DOCUMENT RESUME

ED 100 495 PS 007 583

AUTHOR Lukas, Carol VanDeusen
TITLE Measuring the Extent of Treatment Implementation.
PUB DATE Aug 74
NOTE 16p.; Paper presented at the Annual Meeting of the American Psychological Association (82nd, New Orleans, Louisiana, Aug. 30 - Sep. 3, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Demonstration Programs; Educational Alternatives; Evaluation Criteria; Evaluation Methods; Measurement Goals; Observation; *Preschool Programs; *Program Development; *Program Evaluation; *Research Problems; Standards; Teacher Evaluation; Validity
IDENTIFIERS *Project Head Start Planned Variation

ABSTRACT

This paper reports the problems involved in measuring the extent to which the practice in program sites matches model theory formulated in Head Start Planned Variation. The main focus in studying program implementation is the question of internal validity, whether the children in the HSPV programs actually receive the treatment being evaluated. Several attempts were made to develop an appropriate measure, though none of the measures was found to be completely adequate. Three types of information needed to develop a good measure of implementation were determined: a definitive model description, clear-cut standards of implementation, and judgments concerning comparability within and between models. The problems connected with obtaining this information were described. It is concluded that precise and accurate evaluation of program implementation is difficult and some suggestions are made to alleviate evaluation problems. (SDH)

Measuring the Extent of Treatment Implementation

Carol VanDeusen Lukas

Harvard Graduate School
of Education
Cambridge, Massachusetts

This paper was prepared for presentation at the
American Psychological Association Convention,
August, 1974

Although the primary focus of Head Start Planned Variation is on the effects of the program models on children, the study of model implementation has also been important. To implement means broadly to carry out or to put a policy into practice -- in this case the policy is the installation of educational curricula, or models, in Head Start sites. A number of questions can be raised under the heading implementation. One set, for example, focuses on the process by which the models were brought to the sites: What factors affect that process? Why are the models adopted by some individuals and not others? Another question looks at outcome of the process: to what extent does the practice in the site match model theory? I will limit my presentation to this last question. In doing so, I am interested in implementation as a question of internal validity. The main purpose of HSPV, as we have said, is to test the effectiveness of different educational models. Before drawing conclusions about test results, it is important to determine whether the children actually received the treatment being tested. If they did not, then we cannot assume that differences in outcomes result from the treatments and therefore cannot draw conclusions about model effectiveness. Likewise, we cannot assume that the lack of differences disprove the theory on which a model is based.

In HSPV, the central issue in determining the extent of implementation was to develop an appropriate measure. This is not an area which has received much attention in the past. It is not usually relevant in laboratory experiments where the researcher has control over the treatments and therefore can

Teacher Aptitudes. The second major component of our model is Teacher Aptitude. In our study in California, we focused on five kinds of aptitudes relevant to teacher behavior. They are: (1) verbal, (2) numerical, (3) reasoning, (4) memory, and (5) the kinds of divergent production aptitudes sometimes called creativity but which are probably be better described as fluency and flexibility.

These aptitudes were selected after considering both the literature on cognitive factors presented in various models of the structure of intellect (Carroll, 1974; Guilford, 1967) and also the very limited amount of research which has been done on the relationship between teacher aptitudes and pupil achievement.

Verbal ability is the only teacher aptitude which has received much research attention. Most of this interest stems from the finding of the Coleman study (Equality of Educational Opportunity, 1966) which showed that teachers' verbal ability was one of the characteristics most consistently related to pupil achievement. This finding has been replicated by other researchers (e.g. Hanushek, 1970).

