computer-scored multiple-choice test questions. The “Writing Prompt” scores came from manually-scored writing tests using a rubric. As seen in the previous two graphs, an important pattern developed between these two types of tests. The multiple-choice tests generated a fairly clear distinction between those students who had proficiency (41% T1), and those who didn’t (46% T3). In contrast to the discriminating results from the district tests, was the bell-curved grouping from the
rubric-scored writing prompt. Hartwell would argue that the discrepancy between the prompt score and the district test score would exemplify the queerness of “language in isolation” (125). While partially true in my opinion, the score discrepancy also indicates the difference between entirely different testing measurements, and that these differences must be better understood and appreciated within the context of testing validity.

The district Grammar test located 87% of the total number of student scores in either the T1 or T3, whereas, the Grammar Prompt placed only 64% of their scores in the T1 or T3 categories. Keep in mind that both set of scores were Grammar scores which get averaged together in the report card. Perhaps it was a wise procedure to average two entirely different types of assessment scores. However, it would be an important part of ascertaining test validity if we first established a theoretical basis and understanding for why these scores are generated and used in the manner that they are. To that end, an article like Hartwell’s might prove beneficial. In any event, the point of all these data streams is to demonstrate—not the conclusions from this information—but the trends and possible discussions which could only arise from the analysis of such data.

The Push Toward Proficiency

“Standards-based reporting…can create the undesirable incentive to focus most on the kids who are nearest the standard that counts, to the detriment of others” (Koretz, Measuring 67). With each passing year, the question has been which group of students do we target for interventions? Setting aside the intensive support given to the T3 students in which alternative ELA curriculum has been given as a substitute for the core curriculum, it has been an uncertainty as to who we target and include in interventions.
These interventions have most typically been before or after school groups, although they are now including the types of modified instruction during the school day as well. Before and after school groups have typically ranged between 12 to 20 kids (which some would argue is too large for an effective small-group intervention). Nevertheless, these groups have targeted different sub-skills in reading like Fluency and Comprehension. These groups have been teacher-directed, although the newest 5th grade group is going to be based upon computer software which tests a wide range of reading skills and provides detailed reports. For example, I was recently surprised to find that two of my stronger readers were very weak with “site words.” Here again, teacher-based assumptions and anecdotal observations are supremely benefitted with the supplement of some form of standardized data stream.

At one point, many staff members, including myself, sought to address the needs of the T3 students by trying to bring as many of them as possible over to the Basic category. That approach made sense at the time, being that these students were the most at-risk academically. Meeting the needs of the T3 students would potentially benefit everyone within the school environment if for no other reason, the frustration levels of the T3 students would abate with certain modified interventions, thus improving the overall climate and harmony inside and out of the classroom. More plainly, incidents of classroom disruptions and overall behavioral problems lessen when T3 students feel more successful. My own experience has shown me what other educators already have known, that typically the worst behaviors come from the most frustrated students (not always, but
often). Addressing the frustration level of the T3 students would have a direct impact on
school climate.

However, the more recent thought has been to target the T2 “bucket” students
with the intention of moving them to “proficient” status, T1. The trade off has been that
some groups would be serviced, while others (T3) would be left at least partially
unassisted. The question now becomes tactical: which sub-group provides the greatest
benefit to our school’s test scores, as well as to the overall growth (AYP). If there is an
agreement that we will not be able to address everyone’s needs, then some students are
going to be left out. The question is who? It’s a triage question. One primary teacher
addressed this concern by saying that her data analysis was supposed to provide her “with
areas of weakness in the student’s progress. However, it seemed that the progress was
geared towards the students that could be moved to the [“proficient’] area only.”
Unfortunately, when attention is given to the T2 students ahead of the T3 students, it
becomes a distortion of the triage approach.

The problem is real, and many teachers talk openly of the perverse
incentives that these distortions create for them…teachers have an
incentive to focus their efforts primarily on students near a cutoff point
between standards, because only changes in performance among those
students will register…focus on students who can be moved across the
proficient cut score, and ignore even the other performance standards
because they don’t matter for purposes of NCLB accountability. (Koretz,
Measuring 195)

It feels counter-intuitive to redirect resources away from the neediest T3 students and
focus them on the T2 students instead. An example of directing resources towards T2
students was evidenced in a recent administrative directive in which college tutors were
to be used “in moving the students who are scoring in the basic levels up to proficient.”

Although it may feel counter-intuitive, a strategy has been established to help those students whose test scores help Parker elementary the most.

The problem further complicates itself with the transitory nature of our school. After a certain period of time—relatively early in school year—CBEDS are accumulated and recorded (California Basic Educational Data System). This information forms the census data for our school’s demographic composition, and becomes the imprint for many different ways of identifying our school population. For example, a CBEDS-dependent formula helps determines the extent to which we are in compliance with the class-size reduction component of our QEIA grant (Quality Education Investment Act is a 7-year grant in which Parker receives approximately $300,000 each year, $2 million dollars in total). However, most noteworthy for our discussion of testing pressures is the fact all students who arrive at Parker after CBEDS are reported in October are not counted on our STAR test data. Where those students’ scores end up is beyond my knowing. It’s a bit like grade-school principal who said, “we have children who just disappear from the face of the earth”’ (Kozol 113). All I know is that they don’t count for or against my class, and they don’t count for or against our school. That could certainly cut either positively or negatively depending upon the type of student who arrives after CBEDS. More often that not, families that change schools in the middle of the school year do so because of hardship. It is not that unusual to get an academically successful student in the middle of the school, although that is less likely; it is more common to get students who bouncing from school to school. But in any event, those
students do not factor into the STAR test data for the school. As the pressure mounts to achieve a certain measure of success on AYP, the corrupting influence to redirect resources towards only the most valuable students (in terms of AYP) increases as well.

The implication of all this is that those students hold no strategic value in terms of the STAR test, which in turn means they provide no intrinsic reason for assigning them additional invention resources. Even if a student falls into the T2 “bucket kid” definition, they still hold no advantage for allocating precious intervention resources. Continuing along the same thought, a teacher could say to him or herself that if the student doesn’t count on the STAR test, then he or she doesn’t get any extra attention or concern. An upper-grade teacher said that she had to make a similar decision regarding which kids to include in her after school group; two “bucket kids” were left off the group to make room for others who were here at CBEDS. She said “It’s a moral thing…you get stuck. I hate it.” While this may seem like a hard-hearted response, it would be unreasonable to assume that considerations like these don’t occur occasionally in this era of high-stakes testing.

At the conclusion of my survey for teachers, I asked them to describe to what extent they found the following quotation from Daniel Koretz to be either consistent or inconsistent with their own professional experience. In regard to “predominately poor and minority students” Koretz says:

Where performance targets are very high relative to current scores and where community supports for achievement are comparatively weak, teachers face a far more difficult task and may be more inclined to cut corners in trying to reach their targets. (Measuring 250,1)
Many teachers responded that they “do not cut corner to reach the targets no matter who my students are.” Other teachers even found the quotation to be an affront to their “professionalism.” I wondered whether responses like those might have taken the phrase “cut corners” to imply cheating.

However, a slight majority of responses identified with the quotation, finding it “very consistent with my professional teaching experience [in excess of a dozen years].” Several teachers felt emotionally conflicted as a result of the pressures generated from testing, saying “it breaks my heart that the corners that I’m cutting are: science, social studies, art, music, PE and ELD [English Language Development].” Still others expressed “a feeling of hopelessness and despair at worst and a simple feeling of being overwhelmed at best.” One upper grade teacher clarified the term “cutting corners” by saying that “if by ‘cut corners’ it is meant to sacrifice meaningful and authentic learning in the pursuit of ‘targets’, then I certainly cut corners.” While not necessarily representative of the views of the entire staff, one teacher accurately captured my attitude:

As test scores continue to be a focal point to administrators and the desire exists by some for performance based pay, teachers will reach a point where they may have to consider cutting corners to produce the desired outcomes. A strong cry for support will be heard from the lower performing and lower socio-economic schools as teachers may leave to teach in schools with higher test scores, more support from home, and a lot less stress.

