DOCUMENT RESUME

ED 475 028 HE 035 795

AUTHOR Jordan, Larry

TITLE Accountability Indicators from the Viewpoint of Statistical Method.

PUB DATE 2002-00-00

NOTE 41p.; Paper presented at the Annual Forum for the Association for Institutional Research (42nd, Toronto, Ontario, Canada, June 2-5, 2002).

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Accountability; *Educational Indicators; Graduation Rate; *Higher Education; School Statistics; *Statistical Analysis; Student Surveys; Transfer Students

ABSTRACT

Few people seriously regard students as "products" coming off an educational assembly line, but notions about accountability and quality improvement in higher education are pervaded by manufacturing ideas and metaphors. Because numerical indicators of quality are inevitably expressed by trend lines or statistical control chars of some kind, they are governed by statistical principles of quality control. The principles are fairly simple, but campus groups convened to establish numerical accountability goals and objectives are often unaware of them. This paper provides some examples of accountability malpractice and some guidelines for expressing accountability indicators better. The indicators discussed in detail are: (1) 1-year continuation rates; (2) 6-year graduation rates; (3) percentage of students expressing satisfaction with aspects of their education on a survey; (4) upper division units taken by students who entered the institution as community college transfers, compared to those taken by students who entered the institution as freshmen; and (5) percent of freshmen proficient in mathematics. (Contains 3 figures and 15 references.) (Author/SLD)

Accountability Indicators from the Viewpoint of Statistical Method

Larry Jordan
California State University, Los Angeles

ABSTRACT

Accountability Indicators from the Viewpoint of Statistical Method

Few people seriously regard students as "products" coming off an educational assembly line, one hopes, but notions about accountability and quality improvement in higher education are pervaded by manufacturing ideas and metaphors. Because numerical indicators of quality are inevitably expressed by trend lines or statistical control charts of some kind, they are governed by statistical principles of quality control (Shewhart, 1939). The principles are fairly simple, but campus groups convened to establish numerical accountability goals and objectives are often unaware of them. This paper provides some examples of accountability malpractice as well as guidelines for expressing accountability indicators better.

Accountability Indicators from the Viewpoint of Statistical Method

This paper discusses some of the statistical principles that underlie accountability assessment and the expression of other performance goals that rely upon statistical indicators. The basic theory on how to monitor statistical indicators can be found in the literature on quality control. This paper's title is an homage to Walter Shewhart's (1939) classic monograph, *Statistical Method from the Viewpoint of Quality Control*, which sets forth the basic principles for the design and interpretation of statistical quality control indicators. W. Edwards Deming, at whose invitation Shewhart presented the lectures preserved in his monograph, has provided popular treatments of quality control techniques in several books (1986, 1994) and has shown how these techniques can be used by management to help improve the quality of products.

The language and the best examples of quality control tend to arise from manufacturing applications, where there are usually objective criteria for product quality. Machine parts have to be manufactured within a certain tolerance in order to work; errors of manufacture are bad and we want to keep them low (at reasonable cost); sales are good and we want to keep them high (within the capability of a plant to produce and markets to absorb); and so on.

Few people seriously regard students as "products" coming off an educational assembly line, one hopes, but notions about accountability and quality improvement in higher education are pervaded by manufacturing ideas and metaphors. Whether students are products or not, however, numerical indicators of quality are inevitably expressed by trend lines and statistical control charts of one sort or another, and these are governed by statistical principles of quality control. These principles are fairly simple, but there seems to be little awareness of them in some of the campus groups that have been convened to establish numerical accountability goals and objectives. These groups are often pragmatic in the sense that somebody "wants numerical objectives and so let's give him numbers and get on with our real work". They do not want to get bogged down for hours with minutiae. Thus, in one campus's implementation of accountability objectives required by the California State University (CSU) system office,[1] one finds campus committees doing things such as:

- establishing an objective to increase by 0.5% in two years and 1% in four years a rate in a time series where the size of the 95% confidence interval is about ±2% at each time point; or
- establishing an objective for a rate (percent of students who graduate in six years while taking courses at a pace that does not allow them to graduate in six years) that can only arise from misclassification errors in the data; or
- expressing rates to three significant places for baselines or objectives in subgroups having, on average, fewer than 100 students; or
- establishing campus accountability objectives for matters that are not under campus control (e.g., percentage of freshmen who are "fully prepared" in English and mathematics).

These are examples from the A fundamental problem with the manner in which accountability indicators and goals are framed is the failure to understand and to allow for ordinary imprecision

[1] The opinions expressed in this paper are those of the writer and do not represent official system or campus positions. The examples are from the accountability process as implemented in its inaugural year. Some features of the CSU process are being modified, partly as a result of campus feedback to the central office.

and variability of measures. This tendency is encouraged by the CSU implementation of the accountability process, which sets forth numbers and says, in effect, "Here's where we are; do better." In the accountability process for the CSU, campuses were given "baseline" values, which were the values of the indicators in the most recent year for which data were available. Campuses were then asked to establish objectives at intervals two and four years into the future. The implication is that we will show improvement over the baseline and continue to monitor these values into the foreseeable future for the edification of system administrators, legislators, and ultimately the citizens of the state. The implementation by the CSU encourages the ideas that the baseline numbers are deterministic entities, that we understand the inputs that yield certain outputs, and that campuses can manipulate them at will. This is usually not true. In the first place, we should think of the baseline numbers as sampled values, with some degree of built-in sampling or random error. In the second place, the production functions that relate inputs to outputs are unknown or very poorly specified. In the third place, many factors known or likely to influence the indicators are not under campus or even systemwide control. In the fourth place, it is a misnomer and a mistake to take the most recent point in a time series and call it a "baseline." The notion of a "baseline" is that we literally have a base "line" from which dependable measurement can occur. In quality control analyses, a baseline is the mean or median value in a stable time series. Starting from a single data point, without information as to whether it is above or below the baseline and without any evidence that the time series is stable (or "under statistical control," to use Shewhart's expression), we run the risk of making naïve predictions about what can be achieved.

This paper aims to provide some information about how to assess variability in accountability indicators, with examples from the CSU accountability project that is currently under way. Accountability indicators are often expressed as percentages or averages that arise from a regularly reported time series or from a new series that is generated to support accountability goals. Examples of indicators used in higher education include:

- One-year continuation rates
- Six-year graduation rates
- Percentage of students expressing satisfaction with aspects of their education on a survey
- Upper division units taken by students who entered the institution as community college transfers, compared to those taken by students who entered the institution as freshmen
- Percent of freshmen proficient in math

The idea of monitoring percentages or averages of quality indicators over time is common and appropriate. However, measures of *level* such as percentages or averages must be considered in

the light of *normal variation* of the measures, usually expressed by standard deviations of the raw measures or standard errors of estimate for the level indicator.[2]

Each of the indicators listed above, which will be discussed in turn, raises particular statistical issues. Most of the basic issues will be discussed in connection with the one-year continuation rates, in the first and longest section of the paper. The section on graduation rates discusses the handling of accountability indicators for subgroups. The section on satisfaction rates discusses variability in rating scales and the sample sizes needed to detect significant effects. The section on upper division units discusses variability in indicators defined as differences between differences. Finally, the section on freshman proficiency discusses the systemwide Entry Level Math test and its potential as an accountability indicator. In an effort to improve readability of the paper, some of the more technical passages are relegated to appendices. Although the examples come primarily from the CSU accountability project, the principles apply in other settings where measurable results are expected or compared, including "benchmarking" efforts, assessment outcome research, specification of tangible results in grant proposals, expressing goals and objectives established under departmental work plans and campus strategic plans, and the like.

1. One-Year Continuation Rates

One-year continuation rates for freshmen are expressed as the percentages of students who enter as first-time freshmen in fall and who are still attending the following fall. Similar rates can be defined for upper division transfers, but this paper will focus on the data for freshmen.

With the baseline for one-year freshman continuation rates at 82%, the ad hoc group working on this indicator set objectives of 82.5% two years out and 83% four years out. Apart from the fact that the ad hoc group lacked the resources, responsibility, and requisite power to actually accomplish such improvements, as well as the fact that it is not clear whether an increase in a one-year continuation rate would represent improvement rather than merely change, such objectives completely ignore random variation in rates. Given that the 82% rate is based on about 1,000 students, statistical theory indicates that the standard error of estimate for the rate is about 1.2%.[3] This implies that we have only about two-thirds confidence that the true value at the baseline is within ±1.2% of the observed 82% rate. Thus, the objectives established for next year's freshmen and for the freshmen class two years following that are within the range of random variation for the so-called baseline rate. As we seek measurable objectives for

[2] Level indicators such as percentages and averages are the first things taught in an introductory statistics course. The next things taught are variances and standard deviations, which is where we start to lose people, espccially when we shift from the standard deviations of observations to "standard errors," which are standard deviations of means or other summary statistics from a distribution of observations. Only a few students—those in the extreme tail of the statistical student distribution, as it were—ever fully grasp and internalize notions of statistical significance, power, correlation, regression, reliability, validity, and so on.

[3] See Appendix 1 for a technical discussion of standard errors for proportions.

improvement, it makes little sense to have objectives that will not be significantly different from the baseline.

Unless the causes of variation in continuation rates are well understood and subject to management control—and there is often little reason to believe that they are, except in the most general terms—then the committee's 82.5% and 83.0% objectives essentially commit the University to a throw of dice. For a population where the true proportion is 82% we should not get excited in looking at rates in samples of size 1,000 unless new values lie outside a 95% confidence interval (about 80% to 84%). In this situation, values higher than 84% will occur about once in 40 times due to chance and values lower than 80% will also occur about once in 40 times due to chance.

As mentioned earlier, it is a mistake to refer to the most recent point in a data series as a "baseline," since this point will rarely be on the trend line of the observations, being above this about half of the time and below it about half the time. Thus, one really ought to take the mean of a series of observations as the base "line," to allow for random variation about the line. Figure 1 shows two short time series generated at random, using a mean of 82% and a standard error of 1% for each observation. Note that the right ends of the two time series differ by a little over 1% even though the population percentage of 82% is the same for both. To visualize what an improvement of 2% over the 82% baseline might resemble, imagine that the rightmost time series is raised by 2%, but with the same peaks and valleys. While the low end of the new series will overlap with the high end of the baseline series, the new series would represent genuine improvement. However, multiple observations over time would be needed for persuasive demonstration of this level of improvement.[4]

Sampling, Random, and Non-Random Variation. So far, we have been talking mostly about sampling variation, which is a particular kind of random variation. Where sampling variation is present and the form of the parameter distribution is known, then we can derive confidence intervals on measurements, as described Appendix 1. However, random variation due to imprecision or unsystematic error of measurement may also be present, as well as systematic or non-random variation. W.E. Deming distinguished between special and common causes of variation:

[4] We would infer this because the before and after time series have each been extended to nine data points, effectively reducing the size of the pooled standard errors to one-third the size of the standard error for a single data point. Thus, instead of $t = 2 / \sqrt{2} = 1.4$, representing a **t** test of the difference between just two data points, which is not significant at the 0.05 level, we would have $t = 2 / \sqrt{(2 / 9)} = 4.2$, which is significant at the 0.001 level.

- Special causes (Shewhart called them "assignable causes") are ones that arise from known or identifiable causes. To the extent that the causes of variation arise from known problems and it is cost-effective to fix them, such causes can be removed. Some special causes—changes in system admission requirements that alter yields and characteristics of entering students, fluctuations in competition between universities (e.g., decisions made by the UC can affect the types of students entering the CSU), economic conditions that alter demand for higher education, etc.—can be identified but may not be controllable by a campus.

