Principles for Accountability Designs

Herbert J. Walberg

This chapter describes and illustrates a dozen design principles for school district, school, staff, and student accountability. Although many policymakers and analysts would agree with the need for accountability, the means are neither obvious nor agreed upon. Moreover, technical problems abound, including the lack of universal achievement scales, the matching of achievement tests to goals and standards, the scaling and expression of test scores, and the causal attribution of success to district central-office staff, principals, teachers, parents, and student socioeconomic background and effort.

Yet perfection is the enemy of steady improvement. States, foundations, and districts have made progress in solving these problems. They provide examples of principles that may reasonably be incorporated into accountability systems. Recent test scores for schools, for example, may be compared with their previous scores or those of comparable schools, or they may be statistically equated for fairer comparisons such as value-added metrics, which take into account previous tests scores, student demographic characteristics, and other factors. They may be reported in ways that are readily comprehensible to parents, the public, and legislators.

Well-defined standards can play a central role in accountability. By using them, school progress can be gauged according to the percentage of students that attain various levels such as the National Assessment Governing Board's Basic, Proficient, and Advanced levels for various subjects and grade levels.

Various difficulties, however, can arise: Standards may be too general or detailed, too difficult or easy, or too many or few. Similarly, tests may be difficult to construct. If employed for high-stakes decisions, a test may serve well for the first administration only but require new tests each year, which would result in difficult calibration with old tests to measure progress. Teachers may teach more exclusively to only that subject matter represented on the tests. Finally, it may take time to design curricular materials, lessons, and tests that best reflect the standards.

Even so, various states, foundations, and districts are solving these and other problems. As they gain more experience, they and others can review their progress, improve their programs, anticipate difficulties, and avoid them. This chapter draws upon the experience of such states and districts and sets forth implications for further improvement.

DESIGN PRINCIPLES

Authorities have written many books on accounting, auditing, various aspects of board-management and management-labor relations, and related topics. Accountability in education is far less mature, agreed-upon, and explicit. Yet some principles can be set forth to guide the development of K–12 school accountability systems that have proven workable in education practice. They are set forth in this section, and real-world examples are given in the next section.

A. General Principles

Dictionaries emphasize accountability as liability for being called into account or answerable for an explanation. Either

meaning implies at least two sets of actors—those being called into account and those doing the calling, for example, management and labor, and parents and children. Schools, however, have a long and complex chain of accountability—citizens who elect their legislators to represent their interests, appointed or elected state school board officials and superintendents, local boards, superintendents, and central office staff, principals, teachers, and students. It might be useful to think of each group as accountable to its predecessor in this list.

This linear accountability, however, is oversimplified, since, for example, federal regulations, high school department heads, other system employees, business influences, and others may require consideration. Superintendents may be accountable to the public as well as to teachers, other professionals to their professional associations, students to their peers as well as to parents and teachers, and so on. Still, the important point is to recognize at least the two actors, one accountable to the other. Given this recognition, what principles make for effective accountability?

1. Independence

In evaluating superintendents, school boards cannot rely completely upon their chief executive officer to provide accurate information. In addition to the superintendent's views and information, they should seek by formal and informal means the input of citizens, parents, teachers, auditors, and other third parties. Tests, community surveys, and public hearings are some formal means to complement official board reports and board members' impressions from informal observations. Similarly, legislatures, state boards, and other groups should acquire or require independent information in addition to that routinely reported by those held accountable.

2. Focus on Results

Groups responsible for accountability routinely possess and discuss information on inputs but often are less well

informed about results. For example, school boards routinely discuss finance, spending, class size, and staffing, among other things, but they appear less knowledgeable about where their students stand with respect to standards and rankings against similar or nearby districts and schools. Even educators themselves often have little technical mastery of psychometrics and statistics that would allow them to critically evaluate their students' progress.

3. User Friendliness

Readily understood reporting is desirable. Perhaps even a single number or two may best serve ocasionally. Many colleges, for example, want only two test scores and an applicant's high school grade-point average in making admission decisions. Stockholders and potential investors may first want to know the profit and increase in earnings, then the basis of the calculations, and then other organized numerical and verbal information.

What isn't as useful is a mass of undigested numbers often reported by states and districts in large, unwieldy books of computer printouts. A better system, exemplified in a subsequent section, allows school board members, educators, parents, and other interested parties, even those without technical experience, to design and execute within a few minutes reports with the comparisons and degree of detail they wish. They may then publish the report, and any comments they wish to make, on the Internet.

4. Timeliness

When I served as chair of the Design and Analysis Committee of the National Assessment Governing Board, it took the test vendor about sixteen months following the test administration to release the results. School boards and teachers often get test results long after their time of prime usefulness, namely, immediately. Large, national business firms usually report quarterly results, but some are capable of aggregating daily sales figures. One mark of a good teacher is getting test

results back the next day. Boards, educators, firms, and others need rapid turnaround to make results useful.

The near future looks bright for timely results. Apparently, more than 100 firms are now working on computerized tests administered on the Internet. Such tests can be scored in several seconds; they save printing and mailing costs, can be quickly updated, and may require as few as a third of the testing time and items as the usual tests because they adapt the difficulty of the items to the students' ability, which is better estimated with each successive item. Open to the public, parents and students, independent of schools, could check their progress on demand in any given subject.

