DOCUMENT RESUME

ED 187 729, TM 800 224

AUTHOR Merkel-Keller, Claudia
TITLE Parameters of Quality Control and Decision Making At the State Level.
PUB DATE Apr 80
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Compensatory Education; Elementary Secondary Education; *Models; *Quality Control; School Districts; *School Statistics; State Departments of Education; Test Results
IDENTIFIERS Elementary Secondary Education Act Title I; New Jersey; RMC Models; *Title I Evaluation and Reporting System

ABSTRACT

The recommendation is made that Elementary Secondary Education Act Title I data should be subjected to quality control procedures at local, state, and national levels. An industrial quality control model may provide a useful approach, particularly at the state level. A brief description of the Title I Evaluation and Reporting System is given, including mention of the three RMC Models and the use of a normal curve equivalent scale. The aggregate reporting model is described in which schools report to districts, which report to states, which in turn report to the federal level. Shewhart control charts are recommended for one phase of quality control as a means of identifying statistical outliers. For another aspect of quality control, a list is provided of items to be checked in verifying the data received from school districts by the state education agency. (CTM)

SYMPOSIUM FROM IMPLEMENTATION TO UTILIZATION: TITLE I EVALUATION IN 1980

PARAMETERS OF QUALITY CONTROL AND DECISION MAKING AT THE STATE LEVEL

Claudia Merkel-Keller
New Jersey State Department of Education

Paper presented at the Annual Meeting of the American Education Research Association, Boston, Massachusetts, April 7 - 11, 1980.

TABLE OF CONTENTS

Thanks are expressed to Gerald De Mauro and Margaret Hoppe of the New Jersey State Department of Education for their comments during the course of the preparation of this manuscript.

Background:

Since the inception of ESEA Title I in 1965, Title I programs have been providing annual evaluation reports to their respective states, and the states in turn to the federal government. These reports individually provided information about both the program processes and program outcomes. Yet, collectively the data from these individual reports could not be aggregated at any level.

Of course from a policy and decision-making level, both state and national, the question arises "Is the program working?" that is to say "Are the children learning more as a result of the program?" These questions can be extrapolated to the following budgetary one, "Is the dollar input per program equaling the dollar output of the program?" which again is more succinctly stated "Are we getting the bang for the buck?" Directly or indirectly all of these questions, both pedagogical and fiscal, gave rise to Congress' plea and demands of the United States Office of Education to develop an evaluation and reporting system which would reflect a systematic attempt to formalize data collection and reporting practices across the states. The intent was to ensure that uniform, relevant and meaningful data would be available to educators at the local, state and national levels.

USOE's response to the Congressional mandate was the development of a system (Title I Evaluation and Reporting System-TIERS) which in essence is designed to collect and summarize data on six topical areas covered by the Title I program including: student participation, staffing, parental involvement (PAC), in-service training, cost and student impact. The initial thrust of the system is on student outcome data in projects providing instructions in the areas of reading, language arts, and mathematics.

The system which legally became effective in the fall of 1979 is comprised of three outcome evaluation models or designs:

Model A: the norm-referenced model
Model B: the control-group model
Model C: the special regression model

All three models are designed to be used with any valid and reliable norm-referenced test or criterion-referenced test. Additionally, each of the models requires both pretesting and post-testing and imposes some special conditions and restriction on the testing itself. The three models each provide data on an observed post treatment performance measure and an estimate of what that performance would have been without the program (i.e., without the treatment).

Impact gains are reported in what is termed a Normal Curve Equivalent (NCE) scale which is a 99-point scale tied to a distribution of test scores of a nation-wide representative sample of students and matches the percentile ranks of that distribution at values of 1, 50 and 99.

As designed, TIERS begins data collection at the project (unique combination of personnel, resources, methods and activities that define a particular treatment) level or school level. With Model A some preliminary analysis of impact data occurs at the school level. Data are then aggregated and analyzed at the LEA with the resultant analysis reported to the state. The LEA then in turn aggregates its data and reports it to the federal government.