- Common causes (Shewhart called them "chance causes") are ones that arise from normal variation (including random variation) in each of the components of a system. In a manufacturing setting, common causes may include sampling error due to batch sizes as well as imprecision of test instruments, machine calibration settings, quality differences in raw materials, possibly climate or environmental fluctuations, worker variability, and so on. In an educational setting, common causes of variation may include sampling error from differences in the sizes of entering classes as well as imprecision in measurement constructs derived from analysis files, human reporting error, and so on. Sampling error is useful because it can be quantified based on established statistical principles. However, the other common causes of variation cannot be specified a priori, but must be established empirically. Precision of indicator measurement has to be achieved and not merely assumed.

When special causes have been removed, what remains is common variation, and when a system is subject only to common variation, then it is said to be in "statistical control." While sampling variation is a useful point of reference, it is not the only source of common variation in continuation rates or even the most important one. Figure 2 illustrates one-year continuation rates for fall entering freshmen at our campus, with the vertical lines indicating 95% confidence intervals for the rates. With the striking exception of the data points in fall 1990- to fall 1992, each data point is within or close to being within the 95% confidence interval for the previous data point, suggesting that the adjacent points are not significantly different.[5] What is one to make of this? Something more than random variation is occurring, but our explanations are necessarily ad hoc.

The low point in fall 1988 is partly due to the conversion to a new student information system that occurred in that year. The older system used arbitrary six-digit student IDs, and a great many incorrect SSNs were discovered when we converted to a system that used SSNs as student IDs. The conversion effort also taxed administrative units to the utmost, to keep normal processing on track, and some of the data for fall 1988 entrants may have been faulty when first reported.

The years 1990-94 were marked by a recession and budget crisis in California, during which University budgets were slashed in three successive years. The odd zigzag in trend from fall 1990 to fall 1993 is probably caused by the fact that campus budgets were slashed in fall 1992,

[5] The 95% confidence intervals in Figure 2 are represented by lines equal to 2.0 standard errors on either side of each mean. If the standard errors for two groups are approximately equal, then as a rule of thumb, the group means are significantly different when they are at least 2.8 standard errors apart.

for the second straight year, and fully enrolled classes had to be cancelled. Cancellation of classes affected the continuation rates for fall 1991 entrants—continuing freshmen having the lowest registration priority at our campus—which helps explain the dip in continuation rate for that cohort. The fall 1992 cohort, the smallest we had during the time period of this data, may have been somewhat more selective than other cohorts, and as freshmen successfully entering college in the midst of a recession, they may have had a greater motivation than other cohorts to persevere. For whatever reason, the fall 1992 cohort has had higher continuation and graduation rates than the other cohorts before or since.

Over the period marked by the data, the state's economy went from the budget crisis and anxiety of the early 90's to the "irrational exuberance" (Greenspan) of the late 90's. Freshman fees were increased three times and decreased once during this period, which would also affect continuation rates. In every year there were assorted and mostly unevaluated policy and procedural changes at both the campus and system level that probably affected continuation rates, not always in the same directions. All of these myriad factors *could* have affected rates, and some of them undoubtedly *did* affect rates, but it is not at all clear what caused what to occur.

If we ignore the data points from 1991 to 1994, with the possible rationalization that these were recession years and therefore shouldn't "count," then there is an increase of about 5 percentage points in the level of rates from the 1988-1990 period to the 1995-1998 period. We'd like to think that campus actions contributed to this change. After reviewing results from the 1994 Student Needs and Priorities Survey, there was a lot of attention paid to student satisfaction and to continuation and graduation rates as indicators of satisfactory progress. The campus strategic planning process was revamped, subcommittees and work groups were formed, and there was a lot of heated dialogue about who we are, how we can be better, and so on. The University's June 1998 "Report and Recommendations for Improving Retention and Graduation at Cal State Los Angeles" had four pages of recommendations for the kind of leadership and change in campus culture that would help our students to succeed, would improve student satisfaction, and quite incidentally, would improve retention and graduation rates. One of the University's three self-studies completed in connection with the 1999 WASC accreditation dealt with the assessment of student success. During 1996-2000, the University benefited from a Title III Sustaining Effects grant that enabled us to establish advisement centers for each of our six colleges, provide systematic training to academic advisers, and provide manpower for aggressively monitoring student progress. Thus, there were many activities that might have contributed to a five-percent increase in continuation between 1988-1990 and 1995-1998. It would be nice if student continuation rates improved as a result of the activities and intense self-scrutiny that surrounded them, but they could also have improved in part as a general consequence of irrational exuberance in the economy, changes in admissions selectivity, or other unnoted factors.

Reasons for Non-Continuation. A one-year continuation rate is an equivocal indicator. What does it really mean? High rates are usually considered good, but can they be too high? Is there an upper limit to how high these rates can go? Students are always making personal cost-benefit calculations about whether going to college and succeeding in college is worth their time, money, and effort. There is no way to force continuation on students who do not wish to continue, except perhaps by paying them high subsidies to attend. Why do students not continue beyond the first year? Some of the freshmen who leave are driven out, while others opt out:

1. There are students who fail because they can't or won't master required work in their freshman year.
2. There are students who leave because some graders are too strict.
3. There are students who stay because some graders are too lax.
4. There are students who leave because their financial packages are inadequate for their needs.
5. There are students who stay because their financial packages are adequate for their needs.
6. There are students who go to another campus because they are not happy at our campus, just as there are students who come to our campus because they are not happy at another campus.
7. There are students who opt out of higher education because they are too busy with work, play, family, or whatever to go to school.
8. There are students who have to drop out because their families move out of the area or because they get incarcerated, have medical problems, die, and so on.
9. If the economy is bad, some students stay in school so they can eventually get better jobs.
10. If the economy is bad, some students leave school to help support their families.
11. If the economy is good, some students don't have to work because their families can support them.
12. If the economy is good, some students can get satisfactory jobs without going to school.

And so on and so forth. Some of these factors tend to raise one-year continuation rates and others tend to lower them. We have little information about which of these conditions are present for what percentages of students in given freshman cohorts; about which of them are or can be affected by university actions; about where students who have dropped out or skipped out may have gone or why; or about what actions *should* be taken to achieve particular results. Is it always the case that higher continuation rates are better? Is it reasonable for continuation rates to reach 100%? No: there should be a point at which we say that the level of an indicator is about as high as it can be or should be.[6] Finally, are there costs associated with a given degree of increase and if so, what are they?

[6] In manufacturing applications, quality control indicators are usually specified with upper and lower limits. Even in the case of manufacturing errors, where there is an ideal lower bound in the sense that it is desirable in principle to reduce errors to zero, the cost of this may be prohibitive. It may be cheaper to discard 1% of a batch than to invest in the kind of equipment and process monitoring that would be needed to achieve a lower rate. Thus, an error rate of 5% might be economically unacceptable, while an error rate of 1% or less might be acceptable. The costs for public institutions to achieve very high continuation rates has not been thoroughly studied.

Reasons 1-3 for non-continuation above deal with academic standards and quality. Students who fail to meet academic standards after having reasonable opportunities to do so should not be allowed to continue. It is probably a mistake to assume that academic standards are equal across and within campuses, but it would clearly be a betrayal of our educational mission to lower standards with the narrow aim of boosting persistence rates. Items 4-8 deal with students' own needs and preferences, which can perhaps be partially influenced, but never fully controlled. Items 9-12 deal with general economic or societal effects on higher education participation rates, which are largely out of campus control and not well understood. It would seem then that continuation rates should be at some optimal level for a campus and that year-to-year rates would fluctuate randomly around that level. However, it is not clear what that optimal level should be for a given campus, clientele, and economic climate.

What do we mean by "improvement" in continuation rates? Are the CSU continuation rates in fact poor and in need of improvement? Some comparative information on one-year continuation rates is available from reports by the Consortium for Student Retention Data Exchange (CSRDE), a voluntary organization of universities who have pooled their retention data. CSRDE defines one-year continuation rates based on all freshmen attending full time (12 or more units attempted) in their first fall term. One year CSRDE continuation data for freshman cohorts from fall 92 through fall 98 are shown in Table 1 for institutions in four "selectivity" groups, with selectivity defined in terms of average SAT scores (or equivalent average ACT scores for institutions that use the ACT rather than SAT) for freshmen.

One-Year Continuation Rates for Freshmen by Selectivity Groups – CSRDE

Selectivity Group	1998 SAT Average	or	1998 ACT Average	1992-98 Avg Rate
Highly selective	>1100		>24	86.0
Selective	1045-1100		22.5-24	78.0
Moderately selective	990-1044		21-22.4	73.4
Less selective	<990		<21	70.6

Thus, there is a clear gradient from a 70.6% average rate for less selective institutions up to an 86.0% average rate for highly selective institutions, although there is also considerable variation (not shown) within selectivity groups.

Within the CSU, two campuses were in the highly selective group in 1998, one was in the selective group, two were in the moderately selective group, and the rest were in the less selective group. Using weighted averages for the CSU campuses participating in CSRDE, the CSU as a whole had an average SAT score of 992 and one-year continuation rates of 78% (1993) and 79% (1998). Thus, while the CSU is at the borderline between the moderately selective and less selective institutions in terms of SAT scores, it achieves one-year continuation rates at about

the average for selective institutions. Given the selectivity of the CSU system, based on average SAT scores, our one-year continuation rates are higher than expected.[7]

One of the main reasons for the CSU's relatively high continuation rates is that tuition at the CSU is unusually low among public comprehensive universities. It is also the case that there is no systemwide university policy (as there is at the UC) that requires students to attend full time. We even give students a price break on tuition if they register for six or fewer units. This makes it both economical and convenient for students to continue their studies with a minimal time commitment, which is especially helpful for students who work full time and get educated part time. These factors undoubtedly help keep systemwide one-year continuation rates relatively high, but they also may help keep systemwide six-year graduation rates relatively low. Naturally, campuses should not be accountable for continuation rates to the extent that they are determined by systemwide policies and not under campus control.

Given the fact that the fall 98 "baseline" continuation rate of 82% is relatively high, the fact that the campus has engaged in a series of activities intended to increase continuation rates since 1995, and the fact that these rates have inescapably random components, my accountability goal recommendation was that an 82% continuation rate is reasonable for our campus and we should estimate the rates two and four years out to be about 82% ± 2%. However, this recommendation did not carry the day, the ad hoc group established two and four year goals of 82.5% and 83%, respectively, on the theory that increases were expected and standard errors were not.

2. Six-Year Graduation Rates

Graduation rates in the accountability project pose some interesting problems. These rates use Joint Commission on Accountability Rates (JCAR) methodology, in which separate rates are computed for students on a fast, medium, or slow pace of taking coursework. JCAR methodology was developed on the reasonable premise that students cannot be expected to graduate within four, six, or any given number of years if they do not take coursework on a pace that permits this. For entering freshmen, therefore, a student's pace from entry through either stop out, graduation, or completion of four years of coursework (whichever occurs first), is defined as follows:

---

[7] The persistence gradient in the quartiles formed from average SAT/ACT scores should not be taken to imply that test scores are the real cause of persistence and dropping out. Test scores are famously correlated with economic indicators. If the CSRDE data were to be arranged by quartiles of freshman parental income or of institutional resource measures such as instructional expenditures per student—variables not included in the CSRDE database—we would probably see the same sort of gradient. The fact that such selectivity gradients are present suggests that we can affect continuation rates, perhaps more effectively than by anything we do after students enroll, by changing admission criteria. However, this violates the spirit of the persistence indicator, which is intended as a measure of post-admission program quality and success rather than as a measure of admission selectivity.

Fast pace	On a pace to graduate in four years
Medium pace	On a pace to graduate in six years
Slow pace	Other freshmen

For each of these three groups, the campuses were provided with separate four-year, six-year, and "eventual" graduation rates[8], as shown in the following table:

Summary - Fall 93 Freshman Baseline Rates – Cal State L.A.

