DOCUMENT RESUME

ED 247 978 JC 840 481

AUTHOR Smith, Jean Burr
TITLE Data Collection and Descriptive Statistics for Effective Program Evaluation.
PUB DATE Nov 83
NOTE 16p.; Paper presented at the National Convention of the American Mathematical Association of Two-Year Colleges (Orlando, FL, November 10-13, 1983).
PUB TYPE Viewpoints (120) -- Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Developmental Studies Programs; *Evaluation Methods; Formative Evaluation; Mathematics Anxiety; Program Effectiveness; *Program Evaluation; *Research Design; *Research Problems; Summative Evaluation; Two Year Colleges

ABSTRACT

Following a discussion of various methods of evaluating the effectiveness of developmental programs, this paper presents a research design for the evaluation of a developmental math program. First, the paper examines the objectives and benefits of conducting formative and summative program evaluations. The paper then identifies the single group pre-test/post-test comparison as the most commonly used method of evaluating remedial programs, and identifies the biases to which the method is vulnerable (e.g., test administration, student attitude, instrument, and external learning biases). Next, methods of reducing the influence of some of these biases are suggested. The paper then focuses on the marginally remedial/marginally exempted comparison, the remediated/exempted comparison, norm-group comparison, cross-program comparison, and historical comparison methods. The paper then recommends that program evaluators be objective, informed, comprehensive, pragmatic, political, selective, and prepared to compromise. The last sections describe the methods of data collection and analysis, and the findings of the developmental math program evaluation. The project (1) investigated the number of students who completed the math courses, the number who took a more advanced math course after completing the developmental course, and marks in advanced math courses; and (2) compared results from classes in which math anxiety was recognized and treated with those in which it was not. (LAL)

DATA COLLECTION AND DESCRIPTIVE STATISTICS FOR EFFECTIVE PROGRAM EVALUATION

Jean Burr Smith
Professor of Mathematics
Middlesex Community College
Middletown, Connecticut

Paper presented at the AMATYC National Convention, November 10-13, 1983, in Orlando, Florida.

DATA COLLECTION AND DESCRIPTIVE STATISTICS
FOR EFFECTIVE PROGRAM EVALUATION

I am going to approach this topic in reverse order--looking at the possibilities and pitfalls of evaluation of developmental programs and then the data collection and descriptive statistics.

First of all we must decide whether we are looking for a "comprehensive" or summative evaluation or, more simply, a formative one; one which will identify those elements in the instructional program which contribute to its' effectiveness and thosewhich need improvement. Studies like the latter are quick, simple, and informative. They may be designed to examine any aspect of a remedial program — from texts, to methods of instruction; from testing procedures to exit criteria; from which of the students are learning the content, to what content is being learned. Formative evaluations do not usually bind themselves to generalizations but they are invaluable in putting the program on the right track and seeing that it stays there. Summative evaluations, intended, as the objective indicates, to measure over all changes, may cover a number of areas such as:

1. Appropriateness of Objectives: Are they actually the premise on which the program is based? Are some of them misguided or inappropriate?

2. Appropriateness of Content to Program Objectives: Is drill in fundamental operations of arithmetic necessary in a world of inexpensive hand calculators? Is ability to figure compound interest a key to future success?

3. Appropriateness of Placement Procedures: Whatever the basis — testing, high school records, interviews — what percent are underplaced and what percent overplaced?

4. Effectiveness of Instruction: Are the students learning the remedial content and, if so, is the learning the result of remedial instruction or extraneous factors?

5. Efficiency of Instruction: Can the same learning be provided for less money or more learning for the same?

In practice most summative evaluations focus on number 4 - Effectiveness of Instruction - whether or not or to what degree students are learning the course content and how well they are succeeding as a result in subsequent courses. Particularly important then are pre-program and post-program measures - what the student knew before and what the student knew after participation in the program.

The single group pre-test-post-test comparison is probably the design most commonly used in the evaluation of remedial programs. It is certainly the easiest to implement. You have a group of students who start and finish the course and you compare their knowledge at the end with their knowledge at the beginning. Unfortunately, this design is often of least value. Even if post-program scores are significantly higher than pre-program scores, and they usually are, this change cannot be automatically attributed to effectiveness of the program. This single group pre-test-post-test- comparison is particularly vulnerable to a host of extraneous factors known as biases which distort results and cloud interpretations and may be from the peculiarities of the student reaction to tests and testing procedures, or from learning that takes place but not as a result of the remedial program. These are biases:

1. Test Administration Bias: If the administration of pretests differs significantly from that of postest. The pretest may be a part of a large battery of tests given in a poorly lit auditorium. The posttest may be given in the small, comfortable class group by a sympathetic teacher who answers leading questions, allows extra time or otherwise contributes to exaggerated gains.