Issue:

The system as described, although having a number of rigorous technical and implementation rules, is basically a decentralized evaluation and reporting system that peaks or pyramids in terms of the data which initiates at the project level in a local district later to be captured in a national snapshot of program impact. The issue which looms paramount is that of quality control. That is to say how good are the data which are being aggregated at all levels and hence how valid are the policy decisions which can be derived from that data.

It is useful at this point to examine a schematic of the projected aggregate reporting model:

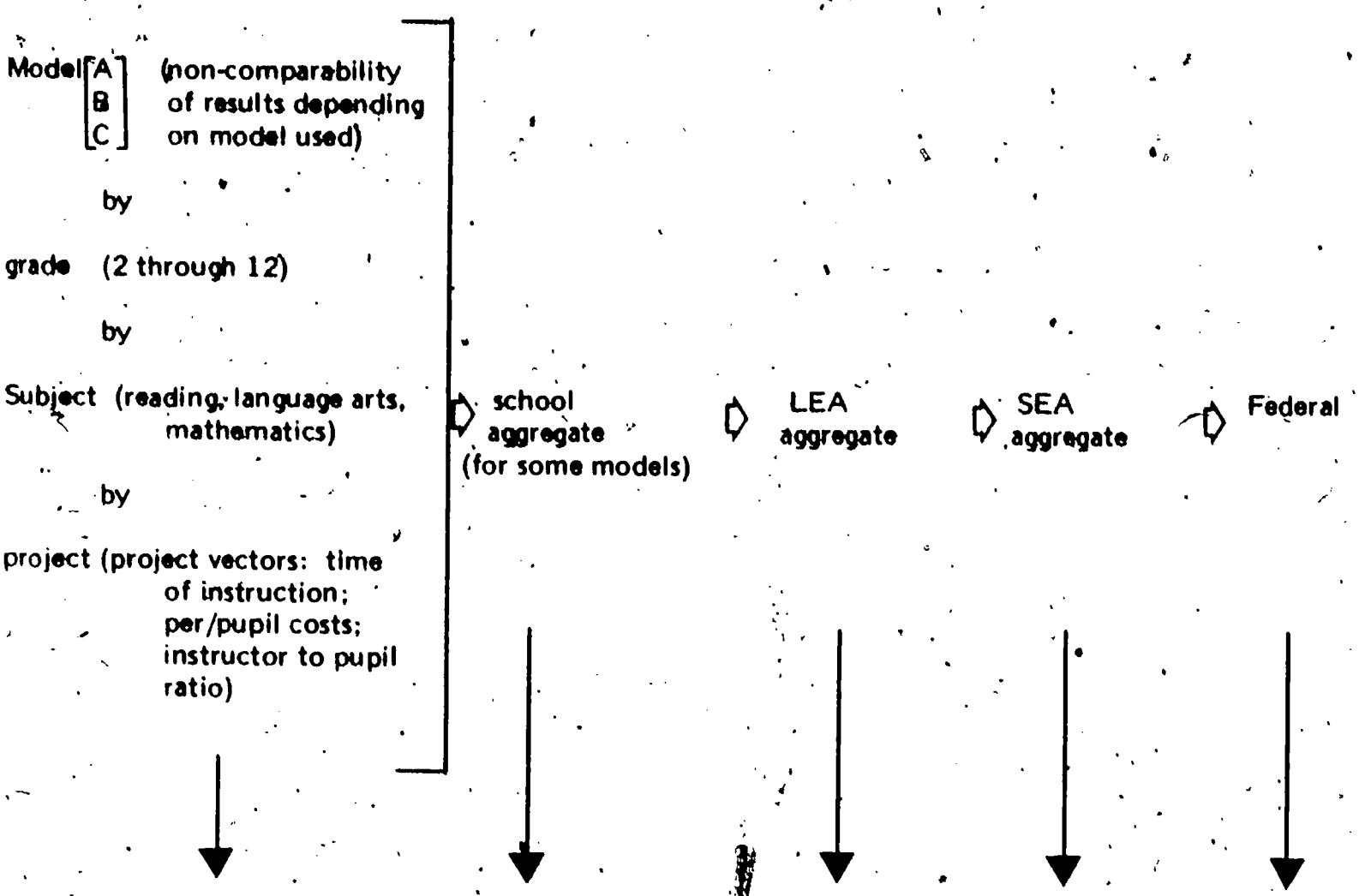


FIGURE I

AGGREGATE REPORTING MODEL

To date little to no work has been done at any level on the issue of quality control and quality assurance of the data which comprise the elements of the reporting system. As can be seen, there are three major data quality control points to be dealt with in the Quality Control System, with each reporting entity needing to grapple with the problem at its various stages of complexity.