The pressure around testing is palpable, overwhelming at times, and exists on a daily basis. Although too sensitive to investigate, a teacher at Parker a few years ago was scrutinized for having potentially in authenticating her students’ STAR test scores. At
the time, I remember feeling very judgmental towards such a possible act of cheating. Today—as immoral as it may appear—I find some degree of sympathy towards the educator who buckles under the relentless pressure of high-stakes testing.

Performance Rewards & Test Prep

As a father of two children, I bristle at the ease at which I can use or overuse the reward of sugar, presents or some visual media in order to procure a specified outcome from my children. It is all too easy to do, especially when much is at stake. My wife and I find these performance rewards occurring much more frequently when certain deadlines must be met, or when the stakes are the high. For example, our daughters are oftentimes awarded hot chocolate in the mornings if they get ready for school expeditiously; sometimes they’ll get a jumbo marshmallow too, if they really hustle.

The same high-end rewards seems to coincide with high-stakes performance or achievement: students receive a homework pass for bringing their parents to Open House, or an ice-cream bar for returning a parent survey upon which grant-funding is dependent. The most notable reward in the eyes of most 5th graders would be the end of the year field trip to the waterslides in exchange for their best efforts on the STAR test. “Best effort” has been measured in the 5th grade based upon a couple of criterion. First, students must be in attendance and on-time for all STAR tests. This becomes problematic in situations in which legitimate illness or personal-necessity occurs. However, the hard-line public expectation of perfect attendance during the two weeks of testing generally produces so few tardies and/or absences that making reasonable
accommodations for legitimate absences does not seem to significantly undermine the authority or execution of the public expectation.

The other criterion for “best effort” is that students must produce evidence of their thinking process on scratch paper for each test item. Specifically, that includes students producing a written record of their consideration of each and every possible test answer, which means the elimination of the distractors. While reading through the STAR testing material and student answers is not allowed, monitoring the fidelity by which students engage the test is something of great interest and importance to me and my principal. That means that scratch paper is collected and observed to make sure that students are applying their maximum effort on all of the test items. The reward for sufficiently complying with both criteria is a day at the waterslides. Is this a good teaching practice? I have thought so. Is it a practice which may affect the performance and outcome of the testing? Without question. That is why it was designed. However, if the authentic outcome was to measure a more natural, more comprehensive assessment of student and teacher performance, then such high-stakes rewards most likely distort the original expectation quite significantly.

Just as performance rewards can occur in response to the pressure of high-stakes testing, the explicit teaching of testable items—Test Prep—exists as a result of the testing climate. Highly problematic in regards to how this practice affects the legitimacy of test scores, Test Prep was an explicit curricular component, costing the school several thousand dollars in workbook materials. Koretz warns that “inferences about improvement are undermined by certain types of test preparation that focus on the
specific sample included in the test” (Measuring 34). Differing staff responses sounded fairly—if not eerily—consistent in their goals for Test Prep:

- “provide content area questions that would be on the exam”
- “help students practice taking a test”
- “give test taking strategies practice…familiarity with the test”
- “increase awareness of the types and varieties of questions that our students would encounter”
- “I wanted my students to get a feel for the atmosphere of testing”

Absent from these responses is any amount of duplicity about the intention of Test Prep. In an environment where the stakes are so high and resources are so precious, none of the staff members I surveyed indicated any interest in equivocating around the purpose and value of Test Prep.

Using myself as an example, I used Test Prep for, among other reasons, to help students develop testwiseness. Having taken many high-stakes tests over the years of my own schooling, I thought my experience would help. An important goal for my Test Prep period was to provide an awareness and skills in eliminating wrong answers (distractors). My intention was to get my 5th graders to eliminate two distractors, thereby giving them a 50-50 chance of randomly selecting the right answer from the two remaining possibilities. I coined this strategy the “Fifty-fifty,” making that an explicit goal for each test question. Rather than targeting the goal of identifying the correct answer to the test item, my direction was to increase the probability of a correct response (either intentional or unintentional) through the elimination of test distractors. Changing the mindset away from mastery towards “Fifty-fifty” produced an entirely different emotional atmosphere.
I felt that the STAR tests were already stressful for students and teachers. Setting the students’ sites on the objective of a “Fifty-fifty” helped students better cope with the rigors of the STAR test, or so I justified.

The conflict around this type of endeavor came from the amount of instructional time devoted to the explicit teaching of testable items. The time allotted for each Test Prep period was 30 minutes each day. As mentioned in Chapter 2, this allocation of instructional time comprised nearly 10% of our total of 1,585 minutes each week. One support staff, however, adequately conveyed the feelings of the majority of staff members in saying:

I am not happy with “the tail wagging the dog” in terms of the narrow range of knowledge that is assessed through this testing, nor am I happy with the amount of time that is being used with test prep and administration. It takes away from areas like science and social studies, and for some of our students it reinforces that they are failing rather than pointing out areas of strength. However, I am even less willing for Parker to continue to take a bad rap for being in Year 5 of Program Improvement despite making steady gains in test results.

Are rewards really rewarding what I want them to reinforce? Are my motives clear in what it is I want to get my students to do? Will these rewards and interventions produce the product I want (student learning) or the product I need (proficient test scores)? Does the system reliably measure and evaluate the student performance I am conducting? It’s hard to know what is really happening.
Also mentioned earlier (Chapter 1) was the fact that the amount of instructional time devoted to testing nearly tripled the amount of testing reported from the District of Columbia Public Schools (DCPS). The DCPS indicated that approximately 21 hours a year went towards testing over the course of their academic year which, like ours, has 990 instructional hours (2%). Our total for ELA, Math and STAR testing totaled nearly 55 hours during the 2008-09 academic year (nearly 6%). 3,317 ELA assessments were administered last year alone to the sixty-eight 5th grade students, averaging 49 ELA tests for each student. Nearly thirty-thousand ELA assessments were given to Parker elementary students last year. If those numbers alone don’t speak to the volume of testing that has overtaken Parker, then add in the amount of time spent on Test Prep: now over 15% of our instructional time—the single most valuable resource any school possess with the one exception being the human resource of the students themselves—is consumed by testing or the explicit preparation for testing. Keep in mind, I am using a very conservative definition of testing and test prep. “How many hours does a teacher spend preparing students for ‘multiple assessments’? That answer depends on the interpretation of the term assessment” (Scherer 5). It could easily be argued that “preparation” for tests occurs all the time within the core of instruction. I am leaving all of that ambiguity out of my definition of testing and test prep. Fifteen percent: that is narrowly defined, includes only the two most prominent curricula (ELA and Math), plus the STAR test, and is indisputably documented in the district’s testing databases (incidentally, I am excluding another state-mandated testing practice in PE, not to mention Science, which is also on the 5th grade STAR test).
Putting the 15% of instructional hours devoted testing into perspective, I want to
draw a comparison to the Visual and Performing Arts. Last year the music program was
in its last year of existence. As is common for many elementary schools, the music
program was available for only the 4th, 5th and 6th graders. So at Parker there were
approximately 60 students who participated in band. While that number is restricted to
upper grade students only, there was no other Performing Arts curriculum available. If
990 instructional hours are multiplied by the approximately 420 students at Parker, a total
of 415,800 student-hours are available each year. Furthermore, if 60 upper-grade
students received an hour of music instruction each week for six months that would equal
1,440 student-hours, which is 1/3rd of 1% of the total instructional time spread across
every student. If those numbers seem unfair, given that I am using primary student-hours
to factor into the total number of student-hours at Parker, then I will isolate the 4th, 5th,
and 6th grade student-hours (however, keep in mind that no other visual or performing
arts curriculum existed for the primary students either; so one could say including their
student-hours was reasonable). 177 4th, 5th and 6th grade students took the STAR test last
year. Multiplying that number of students by 990 hours equals 175,230 upper grade
student-hours available last year, of which 1,440 were spent on music; that still doesn’t
equal 1% of the instructional time available. In a head-to-head comparison, upper
graders spent 1,440 student-hours on music, compared to over 26,000 student-hours
testing. Reconsidering all students at Parker, 1,440 student-hours were spent on music
last year, compared to over 62,000 student-hours of testing on ELA, Math and STAR. As
painful as that comparison seems to those of us who extol the virtues of the arts in
education, mercifully perhaps, the comparison ends there, however, as music was discontinued this year.