JCAR Group	4-year	6-year	Eventual
Fast pace	21.4	68.8	74.7
Medium pace	0.0	31.6	46.1
Slow pace	0.0	2.2	19.4
TOTAL	3.9	35.2	49.5

To begin with, one needs to know standard errors or sample sizes for these rates in order to help readers judge the random sampling variability that is inherent in the rates. If rates are computed for subgroups, then the subgroup rates necessarily have larger standard errors than the group as a whole, because standard errors are inversely related to the square roots of sample sizes. Using formula [1] introduced in Appendix 1, standard errors for the three groups and three persistence indicators are very different from one another, as shown in the next table.

Standard Errors - Fall 93 Freshman Baseline Rates

JCAR Group	N	4-year	6-year	Eventual
Fast pace	154	3.3	3.7	3.5
Medium pace	595	-	1.9	2.0
Slow pace	93	-	1.5	4.1
TOTAL	842	0.7	1.6	1.7

Over two-thirds of the students are in the medium-pace group at our campus, which therefore has the smallest standard errors (except for the anomalous indicator for students who graduate in six years despite being on a pace too slow to graduate in six years). Ninety-five percent confidence intervals for the medium-pace group are roughly 32% ± 4% for six-year graduation rates and 46% ± 4% for six-year persistence rates. Given the sample size for the fast-pace group, 95% confidence intervals for rates in this group should be stated as ± about 7%. For the slow-pace group, with only 93 students in the base year, there were evidently 2 students (2.2%!) who graduated within six years and another 18 students (19.4%!) who were still attending six years after entry.[9] How do slow-pace students (who by definition are not on a pace to graduate in six

[8] Based on CSU studies showing that most students still attending after six years will eventually graduate, eventual rates are estimated as the percentage of those graduated and/or still attending in the fall term following the sixth summer term after entry.

[9] The six-year graduation rate for slow-pace students is a case where the standard error approximation in formula [1] of Appendix 1 is too crude, since we have both a small sample and

years) graduate in six years? They might be freshmen who enter with transferable college credits earned while still in high school; they might be students who start slow with us, take courses at a faster pace somewhere else, and return for graduation; they might be students who have been attending more than one institution during the period when pace groups are defined; or they might simply be data errors. Setting six-year graduation objectives for the slow- and medium-pace groups or four-year graduation objectives for the medium-pace group makes little statistical or practical sense.

As with the one-year continuation rates, a single data point (fall 93 freshmen) should not be considered a "baseline," and we should be looking at points in a time series, partly to improve the combined sample size in order to gain precision and partly to assess the non-sampling variation in the data. The following table provides information for pooled data from the 1990- through 1998 cohorts at our campus.[10]

Campus Averages and Standard Deviations for Graduation Rates
1990-1998 Cohorts – Regularly Admitted Freshmen

	Averages				Standard Deviations			
JCAR Group	N	4-year	6-year	Eventual	N	4-year	6-year	Eventual
Fast pace	186	30.3	66.3	73.2	34	3.1	5.3	6.6
Medium pace	695	0.4	29.6	47.5	79	0.2	3.5	2.0
Slow pace	112	0.1	2.0	21.7	26	0.4	1.0	1.9
TOTAL	993	4.1	33.2	49.2	134	0.7	3.1	2.2
CV[11]					13%	17%	9%	4%

This table reports average cohort sizes and rates over a nine-year period, as well as empirical standard deviations of the averages, for the fast, medium, and slow pace groups. Notice the group sizes (N) and their standard deviations. The group sizes are as variable or more variable than the six-year and eventual graduation rates, making it difficult to simultaneously meet or predict the group targets and the overall target.

---

a rate too close to zero.

[10] These rates are from campus data. We cannot replicate Chancellor's Office JCAR rates, which use systemwide files and make certain adjustments for anomalous data and multi-campus attendance, but the rates should be reasonably close to those in the system data. The sample sizes of the freshman classes are based on all nine years, while the four-year graduation rates are based on the 1990 through 1996 entering cohorts and the six-year graduation and eventual graduation rates are based on the 1990 through 1994 entering cohorts.

[11] The coefficient of variance, **CV = 100% × S / M**, is the ratio of a standard deviation to a mean for a distribution. It is useful for comparing variability in measures having different metrics, since it expresses the standard deviation as a proportion or percentage of the mean. For the medium-pace group, the CV is about 11% for the group size, 12% for the six-year graduation rate, and 4% for the eventual graduation rate. Thus, the group size and six-year graduation rates are considerably more variable than the eventual graduation rates.

Note that the average size of the freshman classes over this period was 993 and that the size of the fall 93 cohort in the "baseline" table was noticeably smaller at 842. As one of the small freshmen classes admitted during the 90-94 recession, the fall 93 cohort is arguably different from current students in some respects, so there is little guarantee that the 35% six-year graduation rate for this group is typical. Accordingly, overall goal for eventual graduation rate should probably be set from a baseline figure of 33% rather than from the 35% obtained for the fall 93 cohort. It should also be noted that eventual graduation rates are largely determined by what happens during the first few years after entry. Thus, it is probably too late to materially affect graduation rates for freshman cohorts that entered in, say, fall 94 through fall 98.

There are valid reasons for wanting to increase our overall six-year and eventual graduation rates above the 33% and 49% levels, respectively, and for trying to get more of the 49% of students we think will eventually graduate to do so at a faster pace. One way to do this is probably to try to move students from the slow-pace to the medium-pace group and from the medium-pace to the fast-pace groups. Above all, we want to keep students attending and making steady progress, but it is not clear how one would establish separate goals and retention tactics for the different JCAR groups. For one thing, JCAR pace groups are never identified anywhere in student information systems, except after the fact. Campuses classify students by discipline and class level, not by graduation pace group. While there are interventions for students on academic probation and while departments may encourage students nearing completion to finish the last few outstanding papers, projects, or courses, it is difficult to imagine tactics specifically designed to change graduation rates within pace groups other than by changing the paces themselves. The JCAR group boundaries are also very arbitrary. Students on a pace to graduate in 5.75 years and those on a pace to graduate in 6.25 years may be, from the student and department perspectives, making equally satisfactory academic progress. The fact that Universities are counting *six*-year graduation rates as a conventional accountability standard is truly irrelevant with respect to the quality of education achieved by students graduating in 6.25 as opposed to 5.75 years. (Clearly the campuses should encourage students who are on track to graduate in 6.25 years to finish within 6.0 years instead, to "improve" their graduation rates, but this is somewhat akin to creative accounting—e.g., by spending money this year rather than next year or vice versa. Annual rates may change, but the overall effect remains about the same.)

Overall graduation rate is the accountability indicator I would personally most want to improve for my campus, and I would want to do it without changing our traditional clientele, many of whom are from neighborhood schools and school districts that serve some of the most educationally needy and economically disadvantaged students in the state. The JCAR methodology is useful in drawing attention to pace as a factor in persistence (how could it not be?) and I would expect persistence rates in the pace groups to improve as a result of continued retention efforts. However, the notion of setting separate objectives for the pace groups is odd and leads in some instances, as in the case of the smaller JCAR categories, to unmanageable tasks.

3. Satisfaction Rates

In the current market economy, where we educate education consumers rather than, strictly speaking, students, there is a great deal of concern for what is called "student satisfaction."[12] This is typically measured with rating scales administered to somewhat randomly selected groups of students. Since some campuses are or will be establishing accountability indicators using satisfaction ratings, it may be useful to review their statistical properties. Appendix 2 provides a discussion of variation as a function of the number of response categories and distributions of responses in the rating scales.

For many years, the CSU has used the following 5-point scale of "general satisfaction" in its periodic Student Needs and Priorities Survey (SNAPS):

> Please mark the one response that comes closest to your feeling about the statement: "I am pleased with my overall experience on this campus."
>
> 5 O Strongly Agree
> 4 O Agree
> 3 O Neutral
> 2 O Disagree
> 1 O Strongly Disagree

As discussed in Appendix 2, five-point scales of this kind usually have standard deviations of about 1.00 or perhaps a little less. With this information, we can construct a table of the approximate sample sizes needed to detect mean differences of a given size in general satisfaction ratings.

Sample Sizes Needed in Each Group to Detect Selected Mean Differences on a 5-Point Scale at the 0.05 Level with 0.80 Power, with Within-Group Standard Deviations Equal to 1.0

Size of mean difference to be detected	0.05	0.10	0.20	0.30	0.40	0.50
Sample size needed	6,279	1,570	393	175	154	99

Thus, to have an 80% chance of detecting a mean difference as small as 0.10, one needs a sample size of at least 1,570 in each of the two groups (or at each of two time points, if we are looking at ratings for successive cohorts).

[12] I alluded earlier to the metaphoric representation of students as "products" of an educational system. An alternative and somewhat contrary representation treats them as "customers" to whom education must be marketed.

As with continuation rates, random variation in satisfaction rates is largely a function of sample size. However, unsystematic effects that are unrelated to group sizes can also arise in empirical research. For several years, we have been using the Student Needs and Priorities Survey (SNAPS) item on "general satisfaction" in other student surveys administered on campus. In 1999, as it happened, we administered three separate surveys of students at our campus—SNAPS, the Noel-Levitz Student Satisfaction Inventory (SSI), and the Higher Education Research Institute College Student Survey (CSS). SNAPS was a systemwide mandated survey, the SSI was administered in connection with some consulting work on campus by the Noel-Levitz organization, and we administer the CSS every year as a matter of campus policy. SNAPS was administered in Winter 1999 and the other two surveys were administered in Spring 1999, using parallel samples of students in intact classrooms, with the results shown in the following table.

General Satisfaction Average from Three Surveys Administered in 1999

Statistic	SNAPS	SSI	CSS
Mean	3.60	3.48	3.74
Standard Deviation	0.90	0.99	0.99
Standard Error	0.03	0.04	0.04
Sample Size	1,023	558	797

All samples were selected in the same manner and the wording of the items and five-point scales were identical, yet the mean values ranged from a low of 3.48 to a high of 3.74. The mean from each survey is significantly different from the means for the other two at the 0.05 level, despite the fact that all three surveys were administered within about a four-month period. The 0.26 difference between the SSI and CSS is significant at the 0.001 level even though the surveys were administered in the same weeks of the same term to parallel samples of students. Although randomness cannot be ruled out, even for results as strange as this, these are rather large mean differences to have been obtained by chance. Causal explanations might be that there were nonrandom differences in the nature of the classrooms selected for the SSI and CSS samples or that there were unsuspected differences in the contexts established by the two survey instruments.[13] The differences certainly do not reflect anything that the campus was doing to change the satisfaction ratings by the students taking the three different surveys.

[13] When ratings at a certain level in one set of items influences the ratings on subsequent items, it is called a "context effect."

The campus takes student satisfaction seriously. One of the three self-studies at our last WASC accreditation was on student satisfaction, a quality service committee was established by the Strategic Planning Coordinating Committee, and a quality improvement committee is a permanent work group reporting to the Enrollment Management Steering Committee. No doubt these activities are having effects on students' attitudes and self-reported satisfaction and we can expect to see evidence of these effects in periodic surveys. However, the finding that different sorts of surveys (SNAPS, SSI, and CSS) may show different levels of satisfaction indicates that we should be careful to keep sampling procedures and survey measures of this as uniform as possible, otherwise instrument variation can be larger than the effects of campus action. It is also important to recognize that there are many random and non-random but unsystematic sources of variation that can interfere with a straightforward evaluation of student satisfaction.

4. Upper Division Units Earned

Sometimes goals are expressed as differences between two measures. This particular accountability indicator is expressed as a difference between differences, since we compare community college transfers with freshman entrants in terms of a difference between units earned when at graduation and when students achieved junior status at the CSU. Differences and differences between differences usually have greater variability than the individual measures whose differences are being taken.