The LEA must question the veracity of its data.
The SEA must question the veracity of the data it receives from its LEAs
The Department of Education must question the veracity of the data it receives from the states.

The reporting model is a cummulative model and hence if strict quality control procedures are not put into action we will cummulate data error to the federal level and be in no better postion to say anything meaningful on the output of Title I programs than we were able to say in 1974 (pre-TIERS).

Brief Literature Review:

Colleagues are beginning to conceptualize and deal with the issues of quality control of data and sources of data error. A scan of the literature related to what colleagues term "Quality control" is scant in terms of documentation and I believe focused more on implementation errors at the local level than on the concept of statistical quality control which will be explaned later.

Crane and Bellis (unpublished, 1979) analyzed some 40 district reports from Illionis Title I districts which used Model A-1. The purpose of their analysis was: (1) to determine the types of error made and to determine if possible the effects on NCE gain estimates, (2) to provide information to design quality control procedures. Their findings indicate high error rates in critical areas of technical implementation of the models as well as problems related to aggregate n's, floor and ceiling effects and the effects of error on NCE gain estimates. From their findings, they see quality control procedures as critical to further reduce the error rates that they found in analyzing their sample data. The Crane and Bellis data base are extremely useful for comparison purposes.

As recently as January of 1980 Hiscox and Deck in an unpublished paper discuss a comprehensive approach to the issue of Quality Control of Table I data. Their approach correctly points out that since the TIERS is a layered reporting system—LEA to the State, and the State to the federal government—that quality control requires consideration and resolution at all of these levels. They suggest a four stage model to pinpoint error—error in planning, implementation, analysis and reporting. The benefit if the work of Hiscox and Deck is in the classification schema of the potential errors by category and the suggested corrective action to be taken respective to the importance of the error type.

View on a Definition of Quality Control

It would seem to me that resolution of error in the various stages of LEA implementation of the evaluation and reporting system will be a bold step forward in establishment of a clean data base. None-the-less at the state level, which is the level concerning the focus of this paper, quality control takes on some added meanings. Given limitations of both time, personnel and resources most states including New Jersey must rely on a set of procedures (call them quality control procedures) to examine the aggregrated data base which they receive from their LEAs. Very few states have the sophistication of an evaluation data audit function or the overall computer capability to manage the summative analysis of individual student data at the state level. Quality control is the issue which confronts us. As I define it quality control techniques are superimposed on a data set after reasonable attempts have been made to satisfy the technical and imprelemnatiation concerns of the Title I Evaluation and Reporting models. I don't view debugging of problems in LEA planning, implementation analysis and reporting in the same light as I do "statistical quality control." Statistical quality control as I would like to apply it to an educational setting in essence is an industrial concept. Quite simply many of the techniques developed by mathematical statisticians for the analysis of data may be used in the control of product quality. The basic foundation of the statistical quality control model applied to industrial data are briefly outlined below.

Statistical Quality Control: The Industrial Model

Statistical quality control should be viewed as a kit of tools which may influence decisions which are related to the functions of specification, production or inspection. There are four separate but related techniques that constitute the most common working statistical tools in quality control. These tools are:

1. The Shewhart control charts for measureable quality characteristics. These are described as charts for variables, or as charts for $\bar{X}$ and R (average and range) and charts for $\bar{X}$ and σ (average and standard deviation).