I found it surprisingly difficult to find other districts reporting the amount of instructional hours they devote to testing. I assumed it was because of the uncertainty at how that term would be defined. However, in this era of accountability, I do not think it will be long before other districts compare their practices with one another. One report from a nationwide survey by Beth Morton and Ben Dalton indicated that “over the last 10 years of the study of 1st – 4th graders, the percentage of instructional hours ‘spent delivering instruction in four subjects’ [increased or decreased as follows]”:

- English +2.6% (+.7 of an hour/week [25.2 additional hours annually])
- Math +.1% (+.1 of an hour/week [3.6 additional hours annually])
- Social Studies -1.9% (-.5 hour/week [68.4 fewer hours annually])
- Science -2.1% (-.7 of an hour/week [25.2 fewer hours annually])

I took the liberty to include the total amount of hours over the course of an academic year and included it in the brackets above based upon their mention of a “36 week school year” (Morton 2,3).

There may be fewer comparisons to draw between my own urban public school and the results found in the study by Michael Zellmer, Anthony Frontier, and Denise Pheifer insofar as they examined the statewide costs of high-stakes testing in Wisconsin. Nonetheless, Zellmer found that the statewide assessment for Wisconsin in 2004-05 required “4.75 to 8.66 hours of administration time annually for each student...with full implementation of NCLB testing, that number will more than double” (Zellmer 43). It
would be impossible to know if “administration time annually for each student” along with “NCLB testing” would compare to the 55 hours Parker spent on ELA, Math and STAR testing; however, they very well could be comparable, as both are statewide accountability assessments. If they are comparable, the amount of testing at Parker more than triples the high end (8.66 hours times 2) of what Zellmer is describing. Perhaps I should not be too surprised at the rate and rigidity with which we test our students. “Schools that serve low-income and minority students are especially heavily sanctioned environments” (Darling-Hammond 169). That is a fact that is reiterated constantly to the support staff, teachers, students and families at Parker.

Administrative Policies vs. Classroom Realities

How a teacher at Parker navigates his or her students through the learning process in light of the substantial pressures applied from outside sources just might be the greatest character trait he or she employs on a daily basis. “Teachers are put in the position of having to follow restrictive policies on the one hand and trying to meet the variable needs of their students on the other” (Darling-Hammond 66). Darling-Hammond calls these types of teachers who negotiate their way through so many high-stakes variables “creative saboteurs” (78). I take some issue with the suggestion that as “saboteurs” teachers would take an interest in accomplishing the ruin of the establishment. However, I understand very well the idea that in order to accomplish certain things, desperate measures are required that would not satisfy the district mandates. Better said perhaps is a quotation the former CEO of Xerox, David Kearns, who in 1988 said that it is a “baleful commentary on the nature of our schools, but the best and brightest
teachers...have to be canny outlaws to do their jobs well” (qtd. in Darling-Hammond 49).

Clearly, I take exception to the idea that teachers are behaving in an underhanded manner. However, there are numerous anecdotes regarding teachers countermanding administrative orders for the betterment of their teaching.

I developed a bit of an infamous reputation for one such minor conflict with the Open Court Reading coaches over the location of my Open Court Sound Spelling Cards. These 6”x18” cards were supposed to be put up along the front of the class for all the students to see. However, the dimensions of my room were such that these cards would have to go around the clock, over the PA speaker, and literally on top of the top edge of the projection screen. This line of cards would finish in crowded little rows underneath the shadow of my wall-mounted TV. A better location in my opinion was the longer and less-cluttered window-lined wall. Putting them over a portion of the window would not only allow the whole series of cards to run uninterrupted, but would have a natural backlight. Being that my students’ desks were situated diagonally towards the front of the room, every desk had a comfortable view of both the front of the room, as well as the window-wall where these cards remained for several years. Different reading coaches and curriculum directors became frustrated at my refusal to comply with a direction which didn’t fit the physical dynamics of my classroom. This tension resulted in a relatively small, but unhealthy pride concerning my “outlaw-ness.” While my rationale for ignoring the district mandate appears to have been well justified, the gratification from working in an outlaw manner was not professional.
Most [teachers] experienced tensions when they tried to reconcile their education goals with those of their school districts. These tensions typically resulted in conflicts with administrators, feelings of being overwhelmed by excessive demands, or a continual sense of guilt at not being fully able to serve two masters. (Darling-Hammond 90)

When another curriculum coach approached me on the subject and asked me in a less dictatorial manner, I was better able to laugh it off and comply with the district wishes. However, a more reasonably minded person than I might determine that a more constructive, harmonious and empowering approach might have been sought by both me and the district enforcement personnel had we interacted more collaboratively and less adversarially. Nonetheless, one-size doesn’t best fit all.

Dictatorial mandates don’t suit the diverse nature of my classroom dynamics. The nature of my classroom is best understood as a patchwork of 20-28 individual entities, many of which carry a much more difficult burden on a daily basis than I ever did as a youngster. Moreover, the time has long since passed since we can overlook the real-life circumstances which factor into the performance of our students. I recently reconnected with an old student of mine named James Randolph (a pseudonym). James was a student of mine when I first began at Parker. Being relatively new to teaching, and to the neighborhoods to which I was hired to serve, I was at a loss to understand James’s predicament. Our outreach counselor had visited James’s home based upon on-going concerns about James’s well-being, along with his mother’s attentiveness to matters such as her employment, sobriety, etc. After our outreach counselor conducted a home visit, the she told me that there wasn’t any food in James’s refrigerator. It wasn’t that I was astonished or saddened by the news; I simply didn’t comprehend the remark. I thought
perhaps she meant figuratively that there wasn’t anything of any real interest to eat (much like I would comment as a teenager: “Mom, there’s nothing to eat”); in fact, her comment much more closely aligned with the literal description of the refrigerator having no food items inside of it. James—as I was later to learn—was much less of a rarity than I had hoped or expected. Many students eat their free breakfast and lunch at Parker, and then go the rest of day until the next day’s breakfast again.

I am neither an apologist for the poor condition of our neighborhoods, nor the sub-par condition of our test scores. My record of 10-plus years of delivering instruction within what some would consider a desperate environment establishes the fact that I am not intent on excusing the reality of student’s lives. However, if, as some have suggested, there exists no differences between the warm comforts of a stable, supportive household and the wanting difficulties of poverty, then shame on the individual whose ignorance or hard-heartedness produces such an abominably warped utterance. For example, in her June 2009 newsletter, Missouri Republican state representative Cynthia Davis questioned the value of free or reduced meals for public school students in school, writing that, “Hunger can be a positive motivator” (“Hungry Kids”). Supporting that claim, Mike Spence, a member of the West Covina school board said, "The government is trying to usurp the responsibilities of the parent. There is a trend to take over aspects of what the family does" (“Hungry Kids”). In my opinion, there is nothing “positive” nor “motivating” about James Randolph having an empty refrigerator. In fact, students living in poverty and hunger is becoming increasingly common. “A growing share of U.S. children (nearly one in four) now live in poverty; many arrive at school hungry,
unvaccinated, and frightened by the violence that surrounds their lives” (Darling-Hammond 25). It is hard to expect student productivity in light of these daily barriers.

Furthermore, it is reasonable and right for some accommodation to be built into a testing system. “Income itself does not directly affect students’ learning, but other factors that are influenced by income, or that determine income, do” (Koretz, *Measuring* 129). Assessment and delivery of instruction are inseparable; and classroom management—which is most heavily affected by the support and structure being offered to students at home—is the leading variable affecting successful delivery of instruction. Such an accommodation might simply include expansion of the “safe-harbor” designation for schools in at-risk neighborhoods. Accommodations should account for the transitoriness of the student population. It is not uncommon for me to turn over 20 students in the course of a year, some students staying only a few weeks at Parker. Those fluctuations not only need to be accounted for, they need to reflect back on the problematic reality of teaching in such a dynamic environment. Add to that equation the number of English language learners, crime rate, etc. and you have some real variables for consideration. I appreciate the fact that researchers like Koretz document the fact that noneducational factors heavily affect the quality of student learning: “The scores of any given cohort of students are shaped not only by the quality of their education but also—and powerfully—by noneducational factors, such as social background” (*Measuring* 69). These noneducational factors have clearly and profoundly affected the educational experience for my students since the day I was hired at Parker.