Community college transfers typically have higher units earned at graduation than students who start at the CSU as freshmen. This may be due to inefficient scheduling of coursework by students, possibly with inefficient or poorly communicated coursework articulation as a contributing factor. Under Trustee policy, we do not want to have transfer students penalized for starting at a community college, so the difference between upper division units earned by junior community college transfers (JCCT) and first time freshmen (FTF) is taken as the statistical indicator to be monitored.[14] Looking at the trend of JCCT – FTT upper division unit differences, we have:

Group	JCCT – FTF UD Difference	Standard Error
96/97 graduates	9.9	2.2
97/98 graduates	4.9	2.1
98/99 (base year) graduates	2.2	1.9
Average	5.6	2.1

This looks like a nice trend in the direction of lower UD differences, with the smallest difference (2.2 units) in the base year of 98/99. However, since the standard error of the differences is a about 2.1 units in any one year, the three observations are close to being within a 95% confidence interval of the overall mean of 5.6. Given three such arguably random observations, they can be ordered in eight possible ways. There is one chance in eight that the observations

[14] See Appendix 3 for a discussion of the standard error of this indicator, which is formally a difference between correlated differences.

would be ordered from high to low, as we see them, but it is equally likely that they could have been ordered from low to high. There is also a three-quarters chance that they could have occurred in one of the other six possible orders. The campus has been given a baseline value of 2.2 units for the indicator, with the expectation that we will show improvement in subsequent years. However, in all likelihood, the 2.2-unit difference between differences is a random low point in a time series that is too brief to serve as an adequate baseline. Thus, 5.6 units is a more appropriate baseline number for establishing a goal with respect to the upper division unit difference.[15]

5. Percentage of Proficient Freshmen

The last indicator to be discussed is the percentage of freshmen proficient at math prior to entry. I have focused on math proficiency rather than English proficiency, in part, because I know more about the assessment of math proficiency. Mathematics proficiency at the CSU is assessed primarily using a test called the Entry Level Math (ELM) test.[16] Students may also establish proficiency by passing AP math, by having sufficiently high scores on ACT, SAT I, or SAT II mathematics tests, or by passing courses that are equivalent to or more difficult than a CSU general education mathematics course. In the fall 2000 cohort, about 55% of entering CSU freshmen were considered proficient in math. The CSU Trustees have established a goal of having 90% of entering freshmen proficient in math by fall 2007.[17]

---

[15] To this writer, a more useful and conceptually simpler indicator would have been the JCCT – FTR difference in units at graduation, which averages about 13 units for our campus, since that represents excess units (roughly an extra quarter of coursework) earned by students who started their baccalaureate at a community college. The difference between differences is considered to provide a better indicator of Trustee policy, which is stated with respect to upper division units earned while attending CSU campuses. However, it should be recognized that differences between differences will generally be more variable than individual components of the indicator.

[16] In 1998, I made a study of the ELM and its scoring and test equating methods, presenting the results at a meeting of the systemwide ELM Advisory Committee (ELMAC) (Jordan, 1998a, 1998b). It should be noted that the writer's views on the ELM do not represent official views of the CSU or campus and may differ from those of the majority of ELMAC members.

[17] Our campus has baseline figures for fall 1998 of 23% of regularly admissible freshmen proficient in math and 21% proficient in English. The group working on this indicator obtained goals for fall 2002 and fall 2004 entering freshmen by linear interpolation between the fall 1998 baseline values and the 90% Trustee target for fall 2007, resulting in goals of 40.7% proficient in math and 41.3% proficient in English by fall 2002 and 61% proficient in math and 62% proficient in English in fall 2004.

One problem raised by this indicator is that the nature of the ELM test has changed over time, so that we have a poorly established baseline and it is somewhat difficult to establish or interpret objectives for change. Unlike the situation with one-year continuation or six-year graduation rates, which are subject to sampling error but little instrumentation error or unreliability, ELM scores are subject to unreliability as well as sampling error, and the test itself has had major content changes after fall 1992 and fall 1998. A technical discussion of the scaling and content of the ELM is in Appendix 4.

I am fond of the aphorism by Ortega y Gasset, "Name a concept and reason flies out the window." "Proficiency" is good and "remediation need" is bad, but one needs operational definitions of these terms in order to discuss them usefully. There is no universal definition of "proficiency" or "remediation need." A national study of freshmen in fall 1995 indicated that 18% of freshmen at public four-year colleges and 9% of freshmen at private four-year colleges needed math remediation (NCES, 1995), with remediation defined as the individual institutions defined it. Based on data available by state, such as average SAT scores for college-bound freshmen, there is little evidence that California students have dramatically worse math preparation or lower math skills than students in other states. Therefore, the CSU standard, which determined that about 52% of fall 1995 freshmen "needed" math remediation, must be one of the strictest in the country. Only 45% of fall 2000 freshmen "needed" math remediation, but this may be due more to changes in the standard than to changes in freshman skills.

For the CSU, the math proficiency and remediation concepts are operationally tied to the ELM. This test changes subtly every year (since no two versions have exactly the same items), it has been changed in major ways several times in the past, and it is currently being considered for yet another revision. Figure 3 shows the percentage of freshmen considered "proficient" or "fully prepared" from fall 1989 through fall 2000, along with Trustee goals for the fall 2001, 2004, and 2007 cohorts. The figure has been annotated to show major changes in the instrument over time.

- The most dramatic change in ELM content occurred after fall 92, when the test was made more difficult by adding Algebra II content. Prior to that time, consistent with the CSU subject area requirement, the ELM tested for content from the first two years of high school mathematics, traditionally Algebra I and Geometry. When the CSU began requiring three years of high school mathematics after fall 92, the ELM was made more difficult and passing rates dropped from around 75% systemwide in fall 92 to below 50% in fall 95 through 98.

- The next major change occurred after fall 98, when content in the areas of data analysis and probability and statistics replaced some of the Algebra II content, and, for the first time, calculators could be used in taking the ELM. To minimize the impact of calculator experience, there was some effort to redesign problems so that calculators did not confer a computational advantage. "Percent proficient" increased after fall 98, but it is not clear to what extent this happened because student skills increased or because the test became easier or both. [18]

[18] The writer's analysis of data from the pilot test conducted when the test content shifted in fall 1999 indicated that 60% of the pilot students came from two large high schools in L.A. County, which happened to be at opposite ends of the ability and socioeconomic spectrum (see Jordan, 1998b). The revised test was significantly *easier* for students in the high-status high school and

- Current changes being considered for the ELM include shortening it from 60 to 45 scored items, substituting some word problems requiring mathematical reasoning for items requiring purely algebraic manipulations, and further diluting the Algebra II content.

From fall 93 to fall 98, the measured proficiency of CSU freshmen dropped from 55% to a low of 46%. Explanations for this trend include:

1. Students were less well prepared by their high schools in fall 98 than in fall 93.
2. There was a change in characteristics of the students enrolling in the CSU over this period.
3. Because of increasingly stringent enforcement of policy that students must be assessed before registration for classes, the tested skill levels in the pool of test takers were lowered because scores for freshmen who might previously have postponed and perhaps dropped out before ever taking the ELM were included in the pool.
4. Difficulty levels of the test were drifting upward, causing more students to fail in fall 98 than in fall 93.

Of these four explanations, (1) is probably ruled out because statewide SAT math scores for California college-bound seniors, even with an increasing percentage of seniors taking the exam, rose slightly over this period, as shown in the following table:

Math Proficiency (CSU Freshmen) & Average SAT Math (CA College-Bound Seniors)

Year	**93**	**94**	**95**	**96**	**97**	**98**
% Proficient-Math	55	52	48	47	46	46
SAT Math	508	506	509	511	514	516

Explanations (2) through (4) have been suggested as being among the reasons for the decline in scores, but have not been definitively evaluated to my knowledge. Explanation (4), that the test difficulty level has drifted over time, could be tested by an experiment in which current students were randomly assigned to take the ELM exams that were administered in 93, 95, and 97, say, to determine whether the passing rates as scored are identical within sampling error.[19]

---

slightly but not significantly more *difficult* for students in the low-status school. Whenever test content changes, it is likely to change which *individuals* are deemed proficient, even when the average difficulty level has been constrained to be equal between two forms of the test. The revisions after 1998 may have widened the gap between students in the low-status schools and those in the high-status schools. Calculator use is a factor predicting higher math scores. In the data for fall 2000 college-bound seniors, students who report that they use a calculator "almost every day" have an average SAT Math score of 541, while those who use it "once or twice weekly" or less have an average SAT Math score of 473.

[19] This would constitute a norming study of the ELM. Present procedures do not adequately ensure that overall difficulty of the test since 1992 is stationary, even without the major changes in content that have occurred. An interesting analysis that would tend to show how or whether ELM difficulty has changed over this interval would be to determine the percentages of students passing the ELM for students with various SAT Math scores. A particularly interesting analysis

To summarize the material in Appendix 4, the ELM is subject to several types of measurement error:

- There is standard equating error (Angoff, 1982) that can accumulate when a series of versions of a test are equated via a daisy-chain model for adding new items. (This could be corrected by the norming study mentioned above, to determine whether alternate forms of the test yield similar passing rates when randomly assigned to current students.)

- There is imprecision due to the fact that, near the proficiency cutpoint, each additional correct answer represents about 3% more individuals, so the percentage on any given version of the test cannot be accurate within more than ± 1.5%. (This could be improved by having a longer test with a somewhat wider range of item difficulties.)

- There is unreliability of measurement due to the fact that individuals taking an alternate form of a test or taking the same test at different times will achieve different scores. (This could be improved with longer tests and better equating.)

- There is sampling error to the extent that individuals assessed may not represent the population of admissible freshmen.

If we estimated and added up the variance from all these sources, it would probably amount to several percentage points. Thus, passing rates in the range from 46% to 48% in fall 1995 to 1998 may not be statistically different. The decrease from 55% in fall 1993 to the 46-48% range in 1995-97 and the subsequent increase back to 55% in fall 2000 are statistically significant changes, but we do not know the extent to which the proficiency changes are due to changes in the students or to changes in the test.

In addition to lack of norming studies of the ELM, there has never been to my knowledge a validity study of the ELM, showing its value in predicting outcome criteria in various fields. Math skills at the level tested by the ELM have been assumed but not demonstrated to be valuable across the board. The need to have arts and humanities majors meet the same entry level math criterion as majors in scientific or engineering fields has not been shown. Clearly, there will be individuals in any field with specializations that require higher math skills, but there is no reason why all students in every field should have to jump a high math hurdle.[20]

---

would be to look at entering freshmen in each year with SAT Math scores near the ELM exemption point (currently 550, previously 560) to determine passing rates on the ELM over time. This would take advantage of the fact that some students with exempting SAT scores also take the ELM, and also assumes that SAT difficulty levels have not drifted over the same time period.

[20] I tried to argue this point with ELMAC when I met with them in 1998, but found only limited support for the ideas that the ELM criterion was most appropriate for students in math-intensive fields and that we needed dual math criteria—one for students in math-intensive fields and another for students in other fields.

Finally, campuses should not be required to set objectives for the percentages of entering freshmen who will need remediation, since this is determined almost entirely by forces outside of campus control. Realistically, there is little that campuses can do to affect the proficiency of entering freshmen other than to apply system policy with respect to admissible students.[21] The goal of having 90% of entering freshmen proficient in math by the year 2007 is somewhat analogous to requiring the average SAT math scores for 90% of entering freshmen to exceed 550.[22] When you consider that most college-bound seniors in California (about half of high school graduates) take the SAT, that only 63,600 of the college-bound seniors (34,300 males and 29,300 females) in fall 2000 had SAT math scores of 550 or higher, that some fraction of these will attend private institutions, that another fraction will (probably at least 20 or 30%) will need remediation for English,[23] and that the UC and CSU between them enroll about 55,000 freshmen, it is not clear where all the high-scoring freshmen are going to come from. Thus, the 90% goal seems mathematically implausible. The only way that we will be able to say that 90% of freshmen are proficient in math is by either raising admission standards dramatically between now and 2007 (with the probable side effect of reducing freshman enrollments), or by lowering the proficiency criterion to something closer to the criteria that are used in other institutions and states. The campuses cannot affect either of these, which would have to arise from systemwide policy changes.