2. The Shewhart control chart for fraction defective (p chart).

3. The Shewhart control chart for number of defects per unit (c chart).

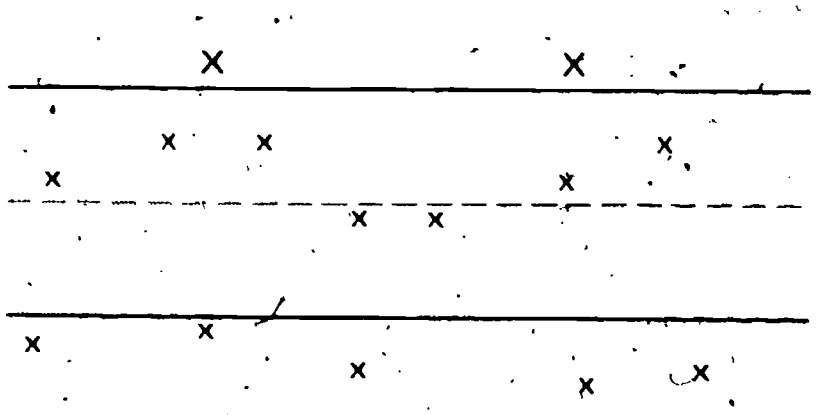
4. That portion of sampling theory which deals with the quality protection given by an specified sampling acceptance procedure.

As explained by W. A. Shewhart (1939, pg. 49); "Measured quality of manufactured product is always subject to a certain amount of variation as a result of chance. Some stable 'system of chance causes' is inherent in any particular scheme of production and inspection. Variation within this stable pattern is inevitable. The reasons for variation outside this stable pattern may be discovered and corrected." It is quite clear that the power of the Shewhart technique lies in its ability to separate out these assignable causes of quality variation. This is done post hoc through an examination of the outlyers which fall outside of the pre-established upper and lower quality control limits (however wide or narrow these tolerances are set).

Not only is the control chart a powerful technique for industrial applications but can also be applied to the educational setting as it specifically relates to the Title I data base.

Statistical Quality Control: The Industrial Model Applied

The concept basically is one of assuming that the range of output performance varies within an "acceptable range" (upper control limit and lower control limit) graphically reprsented as follows:



Performance of course can and does fall outside of the control limits and hence outlyer data points can be noted. Therefore, a control chart can provide the following types of information:

1. Basic variability of the performance characteristics;

2. Consistency of performance; and

3. Average level of performance.

The upper and lower control limits on any control chart can be established by examining the empirical data base and determining how much tolerance or variability one wishes to tolerate in the system. Data falling outside of the upper and lower control limit can then be examined as to why those data are showing up there.

Eg. If data falls above the upper control limit, this may signal (trigger) an examplary program for further review. (One of the established purposes of TIERS.)

If data falls below the lower control limit, this may signal (trigger) problems with a program or intervening variable which require further review.

New Jersey State Department of Education's Proposed Quality Control and Quality Assurance System:

The New Jersey State Department of Education is planning to implement a 3-phase overall quality assurance and quality control approach for its FY'80 compensatory data base (Title I, State Compensatory Education, and those compensatory programs funded by other sources) as follows:

Phase I: All New Jersey Basic Skills Preventive and Remedial Programs must satisfy the technical and implementation requirements of the evaluation and reporting models.

Additional materials and training are being made available to ensure a sufficient knowledge base at the LEA level. Initial compliance with the technical implementation and programatic implementation constructs delineated below will be determined through county office and SEA program monitors.

Interpreting the educational impact gains demonstrated through implementation of Model A^1, A^2, C^1, and C^2, requires consideration of technical and programmatic characteristics before sound decision-making and program planning can occur. In review of LEA data, the following issues must be considered:

1. LEA districts programs that show negative or no impact gain for a grade or subject should be reviewed. To assure gain is not reflecting just inaccurate or improper implementation, the following areas should be reviewed in detail.

Technical Implementation Criteria:

1. Test administration occurred at or near norming date for pre and post evaluation. If administration occurred more than two weeks on either side of the norming date, accurate interpolation procedures were implemented.

2. District mean was aggregated by:

   a. adding all students and then dividing for district mean.

   b. weighting class and building means to obtain LEA mean.