Commoditizing Education
Julia Barrier-Ferreira, in her article “Producing Commodities or Educating Children,” advocates for an educational system which advances the “full range of human talents,” including the emotional and spiritual development of students (139). This was the first time I had thought of the educational experience being reduced to a mass-produced product, indistinguishable from all other like products. With items such as sugar or rice, there is no real difference between a commodity produced in one part of the world from a commodity produced somewhere else. What is left then is a product that holds no real value in terms of quality in comparison to other like items; it only holds a value in terms of its quantity. Its quality is fairly standardized; the location of its production and distribution are insignificant: the only real comparisons to be made would be the amount of the commodity. For Parker, the commodity is the amount of proficient student test scores we can distribute annually. Darling-Hammond connects schools to factories, saying that “like manufacturing industries, schools were developed as specialized organizations run by carefully prescribed procedures engineered to yield standard products” (16). Underwood writes that “these procedures make every individual into a ‘case,’ one that can be compared, measured, and judged in relation to others” (49). I understand that from an observer’s perspective looking at student performance requires a fair amount of objectification. However, when students become “cases” or commodities and remain that way well into the day-to-day practices of a school like Parker, we have lost our way as instructors of individuals.

Likewise, with the commodity of test scores, the comparison between schools is not the quality of proficient students—or the quality of their educational experience—it
is merely the amount of proficient student test scores (notice that I am distinguishing between proficient students and proficient student test scores, as they are not necessarily the same thing). Therefore, the distinguishing characteristic of a commoditized product is that its value is determined by market forces. In the case of a school, those market forces would be financial, temporal, political and human resources. Unfortunately for a school serving a low-income neighborhood, “if children are seen primarily as raw material for industry…market values do not favor much investment in the poorest children” (Kozol 75). Market forces are going to favor investment in the higher performing students and school districts, further disadvantaging the at-risk schools.

Commoditization is a complicated problem with many negative consequences, perhaps the most damaging being the depersonalization of the student. Underwood says that assessment procedures “put a premium on getting good scores in a cost-effective manner without regard to ‘what that effort does to the teacher, the learner, or the curriculum’” (217). One manifestation of the commoditized educational experience can be seen in the Data Wall. Covering two 4 x 16-foot walls, these pocket charts contain basic pieces of student information on individual cards, including assessment data. The cards are then placed according to the students cumulative score in either ELA or Math. All 420 students from Kindergarten through 6th grade are charted in order to see emerging trends or patterns. For example, a large number of students from Kindergarten through 2nd grade scored Advanced or Proficient in ELA (Figure 8). The bottom left of the photo indicates how the majority of students beyond 3rd grade fall into the Basic, BB, or FBB categories.
When discussion commenced at staff meeting last year about the possible hiring of an intervention specialist to assist our students, the data wall pointed to a real need to help transition students around 3rd grade. While an impersonal arrangement of living, breathing students, the data wall does serve an important function in helping to makes trends and patterns more comprehensible to staff, thus providing a more effective implementation of instruction, intervention, staffing and resources. It cuts both ways: the data wall assists in determining RtI; but it also continues this process of objectifying individual and their dynamic experiences.
FIGURE 8. 4X16-FOOT ELA DATA WALL

FIGURE 9. DATA WALL (CLOSE-UP)
Am I suggesting that the current state of our educational system within California in general or Parker Elementary in particular resembles a sugar cane factory producing a nameless, faceless commodity? There are certainly enough comparisons to be drawn to raise an uncomfortable amount of alarm. There is no question that the standardization process of curriculum and assessment both within the state and throughout the nation has deliberately targeted the goal of unifying instruction. It is a noble goal to have equal standards and expectations for all 5th graders from Fresno to Richmond. However, it is quite a disturbing outcome if this emphasis towards assessments creates commoditized 5th graders that bear no qualitative difference be they from Fresno or Richmond.

The question arises as to the source and origination of these educational priorities which seem to point our schools in the direction of commoditization. The priorities for my school are specified in our “School Plan,” its three main objective being:

1. All students will meet annual measurable objectives (as measured by NCLB) in English Language Arts as measured by California Standards Test.
2. All students will meet annual measurable objectives (as measured by NCLB) in Mathematics as measured by California Standards Test.
3. Harmon Johnson will develop and maintain a positive learning and working environment (Johnson 12-17).

The first two goals inevitably receive the lion share of resources and attention, most likely because of their ability to measure and articulate growth. Even more so, the first two goals bear upon the STAR, while the third goal does not. The repetition of “all students…measurable” reminds me a bit of what Koretz cautions when he says that “students’ performance is reported simply in terms of the resulting bins into which their scores fall” (Measuring 181). It is so devoid of anything individual; it points directly to
the quantity-versus-quality, one-size-fits-all mentality. As Thomas Sobol says, “our goal is standards, not standardization.” (470). I am not sure our school plan adequately differentiates between the two.

The third goal looks eerily like the proverbial “third wheel,” connected, but unmistakably different than its counterparts. Truth is, we as a staff are growing less and less accustomed to addressing the issues referred to in the third objective, as our attention and training continually points us in the direction of measurability. As Quellmalz so accurately states, our attention becomes focused on producing “standard, replicable application” within both our assessment and teaching practices (69). “Standard” and “replicable:” words to describe a factory, not a school where I’d want my daughters in attendance.

Some students have come back to visit me after five or ten years, sharing some of their favorite anecdotes they recall from their 5th grade days. Oftentimes, students report to me that our end of the year camping trip was an extremely memorable experience. This is understandable considering that for most of my students they had never been camping before. In fact, when I take students out to a restaurant like Mimi’s Café as a reward for some academic accomplishment, it is common for many of them to report that it was their first trip to a sit-down restaurant of any kind (excluding fast food, like McDonald’s or Taco Bell).

Another likely anecdote to be heard from a visiting student is the name of a song that we’d sing in class like “Who Put the Bomp.” Holding onto an activity like song practice has been increasingly difficult over the years. Unable to justify it as part of
either my “opening or closing” periods, I even resorted to making song practice an activity station within our Physical Education rotations (my rationale being that it was an aerobic skill activity requiring muscle control.) This reasoning proved too flimsy to justify its continued existence, and it too was cast onto the scrap heap of other academic or enrichment activities unable to compete with the highly valued Mathematics and ELA instruction. Those activities include the virtual elimination of Social Studies and Science as stand-alone curriculum. They are embedded within the core ELA curriculum, but can only exist as a double-dipping by-product off the core instruction. Art has literally gone underground and is mentioned with ironic humor and secrecy when included in a day’s events.

Read-alouds, like the novel *Where the Red Fern Grows*, (which I have read to my classes every year for eleven years) have only been possible for us these last few years as an activity that takes place before the district completes its CLEDT testing and fully expects our ELD periods to be operational. Perhaps the least controversial comment in this paper would be to say that these opportunities to connect with my students are the richest and most enduring opportunities for my students and me. “In this assembly-line conception of teaching, relationships matter little” (Darling-Hammond 195). However, in the increasingly commoditized educational environment, these types of opportunities become far less valued from an administrative perspective, as well as far less frequent.

Much can and should be said about the necessity for assessment measures to gauge student/teacher performance, and to help inform proper planning of instruction and intervention. A school like Parker that has struggled near the bottom of an
underperforming district for a number of years must employ every reasonable measure, as well as exhaust all available efforts and resources in order to redirect the performative momentum of the school. However, the measure of a student (or a teacher) is vast.

Many components factor into such an assessment of an individual, including those elements which would be harder (but not impossible) to quantify. Many of those elements which are difficult to measure reside within the domain of the social skills, which include not only the behavioral skills, but also the arts. At present, Parker has no music program. There are very few, if any programs which specifically target the character building of at-risk students (programs which have existed in the past). The library services fluctuate between being closed, to being offered in a stripped-down fashion (no 5th grader has been to library with their class: no time). Interschool programs which encourage students to write letters to fellow student pen-pals exist only as memories to those staff members who have been around long enough to remember them (or to those who still have a metal mailbox in their classroom, like me). The school beautification program (The Green Team) and student choir—two after-school activities that I created and ran—have remained dormant for years. Programs like these enrich and round out the character of students in ways that are not easily discernable in a standardized manner, yet very well may manifest a positive presence within an individual’s life at a later date. The value of such student opportunities truly goes without question. The alternative is what Darling-Hammond describes as “efficient batch processing of masses of children…students move along a conveyer belt from one teacher to the next, grade to grade, and class period to class period to be stamped with lessons
before the move on” (17). I can’t imagine that extreme view of education would benefit anyone.