We do not know, however, why teacher verbal ability is imporant to pupil achievement. A number of reason have been suggested. These include hypotheses ranging from suggestions that verbal ability is simply a proxy for general intelligence to suggestions that teachers with higher verbal ability are more able to negotiate their way into schools where pupils overachieve for other reasons. It seems likely that the size of a teacher's vocabulary may be related to her success in communicating with students. However, this relationship may not be the same for all groups of students. For example, in working with Mexican-American children with a more limited knowledge of English, the larger teacher vocabulary may be a source of confusion rather than provided with multiple avenues to understanding.

assume that the treatments are replicated as intended for all subjects. The issue is relevant in social programs but although people are aware that programs do not always turn out as intended, they have not systematically studied the discrepancy between intention and practice; the majority of evaluations of service programs have dealt only with program impact. In wanting to study the extent of treatment implementation in HSPV, then there was no standard methodology on which to draw.

Several attempts were made to measure the extent to which the models were implemented in HSPV. One approach was to have the sponsors rate the performance of the individual teachers working in their models. Using a common form for all models, the sponsors were asked to rate teachers on a 0 to 9 scale running from unacceptable to outstanding.

A second approach was to observe and rate models components. For each model, a description of the model in the form of a checklist was drawn up by the staff at Huron in conjunction with the sponsors. The checklists were used as observation instruments by consultants hired by OCD to monitor implementation. They rated each component of the model on a scale of one to four, not present to fully implemented.

A third approach, developed at the Stanford Research Institute, was based on elaborate standardized classroom observation procedures. Among other things, the procedures coded interactions within the classroom for five minute periods. An observer recorded who did what to whom and how. These codings were then aggregated to model relevant variables. The same instrument was used for all models and became model specific through development of the variables. It was used by observers trained at SRI to high levels of reliability. The most detailed analyses of these data were done by SRI not by Huron, so we have less knowledge about them than about the data from the other instruments.

Any of these measures can be used in either of two ways to take differences in implementation into account in outcome analyses. One method requires that a cut-off level be set: we would decide what level of implementation represented a classroom situation which was close enough to the model idea to provide for a valid test of the treatment, and only analyze effects in classes which exceed that level. We would essentially be making a simple distinction between treatment and no treatment. A second method uses the extent of implementation as a covariate in outcome analyses. All classes would be used but differences in the extent of implementation would be controlled for.

Unfortunately, the measurement of treatment presence in HSPV was not as straightforward as we had expected,,and none of the measures was completely adequate. We did learn a great deal, however, about what is needed for a good measure of implementation. Our lessons can be grouped under three headings: model description, standards of implementation, and comparability.

The first step in developing an implementation measure is to describe the treatments. Clearly, before we can determine whether treatment is being used, we need to know what that treatment is supposed to be. In HSPV, the sponsor ratings mentioned earlier, for example, did not take this step. They were not based on an explicit definition of the models, but instead relied on the raters' personal conceptions. Because we did not know how the raters defined the model, we did not know if one rater had the same view as another. The problem more complicated, however, than simply failing to make the definition explicit. We discovered in developing the consultant checklists, which do contain explicit descriptions, that describing the treatments is not an easy task. It is difficult, first, because treatments changed over the course of the study. Most models were not fully developed when the experiment began. When sponsors began working with Head Start staffs, they found that some of their original ideas were not viable and that they had not thought of everything relevant to running a classroom. By necessity, then, most models continued their development during HSPV. As a result, model descriptions grew quickly obsolete. Moreover, some sponsors never felt their model reached a final form, and therefore resisted committing themselves to written statements about the characteristics of their programs. While program development is educationally desirable, the models cannot be said to be clearly defined in a conventional experimental sense because they are different at different points in time and because they may evolve to a variety of forms in different sites.

A second problem in describing the treatments stems from differences in the level at which they are specified. Ideally, we would like to have all models described in operational terms -- that is, in terms of specific behavior -- because the presence of a treatment can then be established easily and precisely. In HSPV, however, most sponsors described their models in terms of broad principles rather than specific behavior. Referring again to the consultant checklists, while one model description, for example, contains behavioral statements like "Teachers know format of lesson and look down at and look only for examples," another is described in broad terms such as, "The adult challenges and supports problem solving and coping behavior." These differences would have occurred even if models were well-developed, for in most cases they resulted from the nature of the model rather than from an incomplete theory: While some sponsors closely prescribe the classroom activities and interactions in their model, others offer only general principles because they believe that model teachers should carry out the principles in their own style and according to the needs of the children with whom they work. Again, although such a strategy no doubt is educationally sound as methods which closely prescribe class behavior, a description in broad terms is difficult to work with when the model is being used as an experimental (or quasi-experimental) treatment.