3. Data was aggregated on students who had both pre and post test scores.

4. Out-of-level conversions were implemented properly.

5. Floor or ceiling effects did not occur at pre or post testing periods.

6. Conversions to NCE scores were accurate.

7. Model A^2, C^1, C^2 only. Correlation coefficients were higher than.6.

8. Model A^1/A^2 - Students were not selected on the pre-test.

9. Model C^1/C^2 - A strict cutoff score was utilized.

10. CRT used for Model A^2 or C^2 had demonstrated validity and reliability.

11. Appropriate level and some form was administered at pre and post time.

Programmatic Implementation Criteria

1. Tests administered for evaluation reflected the curriculum of the compensatory project. At least 75% of the items of the instrument measured skills taught in the compensatory project. (content validity)

2. Data was aggregated only on students who fulfilled the following criteria:

   a. Participated in a program more than four months.

   b. Program was fully operating for the period between pre and post testing.

   c. Students attended the program at least 2/3 of the time.

2. LEA district programs that show impact gain above 20 NCE points for any grade or subject should be reviewed. To assure gain is not reflecting inaccurate or improper implementation, all items in Item 1 should be reviewed plus:

   1. Tests administered to the students has not been utilized more than twice for any student.

   2. Test instruments were two or more grades below the grade the student is currently enrolled.

3. Non-Test Data

   1. Attendance, attitudinal scales, or other indicators of student improvement should be available to support test findings.

   2. Aggregation of non-test data must be implemented as follows:

      a. Developmental Project: Indicators should be only aggregated on students who have been the developmental project more than two years.

      b. Compensatory Project: Indicators should be only on students in the project more than four months.

Phase II. This phase is conducted at the state level and consists of preliminary computer edit routines run on the aggregated data received from the LEAs. The computer edit routines will be dsigned to kick out data for the following reasons:

1. Testing not conducted at the norming date

2. Inappropriate form and level of the test used for pre-testing and post-testing

3. NCE gains which are too high or too low (this needs to be defined). For discussion sake the edit routine might be designed to bounce out lets say an NCE gain of 45 and one of -10).

4. Errors of conversion for percentile to NCE.

etc.

Phase III. This phase consists of the application of statistical quality control procedures using the Shewhart control chart: The upper control limits and lower control limits will be established by the state for the achievement data base.

For the state's review purposes, only 5% of the data falling outside of the control limits (2½% above, 2½% below) will be examined.

As was metioned earlier, data falling outside the upper control limit may signal (trigger) an exemplary program for further review just as data falling outside the lower control limit may signal (trigger) problems with a program or some other infervening variable(s) which require further review.

An Application of Phase III Quality Control

Each year New Jersey districts via the State's reporting structure provide annual reports delineating mean student achievement, scores as well as pre and post achievement test scores by grade for students in compensatory programs (Title I, State Compensatory Education, locally-funded programs, etc.)

Over the past two years, the Office of Evaluation has analyzed district scores to determine program effectiveness for compensatory education (Title I, State Compensatory Education, locally-funded programs, etc.) populations.

Analyses are conducted in Normal Curve Equivalence (NCE) scores. This type score enables computational procedures that could not be conducted with percentiles or grade equivalents. The scores offer the advantage of estimating the relative performance of children based on performance of their peers. As a result, no program expectations for post test achievement can be made based on the scores of children on the pre tests, with reference to the scores of their peers.

Over the past two year, almost all New Jersey districts have demonstrated achievement gains in the compensatory education populations above the expected gains. Based on the distribution of gains made by districts in each grade level, a projection may be made of which districts are gaining most and which are gaining least.

Also based on the distribution of gains, statistical quality control criteria may be utilized to determine both the positive and negative outlying districts. Because these gains vary so much by grade level, separate computations should be made for each grade. The data from each year, collected in late summer, should be used to establish these criteria for achievement performance for the following year.