---

[21] Guidelines for this indicator, which fall under relations with K-12, indicate that "CSU faculty have been collaborating with high school teachers on expectations, diagnostic testing, and instructional approaches, as well as improving the quality of training provided to new K-12 school teachers in the state." We do have math faculty making contact with high school math teachers in our service area and holding workshops on for math teachers on campus. This could eventually have a positive effect on student test scores, but it is unlikely to result in 90% math proficiency in freshmen who are regularly admissible under current standards.)

[22] Because 550 on the SAT Math test exempts students from taking the ELM, it may be useful to look at the groups achieving an average score this high among the California college-bound seniors for fall 2000. Students reporting three years of high school math had an average SAT Math score of 486, those reporting four years of high school math were at 530, and those with more than four years of high school math were at 570. The average for students where the highest level of parent education is a baccalaureate was 542, while the average for students where the highest level is an associate degree was only 499. The average for students in suburban schools was 549, while the averages for students in urban and rural schools were 499 and 495, respectively. Thus, for achieving SAT Math scores of 550 or higher, it helps to have a more intensive math curriculum than the CSU requires, to be a second generation college student, and to live in the suburbs.

[23] According to data also published by the College Board for the fall 2000 college-bound seniors, the SAT Math difference between men and women has averaged about 39 points between 1972 and 2000, with a range from 35 to 46. Thus, unless we can eradicate the long-standing gender difference in tested math ability, meeting the Trustee goals will entail reducing the numbers and percentages of women eligible to enter the CSU. Raising a hurdle does not make anyone a better jumper, but it does keep some citizens from getting to the other side.

6. Goal Setting and Quality Improvement

This paper is not written to argue that indicators of campus quality cannot be changed by deliberate action nor that the quality of education provided to students cannot be improved and ultimately measured. Rather, it is intended to argue that, in establishing quality goals, one needs to understand and allow for random and systematic variation in the quality indicators. One has to know and appreciate the precision of one's instruments. Micrometers can measure within a fraction of a millimeter, yardsticks to within a quarter of an inch, satisfaction surveys to within one or two standard errors, and so on. It is a poor workman that blames his tools, according to the proverb, but quality indicators, like other measuring instruments, can only be as precise as their design allows.

In Deming's view, the task of management is to study and understand the causes of variation—in effect, to identify special causes and remove them where feasible, so that one is left with common causes that are essentially random variations around a level of an indicator that reflects what the system is capable of producing at a given time. Once an indicator achieves this level of statistical control, then by changing processes or components of the system, one can change the level of the indicator in one direction or another. Deming is skeptical of goal setting that precedes a thorough understanding of a complex system. As he says:

> "A numerical goal accomplishes nothing. Only the method is important. By what method? A numerical goal leads to distortion and faking, especially when the system is not capable to meet the goal."
>
> —W.E. Deming, *The New Economics for Industry, Government, Education.*

This is perhaps a surprising statement from a statistician, but it arises from Deming's considerable consulting experience with organizations that expect to have numerical quality improvements without doing the difficult homework and teamwork of figuring out how quality within the organization is or can be achieved. Before trying to change a system, one must study it over time to determine how much random variation exists in the system and whether it is a stable system that is producing similar results year after year. Otherwise, one runs the risk of "mistaking coincidence with cause and effect" and of rewarding those who get heads and punishing those who get tails in a particular year.[24] If Deming is correct, then actions and methods are more important than goals per se. As he argues in *Out of the Crisis* and the posthumous *New Economics*, it is demoralizing for workers to be expected to meet goals that an existing system cannot achieve, a situation that can lead, as noted in the citation above, to distortion and faking of data.

---

[24] The State of California is currently making cash awards to K-12 schools for increases in student test scores that may be partly upward regression to a mean. (Regression to the mean refers to the tendency of low-scoring units to increase and high-scoring units to decrease, when measured a second time, simply and solely because of the extent to which they were randomly above or below a regression line when they were measured the first time.)

There is a lively statistical literature on what are called "education production functions," typically involving regression analyses of the determinants of educational variables having some policy relevance. The failure of such predictive models to yield information that policy makers can use has lead to what Monk has called an "outcomes-as-standards" strategy, in a passage worth quoting at length:

> On the one hand, there is consensus that existing education production research has been largely unsuccessful at revealing the schooling inputs that dependably contribute to enhanced learning gains of students. There is no shortage of pessimistic assessments of what this literature has contributed. . . . On the other hand, there is a drive toward raising the level of educational production (sometimes coupled with concerns over improving efficiency), which is strong, and probably growing, and presupposes a nontrivial store of knowledge regarding the ability of state, district, and school officials to enhance productivity. . . .
>
> An important policy response involves having a centralized authority—typically although not necessarily a state—focus on the outcome side of the production function, set minimum standards, and hold constituent units, be they school districts or schools, accountable for meeting the standards through a system of positive or negative incentives. . . . Numerous states and some school districts have implemented reforms containing outcome-based incentives. In so doing, the more centralized authority sidesteps having to spell out the ingredients of education success and can sit back and act as judge and jury of those with more immediate responsibility of producing the desired results.
>
> This policy response can be viewed as a strategy, perhaps even an ingenious strategy, that successfully finesses the ignorance that characterizes our knowledge of the underlying education production function(s). Ingenious though this "outcomes-as-standards" response may be, there are serious deficiencies that are not sufficiently well appreciated (Monk, 1992, pp. 307-308).

As Ernest House (1998, p. 75) remarks, "If one cannot define the appropriate inputs [to an educational production function], why not specify the outputs and demand that educators produce them however they can?" However, this violates sound principles of quality management and improvement, at least as espoused by Deming. It is premature to expect campuses to "improve" indicator rates without understanding whether the rates are in fact poor under the circumstances, which determinants of the rates are under campus control, or how changes in campus practices might change these rates upward or downward. This criticism is not just a matter of routine academic foot-dragging, but an insistence that quality principles and appropriate statistical methodology should be applied to the measurement and control of quality, which is what accountability is supposed to be about. Statistical discipline should be applied to quality control management.

Deming, who was an astute and rigorous statistician, also makes the surprising statement that "the most important figures needed for management of any organization are unknown and unknowable." He elaborates with examples suitable for a manufacturing business, but analogous examples can be found for the University:

> One cannot be successful on visible figures alone. Now of course, visible figures are important. There is payroll to meet, vendors to pay, taxes to pay; amortization, pension funds, and contingency funds to meet. But he that would run his company on visible figures alone will in time have neither company nor figures.
>
> Actually, the most important figures that one needs for management are unknown or unknowable, . . . but successful management must nevertheless take account of them. Examples:
>
> 1. The multiplying effect on sales that comes from a happy customer, and the opposite effect from an unhappy customer. . . .
> 2. The boost in quality and productivity all along the line that comes from success in improvement of quality at any station upstream.
> 3. Improvement of quality and productivity where the management makes clear that the policy of the company will henceforth be to stay in business suited to the market: that this policy is unshakeable, regardless of who comes and goes.
> 4. Improvement of quality and productivity from continual improvement of processes; . . . also from elimination of work standards, and from better training or supervision. . . .
> 5. Improvement of quality and productivity from a team composed of the chosen supplier, the buyer, engineering design, and sales customer, working on a new component or redesign of an existing component.
> 6. Improvement of quality and productivity from teamwork between engineers, production, sales, and the customer.
> 7. Loss from the annual rating on performance.
> 8. Loss from inhibitors to pride of workmanship of employees (Deming, 1994, pp. 121-122).

Deming taught that quality is the important thing. He was not in favor of the mechanistic application of the indicator-monitoring programs that have been promoted historically under the banners of TQM and CQI, sometimes promoted in his name, but seldom conducted in his spirit. If we pay attention to quality, if quality is rewarded, if every worker is imbued with desire to be in and to improve a quality organization, then goals will take care of themselves. Deming is passionate about and often gets preachy on the topic of worker pride as the most valuable single asset of a quality organization. He taught that establishing numerical goals was an empty, wasteful, and often destructive exercise unless we understand what it is about a system that produces particular results and what methods can be used to change the system to bring about results that are in some definable sense better. Understanding, he believed, should precede goal setting and not vice versa.

While the drive for accountability from legislators will probably not go away, I wish we could refocus our efforts to emphasize "assessment" rather than "accountability." The standard text on assessment is Alexander Astin's *Assessment for Excellence*, which states:

> . . . I have repeatedly suggested that our assessment practices in higher education will serve our talent development mission much more effectively if we do less meritocratic assessment (for purposes of ranking, screening, and certification) and more assessment designed to benefit the assessee (i.e., to provide feedback to the learner and to enlighten the practitioner).

This is the spirit in which the CSU accountability project is or should be framed. Astin stresses that "assessment" has a dual aspect. It entails measurement and is intended to improve the functioning of students, staff, and institutions. Words matter, however, and accountability and assessment are not the same thing. "Accountability" emphasizes counting for purposes of answering to a higher authority, while "assessment," at least as defined by Astin, refers to information gathered with the aim of self-improvement for the sake of excellence. Excellence arises from within; it is seldom imposed from without. Accountability leads to a priori and sometimes questionable specification of indicators for measuring broad organizational trends, while assessment is or should entail ongoing exploratory and ad hoc analyses of what works and does not work for particular operations and in particular departments.

But I digress. The main point of the paper is that the statistical principles developed for quality control indicators apply as well to indicators developed under accountability plans. There is a definite statistical discipline that applies to the monitoring of quality indicators, but in the implementation of the CSU accountability process, it is not always being understood by campuses nor being respected by the system. At the very least, baseline levels of accountability indicators should be reliably measured and their variability should be reported along with their baseline levels. Taking the variability into account, measurable objectives for improvement should be detectably different—i.e., different in the sense of representing statistically significant changes—from the baseline levels. Because of the unavoidable variability of accountability indicators, it may not be possible to show statistical improvement in a span of one or two year intervals, particularly in campus data, where the sample sizes are so much smaller and statistical power is so much lower than in systemwide data. However, incremental improvements should result in changes in the level of an indicator time series in the long run.

Appendix 1. Standard Errors for Proportions

The average number of regularly admitted freshmen is about 985, and any given class should be considered as a more or less random sample of all the possible classes of size 985 that *might* have attended. The fact that a *particular* class attended in a given year is important, but we should think of it statistically as a sample from a population of all the possible students who meet our criteria and might have attended in that year.

With the average size of freshman cohorts at 985, the standard error **SE(P)** for a rate **P** in samples of size **N** from a population in which the true rate is 82% is about

$$\begin{aligned} \mathbf{SE(P)} &= \sqrt{(\mathbf{P} \times \mathbf{Q} \;/\; \mathbf{N})}, \text{ where } \mathbf{Q = 100\% - P} \\ &= \sqrt{(\mathbf{82\% \times 18\% \;/\; 985})} \\ &= \mathbf{1.2\%.} \end{aligned} \qquad \textbf{[1]}$$

Under normal theory, a 68% (or roughly two-thirds) confidence interval for a percentage is estimated as the population value plus or minus the standard error—i.e., only about one third of observations from this population would lie *outside* the interval from 80.8 to 83.2%, with about one sixth on either side of the interval. Thus, through purely random factors from this population, the next point in the series would exceed 83.2% about one in six times ("meeting" the 83% four-year goal). However, the next point will also be randomly smaller than 80.8% (not only failing to "meet" the goal but also dropping by a full percentage point below the baseline) about one in six times. Because a 68% confidence interval implies that nearly one-third of random values will lie outside the interval, it is usual instead to use a wider interval, often a 95% confidence interval, which can be estimated as the population percentage plus or minus two errors.