5. Incentives

Simply publishing results appears insufficient for progress. People and groups responsible for accountability should be able to offer incentives and sanctions for performance. Praise and recognition may go a long way, but money talks. The prospect of being hanged in the morning, wrote Samuel Johnson, concentrates the mind. There is much interest in superintendent bonuses for results, "merit pay" for teachers, and even payments to students. Schools have been closed for repeated failure; more students are being held back a grade because they haven't met standards. Schools of choice risk closing if they attract no students. Analogous thinking dominates much of the rest of society. Why not schools?

B. Examination Principles

As pointed out above, outcome information should be central for accountability, and test results provide the best indicators in several respects. Because multiple-choice tests are increasingly used for this purpose, and because many theorists, educators, and even psychometrists have criticized them, this section offers reasons why they should prevail.

1. Objectivity

Subjective impressions and reports about student work may be valuable, but they cannot be substituted for objective information, particularly quantified information. The most efficient, and perhaps most objective, indicators of outcomes are results on multiple-choice examinations because they are relatively cheap and require little subjectivity. Though useful for teachers in evaluating students' classroom work, essay examinations, laboratory exercises, oral reports, and similar "authentic examinations" are often highly subjective and lack technical adequacy, and they usually add very little information to what can be quickly assessed with objective procedures.[1]

2. Fairness

Objectively scored, often machine-scored, multiple-choice tests can be the fairest of all examinations in several senses. Teachers may be biased for or against some of their students. They may favor their own students or those of fellow teachers when results have high stakes. For this reason, other countries remove identification information from high-stakes examinations and employ teachers other than the students' own to make examinations fair.[2]

Multiple-choice tests are fair and defensible in another sense. In a small amount of time, they can sample a variety of parts of the subject matter and a range of "cognitive processes" from factual knowledge to "higher-order skills," such as analysis and synthesis. In contrast, a single essay question given in the same amount of time may arbitrarily give a huge advantage or disadvantage to a student, depending on whether or not an individual student had concentrated

[1]Herbert J.Walberg, Geneva D. Haertel, and Suzanne Gerlach-Downie, *Assessment Reform: Challenges and Opportunities,* Bloomington, IN: Phi Delta Kappa, 1994.

[2]John H. Bishop, "The Impact of Curriculum-Based External Examinations on School Priorities and Student Learning," *International Journal of Educational Research,* vol. 23, no. 8, 1996: 653–752.

study on the particular subtopic. A single essay question, moreover, can easily "leak out," and students who happen to find out may have an unfair advantage. Good writers who actually haven't mastered the subject matter can overly impress some graders. Although essay examinations, laboratory exercises, and problem solving should have a major place in the classroom, they often entail special difficulties in large-scale accountability systems.

3. Value-Added

Students who achieve well one year are likely to do well the next (and the likelihood increases with age). The same is true of schools, districts, and states. Further, how well they do is substantially determined by socioeconomic status and related factors, but schooling is another cause. To indicate the school's contribution to achievement, we can calculate value-added scores by subtracting the percentage of students attaining a given score or standard this year from last year's percentage.[3] Policymakers increasingly recognize that value-added scores better indicate the school's or teacher's contribution to achievement than do test scores at a single point in time. The apparent success of suburban schools, for example, may be substantially attributable to their socioeconomic composition rather than their efficiency.

Unadjusted, non-value-added scores, however, can complement value-added scores, and together they give policymakers more information and are less misleading than either one alone. For some purposes, moreover, status scores can serve alone. If schools are similar in socioeconomic and other

[3]Other ways to make value-added calculations are more complex, such as achievement residuals from regressions on previous scores, socioeconomic status indicators, and other measures, possibly for separate groups. In employing regression, mixed models, and their variants, we give up transparency or comprehensibility, especially to citizens who pay for schools. We also put ourselves in the hands of statistical experts who may agree that everything employed or proposed is defective but can agree on little else. Is it worth it?

advantages, they may be more validly compared on status. Nearby districts and schools may be of great interest to boards and parents, which may increase interest and justification of status scores. Some theorists and educators say that all children can learn to the same degree. Though value-added scores may best measure the progress of children or schools most in need of catching up, their ultimate interest should be status scores.

4. Balance

Balanced accountability systems require tests of multiple subjects, including science, history, geography, civics, and other subjects rather than the usual mathematics and reading tests alone, even though we may consider reading and mathematics to be foundational and pervasive. Boards need to consider ways to weigh or otherwise combine scores for an overall accountability index as well as providing desired detail.

Though multiple-choice tests are exceedingly efficient and cheap compared to other parts of educational programs, educators and students may face a considerable amount of testing, including the National Assessment of Educational Progress tests, national commercial tests, and state tests as well as special district, school, and frequent classroom tests. Some duplication of subject matter assessment may be desirable, but responsible boards and educators need to think through the entire testing program.

5. Score Expression

Scores on tests may be reported in a variety of ways such as percentiles and normal-curve equivalents. Because they are readily understood, methods of estimating the percentages of students attaining a given judgmental standard (or a national percentile such as the fiftieth or ninetieth) are gaining considerable ground. Still, such simple, concise indexes may cause distortions in educational programs. Strong incentives for getting the maximum number of students past, say, the

fiftieth percentile may cause educators to neglect students who can easily pass this threshold and those who have little chance to make the cut. For this reason, the average of all students may be a better single representation of either status or progress.