So I return to my analogy of the boys’ bathroom sink which has twice fallen off the wall: the first time because it was fastened with nails; the second time, multiple screws were driven into the same hole. I see our educational process as possessing some of the same haphazard qualities as the sink repair, trying to drill multiple students into the same commoditized experience. One size does not fit all; neither do all multiple kids fit conveniently—or even accurately—into the same designation.

“Schools are consequently almost always aboil with some kind of ‘change,’ but they are only rarely involved in any deliberate process of improvement, where progress is measured against a clearly specified instructional goal” (7).

“Building a New Structure for School Leadership” by Richard Elmore, Ed.D. Harvard University Graduate School of Education
Chapter 4

RECOMMENDATIONS AND CONCLUSIONS

In the previous chapters I have described the narrowing effect that over-testing has had upon the educational experience at Parker Elementary. I have also placed the case study of my school within the larger context of scholarship from the field of Education, particularly the topic of assessment reform. Now I want to return to Messick’s definition of validity in an attempt to apply this evaluative device specifically to Parker’s testing procedures.

Testing Validity

Messick says that “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (5). The quotation can be reformatted into two criteria statements: 1) an assessment protocol should have some identifiable rationale as its basis, grounded in theoretical assumptions and empirical classroom evidence; 2) the testing data should generate inferences about student and teacher performance, along with an array of interventions to support student learning. Both criteria must be satisfied in order to establish the testing program as being a valid measurement procedure. Remembering Lee Cronbach’s mention that the search for validity never ends, it is important to develop an understanding of these two criteria in a way that extends into the future; considering validity as a one-shot endeavor misses the essential nature of this examination. Koretz summarizes the need for validity:
Using the definition of validity, a standard can be established determining to what extent my school meets this stated goal. Considering that NCLB specifically requires assessments to be “valid and reliable,” the question of validity is not merely a question of prudence, common sense and good stewardship; it also indicates legal compliance. (Measuring 215)

I can’t help but find the legality of validity to be a bit humorous. I would hate to see us all cuffed and carted away for not evaluating our testing protocol. In all seriousness though, there does exist a moral imperative within understanding of every educator to do what is best for his or her students. To that point Cronbach could not be more correct: validity never ends.

As we shall see, my school’s current capacity for satisfying the validity criteria is mixed. Some components are in existence. As an administrator said in his survey response, “Pieces of this plan are already in place. We have teachers working with groups of students but the work is not focused or systematic enough to be evaluated for its efficacy.” However, as mentioned, a general lack of focus and the absence of a systematic approach will be some of the areas of need for Parker to advance towards a level of testing validity.

The first criteria states that our testing practices needed to be based upon two things: a theoretical foundation and empirical evidence as gathered from classroom experiences. As a teacher-researcher I can appreciate both aspects, although I strongly favor the empirical evidence as that cannot be disputed as anything other than reality at Parker elementary. Nonetheless, both are important, especially considering that the theoretical foundation helps elevate the empirical evidence as being more than just the teacher “lore” that Stephen North derides in his book The Making of Knowledge in
Composition: Portrait of an Emerging Field. Unfortunately, the theoretical basis for Parker’s testing program is consistently absent from the awareness of those teachers surveyed. One teacher responded to the question—“What empirical evidence or theoretical rationales (if any) is used either in support for or against our testing practices?”—by saying, “I have no idea what evidence or rationales exist for the way we test.” All other teachers said nothing to the mention of a theoretical rationale, leaving me to conclude that they knew nothing of such a basis. I certainly am unaware of what theory is going into our testing practices. The staff has been given a photocopy of a Marzano article, but I’m not sure how that factors into the testing protocol established by the district.

One thing is certain, however: the district is using the STAR tests as a basis for our own testing program. One teacher said that, “Everything we do is to help the students score proficient on the STAR tests.” Another said, “If a student performs well on the [district] tests, they perform at a higher level on the CST [STAR test].” As a basis upon which to build a testing program, using the STAR test as a model is clearly the most practical of criteria, at least in terms of producing proficient test scores. However, we fall back into the trap of the “tail wagging the dog” as one support staff said earlier; our testing program is based upon its emulation of the state testing program, rather than some theoretical concept of sound educational practice. A teacher indicated in response to the question about the Test Prep period that, “I wanted my students to get a feel for the atmosphere of testing, and the length at which they would be expected to read passages.”

In essence, the culture of testing is self-validating. The STAR test is the new context and
rationalization for our testing procedure. In other words, from a teacher’s perspective, we have no stated theory upon which to explain and justify our testing program. The STAR test—while constructed from theoretical and educational considerations—is not a theoretical rationale; it is a social and political measurement of student, teacher and school performance. Wayne Au confirms what our teachers are saying: “Content is increasingly taught in isolated pieces and often learned only within the context of the tests themselves” (“High-Stakes” 263). Perhaps most discouraging is a support staff’s response to the same question. Instead of mentioning some educational theory to reference our testing practices which consume such a massive amount of instructional time and resources, his answer was that, “The rational is that assessment should inform instruction.” That is the same default response that most of us teachers have, which basically amounts to a blind faith in an untested, unexamined and theoretical-deficient program.

In fairness, I should say that I am confident that a foundation of educational research and theory for the district’s testing procedures—as well as for the STAR test—does in fact exist, although those theoretical premises were never made explicit to teachers. However, in order for Parker elementary to satisfy the definition of validity, the people in whose hands these assessment tools have been entrusted need also to be in possession of the theory behind the development of such tools. Without an understanding of the educational principles which comprise our testing program, we are operating an enormous operation in total ignorance and—worse yet—without questioning the soundness of those educational principles. “Superintendent Honig repeatedly expressed
the view that test validity is a social construct” (Underwood 205). Regrettably, within the social construct of Parker elementary, this element of testing validity—in which a theoretical basis is understood by staff members—does not exist.

The second provision for testing validity—that data should support inferences and actions which lead to improved student learning—is one which is in evidence at Parker. Underwood said that while he worked as a teacher he “encountered no evidence in the field record to support the claim that teachers acting on their own behalf actually did look at test scores as the impetus for specific instructional behaviors” (217). Times have changed. Teachers are presently using data to drive instruction, and have been doing so for the last few years. The Data Wall may be the best representation of putting testing data in a position where it can be analyzed and utilized to help students. However, many other examples of data-driven interventions exist. Here is one response from a Parker teacher regarding how test data informs instruction and/or interventions:

Currently I utilize Lexia [a computer-based ELA program] fully, work with students at workshop based on their level (again, not as strong w/ this as I would like due to vast needs and time constraints). I am beginning before and after school groups on Lexia. I also utilize my bilingual paraprofessional [BP] to work with some EL students that are struggling. Here we are looking at what their skill level is, and working there, versus just having her help with class work.

Her comments indicate how test data is being used in several different ways. Instead of having her BP just “help with class work,” this teacher uses testing data to better locate areas of need, thus increasing the efficacy of the intervention. This teacher is also using Lexia—both in class and before and after school—as an assessment tool to identify student proficiency, as well as an intervention for students needing additional support.
One teacher is piloting a new technology in which students have hand-held devices (“clickers”) to which they can electronically submit their responses. She says, “I start every [Math] lesson with a review using the previous lesson’s “online Quiz,” “Daily Spiral Review”, and “Quick Check”. My students use response clickers for the online quiz so I can get immediate feedback on skills they are struggling with.” The outlook for technology-based instruction looks promising as curriculum, assessment methods and teacher training converge. It is unclear how far away our staff is from the confluence of these three elements, although we are clearly moving in that direction. More needs to be done, however. A support staff member said that, “In the past we spent more time (complete days) going through data and making plans by grade level. I think it is unfortunate that we have discontinued that process and focus. Most, if not all of the inferences about student learning that are based upon testing data are done in isolation by individual teachers without the support or structure of an organized staff consensus. A campus-wide commitment to testing validity will unify staff efforts to increase student learning. It will also accelerate our arrival to the type of instruction that is responsive to trends in assessment indicators, while employing current technologies to heighten student-teacher interaction.