We should mention several statistical caveats about formula [1], which is a large-sample approximation. The formula assumes that the distribution of observed percentages in samples is approximately normally distributed around the population percentage. The assumption breaks down with small samples, with non-random samples, and with percentages that are very close to either 0% or 100%. In the case of small samples, the discreteness of the data prevents the measures from taking on a normal distribution. If samples are not strictly random and of the same size, as in observations on successive freshman classes, then variability in the size and characteristics of the freshman class may do more than campus practices to affect continuation rates. Finally, extreme percentages tend to arise from situations in which the normal distributions do not apply. In a manufacturing situation with error rates on the order of 1 or 2 per 1,000, for example, where rates can theoretically go as low as zero (but never lower) or as high as 100% (but never higher), observations averaging close to the lower or upper limits will follow a Poisson or other asymmetrical distribution.

We can work the formula for the standard error of a rate backward to determine the sample size that will be needed to achieve a confidence interval of a given size. To reduce the 68% confidence interval to ± 0.1% where the population rate is 82%, for example, we would need

$$\mathbf{N = P \times Q \,/\, SE^2(P) = 82 \times 18 \;/\; 0.1^2 = 147{,}600 \text{ freshmen}}, \qquad \textbf{[2]}$$

which is somewhat larger than our usual freshman class. We would need over 100 years of campus data to achieve this sort of precision for the baseline, let alone for meeting the objective. Obviously, we cannot do it. To reduce the precision to ± 0.5%, however, we would need only 5,904 observations, or about six years of data. Because standard errors vary as the square root of sample size, we can establish rules of thumb such as the following: We need four times as many observations to reduce the standard error to one half of its former size; we need one hundred times as many observations to reduce the standard error to one tenth; and so on.

In systemwide analyses of pooled campus data, which have perhaps 20 times as many observations as we have at our campus, standard errors will be about 22% or $1 / \sqrt{20}$ as large as standard errors at the campus. As this may suggest, accountability indicators are imperfectly scalable. At a statewide or systemwide level, a one percent change in an indicator may affect hundreds or thousands of individuals and have considerable economic impact (provided, of course, that we have measured accurately). As the indicator is divided among smaller organizational units, however, precision of estimate varies widely and percentage indicators become more variable and subject to random factors. To take an extreme example, consider a small department at a given campus, which may get only five to ten freshmen a year (some of whom will be undeclared or undecided), where it would be statistically silly to talk about achieving "one percent" changes in freshman persistence.

Appendix 2. Standard Deviations and Standard Errors for Rating Scales

Unlike percentage indicators, where standard deviations are a function of the sample sizes and the percentages themselves, the standard deviations of rating scale indicators depend on factors such as the number of response categories, the labeling or "anchoring" of the categories, and the empirical distributions of responses in the categories. The following table shows the standard deviations of means from symmetrically distributed responses to scales of various sizes, for the cases where response rates for distinct scale categories have a binomial (the nearest approximation to a normal distribution with chunky or discrete data), uniform (equal numbers of responses in each category), or extreme (having equal numbers of responses in the two extreme categories. For any scale, the minimum possible standard deviation is 0, which occurs in the rare case when all responses fall in a single category, and it can be shown that the maximum possible standard deviation occurs in the extreme case with equal numbers of responses in the two extreme cases. Thus, for a three-category scale with responses coded 1, 2, or 3, the standard deviations can range between 0.0 and 1.0; for a four-category scale with responses coded 1, 2, 3, or 4, the standard deviations can range between 0.0 and 1.5; and so on.

Standard Deviations of Symmetrically Distributed Responses to Scales of Various Sizes

	Items in scale				
Distribution Type	**3**	**4**	**5**	**6**	**7**
Minimum possible	0.00	0.00	0.00	0.00	0.00
Binomial	0.71	0.87	1.00	1.12	1.22
Uniform	0.82	1.12	1.41	1.71	2.00
Maximum possible	1.00	1.50	2.00	2.50	2.62

Thus, uniformly distributed responses yield larger standard deviations than binomial ones, and longer scales yield larger standard deviations than shorter ones. The minimum possible standard deviation is 0.0 (with all responses in a single category) regardless of the number of categories, but the maximum possible standard deviation is a function of the number and scaling of the response options. We can make the following statistical generalizations about five-point scales, which are among the most commonly used scales for measuring satisfaction and attitudes generally.

- Standard deviations of five-point scales can be as small as 0.0 (all responses the same) or as large as 2.0 (equal numbers of responses in the extreme categories 1 and 5), but one seldom sees these extremes in real data.
- With binomially distributed data (which are approximately normally distributed) and a five-point scale, the population standard deviation is exactly 1.0.
- For a uniform distribution, where all five response options are chosen equally, the standard deviation will be higher than 1.0.

- For responses skewed in one direction or the other, the standard deviation can be either larger or smaller than 1.0.[25]
- Scale types and the words used to label or "anchor" scale categories affect distributions of responses and therefore standard deviations as well. In the systemwide 1999 SNAPS survey, where there are four major groups of five-point items using three types of anchors, we have the following:

Minimum, Average, and Maximum Standard Deviations Found for Selected Five-Point Scales, SNAPS 1999

Type of Item	Low anchor	High anchor	Min	Avg	Max
Agreement	Strongly disagree	Strongly agree	0.82	0.90	1.03
Importance	Not important at all	Very Important	0.94	1.31	1.59
Quality-Academic	Very poor	Excellent	0.70	0.85	1.11
Quality-Services	Very poor	Excellent	0.84	0.96	1.21

Thus, the "agreement" items (0.90 average standard deviation) are less variable than the "importance" items (1.31). Ratings of academic quality are somewhat less variable (0.85) than ratings of service quality.

In my experience, one can usually take 1.0 as a reasonable and slightly conservative estimate of the standard deviation for five-point quality or agreement scales. This should always be checked against actual data, but when it is true, it enables us to establish rules of thumb about standard errors that then depend only upon sample sizes.

With a population standard deviation estimated to be σ for a scale, the standard errors for means and differences between independent means are approximately:

SE(Mean) = $\sigma / \sqrt{N}$, where N is the sample size;
SE(Mean Difference) = $\sigma \sqrt{(N_1 + N_2) / (N_1 N_2)}$, where N_1 and N_2 are sample sizes for the two groups; and
SE(Mean Difference) = $\sigma \sqrt{(2 / N)}$, where the sample sizes for the two groups are equal.

[25] Faculty and coursework ratings generate some of the most skewed ratings I have seen, with means hovering around six on a seven-point scale. I suspect that this is because students do not really believe that there is no relation between their ratings and the grades they will receive. Empirically, the students are correct. The last time I looked at it, the correlation between average grades and average course ratings was about 0.32. As we have been taught, however, correlation does not imply causation.

If the means come from dependent or correlated measures of the same students, we should reduce the standard errors for differences obtained by the above formulas by about one-half the correlation between the variables being compared. For example, we should reduce the standard error by one-fourth if we think that the two sets of data have a correlation as high as 0.50. If the empirical correlation is zero, of course, the formulas for independent groups apply.

Using the formula for the standard error of a mean difference with independent groups allows us to construct a table of the sample size needed to have an 80% chance of detecting a population mean difference of a given size, at the 0.05 significance level:

Sample Sizes Needed in Each Group to Detect Selected Mean Differences in 5-Point Scales at the 0.05 Level with 0.80 Power where the Within-Group Standard Deviations Equal 1.0*

Size of Mean Difference to be detected	0.05	0.10	0.20	0.30	0.40	0.50
Sample Size ($\sigma = 1.0$)	6,279	1,570	393	175	154	99

*See Cohen (1977) or http://www.health.ucalgary.ca/~rollin/stats/ssize/n2.html.

Thus, to detect a mean difference as small as 0.10, one generally needs a sample size of at least 1,570 in each of the two groups.

Appendix 3. Standard Errors for Differences in Upper Division Transfer Units Earned

This accountability indicator is based on the differences in *upper division* units earned by the two kinds of students, defined as

$$M_{diff} = (M_{gt} - M_{jt}) - (M_{gf} - M_{jf}),$$

where

M_{gt} represents average cumulative units earned at graduation by junior transfer entrants
M_{jt} represents average cumulative units earned prior to entry as a junior transfer entrant
M_{gf} represents average units earned at graduation by freshman entrants
M_{jf} represents average units earned prior by freshmen prior to becoming CSU juniors

Thus, the indicator is a difference between two differences between means. Since the mean units earned at graduation and at the first CSU term as a junior for the same students are considered correlated means, while the means for transfer and freshmen entrants are uncorrelated, the standard error for the difference between differences is estimated as

$$SE_{diff} = \sqrt{(\sigma^2_{gt} + \sigma^2_{jt} - 2\,\sigma_{jt,gt}) + (\sigma^2_{gf} + \sigma^2_{jf} - 2\,\sigma_{jf,gf})},$$

which is the square root of the sum of squared standard errors for each of the four components, minus correction terms that arise from the correlations between mean units at junior level and graduation for the two groups. However, the two correlations between mean units are quite small, so as a practical matter, the standard error of the indicator is approximately the square root of the sum of the four variances of estimate. What does all this mean in terms of real data?

Call the two groups (freshmen and junior transfers) FTFs and JCCTs. Over the short three-year times series available, we have the following:

- At graduation, FTFs average 216.0 units and JCCTs average 229.1 units, a difference of 13.1 units.
- At the time that the students became juniors at the CSU, the FTFs averaged 103.7 units and the FTFs averaged 96.3 units, a difference of 7.3 units. (Because students become juniors in the term that they have earned 90 or more quarter units and undergraduates average about 12 units per term, native freshmen will have about 96 units when they formally become juniors.
- It follows that the FTFs have about 119.8 upper division units and the JCCTs have 125.4 upper division units, as defined.
- The average value of the indicator, the difference between JCCT and FTF upper division units, therefore, is
  125.4 – 119.8 = 5.6.
- These statistics are based on an annual average of 586 FTF students and 329 JCCT students.

Appendix 4. Scaling of Entry Level Math (ELM) Scores

The ELM is currently based on 65-item tests with 60 scored and 5 experimental items, but scores are reported on a 51-point scale, ranging from 200 to 700 by 10s. The rescaling from raw scores to scaled scores is intended to allow scores on each new version of the ELM to be equated with scores on older versions at the selected remediation cutpoint (currently a scaled score of 550). The test equating is performed by Educational Testing Service (ETS), the vendor that develops, publishes, scores, and equates successive versions of the ELM under direction of the ELM Advisory Committee (ELMAC) of the CSU.[26] A primary text on test equating was edited by Holland & Rubin (1982) and contains the collection of technical papers presented at a 1980 ETS Invited Conference on Test Equating.