It may be useful also to examine the percentages of students attaining quartiles or judgmental standards such as the National Assessment Governing Board's Basic, Proficient, and Advanced levels. Such detailed reporting allows a better understanding of where progress is and is not being made.

6. Disaggregation

The National Assessment of Educational Progress (NAEP) and some states and districts report the scores of boys and girls; African American, Hispanic, and white students; and poverty and nonpoverty students separately, thereby allowing a detailed review of each group's progress. The differences among these groups have also been reported as "the race gap" and "the poverty gap." Like quartile and multiple-standards reporting, such indexes allow close analysis of status and progress. In cases of sampling such as NAEP, however, small sample sizes may result in inaccurate estimates of the subpopulation figures.

7. Supplementary Opinion Surveys

Because examinations cannot capture all outcomes of schooling, supplementary information is useful. In principle, elected school board members represent the interests or views of their constituents, which they glean from daily life in the community and in special hearings. But surveys provide systematic evidence about changing views in their communities. Freely given information on possibly discrepant views of staff, parents, and students may give them a better understanding of the schools' problems and possible solutions. Public Agenda and Business Roundtable surveys, for example, show that the public, parents, teachers, and

students support accountability and agree on the need for more rigorous standards.[4]

PRINCIPLES IN PRACTICE

Given the present state of accountability, the foregoing principles are somewhat idealistic. Yet we can consider successful instances of each in a limited number of districts and states, most of which I know from personal experience. This section discusses exemplary applications of the principles and illustrates them with actual accountability evidence.

A. Consumer-Designed Reports

Available for California and Texas thus far, K–12Reports[5] is an Internet program that allows school board members, educators, parents, and citizens to analyze their school's, district's, or county's achievement standings. Without acquiring massive state databases and without spreadsheet skills, they can report scores in a variety of ways and publish comments on their findings. This section illustrates the displays an interested user can generate after investing several minutes in learning how to specify analyses.

1. California County, District, and School Ranks

Table 1 shows the opening screen for California. The counties are sorted from highest to lowest according to reading scores. Marin County is highest, and Merced County is lowest. The weighted average for the state is lower than the median district due to the fact that Los Angeles County has many low-scoring students and the higher scoring districts tend to have smaller numbers of students. San Diego, Sacramento, and San Francisco counties score substantially higher than Los Angeles County.

---

[4]See the Internet sites <www.brt.org> and <www.publicagenda.org>.

[5]See Internet site <www.K-12Reports.com> for additional reports or to specify custom displays and to publish reports.

TABLE 1
Percentages of Students in All Grades Scoring At or Above the Fiftieth National Percentile on the Stanford Achievement Test, Spring 2000

County	*Reading*	*Math*	*Language*	*Spelling*
Marin	**73.7**	**76.9**	**77.3**	**65.9**
Nevada	65.2	70.2	66.3	54.3
Placer	64.6	68.2	68.0	60.5
El Dorado	62.7	66.7	65.9	54.5
Amador	60.9	61.2	63.9	53.0
San Luis Obispo	60.4	66.6	65.3	53.7
Tuolumne	58.8	61.9	60.3	49.9
Sonoma	57.7	60.7	62.0	50.9
Mariposa	57.5	63.3	61.0	47.9
Calaveras	56.6	56.6	57.6	46.3
Humboldt	56.5	59.5	58.6	48.3
Trinity	56.5	61.8	56.4	46.7
Plumas	56.0	56.9	57.6	50.8
Alpine	55.2	61.2	55.9	45.6
Sierra	55.0	59.5	58.7	50.5
Contra Costa	54.9	59.8	59.7	55.6
Santa Clara	54.0	63.1	61.6	56.5
Mono	53.9	56.5	59.2	45.4
San Mateo	53.8	60.2	61.0	56.5
Siskiyou	52.8	58.6	53.8	43.0
Ventura	52.4	58.4	59.0	51.6
Lassen	51.6	56.4	52.3	43.9
Inyo	50.5	55.2	52.9	43.0
Yolo	50.4	56.5	54.6	46.9
San Diego	**50.3**	**59.4**	**56.6**	**50.8**
Shasta	**50.3**	**56.2**	**53.0**	**46.7**
Napa	50.3	57.8	55.7	43.4
Orange	49.5	60.2	57.8	52.1
Alameda	48.9	55.7	55.7	50.6

continued on next page

TABLE 1 (*continued*)

County	*Reading*	*Math*	*Language*	*Spelling*
Solano	48.2	53.8	53.8	50.3
Butte	47.7	52.0	51.9	40.6
Santa Barbara	47.6	55.0	53.2	45.8
Sacramento	**47.4**	**53.5**	**52.9**	**50.6**
San Francisco	**46.9**	**60.1**	**56.4**	**54.6**
Santa Cruz	46.6	53.2	50.1	39.1
Tehama	46.5	53.9	48.2	41.2
Modoc	46.5	55.2	49.8	42.2
Mendocino	**45.9**	**49.6**	**49.3**	**37.8**
Del Norte	44.2	53.4	48.1	41.0
Stanislaus	43.4	52.1	49.0	42.1
Lake	**43.2**	**48.2**	**46.7**	**36.7**
California	**43.1**	**51.2**	**49.9**	**44.7**
San Benito	**41.9**	**48.9**	**45.5**	**38.3**
Sutter	41.2	49.6	47.1	39.7
Glenn	40.6	47.3	46.0	41.6
Riverside	39.5	48.0	46.4	40.1
Yuba	39.1	44.2	43.5	39.3
San Bernardino	37.4	45.4	45.0	39.5
San Joaquin	36.8	45.6	45.0	40.1
Fresno	36.5	45.9	43.0	39.4
Kern	36.2	45.0	42.8	39.8
Madera	35.7	45.6	41.4	36.6
Kings	35.0	39.3	40.6	37.7
Los Angeles	**34.4**	**43.3**	**43.1**	**39.1**
Monterey	34.0	41.2	40.1	33.4
Tulare	30.9	40.4	37.6	31.0
Imperial	29.5	39.8	38.9	36.5
Colusa	29.3	40.6	35.7	30.5
Merced	**29.0**	**39.6**	**37.5**	**31.1**