Many factors can make positive contributions towards this goal of an evolved learning environment based upon valid testing practices: administering fewer district tests, safeguarding test results, using contextualized assessment tools, and eliciting local input. At the top of the list is a reduction of the volume of district tests, particularly in English and Language Arts. As discussed in detail in Chapter 3, the ELA testing
consumes enormous amounts of instructional time and resources which could be better spent in other ways. With great measure and reasonableness, one of my colleagues said, “I know that one rationale is that it is important to have consistency in the content, and type of testing in order for it to be true reflection of achievement. However, to be honest, I am not aware of the empirical evidence to support the frequency in which we test.” I strongly propose a 1/3 reduction in ELA testing which would emancipate 1% of the total number of instructional hours available each year; it would slow the frequency of ELA from every 3 to 4 week, to every 5 to 6 weeks. While a 1% return of instructional time will not undo the commoditization of the educational experience at Parker, it is definitely a step in the right direction. Furthermore, any measure to help abate the narrowing effect must be considered a general trend towards improving student learning, and therefore, increasing valid. How would the reduction in ELA testing affect our ability as teachers to monitor student performance? Another teacher at Parker said:

Less testing would take some pressure off my students. I have many students below grade level and that isn’t going to change drastically in a 3 week period. So telling my student every 3 weeks they are still below grade level is very frustrating and discouraging for them. If we were able to test less it would allow me to help my students…re-write a writing test they may have struggled with. Instead, I feel like we are constantly moving forward and teaching to the next test instead of being able to really focus on areas of need.

The volume of testing is massive by every indicator I could find, including the fact that DRUSD eclipses all other school districts I researched. From the number of hours I have spent educating myself on this particular consideration, a reduction in testing appears to have only benefits for our school. Brian Huot echoes this sentiment in regards to college-level writing tests: “This reduction in the number of times teachers have to grade can not
only free them to do more teaching, but it can also alter their roles within the classroom” (73). Although Huot is referring particularly to writing assessments, the same end-results can be expected with a reduction in any number or type of assessments, which would improve educational quality.

As evidenced by the research of Kathleen Rhoades and George Madaus, testing error has to be considered as a real threat to all types of standardized measurements, none more so than measurements like STAR which account for a school’s AYP. Therefore, as unlikely as it may be from reaching fruition, I am suggesting some type of audit test after STAR testing is completed. Something as simple as allowing students to retake the STAR test would significantly abate the problems we know exist as a result of testing error. Koretz, who discusses in his book with great detail the various types of testing errors inherent in these types of NCLB assessments, says that:

Given the current high-stakes use of tests, we can be confident of the validity of inferences about improvement only if we have an additional type of validity evidence: a comparison to a second measure less threatened by the possibility of corruption (often called an audit test). (Measuring 247,8)

I realize that requesting an audit test on the heels of suggesting a reduction in district testing may appear counter-productive. On the contrary, an audit of high-stakes testing better ensures that our STAR testing is reliable (in the colloquial sense that it can be trusted). In addition, moving instructional time from one area of testing that is vastly over saturated, to another area that need further validation helps restore balance to our testing practices. “A common response to measurement error is to allow students who fail the test a second chance to take it, to lessen the probability that students will fail only
because of measurement error” (Koretz, *Measuring* 178). If it is such a “common response”—being that it makes common sense to protect against measurement error when the stakes are so high—it would be all the more reasonable to enact such a practice at Parker, if not statewide.

Much has been written on the subject of authentic assessments, perhaps as much by anyone as by Grant Wiggins. In the article “An Exchange of Views on ‘Semantics, Psychometrics, and Assessment Reform: A Close Look at 'Authentic' Assessments” Fred Newmann, Ron Brandt, and Grant Wiggins discuss the pros and cons of authentic assessment, among which they say, “Balance is essential in teaching as in testing, but that conventional testing is guilty of a woeful imbalance on the lower end of the taxonomy” (21). Similarly, the advocacy of portfolio use—as well as other curriculum events—must be considered as a viable supplement to high-stakes standardized testing. Too much has been documented to discount the viability of using such countermeasures for state wide reporting, including their use in states like Vermont and Kentucky. Huot says concisely that, “A primary consideration for portfolios is that they help us to see assessment in a new light, one that connects teaching and assessment” (71). That “connection” between teaching and assessment is the contact zone of student learning; of course a school’s testing procedure would seek to target that type of connection as an essential component of their validity criteria. “Meaningful performances in real-world contexts need to become both the stuff of the curriculum and the focus of assessment events” (Darling-Hammond 115). For the purposes of my project, I refer to all types of authentic assessments (performative assessments, portfolio projects, assessment events) as
“contextualized assessments,” insofar as they aim to measure students within a more real-world context. How our school incorporates some measure of contextualized assessment can be deferred to our school’s leadership committee. Nonetheless, it needs to be considered as a means of counter-balancing the stripping away of the rich context of learning via high-stakes testing. This brings me to my final recommendation in terms of testing validity.

No definition of testing validity could be considered legitimate unless it incorporated the concerns of its stakeholders. “If we are committed to assessments that promote teaching and learning, then we must listen primarily to the voices of educators and their students” (Huot 56). I resolutely believe that testing validity does not exist without local input: putting authority and responsibility back in the hands of the teachers. Teachers need “flexibility to teach adaptively…relationships with students for knowing them well and motivating them” is at the core of a teacher’s mandate (Darling-Hammond 71). When asked if a line of communication existed in which site-based feedback factored into decisions about district testing protocol, answers from the four responding support staff were split. Two discussed the manner in which feedback passed back and forth from the district to the teachers and back again. However, the other two indicated that they had “heard of no instances where site-based feedback factors into decisions about district testing protocol.” The split in opinions within the administration should indicate misalignment at the very least, if not a full blown breakdown in communication between teachers and district. The disconnect between teachers and the development of assessment tools further illustrates how abstract these measurements are, and how blindly
and loyally we accept and implement them. One teacher restated my survey question in her response by saying that she sees no “empirical evidence-based on observation and experience” in our testing practices, particularly the STAR test. “If one accepts this meaning, the answer to question 6 is: None. I see the STAR test for this year’s students next year!” In that regard, teachers are fully disconnected from the testing event which most heavily evaluates and impacts our school.

At issue is control, pure and simple. Do districts and administrators trust and authorize teachers to evaluate student needs? I’d say that the degree to which a district restricts teacher authority and enforces top-down mandates is the extent to which its trust in its teachers has atrophied.

One does not “control” improvement processes so much as one guides them and provides direction for them, since most of the knowledge required for improvement must inevitably reside in the people who deliver instruction, not in the people who manage them. (Elmore, “Leadership” 14)

To state it as bluntly as I can, a district like mine which goes to such great lengths to enforce tight controls on the delivery of instruction (“fidelity” it is euphemistically called), and adherence to a brutally rapid testing schedule can only be seen from my perspective as a district which has effectively “teacher-proofed” its schools.

Being a father of two little ones, I know that the motivation involved in child-proofing a home is beyond questions. Therefore, I don’t seek to analyze or question the motives of my district for the steps it has taken to teacher-proof instruction. I fully accept and defend the position that SUSD has created curriculum and assessment protocols which they believe will increase student performance. Nonetheless, the notion that teacher-proofing
instruction supports student learning is a documented and categorical fallacy. When discussing “teacher-proofing tactics as a major school improvement strategy,” Linda Darling-Hammond says that, “The evidence suggests that highly prescriptive curricular mandates do not improve student learning, especially if they effectively control teaching” (53). I am not anti-district, nor do I find the previous quotation to be in conflict with viable district protocols. I do believe unequivocally that my district is working within an outdated model of management.