The ELM proficiency criterion is not based on number of items correct—so many correct answers out of 60 scored items, for example—but as noted above, the ELM raw scores are mapped to scaled scores ranging from 200 to 700. In a version of the test being used in 1998, for example, the scaled scores (S) were obtained from raw scores (R) as

$$\mathbf{S = 183.27003 + 10.37676\ R\ ,\ rounded\ to\ nearest\ 10} \qquad \mathbf{[3]}$$

where R represented the number of correct responses by a given student (without correction for guessing) and the result was rounded to the nearest 10 points. Thus, a score of 34 correct corresponded to a scaled score of 540 (needing remediation) and 35 correct to a scaled score of 550 (proficient). Near the 550 cutpoint between needing remediation and proficiency, approximately 3% of students are at successive scale points (540, 550, 560, etc.). Thus, a shift of the cutpoint by one raw score point in one direction or the other will result in about 3% of the students being classified differently, and the ability of ELM scores to discriminate between students needing and not needing remediation is limited by the coarseness of the scale at the cutpoint.[27] The problem of precision is rather like having a digital readout of weight to the nearest pound. With such a scale, it is never possible to infer anything about inches, so the margin of error is up to one-half pound because of the limits of precision on the scale. In the case of the ELM, the precision of the percentage needing remediation cannot be inferred to within more than the 3% precision (plus or minus 1.5%) with scores from any given version of the test.

In setting goals for remediation need, one should also allow for the fact that a given ELM cutpoint detects varying degrees of math ability at successive test occasions. This is a problem of accuracy or reliability. As a homely example, when I weigh myself 20 times in succession on

---

[26] References on text equating include the collection of technical papers presented at a 1980 ETS Invited Conference on Test Equating and edited by Holland and Rubin (1982) and a text by Kolen and Brennan (1995), who are affiliated with ACT.

[27] Formula [3], with coefficients stated to five decimal places, has a great deal of misplaced precision, since students taking this version of the test will answer either 35 or more items correctly and be considered fully prepared, or will answer 34 or fewer items correctly and be considered to need remediation.

my bathroom scale, which I have done to get data for a teaching exercise, the weights have a standard deviation of about 2 and a range of 5 or 6 pounds. The scale has a digital readout to the nearest pound, so it has a precision of ± one-half pound, while the overall accuracy against a more accurate standard is ± 2 pounds. Admittedly, the bathroom scale is a cheap, spring-loaded device and there are better scales available. However, measuring a person's weight is conceptually much simpler than measuring a person's math proficiency. There is a great deal of subjectivity that goes into the selection of ELM items, content, and equating methods from one test administration to the next, and it would surprising if the accuracy of the percentage of students needing remediation could be determined to within more than 2% or 3% against an ideal standard.

Notice that imprecision and inaccuracy are features of the scale, and have nothing to do with sampling error arising from the fact that different groups of students are taking the test in different years. At a systemwide level, sampling error may be negligible because the numbers of entering freshmen are so large, but at a campus level, with an average of about 1,000 freshmen entering each year, sampling error is an additional error factor on top of imprecision and inaccuracy of the scale.

Test equating and difficulty levels. In the 1992 revision of the ELM, when Algebra II content was added, *items* were specified to have an average passing rate of about 60%, which means that 60% of students in the standardization group would be likely to answer the *average item* correctly.[28] In addition, the remediation criterion was set at about a level such that passing about 60% of the items would indicate proficiency (e.g., in the equating formula for the version of the ELM cited above, 35 out of 60 represents about 58% of items passed.) These two facts taken together—that 60% of students will pass the average item and that passing 60% of the items is the cutpoint for proficiency—imply that about 50% of students in the standardization group will be deemed proficient and the other 50% will be deemed to need remediation. This is what happened and what should have been expected to happen after the 1992 ELM revision. The standardization sample (presumably representing a more or less random sample of high school graduates) was passing the ELM at about a 75% rate prior to 1992, and with the test being made more difficult, we would expect the passing rate in the standardization group to drop.

---

[28] Item "difficulty" is a technical item attribute that is independent of item content. It is entirely possible to write "arithmetic" items that are more difficult than "algebra" items. One can theoretically imagine a situation in which a test measuring pure math content has one-third of the math items replaced with items measuring something unrelated—perhaps vocabulary knowledge—but at the same level of difficulty, so that after doing this three times one would have a vocabulary test that fails as many individuals (though probably not the *same* individuals) as the original test did. Because item difficulty is independent of content, items designed to measure "exponentiation" knowledge, for example, can be either easy or difficult:

Easy: 2^0 **Difficult:** $(x^2y^{-1})^2\,(2xy^2)^0$

It is obvious that the first item is easier to evaluate correctly than the second item is, not necessarily because of conceptual understanding of exponents, but because the second item provides more opportunities for simple computational or combinatorial errors. By varying item

A recent LAO (2001) report on remediation has a section labeled "Almost Half of Regularly Admitted CSU Students Arrive Unprepared in Mathematics," or as an *L. A. Times* (March 27, 1998) headline stated, "Almost Half of Cal State Freshmen Lack Skills." The headline should probably have read, "Cal State Designs Tests That Fail Half its Freshmen." As the LAO explains it:

> In fall 1989, 23 percent of regularly admitted freshmen were unprepared for college-level work. In 1992, CSU increased its admission standards to require three years of college preparatory mathematics. To reflect this change, CSU made the ELM *more difficult by including questions on data interpretation, counting, probability, and statistics.* [This is not correct. These are changes that actually occurred after 1998, when these content areas were enhanced and the test was made easier.] Likely as a result of these changes, the number of regularly admitted students needing remedial classes in mathematics jumped from 26 percent in fall 1991 to almost 40 percent in fall 1992. From 1992 to 1998, the unpreparedness rate continued to climb, reaching 54 percent in fall 1998. (LAO, 2001, p. 4; my italics).[29]

The test was made more difficult, independent of content, because of two facts: First, the contract with ETS stated that they were to produce an exam with an average item passing rate of about 60%; second, mathematics faculty went through an exercise that ensured that students would need to pass 60% of the items in order to be deemed proficient. As noted above, it was the confluence of these two facts that ensured that about 50% of freshmen in the standardization group would fail.[30] In the absence of changes in student skills or CSU selectivity, the raising of the standard ensured that about 50% of future freshmen would fail as well.

---

details, exponentiation items can be written to almost any desired level of difficulty. The difficult exponentiation item is an actual ELM item, published as part of a 30-item test representing ELM domain content over the period fall 93-fall 98 and used for equating in the pilot test for the revisions that occurred between fall 98 and fall 99 revisions.

[29] The LAO paper argues that the CSU has a "perverse incentive" to offer remedial courses, since they are cheaper than some other courses, but are funded at the same level. This is somewhat like saying that there is a "perverse incentive" to offer undergraduate rather than graduate courses, to offer GE courses rather than major courses, or to offer any less expensive courses rather than more expensive ones. Campuses should have latitude in balancing high-cost instruction with low-cost instruction in ways that enable them to meet the needs of their students.

[30] Technically, ETS was asked to "assemble [ELM items] to a Mean Equated Delta in the range of 11.9 to 12.3 with a standard deviation of 1.9 to 2.2." Translating the obscure metric of "mean equated deltas" to percent passing, the contract required ETS to modify the ELM to have an average item-passing rate between 58% and 61% (which is why I say that we told ETS to prepare a test with an average item passing rate of about 60%) and with about two-thirds of the items having passing rates between 38% and 78% (which is a fairly narrow range, consistent with a test that is optimized for discrimination at around the 60% passing rate). A second thing that occurred is that ETS had mathematics judges estimate the percentage of skilled students who "should" pass each item, in what is called a "Modified Angoff cut-score study," sometimes

I have referred several times to a "standardization group." When the ELM was launched in 1988, there were in fact samples of students used to standardize the test, by determining item passing rates and test score cutpoints that would indicate "proficiency." All subsequent versions of the ELM are referred back to the range of skill levels in this original sample. It is important to understand that test equating works because only about one-third of items in any given version of the ELM are new, so that when a new version is administered, the 20 new items can be statistically equated so that their average difficulty is about the same as the average difficulty level for the 40 old items.[31]. The hope in changing the ELM standard from a 75% passing rate to a 50% passing rate was that the population of students would change its behavior so that new students would do what they need to become more proficient. However, if the students do *not* change their behavior or average tested skill levels, then given the way the ELM was changed after 1992, it is guaranteed that we will continue to find that only half of students are proficient. By requiring three years of high school math instead of just two years for regular admission, those students taking fewer than three years of math get dropped from the pool, and this should raise the average tested ability level. However, the original standardization group had a great many students taking three and four years of math already, so it is not as if we had a population of students taking only two years of math and we are now testing in a population of students taking at least three years of math.

---

loosely referred to as "norming" the test, and this exercise determined that their collective opinion was that mathematically prepared students would pass about 60% of the items. A third decision was to equate this 60% passing rate with a numerical ELM score of 550, which is also the proficiency criterion. The confluence of these three technical decisions ensured that about 50% of existing students—those in the pre-1992 standardization group—would now fail the ELM. Since changing criteria on a proficiency test has little effect on actual proficiency, we are still failing close to half of entering freshmen. In 1997, 85% of students who took the ELM were failing it. More students are deemed proficient by exemption than by taking the ELM and passing it, so that the students who are required to take the ELM tend to be less proficient than average. The ELM is a *difficult* test.

[31] The equating process is not error-free, so that Angoff (1982) refers to "standard equating error," which is cumulative error that arises through the "daisy-chain" of equated tests. Normative studies are needed to guard against standard equating error, which should involve testing randomly selected students with alternative versions of the test. Periodically, math faculty trained by ETS go through an exercise in which they judge the percentage of students who *should* pass each item on a given version of this test, and this is called a "norming study" by ETS and the participants, but this exercise does not logically entail that the same percentages of students in a given population *will* pass the items.

References

Angoff, W.H. "Summary and derivation of equating methods used at ETS." In Holland & Rubin, *Test Equating,* pp. 55-70.

Astin, Alexander W. (1993) *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education.* Phoenix: American Council on Education, Oryx Press.

Birnbaum, Robert (2000) *Management Fads in Higher Education: Where They Come From, What They Do, Why They Fail.* San Francisco: Jossey-Bass.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Rev. ed.) New York: Academic Press.

Deming, W. Edwards (1986) *Out of the Crisis.* Republished at Cambridge MA: MIT Press, 2000.

Deming, W. Edwards (1994) *The New Economics for Industry, Government, Education* (2nd ed.) Cambridge MA: MIT Press.

Holland, P.W., & Rubin, D.R. (eds.) (1982), *Test Equating*, New York: Academic Press.

House, Ernest R. (1998) *Schools for Sale: Why Free Market Policies Won't Improve America's Schools, and What Will*, New York: Teacher's College Press.

Jordan, L. (1998a) "Technical Notes on the Operating Characteristics of the Entry Level Math Test." Paper presented at a meeting of the CSU ELM Advisory Group, August, 1998.

Jordan, L. (1998b) "Spring 1998 ELM Pilot Test—Comments and Analysis." Paper presented at a meeting of the CSU ELM Advisory Group, August, 1998.

Kolen, M.J., & Brennan, R.L. (1995). *Test Equating: Methods and Practices.* New York: Springer

Legislative Analyst's Office (2001). "Improving Academic Preparation for Higher Education," Sacramento: LAO. (http://www.lao.ca.gov/2001/remediation/020801_remediation.html).

Monk, D.H. Education productivity research: An update and assessment of its role in education finance reform. *Educ. Eval. & Pub. Pol.*, 1992, *14,* 307-322.

National Center for Educational Statistics. *Remedial education at higher education institutions in Fall 1995 (NCES publication 97-584, by Laurie Lewis & Elizabeth Farris; Bernie Greene, project officer).* Washington DC: NCES, 1995. (Summarized in the National Center for Education Statistics, *The condition of education, 1997 (NCES publication 97-388).* Washington DC: NCES, 1997, pp. 102-103.

Shewhart, Walter A. (1939) *Statistical Method from the Viewpoint of Quality Control.* Edited and with a new foreword by W. Edward Deming, New York: Dover Publications, 1986.