Note: Counties referred to in the text are bolded.

On the Internet version of the table, clicking on the italicized words at the top of the table (URLs, or "universal resource locators," in Internet jargon) sorts in several seconds the counties by another subject. Clicking on any county in the left column displays the similarly ranked district scores within the county. Clicking then on any district URL displays the ranked schools within the district. With an additional click, county, district, and schools can be ranked by means or twenty-fifth, fiftieth, and seventy-fifth quartiles on any subject for all grades together or for separate grades.

The schools can also readily be ranked by an index of value added, that is, in this case, the percentage difference between any grade and the previous grade. From the second to the third grade, for example, only four of the fifty-eight counties in the table, San Benito, Mendocino, Shasta, and Lakes, showed gains in the percentages of students at or above the fiftieth percentile. In the state as a whole, 5 percent fewer third than second graders met this national criterion. For the 4,078,575 students with reading data in the state, the average drop in percentage attaining the fiftieth national percentile is 1.4 percent per grade level, suggesting that the longer students are in California schools, the worse their national rank.

2. *California Poverty-Gap Analysis*

In addition to analyzing the scores of all students, K–12Reports can rank the scores of the following groups within the state, counties, and districts:

- Limited English Proficient, non-LEP, and the difference between them
- Boys, girls, and the difference between them
- Special and nonspecial education students and the difference between them
- Economically advantaged, disadvantaged students, and the difference between them

As an example, Table 2 illustrates the differences among California counties with respect to the poverty gap (or difference between Title 1 and other students). Illustrating the pervasive effects of poverty, every district shows a gap. But the range of differences is huge: In contrast to San Francisco County, San Diego and Sacramento counties have huge poverty gaps. Humboldt, Nevada, Del Norte, and Sierra counties have the smallest gaps, and Marin County has the biggest poverty gap. The service enables users to "drill down" to examine which districts, schools, and grades do well not only on average but also with respect to gaps and reducing gaps among groups from one grade to the next.

Educators and others can also post comments and questions about the tables. They may, for example, venture hypotheses about the results, provoke further analyses, and suggest constructive actions based on evidence selected and analyzed in accordance with their concerns. They can engage in dialogues with others who can readily publish other analyses and commentary on the publicly available Internet.

Thus, K–12Reports illustrates several of the design principles discussed in the previous section, including user friendliness, independent analysis, focus on results, timeliness, value added, comprehensiveness, score expression, and disaggregation. It might be argued also that such readily executable and publicly publishable reports will lead to greater use of incentives for superintendents, principals, and other educators because they provide a better basis of evidence than do underanalyzed data, which are less accessible and analyzable by the public, parents, and legislators.

B. Exemplary District Accountability

Most districts around the country employ a variety of national, state, and district tests. Even so, they rarely analyze tests in ways that are optimally suited for accountability. Though typical in this respect, tiny Butler District 53 in Oak Brook, Illinois, a western suburb of Chicago, is distinctive in

TABLE 2

Gap in Percentages between Advantaged and Disadvantaged Eighth Grade Students At or Above the Fiftieth National Percentile on the Stanford Achievement Test, Spring 2000

County	*Reading*	*Math*	*Language*	*Spelling*
Marin	**54**	**48**	**51**	**43**
Santa Cruz	47	37	41	31
Contra Costa	43	42	40	30
Fresno	42	35	37	27
San Mateo	42	32	35	29
Madera	41	38	37	23
Orange	41	37	36	31
Ventura	40	37	37	32
Santa Barbara	40	33	35	30
Tulare	38	29	34	25
Napa	37	31	38	29
Kern	37	30	33	24
San Diego	**37**	**35**	**35**	**28**
Sacramento	**36**	**29**	**34**	**24**
San Benito	36	27	36	22
Sonoma	36	27	36	25
Monterey	36	31	31	28
California	**36**	**32**	**32**	**27**
Santa Clara	36	31	33	29
Colusa	36	31	39	33
Alameda	34	28	31	23
Imperial	33	28	32	28
Sutter	33	30	31	18
Los Angeles	**33**	**30**	**30**	**27**
Glenn	33	19	24	19
San Luis Obispo	33	31	32	27
Yolo	33	30	30	24
Modoc	32	35	32	23

continued on next page

TABLE 2 (*continued*)