2. Student Attitude Bias: Students may underestimate importance of pretest and do less than their best work so that program effectiveness is over estimated or be over anxious on posttest (which may be final exam) and perform poorly, so that program effectiveness is underestimated.

3. Teaching to the Test: Usually instructors are familiar with content of posttest, therefore unconsciously or consciously they may stress topics included, one type of algebra problem over another, for example. The outcome, then, is an artificial increase in scores.

4. Practice Effect: The mere experience of taking the pretest may prepare students to do better later on; they have had experience with the particular format, the use of the answer sheet, allocation of time, . . . The outcome, again, may be an artificially high post-test score.

5. Instrument Bias: Pretest and posttest must be valid:— content and minimum proficiency level must be appropriate, and reliable:— results must be consistent.

6. Hawthorne Effect: Students performance is likely to improve simply because they are receiving special attention. These gains, which are genuine for the experimental group, may not be sustained for subsequent populations.

7. Drop-outs: The bottom of the class is sifted out rather than taught and post-test scores for these dropouts are seldom included in the statistical analysis.

8. Regression Toward the Mean: Those who initially scored at the extremes will, when retested, tend to score toward the middle.

9. <u>External Learning Bias</u>: Students often improve their basic skills for reasons having little or nothing to do with remedial instruction - The sheer excitement of being in colleor, particularly for the non-traditional students.

10. <u>History Bias</u>: This bias concerns the possible effect of accidental or unpredictable external events on the program under evaluation -- a one-year grant resulting in smaller classes, a strike or a crippling snow storm which reduces class time.

The pre-test-post-test is not necessarily hopeless if we recognize these biases and take steps to reduce their influences: For example:

<u>Test Administration Bias</u>: Set pre test in more relaxed atmospheres and avoid having classroom teachers administer post tests.

<u>Student Attitude Bias</u>: Persuade students of the importance of doing well on placement exam.or give pre and post tests separately -- neither as placement or final exam.

<u>Teaching to Test</u>: Do not allow instructors to see test -- From a bank of tests randomly chose the one to be used.

<u>Practice Affect</u>: Use alternate forms of test (Cover same basic material but in different order).

<u>Instrument Bias</u>: Use tests of established validity.

<u>Hawthorne Effect</u>: Conceal from students fact that program is being evaluated.

<u>Dropout Bias</u>: Use pre-test scores only for those students who also take post test. Do a separate analysis of the drop-outs.

To compensate for these biases an alternative is to use a control group -- a group of students initially comparable to those entering the program who receive no remedial instruction or an alternate form of remedial instruction. Here we can assume biases affect both groups equally and can therefore be disregarded.

Looking at the control group who receive no remediation at all, the remediated - unremediated comparison. The remedial population is divided randomly into 2 initially equivalent groups. The average pre-program measures are compared to check initial equivalence. The extent to which post-program measures differ is the gauge of program effectiveness.

The major disadvantage of this comparison is that the deliberate withholding of remediation is ethically questionable, however until the effectiveness of a remedial program is clearly demonstrated the ethical questions may be somewhat premature -- the program may in fact be a down-right waste of time. Although no evidence exists regarding effectiveness . . . colleges offering remedial programs have been generally reluctant to randomly exempt from remedial work a portion of those students identified as being in need of remediation. This reluctance has prevented the conduct of experimental research which might enable educators to determine which remedial techniques are effective, for whom, and under what conditions.

2. The Marginally Remedial, Marginally Exempted Comparison: The marginally remedial group are those who narrowly fail the pre-test and receive the remediation, the marginally exempt are those who narrowly pass. The assumption is that the two groups are so close as to b considered equivalent. Again the difference in post-program measures is the indication of program effectiveness. This design avoids the moral dilemma of withholding remediation but it measures the effectiveness of the program only with the best of the remedial ones and we can be sure of a measure of success only if remediated ones surpass exempted ones in the post test. If the exempted surpass the remedial it is extremely difficult to decide whether the program has some value or whether it is altogether useless and thus the results are inequitable.

These designs h[illegible] measured remedial effectiveness, not evaluated it.

The following are designs which evaluate learning:

1. Remediated-Exempted Comparison: Here the pre-program measure is usually the original placement score, and the post-program measure a long-term one, - the GPA or performance in college level courses. The comparison is between post-program measures for the two groups. If the remedial group surpasses or even matches the exempted one we have strong evidence the program is successful. However, if the remediated group continues to lag behind no conclusion can be made with certainty and so this comparison is inequitable, also.