Restructuring Leadership

Perhaps it can be said of any organization. It is certainly true of my school system: the purest of intentions, the most erudite commentary and theory on teaching practices, the best resources and technology, the most effective assessment measures available, solid parent support, and devoted, capable educators—none of those elements can be fostered, sustained, and utilized to their fullest potential without an evolved leadership structure suited to this era of accountability. In this final section of my conclusion I am going to advance criticism of all the basic school system positions, including my own. From what I have read and taken to heart, public schools face the difficult challenge of evolving to match the public and political expectations of accountability (Au; Darling-Hammond; Koretz; Pedulla, Abrams, Madaus, Russell, Ramos, and Miao; Sizer). The primary resource for my recommendations regarding leadership change is Richard Elmore’s article “Building a New Structure for School Leadership.” Circumstances warrant changes at all levels of the school district. However, perhaps it is for the best that the sweeping overhaul that I am suggesting be
done in a shared manner, for no other reason than this: schools have always been locations of shared experiences. Why would school reforms be any different?

Again, the basic thread that I am taking from Elmore and weaving through my recommendations about a restructured leadership is the premise that accountability measures and all their attendant expectations from families and school boards and media members and real estate agents and politicians…these dynamics are not going to go away. Therefore, my school district must adapt in order to accommodate the level of transparency, accountability and responsiveness expected of us. Starting from the bottom of the current bureaucratic system, the teachers must evolve from this antiquated system founded upon teaching as an idiosyncratic art. The premise that teaching is a mysterious practice that some do well and others to a lesser extent—that some children can be taught inexplicably while others less effectively—those days are behind us. Teaching is an activity that can be modeled, practiced and transferred between those who do it well and those who have yet to acquire such training and practice. Teachers like myself may balk at such suggestions “respond[ing] with well-known arguments that conjure up the mystery and inviolability of the unique relationship between each student and teacher and its need for distance from bureaucratic or policy controls” (Elmore, “Leadership” 9). Nonetheless, those arguments must be set aside for a renewed commitment to the craft of sound teaching practices. For example, it has recently passed around my campus the idea that randomly selecting students for response (as opposed to accepting raised-hand respondents) is a practice which will heighten student accountability and engagement. However, not every teacher at Parker has developed a system for randomly calling on
students, although many have. A colleague recently modeled how she uses the computer to project student names on the board for response. Simple practices like these can and must be shared and implemented.

Moving from individual teachers to cohorts of teachers, Elmore advances a notion of “comparative advantage” in which teachers collaborate and assume responsibility for teaching particular curricula “consistent with the comparative expertise of their roles and avoid activities that are beyond their expertise” (“Leadership” 23). This strategy of distribution includes the parceling out particular instructional domains of interest or expertise within a grade level team. The misapplied lay-term “team teaching” approaches this concept, although only vaguely. What is being suggested is a model of grade-level collaboration in which certain teachers deliver certain types of instruction. Ideally this would occur as a result of the teacher’s background or expertise, although having collaborated like this with my wife and grade-level partner for a number of years, I can attest to the fact that sometimes we accept certain curricular responsibilities out of boredom, or the need to avoid boredom. I have seen grade-levels take turns teaching the “high students” or the “low group,” out of fairness and as a means to avoid fatigue by a particular teacher. In any event, the benefits are substantial: teachers and students retain a desperately-needed element of freshness, as instruction is delivered with some degree of volunteerism; student accountability ceases to be secluded to a single teacher and is spread amongst a cohort of teachers; isolated and idiosyncratic teaching practices become unsustainable in a collaborative environment; sharing of “best practices” becomes an embedded element in grade-level discussions. Darling-Hammond says that grading
collaboratively (such as writing prompts) “rather than grading students in isolation helps teachers make their standards explicit, gain multiple perspectives on learning, and think about how they can teach to produce the kinds of student work they want to see” (237). This is just one of numerous benefits to be gained from enhanced grade-level collaboration.

Another recommendation that Darling-Hammond makes—one that I have resisted in the past—is the concept of “looping.” She advocates:

place teachers with the same students for two or more years and that creates teaching teams with distributed expertise…result[ing] in greater knowledge of the child and family situation, more time to build useful teaching strategies for individual students. (188, 9)

Working collaboratively with my wife as my grade-level partner for eleven years makes me a well-qualified advocate for stronger grade-level collaboration. However, other teachers at Parker have worked closely longer than we, and there appears to be many distinct advantages to those arrangements. Whether desirable or not though, the future direction for elementary teaching points to increased accountability to which grade-level collaboration is arguably the most potent remedy.

From individual teachers and grade-level cohorts, the next stratum of staffing includes “support staff.” A loosely-defined term which includes nearly all other credentialed personnel who interact with students on campus, support staff have been included with administrators as “nonteaching” personnel. A compelling argument that Darling-Hammond documents with comparisons to other school systems overseas is that the American school system has recently experienced rapid and disproportionate growth in nonteaching positions. Her contention is that schools systems have become top-heavy
with nonteaching personnel, diverting teaching resources away from classrooms and students. She suggests, “reducing the number of specialists and nonteaching staff and increasing the number of classroom teachers” including more “pushing in” instead of “pulling out” (188). While I will abstain from recommending the elimination of specific positions to which I can place the name and face of my friends and coworkers, I do concur with the general aim. Shifting the emphasis from nonteaching personnel towards increased teaching personnel helps children.

However, one newly emerging support staff position is that which works as an intermediary between administrative concerns, the collection and dissemination of testing data, current trends in teaching practices, and classroom teaching itself. This position can be considered by many titles, but at Parker is referred to as the “curriculum coach.” While all positions from the teacher to the superintendent need to evolve towards an intimate understanding of the effective delivery of instruction, the curriculum coach is the axis of these convergent positions. Underwood refers to this job as the “observer-participant,” which may not have existed as it does its present incarnation.

Nevertheless, his description fits the role of the curriculum coach:

> Every school ought to have on staff an observer-participant whose job it is to collect and interpret data for the local school community with reference to whatever assessment systems are applied at a site. The creation of this position as a routine part of the school would acknowledge the power of assessment and would provide the capacity of the school to understand and mitigate the damage done to children by any assessment system. (16)

To further clarify this evolving position of curriculum coach Darling-Hammond warns us that “instructional problems [can not] be solved by inspectors who make occasional
forays into the classroom” (67). This first-line-of-support position to interact with testing data and teaching practices needs to be an integral cog in the classroom machinations. The curriculum coach must know teachers’ practices and students’ background and abilities in order to provide usable feedback with regard to instruction. Lingering too long at either spectral-end of the job description—be it on the data side, or the classroom side—would render that coach ineffective. The coach must straddle a line between current strands of assessment data and changes in classroom practices, in order to provide punctual information and advice. Furthermore, curriculum coach must stay abreast of emerging trends in student performance in a broad sense. Being on top of the numbers is one thing. Having an understanding of the students and the various elements which factor into their productivity is equally important. One teacher showed the current lack of efficacy in regards to our data collection and program implementation when she said, “I know that at one time we were gathering data school wide, and that several staff members put in a great deal of time assisting us in getting that information organized. However, I am not sure what has become of that information.” Another teacher remarked that a quick retrieval of testing data for the purposes of directing future instruction “almost never happens 😞.” Teachers, students and families need to be in the know all the time regarding student achievement. That task must be spearheaded by the curriculum coach.

A compelling thought put forth by Darling-Hammond would be to locate two curriculum coaches within the teaching ranks. By giving a half-time coaching position to an upper-grade and a primary teacher, those teachers could remain in their classes, keep
their pulse on student performance, while interacting with the administrator and executing the ancillary tasks of organizing interventions, training fellow teachers, etc.

The two teaching halves remaining from this contractual arrangement could go to a single teacher working in both grade-levels, which provides yet another opportunity for communication and cross-over between the otherwise disparate grade levels at Parker.

