DOCUMENT RESUME

ED 104 856 SP 009 094

AUTHOR Poynor, Hugh
TITLE To Aggregate Is to Aggravate.
PUB DATE Mar 75
NOTE 8p.; Paper presented at the National Convention of the American Educational Research Association (Washington, D. C., March 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Data Analysis; *Educational Research; *Evaluation Methods; *Research Problems
IDENTIFIERS *Aggregation; Individual Cognitive Demand Schedule

ABSTRACT

Aggregation of data is, by definition, an obscuring of details for the sake of achieving a summary. It is, therefore, potentially harmful to accuracy. Attention should be given to the aggregation that scores undergo prior to statistical tests. Two familiar research designs where this is important are a) in two groups/one measure cases, and b) in two groups/pre-post measurement cases. Another problem for researchers develops when incomplete and missing data are encountered for identification codes, as well as for score values. Unless each data record contains all identifying codes, it will be excluded from one or more aggregates and results at different levels will change depending on the data values for those records which are missing identification codes. Observations recorded on the Individual Cognitive Demand Schedule can be examined as an illustration of the problems with aggregation. Although aggregating the data is generally beneficial in this case, it leaves out considerable evidence about the classrooms. Thus, the value of these data is greatly enhanced by leaving the observations unaggregated. (PB)

To Aggregate Is to Aggravate *

Hugh Poynor

The University of Texas at Austin

Many people in educational research use data as evidence. If their interest is in pupils, they collect evidence from pupils as grounds for their ideas about the way pupils behave. If their interest is in teachers, they collect evidence from teachers as grounds for their ideas about teacher behavior. If their interest is in the impact of school district policies, however, they must always be content to use indirect evidence by collecting data from pupils or teachers and aggregating the details up to a suitable summary level. The process of aggregating typically involves averaging over some score units. For example, pupil scores may be aggregated to the classroom level, and then averaged to represent the class as a whole.

Potentially at least, aggregation is harmful to accuracy because the process of aggregating, by definition, obscures details for the sake of achieving a summary of the data set. Some attention should be given to the aggregation that scores undergo prior to statistical tests such as the familiar t test, since the quality of these scores as evidence is an important issue.

In a district policy research problem, as mentioned earlier, pupils' scores may be aggregated to the teacher level, thereby providing the researcher with as many means for analysis units as he has teachers in each of two policy groups. Then the researcher aggregates these mean scores in order to contrast the impact of two different policies, employing the t test. Many of us have

* Paper presented at AERA National Convention in Washington, D.C., March, 1975.

seen this method used to compare not only policies but also educational programs and products. In such two-group/one-measure cases, the distribution of means for each group, when plotted, will be replete with research evidence, as will be the within-class standard deviations when they are plotted for each group. In both cases, this information yield would not be realized if the simple group mean aggregates were only contrasted with a t test. The main point to be made in illustrating this familiar design is the rather automatic and perhaps thoughtless obscuring of details by researchers in the process of aggregating their data.

Another familiar design employed by the researcher is a two groups, pre-post measurement design, which in the writer's opinion many times yields tremendous insight into policy, program, or product differences. While the design is straightforward, the resulting yield, which is phrased in terms of pupil aggregates, has the clarity of a mudhole. Take a quick look at the kind of statements we are led to make as a result of a pre-post analysis of means: "The average pupil in program A out-gained the average pupil in program B by 4 points," or "On the average, program A pupils exceeded the gain of program B pupils by 4 points." Now, in the first place, no attempt was made to analyze the "average pupil," nor did a search party attempt to locate him.

In defense of averages it should be pointed out that if both groups pre- and posttest distributions were symmetric, then the average pupil could be identified by an arithmetic mean. Even with symmetric distributions the interpretation of pre-post analyses of means is unclear, because gain score mean values can be heavily influenced by only a few pupils in the sample. That is, even though extreme individuals are rare, their influence upon the mean gain value is pronounced, and it is quite possible in a study for a few scores to swing the results for the entire group.

Oftentimes, researchers aggregate pupil scores to the teacher level because pupils may not be independent sources of information. In theory, this is because of the controlling influence teachers have on pupils and their classmates. This is often referred to as the issue of the "proper unit of analysis," as discussed by Poynor (1974) and Glass & Stanley (1970). In order to achieve a proper unit of analysis, these sources explain to us that pupil scores may have to be aggregated within their respective classroom boundaries to yield a single average for each classroom teacher. The researcher who is concerned with this issue can test the independence of pupil responses (Poynor, 1974). If responses prove to be independent, he need feel no pressure to aggregate on this account.

Given independent pupil responses, what <u>should</u> be done with data from a traditional pre-post design? First, a complete analysis of covariance such as suggested by Ward & Jennings (1973), employing a homogeneous slopes test and a groups difference test should be performed. Second, a head count should be made of pupils having no gain, those having positive gain, and those having lost score points from pre to post. These counts should then be converted to proportions. In addition, inspection of the prescores for these three groups could reveal ceiling effects in the test instruments.