County	*Reading*	*Math*	*Language*	*Spelling*
Butte	32	28	29	18
Inyo	32	33	26	19
San Joaquin	31	25	28	18
Amador	31	33	16	17
Merced	31	21	24	17
San Bernardino	30	27	27	20
Stanislaus	29	20	24	19
Lake	28	28	25	12
Riverside	28	26	26	20
Kings	26	19	25	18
Lassen	26	30	20	18
Mariposa	26	16	16	16
Mendocino	25	25	25	18
Solano	25	20	25	19
Siskiyou	25	14	17	8
Calaveras	24	17	22	16
Tehama	24	20	19	8
Yuba	23	16	19	11
Tuolumne	23	18	15	20
Placer	23	21	20	18
Plumas	23	21	19	7
El Dorado	23	25	23	20
Shasta	23	26	23	18
Mono	23	32	23	15
Trinity	21	21	26	14
San Francisco	**19**	**7**	**16**	**15**
Humboldt	**18**	**14**	**17**	**8**
Nevada	**14**	**20**	**18**	**16**
Del Norte	**10**	**5**	**8**	**10**
Sierra	**1**	**–1**	**12**	**27**

Note: Counties referred to in the text are boldface.

being one of the highest spending districts in Illinois and one of the most affluent areas in the country.

By several measures, Butler students appeared to rank far less well than might be expected from spending and economic status. The board wanted a long-term accountability design, an initial assessment of all available data, and a plan to compensate the superintendent for future accomplishments. The board appointed two of its members, the administrative staff, and a consultant to carry out these tasks. The initial report further exemplifies several of the design principles described in the previous section and illustrated below.

1. National Standings

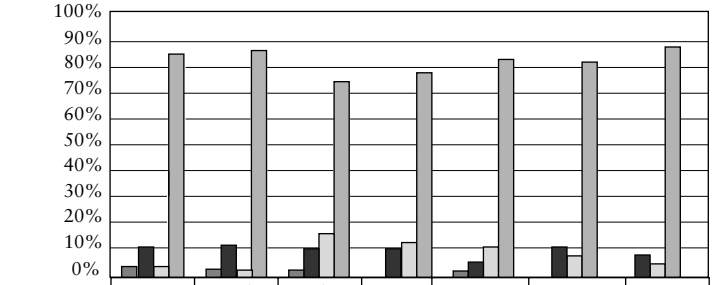
The EXPLORE academic assessment yields achievement information to educators, parents, and students for high school and career planning. The following chart shows how Butler District eighth graders compared with national norms in seven subjects. As expected, the students were concentrated in the first quartile and very few scored below the fiftieth percentile.

The chart also illustrates a common pattern among test data for students, schools, and districts: Contrary to the common assumption, those who do well on one test usually do well on others.

2. Value-Added Analysis

The California statewide, value-added analysis described in the previous section employed "synthetic cohort . . . raw gains," that is, the simple differences in means at a single point in time between one grade and the previous grade. Other things being equal, more complex analyses can be advantageous, especially in tracing individual students over one or more years.[6] Employed in the Butler District is SAS in School, led by William Saunders, who first gained national prominence for such analyses in Tennessee. Data were available for

[6]Experts lack agreement on the best way to calculate these, and the basis and method of some calculations are not completely explicit. Such scores lack transparency and comprehensibility for all but a few people with specialized technical skills.



	English	Math	Reading	Science	Composite	Mechanics	Rhetoric
01–24%	3%	2%	2%	0%	2%	0%	0%
25–49%	10%	11%	10%	10%	5%	11%	8%
50–74%	3%	2%	16%	13%	11%	8%	5%
75–99%	84%	85%	73%	77%	82%	81%	87%

Butler District Assessment

the last three years in nine subjects from California Test Bureau/McGraw-Hill's California Achievement Test data. Grades 2, 4, and 8 could not be computed because test score gains were not available for the three most recent years.

The next table provides a compact summary of the SAS in School results by subject for the years 1997 through 2000. As on traffic lights, a green or favorable light is represented by G. Similarly, yellow (Y) suggests caution and red (R) suggests stopping for a closer examination. Ultra-Red, or R*, suggests value-added gains considerably below those of other schools in the United States.

Because affluent districts score better than state and national averages (often attributable to socioeconomic factors), few or perhaps none have undergone rigorous value-added analysis. For the Butler District, the grade 3 results are mixed, grade 5 made progress in five of ten instances, grade 6 in eight of twelve instances (though its performance was distinctly below the national average in two cases (R*)), and grade 7 made good progress in nine of twelve instances.

In sum, Butler District 53 students made better than national progress in twenty-two, or 61 percent, of thirty-six

Subject	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Language Total				G	G	G	
Math Total				G	G	G	
Reading Total				Y	G	G	
Science		G			G	G	
Social Studies		R			R*	G	
Language Expression				Y	G	G	
Language Mechanics				G	G	G	
Math Computation				Y	G	Y	
Math Concepts and Applications				G	R	Y	
Reading Comprehension				R	G	G	
Reading Vocabulary				Y	Y	G	
Spelling				G	R*	Y	

SAS in Schools

instances during the most recent three-year period. Caution is indicated in 22 percent of the instances. A closer examination is suggested in 8 percent of the cases, and performance is considerably below that of other schools in 6 percent of the instances. This analysis suggests more urgent priorities than the status results almost universally reported and exemplified in the previous chart.