2. Norm-Group Comparison: Pre-program and post-program measures consist of scores on standardized tests, - The improvement of the local remedial population is compared with the corresponding national population on which the test was normed. This design is relatively simple, the comparison is based on information readily available and we can draw conclusions about the value of the program whether local gains are higher or lower that those of the norm group. The negative side is the disparities in the make-up of the norm group and the local population — age, sex, socio-economic status.

3. Cross Program Comparison: This comparison employes as standard of success the achievement of a comparable remedial program, typically at another college. It is assumed that the two remedial groups are initially equivalent, that the two programs have comparable objectives, content, and placement procedures and that the colleges have agreed on common pre-program and post-program measures. The major advantage here is that the results have formative as well as summative values. It not only compares the general effectiveness but shows places where weaker program should be modified. The cross-program comparison is seldom used, however because of the difficulty of locating matching populations, objectives,

contents, and placement procedures. There may also be problems of implementations if the staffs of the two colleges, feeling themselves in competition, are reluctant to cooperate.

4. Historical Comparison: The study is conducted at a single college the comparison being between different semesters. This is particularly valuable when there have been deliberate changes either in remedial program or in the college environment.

Whatever the evaluative design, it must be closely related to the objectives of the remedial program and its merits must be gauged to answer the following questions:

1. Is the design relatively free from the effect of biases?

2. Is it equitable: Does it provide evidence equally well of success or failure?

3. Is it comprehensive: Does the sample reflect the entire remedial propulation?

Hecht and Akst conclude their chapter on program evaluation with the following recommendations.

1. Be Objective: Objectivity is gauged by the extent to which it is based on concrete, appropriate data. Specific criteria for success or failure must be agreed upon in advance.

2. Be informed: Familiarity with range of design options is critical in planning the study. (Two books from U.S. Office of Education , Horst, Tallmadge and Wood 1975, Tallmadge & Horst, 1976.)

3. Be Comprehensive: Assess not only the overall effectiveness of remedial instruction, but also cost effectiveness of program, realiability of placement procedures, . . . In other words, use formative procedures, particularly early in the program.

4. Be Pragmatic: Do only what is possible, taking into account local limitations of resources, equipment, staffing, and time. The ambitious study unfinished in worth much less than the modest one completed.

5. Be Political: In planning and implementing your study be alert to the policy of the college and the misgivings of the staff. Keep all parties informed of your procedures and avoid strategies which may discourage cooperation.

6. Be Selective: Although the study should include all significant areas, choose the type of data to be collected with care. Great massess of data collected without purpose or direction is of no use. Identify precisely the necessary data, draw up a collection schedule so that data will not be lost. In the final report, highlight major results so that they will not be lost in a sea of peripheral information

7. Be Prepared to Compromise: If a particular direction seems to be treading on toes, go another way. Be prepared to compromise between the attainable and the ideal. atthe same time gauging the extent to which such short comings may result in misleading conclusions.

Now, with all the things in mind that I could and may have done wrong, I am going to talk to you about my particular study, and then we will look at studies you want to do.

As anxiety reduction techniques found their way into developmental math classrooms across the country serious questions were being raised as to the effect this was having on the mathematics being taught in these classrooms. "Is the mathematics in these classes being 'watered down'?" "Is the subject matter being replaced by psychological procedures?" "Are the students being 'spoon-fed'?" "How do they survive when they get into a 'real' math class?"

In an attempt to answer these questions I set up the following research project to look into 3 areas:

1. Numbers who successfully completed our developmental classes.

2. Numbers who took a more advanced math course after completion of our developmental one.

3. Marks in these more advanced courses: I compared the results from classes in which math anxiety was recognized and treated with those in which it was not.

I am a member of a 5-person mathematics department in a 14 year old community college in Connecticut. All of us share an enthusiasm for mathematics and a concern for our students but I am the only one who feels the need for students to feel comfortable with themselves in a math class in order for real learning to take place, and, therefore, the only one who actually uses anxiety reduction techniques in the classroom.

Our developmental course, Math 99, reviews basic arithmetic and covers as much algebra as seems individually feasible. We have no set syllabus, nor do we use the same book, and we give different final examinations. The course carries 3 credits towards graduation and, if the student transfers after graduation, he/she receives 3 general elective credits at our state colleges and university. All full-time students are required to take a math placement test, one which we have developed. Students may also place themselves in the course, and many older students do.