Many years ago the aggregation issue and the problematic loss of individual behavior in aggregated data was popularized by Guthrie who studied the learning curves of individuals, rather than averaged group learning curves. Individual curves in his learning experiments were abrupt and single step in nature, and so led him to conclude that one-trial learning was taking place. Had Guthrie limited his analysis to the smooth exponential curve produced by aggregating many of these single step functions, he no doubt would have arrived at different conclusions.

A related problem is encountered by those of us in educational research in the schools, where incomplete and missing data are often encountered for identification codes, as well as for the score values. Making a good aggregate with some missing data is no problem because the missing values may simply be ignored. The problem occurs when multiple aggregations are to be made on the basis of various identification codes, such as an aggregate of posttest scores on teacher codes, on school principal codes, or on school district codes. Unless each data record contains all identifying codes, it will be excluded from one or more aggregates and the results at different levels will change depending on the data values for those records which are missing I.D. codes. Although it is difficult to believe, this writer has seen data sets reported where the differences between groups shifted back and forth in favor of one group, then the other group, only because of changes in the level of aggregation, where there was some missing identification data.

Aggregation of PRIME Observations

Project PRIME has collected 16,600 hours of classroom observation data using four observation systems. While the analysis of this much data may be considered a challenge, aggregating the data in a thoughful and meaningful analysis strategy is difficult. For the remainder of this paper, attention will focus on observations recorded on the Individual Cognitive Demand Schedule (Lynch & Ames, 1972) and the aggregation issues encountered with this schedule.

First, a quick inventory of the volume of computer records presently available, after transcribing the raw observations onto a 94,000 record computer tape:

6182	hours of observation,
1030	school days (6 hours each),
5.5	school years (186 days each).

The observation period required only two calendar months to complete, during which time many observers were involved with many pupils and teachers in several school districts.

The desirability of summarizing this much data seems perfectly clear, and, of course, summarizing the data necessitates aggregating the 94,000 data records according to several hierarchical aggregation levels, namely the I.D. code levels corresponding to pupil, teacher, and school district. As mentioned earlier, our research interests in pupils, teachers, and districts are usually satisfied with empirical evidence collected from them directly or collected from lower hierarchical levels and aggregated to the appropriate level of interest.

It is true with the Individual Cognitive Demand Schedule and all sign systems that events are coded in a binary fashion, they occur or do not occur, during the observation interval. For ICDS this was a four-minute time segment. Summing these binary events for a single pupil and then dividing by the number of records with complete data regarding the event for the pupil produces a proportion score. In this manner, raw observation records are aggregated, and the dependent variables become proportions with values ranging from 0 to 1. Proportion scores for pupils include curriculum activity, classroom structure, teacher task, pupil task, and seating arrangement. By averaging these proportions across all pupils we are able to determine the average proportion of time spent by pupils in various activities under various conditions.

These aggregates fail to satisfy our research interest in the nature of classrooms, however, because averaging leaves out a considerable amount of evidence about the classrooms. For instance, these proportions are created for single variables such as reading activity, small group seating, and teacher drilling. Even though many aspects of the classrooms were recorded, the proportion aggregates provide only a one-dimensional view of the events which took place. We are able to look at many aspects of a classroom in turn, but the joint occurrence of these aspects is lost. For example, we cannot know if reading is taught by using drilling with pupils in small groups, or by lecturing to large groups because these aspects of the class were aggregated independently of each other and cannot be disentangled.

Because multidimensional evidence is richer than uni-dimensional aggregates, much of the research work with the Cognitive Demand Schedule has been accomplished with unaggregated data records. The unit of analysis in this work is a four-minute time segment and the total N is 94,000. Having a time segment as the unit of analysis has great advantages to us when performing cross-tabulations, intercorrelations, and factor analysis work. The joint occurrence of classroom events produces multidimensional patterns or profiles through the use of these statistical methods that would go completely unnoticed if the methods were employed on pupil or teacher level aggregates. Thus, the value of these data as evidence is greatly enhanced by leaving the observations unaggregated.

References

Glass, G. V. & Stanley, J. C. Statistical Methods in Education and and Psychology, Englewood Cliffs, NJ: Prentice-Hall, 1970.

Lynch, W. W. & Ames, C. Individual Cognitive Demand Schedule, Center for Innovation in Teaching the Handicapped, Bloomington, IN: Indiana University, 1972.

Poynor, H. B. Selecting units of analysis, in Borich, G. (Ed.) Evaluating Educational Programs and Products, Englewood Cliffs, NJ: Educational Technology Press, 1974.

Ward, J. H. & Jennings, E. E. Introduction to Linear Models, Englewood Cliffs, NJ: Prentice-Hall, 1973.