This interaction with curriculum and classroom practices will probably hit hardest upon the position that has historically dwelled upon the peripheral outskirts of instruction: the principal. Elmore said that the traditional role of the principal was that of a buffer between the idiosyncrasies of classroom practices and the expectations of the public and/or political interests. That role has outlived its purpose. The new purpose of the principal is to be deeply knowledgeable and astute at providing teachers with practices intended to increase student learning. In this era of accountability principals are “recruited, evaluated, and retained or dismissed based on their ability to understand, model, and develop instructional practice among teachers and ultimately, on their ability to improve student performance” (Elmore, “Leadership” 29). A support staff member at Parker captured this sentiment well when he said that administrators “need to be able to look at data to decide if our teaching practices are effective.” I suspect the growing pains associated with that shift in responsibility will be felt acutely by principals everywhere, as I have seen or heard of very few such principals that would closely fit that description. Elmore continues:

One thing is clear: schools and systems that are improving directly and explicitly confront the issue of isolation embedded in loose-coupling. Administrators—both system-level and school-level—are routinely engaged in direct observation of practice in schools and classrooms; they
have mastered ways of talking about practice that allow for non-threatening support, criticism, and judgment. ("Leadership" 32)

To me, Elmore seems to be describing an entirely new skill-set for a principal, one that includes foremost the ability to talk to teachers about their teaching (a master-teacher of sorts). In my professional life, I have been the most invigorated as an employee when I have worked underneath the guidance and tutelage of someone who deeply understands the craft. The difference between an evaluation by a principal lacking this understanding of the teaching craft, and the coaching and mentoring of an expert in sound teaching practices would not only be profound, in my opinion, the later would be sublimely rewarding and warmly received.

Underwood mentions one more key component to the emerging responsibilities of the principal. If we accept the fact that site-administrators need to be intimately involved in classroom learning, then that requires that position to explicate and communicate the specific educational goals and practices at that school. Underwood says that there must exist a “tight linkage between learning outcomes and school plans” (7), which unfortunately does not exist at Parker, in my opinion. Borrowing phrases from other schools’ plans such as, “All students will meet annual measurable objectives,” our school’s plan sounds cliché and aloof (Johnson 12). It doesn’t resonate with the voices of our school or our community (again, in my opinion). Elmore says that in order for our school to improve student learning we must “create and nurture agreement on what is worth achieving, and they set in motion the internal processes by which people progressively learn how to do what they need to do in order to achieve what is worthwhile” ("Leadership" 25). Articulating practices to improve student learning,
building consensus around those practices, and then monitoring and supporting compliance with those stated expectations—all with an understanding of effective instructional implementation—those are the attributes of the newly evolved principal.

The final stratum of the leadership structure is the superintendent. Elmore says that historically, the superintendents’ role were “based on their capacity to maintain a working majority on a relatively unstable elected board, rather than on their capacity to focus the institution on its core functions and make steady improvements over time” (“Leadership” 7). They too, like the principal, run the risk of becoming a school system appendage, lest they adopt an active role in curriculum and classroom practices. Elmore says that within restructured school districts:

Their superintendents were knowledgeable about, and the key initiators of, changes in curriculum and teaching strategies…were active in monitoring curriculum and instruction in classrooms and schools, as well as active in the supervision, evaluation, and mentoring of principals. Superintendents in high-performing districts were also more likely to dismiss principals on the basis of their performance. These districts showed a much greater clarity of purpose, a much greater willingness to exercise tighter controls over decisions about what would be taught and what would be monitored as evidence of performance, and a greater looseness and delegation to the school level of specific decisions about how to carry out an instructional program. Despite strong leadership, these districts were less bureaucratic than their counterparts. (“Leadership” 26)

I am struck by the revolutionary description of this position. We are fortunate to have a superintendent in SUSD that has both teaching experience and an observable interest in the teaching craft. However, I can’t imagine it too blasphemous a statement to say that our district is fundamentally and cosmically distant from the ideals expounded upon by
Elmore. I say this with the utmost respect because I can’t imagine too many districts that are as progressive as that found in Elmore’s description.

However, the template for our future management structure has been designed and provided, along with ample evidence supporting this eventual transformation. It is again, worth repeating what I consider to be good fortune that all aspects of our school system face a fundamental and shared overhauling. Schools are repositories of shared experiences; Parker is no different. I hope that both my school site and my district will adapt along the lines Elmore and the other critics describe to meet not just the changing face our student population, but also the expectations of an era of accountability and meaningful learning.
Conclusion

Testing practices at Parker elementary are voluminous, if not burdensome and out-of-balance. The result of such an overly zealous assessment schedule is that it consumes far too many instructional resources, while effectively teacher-proofing instruction. Conservative calculations place the amount of testing at Parker at approximately 6% of our total number instructional hours annually, triple the amount of another urban school district in Washington D.C. Test Prep pushes that percentage even higher (15%), leaving less time for other curriculum like Science and Social Studies. The narrowing effect is evidenced in curriculum decisions, in teachers’ responses to stress and testing pressures, and is chronicled in anecdotes about activities set aside in pursuit of test scores. The net result is a commoditized educational experience, usurped of authority and initiative from the teachers, and bereft of richness and variety for the students. The business of education is about producing the commodity of proficient test scores, not necessarily proficient students.

Staff at Parker seems acutely aware of some of these factors, and unaware of others. Classroom learning has already begun to incorporate testing data to better address student needs. All the stakeholders at Parker understand the importance of student performance and are galvanized to that end. However, acceptance of a testing protocol to which none of the teachers seem to possess its educational and theoretical underpinnings renders our testing practices invalid. Also accepted without question is the one-size-fits all approach which promises standardized accuracy—a promise that it has failed to deliver. Furthermore, until issues such as testing volume and local input are addressed,
our practices will continue to dwell outside the definition of validity. Vigorous restructuring of management needs to occur immediately at all levels from the teachers to the superintendent, or else we will be unable to keep pace with the public expectations of accountability, much less manage the atmospheric stresses placed upon teachers, students and families as a result of high-stakes testing.
APPENDIX A

Teacher Survey

Commoditizing Education: The Validity of Testing Practices at an Urban Elementary School
By Craig Seale

Thank you for your time in assisting me with this project. The examination of our testing procedures is of primary interest to me as an educator at our school. It is my intention that my questionnaire will be seem consistent with the care and discretion demonstrated within our profession. If the questions seem in any way contrary to a respectful discussion of our testing practices, please feel free to withhold responding. You are under no obligation to participate. I greatly appreciate your help. Thank you.

1. What were some of your intended purposes for last year’s Test-Prep period (for example: to recreate the testing experience; to provide test-wise awareness and skill; or to provide content area questions presumed to be upcoming on the exam)?

2. How well would you rate the district test questions as being an accurate measurement of the content being taught (1-worst to 10-best)? If possible, give a separate rating for Math tests, and ELA tests.

3. On average, how many minutes a week are spent on Science?

4. What kind of interventions and/or individualized instruction is currently occurring at your grade level? Within your class?

5. Are students organized by ability or performance level? If so, describe.

6. What empirical evidence or theoretical rationales (if any) is used either in support for or against our testing practices?
7. What types of inferences and actions (if any) are formed from the data generated from our tests?

8. To what extent do you find the following quotation to be either consistent or inconsistent with your professional experience. Please elaborate:

“Where performance targets are very high relative to current scores and where community supports for achievement are comparatively weak, teachers face a far more difficult task and may be more inclined to cut corners in trying to reach their targets.”
APPENDIX B

Support Staff Survey

*Commoditizing Education: The Validity of Testing Practices at an Urban Elementary School*
By Craig Seale

Thank you for your time in assisting me with this project. The examination of our testing procedures is of primary interest to me as an educator at our school. It is my intention that my questionnaire will be seen as being consistent with the care and discretion demonstrated within our profession. If the questions seem in any way contrary to a respectful discussion of our testing practices, please feel free to withhold responding. You are under no obligation to participate. I greatly appreciate your help. Thank you.

1. What is the status of our RTI plan?

2. Which numbers of students fit into the three tiers?

3. Does a line of communication exist in which site-based feedback factors into decisions about district testing protocol?

4. What empirical evidence (if any) is used either in support for or against our assessment methods? Describe how this evidence is factored into our assessment plan.

5. What theoretical rationales (if any) form the basis of our testing practices? Describe how these rationales are factored into our assessment plan.

6. What types of inferences and actions are formed from the data generated from our tests?
WORKS CITED


FairTest.org. “’No Child Left Behind’ After Six Years: An Escalating Track Record of Failure.” ---. “How Standardized Testing Damages Education.”


*Harmon Johnson School Plan for Student Achievement, 2009-2010.*