For my study I took all sections of our developmental course from Spring semester, 1977, through Fall 1980, 1074 students. I had a 3 X 5 file card for each student on which I recorded: Name, M or F, semester, Math 99 and other quantitative courses taken afterwards with marks and dates of taking these courses. (I wanted age but was not able to get it). At this point I made the decision as to which statistic was to be used and I chose the significance of difference between percentages as interpreted by the normal curve because the data is translated into scaled data rather than category. and it is a more powerful statistic that those dealing with nominal data. However,, many people have asked about x^2 which seems to be more easily understood so to deal with my first set of comparisons I will use both.

H_o : Number of completions same for both groups.

H_a : Number of completions greater for Group with Anxiety reduction.

First using χ^2

at d = .01

one- Failed Test

D.F. = 1

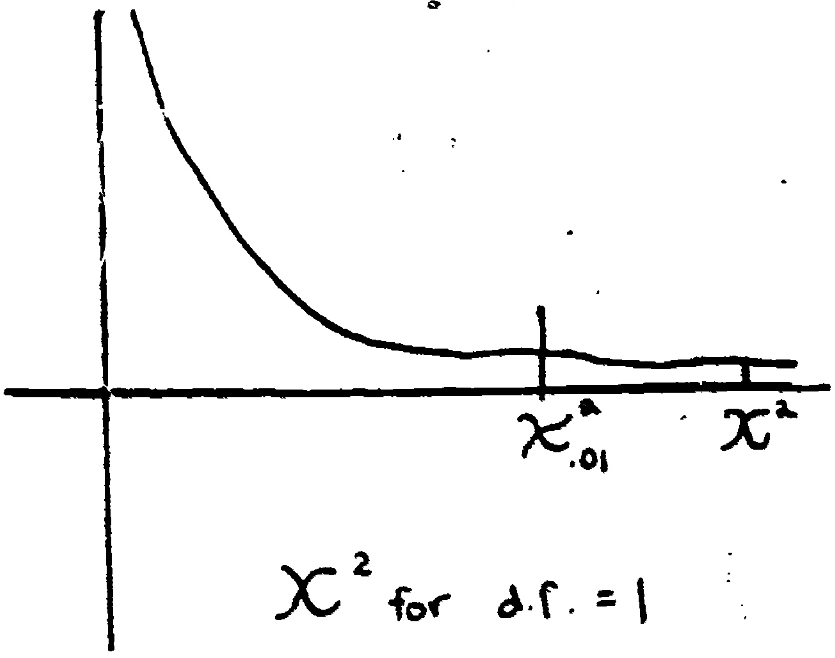
$\chi^2 = 6.6$

	Completed	Did not Complete	Total
With Anxiety Reduction	255 (227.6)	85 (112.4)	340
Without Anxiety Reduction	467 (491.4)	270 (242.6)	734
Total	719	355	1074

To get "expected": for example: $227.6 = \frac{340 \text{ X } 719}{1074}$

"Did not complete" was sum of Failures, Withdrawls, and Incompletes.

$\chi^2 = 3.30 + 6.68 + 1.5 + 3.09 = 14.57$



$\chi^2 > \chi^2_{.01}$

∴ H_o rejected

The overall picture of descrepancies shows that more completed with anxiety reduction and more failed to complete without.

Checking of the significance of defferences between particular percentages is more exact. Using the same data I set up the hypothesis that there was no significant defference between the percentages of those in classes who completed course with anxiety reduction and those without.

H_o: % completions with anxiety reduction = % completion without

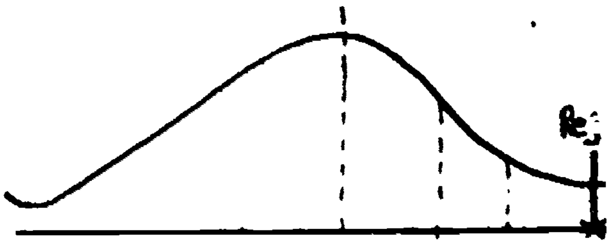
H_a: % with > % without. (at .01 level becau· I expected a large difference)
$z_{.01} = 2.58$

$$p = \frac{734\ (.642) + 340\ (.75)}{734 + 340} = .676$$

$$q = 1 - .676 = .324$$

$$= (.676)\ (.324)\ \frac{1}{734} + \frac{1}{340} = .031$$

$$z = \frac{175 - .632}{.031} = 3.81$$



∴ H₀ rejected: (Probability = .999 or 99.9 % . It is very unlikely that this could happen by sheer chance so H_o is rejected and the difference significant at the .001 level or .1 %)

Second: Using z scores

H_o: % taking another math course after anxiety reduction Math 99 = % taking regular Math 99.