3. Student Progress by Initial Test Score

The next table summarizes progress in grades 3 through 8 for students in five national quintiles during the previous year. Though some statisticians may prefer a three-year assessment of schools, grades, or teachers, their preference or insistence trades immediacy for accuracy. A teacher may, for example, be rapidly declining from excellent to poor to unacceptable, but a rolling three-year average may not signal a sharp warning until the end of the fourth or fifth year, in which case children would have suffered a severe multiple-year setback. So the three-year preference is merely a trade-off to be made by responsible agents such as conscientious board members

Subject	*1 Lowest*	*2*	*3 Middle*	*4*	*5 Highest*
Language Total	++++	+++	++++	++++	+++++
Mathematics Total	+++++	+++	++++	+++	++++
Reading Total	+++	++++	++++	++++	+++
Science	++++	++++	++++	++++	++++
Social Studies	++++	+++	+	++	++
Language Expression	+++++	+++++	+++++	+++	++++
Language Mechanics	++++	+++	++	++++	++++
Math Computation	+	++	+++	+++	+++
Math Concepts and Applications	+++++	++++	++++	++++	+++++
Reading Comprehension	+	++++	+++	+++	++++
Reading Vocabulary	+++++	++++	++++	++	+++
Spelling	+++	+	++	+++	+++

Progress

rather than on moral or "high science" grounds. In addition, teachers who excel in the past year should hear about it and be rewarded quickly, even if the value-added gain measure is somewhat less precise than a three-year or career average.

In any case, each plus in the table indicates that one grade made better than average national progress. The fifth, or highest, quintile under Language Total, for example, has five pluses, which indicates that students in this quintile made better-than-average progress in five of the six grades.

The chart shows that by this criterion Butler students performed well: In more than two-thirds of the cases (207 out of 300), their progress exceeded the nation's. Progress was especially consistent in Science, Language Expression, and Math Concepts and Applications, with twenty or more pluses across quintiles. Except perhaps for the lowest quintile in Math Computation and Reading Comprehension, there was little tendency for top, bottom, or middle quintile students to make greater progress than other students in the district. Still, progress was only average or less than the na-

tion's in Mathematics Computation and Spelling in half or less than half the grades. A board member might ask why there wasn't above-average performance in every subject in every grade. She might also ask for the specific results for each teacher.

4. Grade-Level Analysis

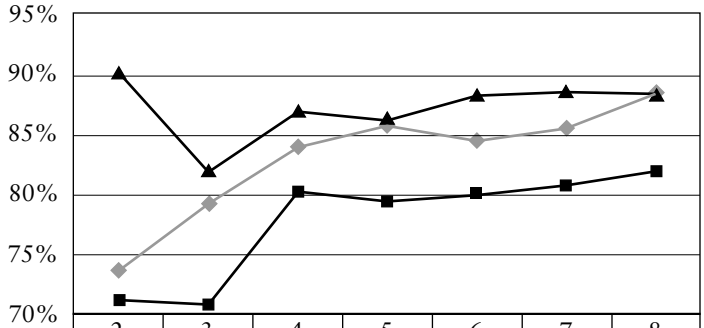
The data discussed in a previous section show that the longer California students are in school, the more likely they are to be below the national median. The percentage above the median in reading, for example, declined from 49 to 34; Language Arts declined from 52 to 40.

On the other hand, the Butler District 53 board had instituted a rigorous test published by the Educational Records Bureau (ERB), which is more often used in private and elite suburban schools, and wanted a similar indication of students' relative progress through the grades even though only one round of test data was available.

The next chart shows the median ERB percentile composite scores (Word Analysis, Reading Comprehension, Mathematics, Writing Mechanics, Verbal Ability, Vocabulary, and Quantitative Ability) for District 53, suburban schools, and independent schools. This chart shows that District 53 and the national sample of suburban schools can be found around the seventy-second percentile; they both generally pull further ahead of national norms with each higher grade. District 53, however, moved ahead faster and actually caught up with independent schools by eighth grade.

5. Peer District Comparison

Given Butler District 53's affluence and school spending, some parents and board members would not be content with exceeding other national samples of suburban schools and catching up with independent schools. Close to Oak Brook is Naperville, with claims for the best science and mathematics scores among elite suburban schools in the nation



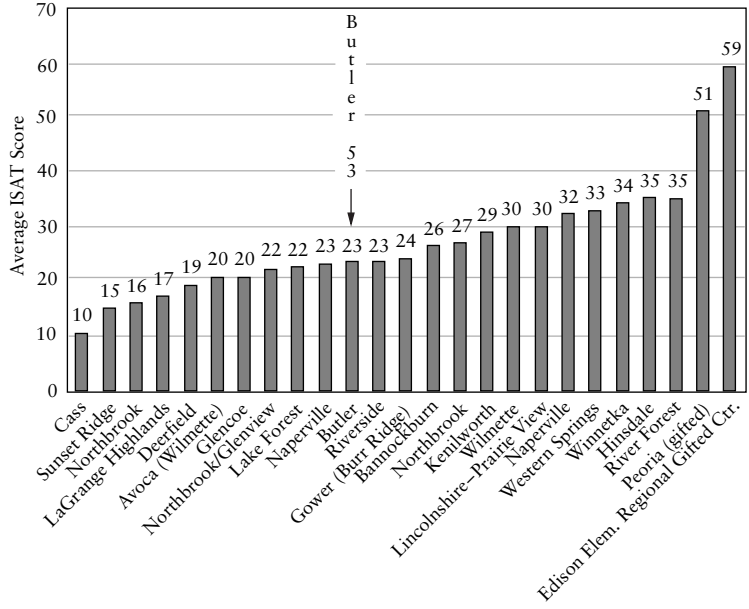
	2	3	4	5	6	7	8
District 53	74%	79%	84%	86%	85%	86%	89%
Suburban	71%	71%	80%	80%	80%	81%	82%
Independent	90%	82%	87%	86%	88%	89%	89%