Dispersion of mean pre to post test differences are analyzed by means of the standard error of differences. At each grade, the mean gains of 95% of all districts will fall between + or — 1.96 times the standard error of differences (S_D) from the mean gain at that grade level. For example, if all districts averaged a gain of 11 NCEs in Computation at Grade 9 and the standard error of differences was 2 ($S_D = 2$), then 95% of all districts gained between 7.08 NCEs and 14.9 NCEs. By formula, this was computed as follows:

$$S_D = \sqrt{\frac{(\Sigma D^2) - \frac{(\Sigma D)^2}{N}}{N(N-1)}}$$

In this way, the state data for each grade may be analyzed each year to determine the top 2.5% and the bottom 2.5% of districts in achievement gain at each grade level. The programs in these districts can then be examined to determine why they fall outside of the upper or lower control limit.

An example utilizing the concept of statistical quality control follows. The data are based on achievement scores for New Jersey students receiving compensatory program for the FY'79. The data did not undergo Phase I of the proposed quality control procedure but did undergo Phase II of the procedures (computer edit routines were run). These data are provided as an illustrative example of the Quality Control model proposed and should not be used to make any statements about New Jersey compensatory programs.

Example I
State Summary
Computation Programs (Title I, State Compensatory Programs, locally funded programs, etc.)

Mean Gain (NCE)	Grade	1.96 x S_D	Lower Control Limit (NCE)	Upper Control Limit (NCE)
17.811	1	5.968	11.844	23.779
12.366	2	2.084	10.282	14.450
13.136	3	1.600	11.536	14.736
11.241	4	1.460	9.781	12.701
9.846	5	1.639	8.206	11.485
10.128	6	1.778	8.350	11.906
8.655	7	1.641	7.014	10.296
6.441	8	1.139	5.302	7.579
5.902	9	1.969	3.933	7.8710
6.807	10	1.688	5.119	8.495
7.343	11	1.541	5.802	8.884
4.811	12	1.207	3.604	6.017

Example II
State Summary
Communication Programs (Title I, State Compensatory Programs, locally-funded programs, etc.)

Mean Gain (NCE)	Grade	1.96 x S_D	Lower Control Limit (NCE)	Upper Control Limit (NCE)
13.685	1	3.454	10.231	17.139
11.158	2	1.526	9.632	12.684
10.384	3	3.829	6.555	14.213
7.346	4	1.135	6.211	8.481
6.993	5	1.370	5.623	8.363
4.001	6	1.611	2.390	5.612
6.981	7	1.015	5.966	7.996
3.953	8	1.015	2.938	4.968
5.559	9	1.115	4.444	6.374
5.393	10	1.677	3.716	7.070
3.472	11	1.267	2.205	4.739
4.320	12	0.999	3.321	5.319

It must be stressed that any gain in NCEs indicates program effectiveness. Those gains above the "high" levels specified, however, may be interpreted as significantly greater than the average of those compensatory education programs that report their results correctly.

This preliminary data set show interesting results:

1. Differences in performance between achievement in computation and communication
2. Higher NCE gains in the lower elementary grades
3. Wide tolerance/control limits in the lower elementary grades (more variability in performance)
4. Smaller tolerance/control limits in the higher grades (less variability in performance)

The ultimate purpose of the data set, however, is to demonstrate quality control concepts with an actual data set. The next step in this state analysis would be 1) to array all of the data from individual districts by grade, 2) to superimpose the upper control limit and the lower control limit and 3) to identify the outlyer data points for further analysis.

Conclusions

This paper has presented an industrial model of statistical quality control for application at the state level on data generated by the Title I Evaluation and Reporting System (TIERS).

REFERENCES

Crane, L. R. and D. Bellis. Preliminary Results of Illinois Pilot District Edits (unpublished paper). Evanston, Illinois: Educational Testing Service, 1979.

Hiscox, S. B. and D. D. Deck. Quality Control: A Comprehensive Approach (unpublished paper). Portland, Oregon: Northwest Regional Educational Laboratory, 1980.

Grant, E. L. Statistical Quality Control. New York; McGraw-Hill, Inc., 1964

Shewhart, W. A. (edited by W. E. Deming). Statistical Method from the Viewpoint of Quality Control. Washington, D. C.: U. S. Department of Agriculture, 1939.