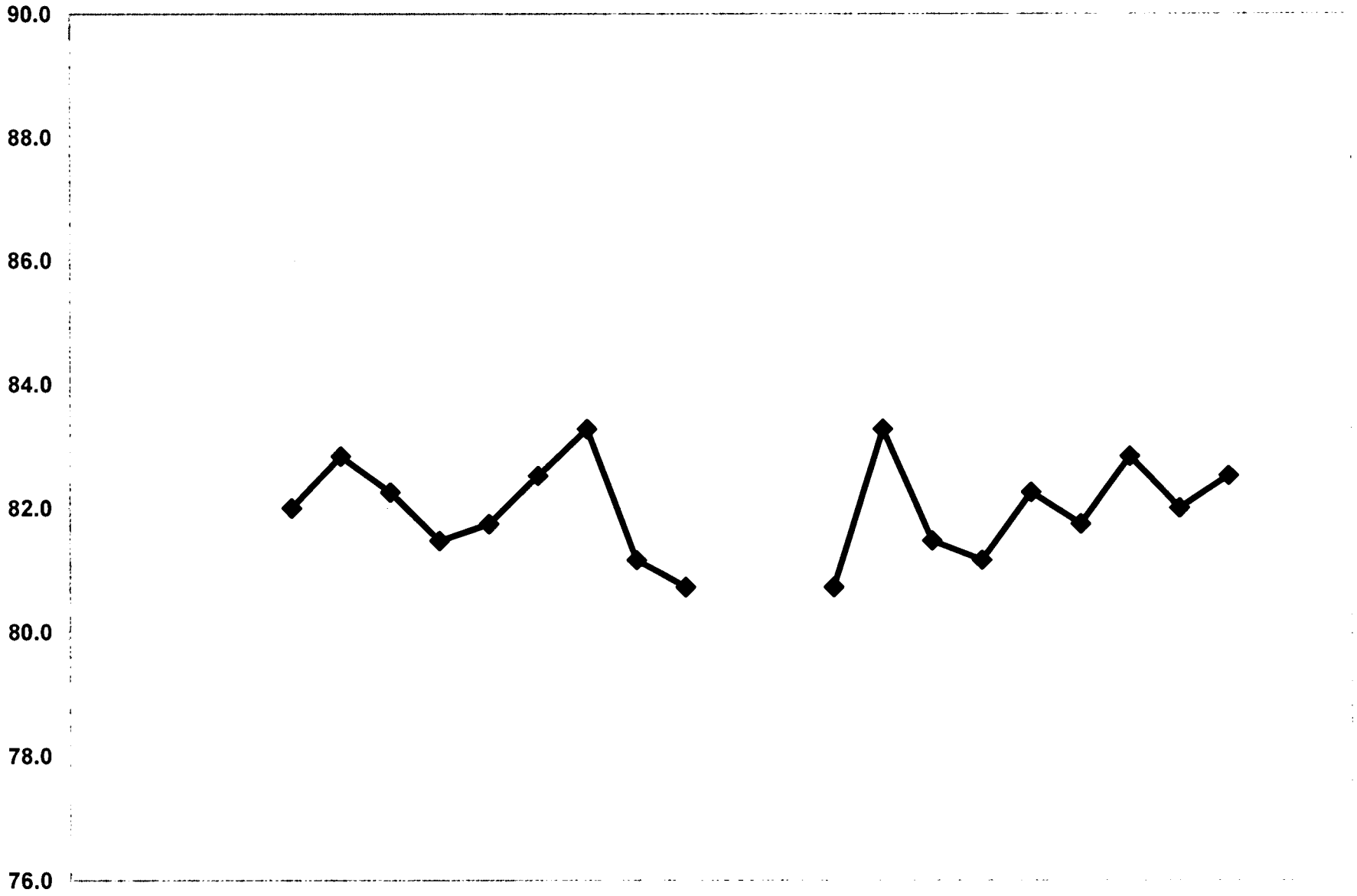


Figure 1. Two random time series generated with a mean of 82% and a standard deviation of 1%

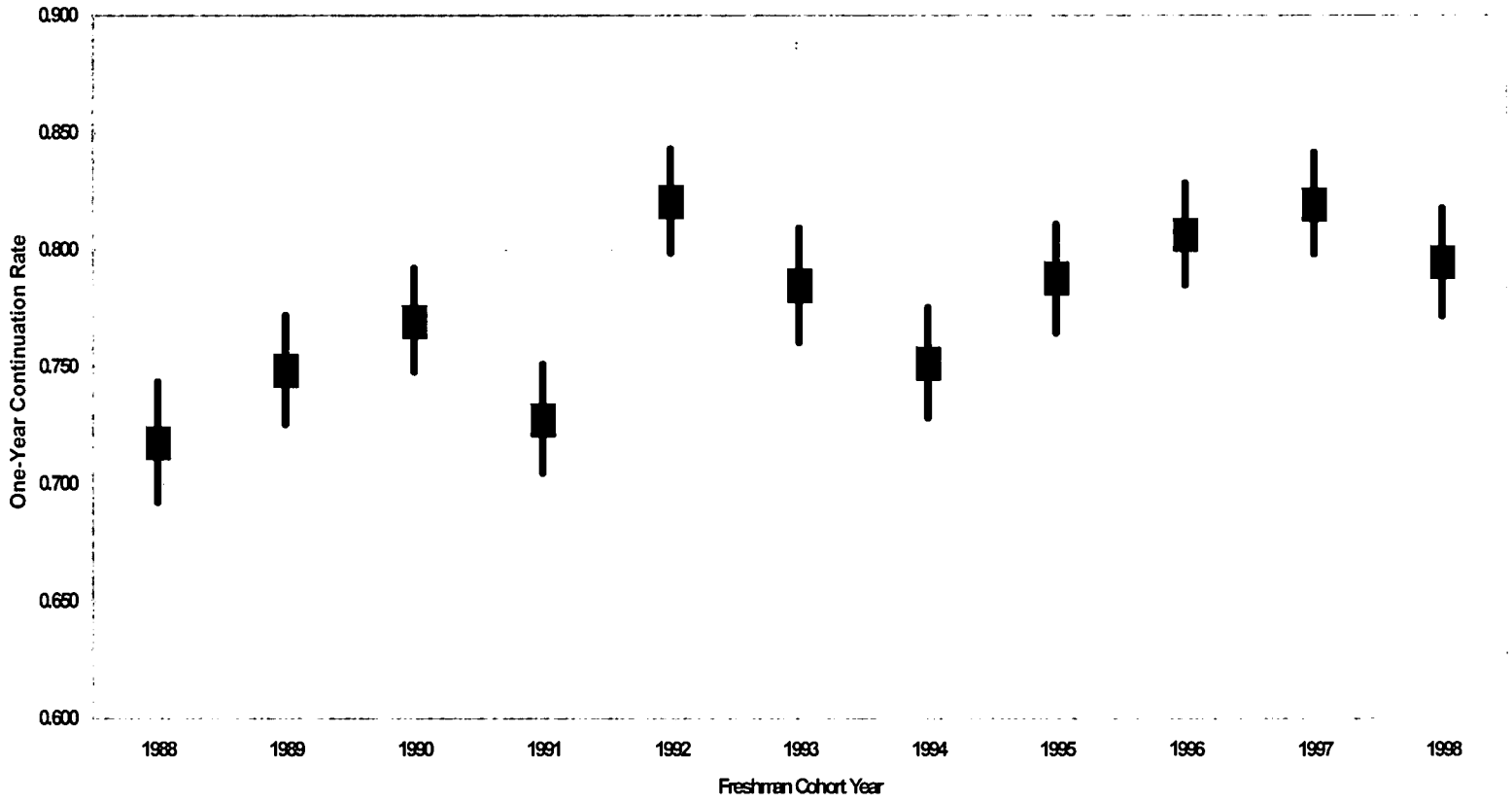
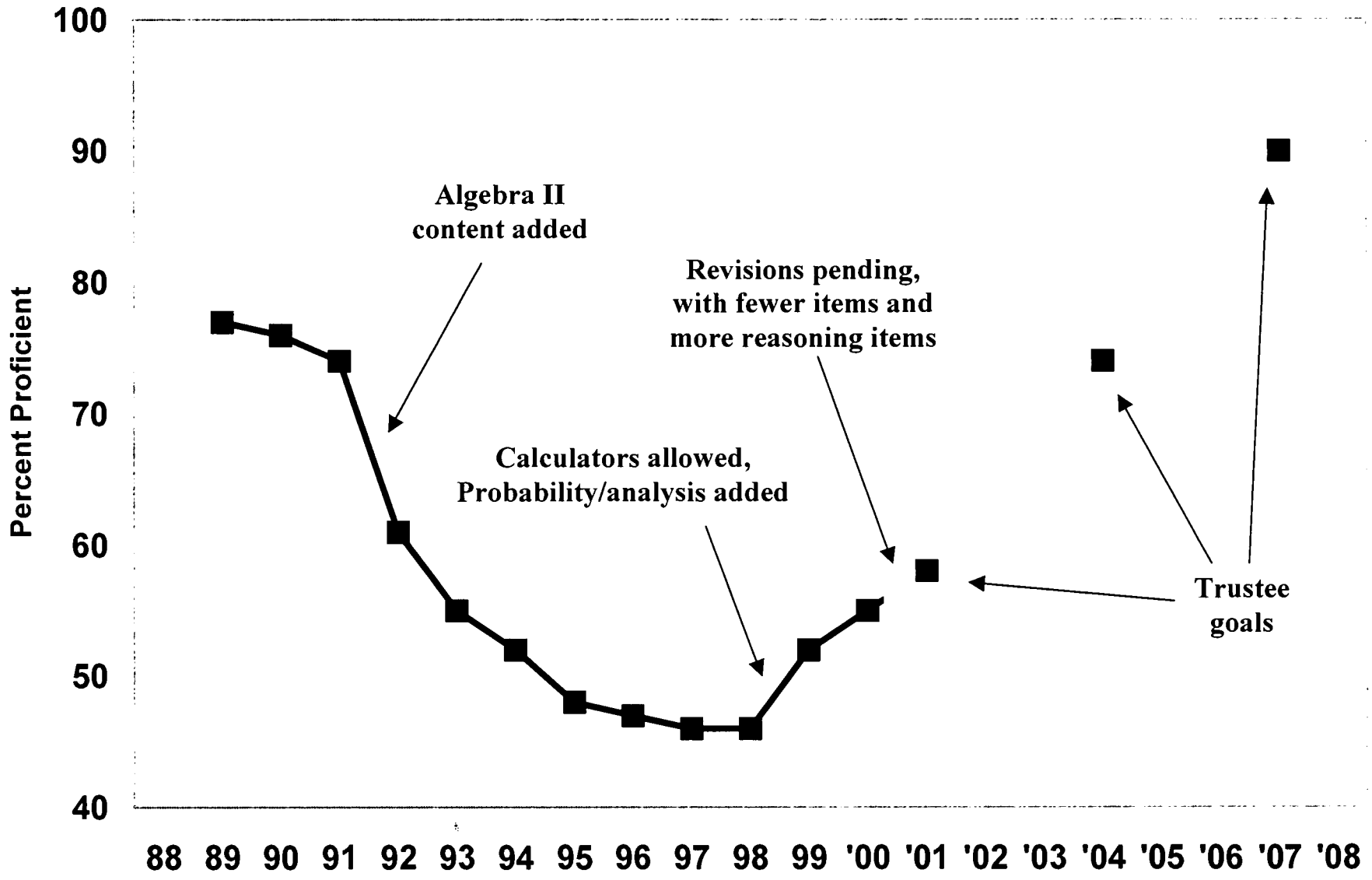


Figure 2. One-Year Continuation and 95% C.I.s, Freshmen, Fall 88 to Fall 98 Cohorts

Figure 3. % Freshmen Needing Math Remediation, Fall 89 to Fall 00 with Goals for Fall 01, 04, & 07.



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").