Median Percentile Scores by Grade

when compared on the Third International Mathematics and Science Study examinations. The Illinois Standards Achievement Tests (ISAT) afforded a comparison of elite Chicago area suburban districts. Because Butler District 53 has only a primary and a middle school, the best-scoring school in each elite district was sampled using composite ISAT scores for reading, math, science, social science, and writing for such schools.

The next chart shows Butler 53 placed fifteenth out of twenty-five elite schools. The top-scoring Edison Elementary and Peoria are schools for gifted children. If these two are excluded, District 53 outperforms almost half the schools in this elite pool.

The foregoing analyses exemplify several design principles described in a preceding section, including value added, comprehensiveness, score expression, and disaggregation. They show that the usual assortment of various test data in school district files can be marshaled to give a fuller picture of the district's accomplishments and needs for improvement. From these and other analyses, the Butler District 53 board members felt confident in setting specific merit-pay goals for the superintendent.



Year 2000 Average ISAT Performance by District

C. Student Accountability and Incentives

Just as much as educators, students need better accountability and more explicit incentives, as they themselves agree. A 1996 Public Agenda national survey of high school students showed that three-fourths believe that stiffer examinations and graduation requirements would make students pay more attention to their studies. Three-fourths also said students who have not mastered English should not graduate, and a similar percentage said schools should promote only students who master the material. Almost two-thirds reported they could do much better in school if they tried. Nearly 80 percent said students would learn more if schools made sure students were on time and did their homework. More than 70 percent said schools should require after-school classes for those earning Ds and Fs.[7]

[7]Herbert J. Walberg, "Incentivized School Standards Work," *Education Week*, November 4, 1998, p. 37.

1. Learning Consequences

Experimental research in the classrooms corroborates students' common sense and insight. Among dozens of teaching methods subject to meta-analysis (statistical syntheses of studies), frequent testing substantially benefits learning because it encourages students to be prepared and provides information on their progress to both students and teachers. Positive teacher feedback about students' good accomplishments is among the most powerful teaching methods.[8]

Surveys also provide support for better student accountability. In his classic 1961 study, *The Adolescent Society,* sociologist James Coleman showed how teenage concerns with cars, clothes, and dating precluded long, hard study.[9] Since then, television has taken a larger share of students' time and consumes nearly as many weekly hours as they spend in classes.

Economist John Bishop has long studied examination effects on learning. From large-scale survey data, he analyzed the effects of examinations of the (U.S.) Advanced Placement program, the New York State Regents, Canadian provinces, and European ministries. These examinations have the common elements of being externally composed and geared toward agreed-upon subject matter that students are to learn within a country, state, or province. They are often given at the end of relevant courses. Most important, they have substantial positive effects on learning.[10]

The largest, and most sophisticated, international comparative analysis of national achievement yet conducted corroborates Bishop's findings about constructive effects of

---

[8]Herbert J. Walberg, "Meta-Analytic Effects for Policy," in Gregory J. Cizek, editor, *Handbook of Educational Policy*, San Diego, CA: Academic Press, 1999, pp. 419–454.

[9]James S. Coleman, *The Adolescent Society,* New York: Free Press, 1961.

[10]John H. Bishop, "The Impact of Curriculum-Based External Examinations on School Priorities and Student Learning," *International Journal of Educational Research* 23 (8) (1996): 653–752.

external curriculum-based examinations.[11] America's lack of such exams, the short school year, and limited homework requirements are three of the major reasons why U.S. students come in last in value-added achievement gains.

On-the-ground anecdotal reports from experienced observers bear out the harm of slack accountability and even complicity. The title of Sizer's book *Horace's Compromise* conveys the too-frequently-implicit contract: teachers give good grades without academically challenging their students; students, in return, don't cause difficulty.[12]

Most economists and some psychologists subscribe to the idea that people rationally choose what they perceive maximizes their benefits and minimizes their costs and risks. Parents would like their children to work hard for future gains, but many adults are held accountable in their own lives, yearly, quarterly, and sometimes much more frequently as the appropriate occasions arise. Yet educators and parents seem to expect that their charges will work hard without feedback for vague, very long-term goals such as gaining entrance to a good college and enhancing their marital and career prospects. To children and teenagers such benefits may appear intangible, uncertain, and in the far distant future. High achievement, moreover, requires time and energies that could go into the pursuit of other fascinating opportunities offered by their peer culture.

What would happen if they were challenged and incentivized? The O'Donnell Foundation of Dallas tried it out and asked me to study the program. The Foundation paid students $100 for each passing score on the Advanced Placement (AP) examinations in English, calculus, statistics, computer science, biology, chemistry, and physics, plus a reimbursement for the cost of taking the exam. The

[11]Ludger Woessmann, "Why Students in Some Countries Do Better," *Education Matters,* Summer 2001: 67–74.