H_a: % taking another math course after anxiety reduction Math 99   % taking regular Math 99. (because of reduction or filtering out of students in regular Math 99's I did not expect much difference). $\alpha = .05$; $z_{.05} = 1.96$

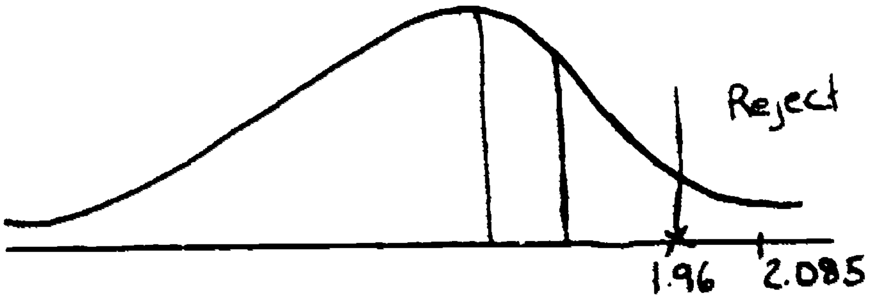
	Total finishing 99	Took another course	%
Without M.A.	354	166	46.89
With M. A.	164	93	56.7

$$p = \frac{(.567)\quad(164)\quad+\quad(.469)\quad(360)}{360 + 164} = .496$$

$q = .504$

$$\sigma = \sqrt{.24998)\quad(.00278 + .00610)} = \sqrt{.0022198} = .047$$

$$z = \frac{.567 - .469}{.047} = 2.085$$



∴ H_o Rejected. Again, not likely that this would happen by sheer chance

3. Comparison of marks in follow-up math courses: I recorded the marks A = 4; B = 3; C = 2; D = 1; F, W, I = 0 and checked first the F ratio for homogeniety of variances to show that they were samples from the same population and then proceeded to test the significance of the difference between two means.

WITH ANXIETY REDUCTIONS

	X	f
A	4	13
B	3	18
C	2	19
D	1	10
F,I,W,	0	25
N =		85

WITHOUT ANXIETY REDUCTION

X	f
4	34
3	39
2	37
1	11
0	52
N =	173

$$\overline{X} = 2 - \frac{16}{85} = 1.81$$

$$\overline{X} = 2 = \frac{2}{173} = 2.01$$

$$S = \sqrt{\frac{180}{85} - \left(\frac{16}{85}\right)^2} = \sqrt{2.0822} \quad 1.4430$$

$$S = \sqrt{\frac{383}{173} - \left(\frac{2}{173}\right)^2} = 1.4879 \quad \sqrt{2.21}$$

Not Significant

F - Test (F - Ratio) $\frac{2.21374}{2.0822} = 1.06$

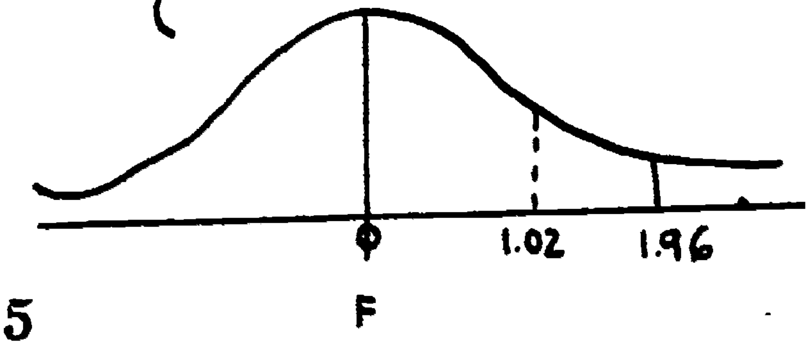
$F_{.05} = 1.58$

∴ Homogeneous

$$\sigma = \sqrt{\frac{(173 + 85) \quad (2.21374 + 2.0822)}{(173) \quad (173 + 85 - 2)}} = .1959$$

$$z = \frac{2.01 - 1.81}{.96} = 1.02$$

∴ } Means not significantly differen



In conclusion then in my study of those students who have been helped to use anxiety reduction techniques in their developmental classes as compared to those who have not, a significantly larger percent finished the class and of those who finish significantly more took another math course and show no significant diff·rence in ability to cope with traditionally taught higher level mathematics courses as indicated by their marks.
That is they did as well as the much smaller and more select group who had been taught in the traditional fashion.