[12]Theodore R. Sizer, *Horace's Compromise: The Dilemmas of the American High School,* Boston: Houghton-Mifflin, 1984.

program also provided a $2,500 stipend to each teacher undergoing training to teach advanced courses in those subjects. They also received $100 for each passing AP examination score of their students.

In the nine participating Dallas schools, sharply increasing numbers of boys and girls of all major ethnic groups took and passed the AP exams. The number rose more than twelve-fold, from 41 the year before the program began to 521 when it ended in 1994–95. After its termination, the program continued to have carry-over effects: In the 1996–97 school year, two years after the program ended, 442 students passed, about eleven times more than the number in the year before the program began. Despite education theory, incentives appear to work in schools as they do in other human activities.[13] To work, however, rigorous, clear standards and significant benefits are required. Otherwise, as some economists maintain, students would be irrational. If we think they are, we may not realize their perceived benefits and costs.

2. Student Accountability Benefits

In short, there is much consistent evidence that accountability and incentives work to improve achievement. Are there other benefits?

1. Higher achievement in high school also increases the probability of admission to college. During the past fifteen years, the payoff for college attendance has more than doubled. Higher achievers are also more often admitted into potentially lucrative majors such as engineering and premedicine. Higher achieving high school students tend not only to be admitted but to graduate from better colleges and to enter graduate and professional programs.

2. As measured on objective examinations, achievement in rigorous high school courses tends to be rewarded as better

[13]See Walberg, 1998, in the work cited previously.

pay for graduates. As Bishop points out, the premium employers pay for graduates with higher mathematics achievement has increased substantially. Front-line workers are increasingly assuming responsibility for functions formerly carried out by engineers and managers.

3. Higher achievement also has broader spillover effects. Parents and communities may derive honor and prestige from high-achieving youth. High achievers raise national income and contribute more to their local economies. They pay more taxes and, as informed voters and citizens, may raise the quality of civic and community life.

4. Achievement information yielded by better accountability systems would be valuable to employers to make better hiring decisions. To the extent that employers pay higher achievers more, they make their workforce more efficient and increase student incentives to do better. Relying on such information would help eliminate subjective racial, sexual, and other bias and the inconsistencies of interviews.[14]

5. Fostered in school, reading proficiency is also of huge economic and social significance. Bormuth's careful survey of about five thousand people aged sixteen and over showed that 87 percent of those employed reported that they had to read as part of their jobs. Typical working people read for 141 minutes per day as part of their jobs, or about 29 percent of the workday. Because the national wage bill in 1971 was $859 billion, Bormuth estimated that U.S. workers earned $253 billion for on-the-job reading. Because there are more workers today, because they undoubtedly read even more, and because their hourly wages have increased, the amount paid for on-the-job reading must be substantially greater today.[15] Arguably, U.S. citizens are paid more for reading than any other activity.

---

[14]See the work cited by Bishop for evidence and further arguments supporting the first four points.

[15]John R. Bormuth, "Value and Volume of Literacy," *Visible Language* 12 (1978): 118–161.

6. Accurate information on applicants would allow colleges to provide merit scholarships and allow advanced students to graduate early. In the 1950s, President Robert M. Hutchins of the University of Chicago designed a program to provide early admission to qualified high school students that allowed them to graduate as young as age eighteen. Many went on for graduate and professional degrees.

The results of less substantial, but carefully evaluated, recent programs show that qualified students allowed to skip to advanced courses learned far more than others who were similarly qualified. Enacted again, "Hutchins degree" programs would save students' time and allow them longer careers. Families and taxpayers would save money.

Grades, however, cannot provide the accurate, objective information required for all these purposes. Teachers vary enormously in what content they teach, the rigor of their examinations, and in their grading policies. About 80 percent of the questions on high school teachers' tests concern factual information rather than analysis, synthesis, and evaluation of ideas. Student ranks in their classes are no better because they are based on averages of grades.

Some high school students can pass examinations for advanced college work in ancient history, calculus, physics, and Japanese. Some college seniors cannot pass freshman high school examinations. American education lacks objective standards. Diplomas and degrees are awarded not for proficiency but for seat time.

Japan and most advanced Western countries employ examinations that overcome these comparability problems. Though there are variations in their design, the examinations are composed for courses in the arts, languages, and sciences offered in an entire nation, province, or state. Though the scope of each examination is well known, they are often graded or checked by educators other than the students' own teachers.

Because the exams and courses are uniform, teachers need not figure out what content to teach and subsequent teachers can depend on what students have been taught. It is useless for students to contest their teachers about standards because they are externally imposed. Rather, students and teachers become coworkers in trying to meet the standards.

CONCLUSION

Accountability works in schooling as it does in other constructive activities. Experience with accountability systems suggests a dozen accountability principles, including a focus on results, user friendliness, independent assessments, timeliness, and value-added indexes. As illustrated in this chapter, examples of their success can be found in states, districts, schools, and classrooms.

The present danger is letting the perfect defeat the better. The schooling establishment and its status quo defenders resist examinations, accountability, and standards because they claim they haven't been tried. As shown here, they have in fact been tried and found successful in this country and overseas. They are pervasive not only in sports and other leisure pursuits but in occupations and professions as well. The big accountability exception is American schooling, which may account for its poor and declining productivity and students' poor preparation for college, work, and citizenship.