Thus, the first step in obtaining an adequate implementation measure was not taken in HSPV. In many cases we know what the treatments were only in the most general terms. Moreover, since the difficulty in obtaining operational definitions resulted primarily from the fact that all models changed and many tended to operate on a level of general principle, we would expect this difficulty to recur in other studies of these models and more generally, in studies of most broad social programs.

A second concern in measuring the extent of implementation is establishing standards of full implementation--or what a class should look like in order to be considered a model classroom. In HSPV, not much thought was given this issue originally and no standard was established because people seem to have expected that a fully implemented class would be easily recognized.

One standard which was not explicitly agreed to but which was a logical one because it is used in laboratory experiments, is exact replication: a fully implemented treatment looks like the model as described and like all other examples of the treatment. This standard proved inappropriate in HSPV, however, for several reasons. For one thing, it was too stringent and was rarely, if ever, met. We found variation in the extent of model use in all models on all measures. Analyses of the classroom observations, for example, revealed that even where classes within a model showed high frequencies on variables selected to reflect model objectives, the consistency on the variables among classes was low--that is, even though two classes showed a high frequency on variable X, the level of frequency in the two classes was not the same. If we applied the standard of exact replication and used only those classes which met the standard in the analysis of the outcomes, we would have lost most of the sample.

The standard of exact replication was also inappropriate because classes probably would not be replicated in the sense of one looking just like another even if they were all at the top of any scale we might devise. This is true for the same reasons that it is difficult to obtain operational model descriptions: most models were not detailed enough to make replication possible either because the models were not fully developed or because some models encouraged individual adaptation of the means of carrying out the model principles. Replication was also inappropriate in view of the conditions under which the study was conducted. Given the complex nature of the situation in which the sponsors were intervening--day to day interactions among people, both children and adults--and given the fact that the treatments were implemented by people in the field and not the researchers or even the model sponsors, it seems clear that no class, even in those models which do not encourage variation, would ever be exactly like another.

But while replication was not the appropriate standard, no alternative was offered in its place. Consequently, the implementation data were difficult to interpret. The sponsor ratings, for example, were done on a scale of 0-9. Since 7, 8, and 9 were labelled as outstanding performance, we can be fairly assured that the teachers who received these ratings were doing a good job in the model. While we do not know from this measure what characteristics these teachers had that caused the sponsors to rate them so highly, nor do we know how a rating of 7 differs from a rating of 9, we would guess that a sponsor would say that the classes of these teachers would be considered model classes. The ratings in the middle of the scale--in the category labelled average--are more problematic, however, because we have no idea of how the raters interpreted average. They could have used it to mean that a teacher, although not doing a brilliant job, was implementing the model adequately enough to be called a model teacher. They also could have used it to mean that the teacher was only partially implementing the model. Each interpretation leads to quite different conclusions. Moreover, even if we knew that these middle ratings indicated partial implementation, without any standards we do not know how large the differences among the classes are or how far they are from being fully implemented.

Thus, we need clear standards of implementation which allow us to determine when a class is close enough to the model to be considered an acceptable example of the model. The first step in doing this is to establish a workable definition of full implementation. Since we have rejected the notion of exact replication, it follows that we must allow for some variation among classes. The critical issue, then, is to determine how much variation is acceptable: to what extent and in what ways can classes differ and still be considered to be be examples of the same treatment? We cannot simply say that, having removed the demand for replication, any thing that happens is a legitimate variation of the model. At some point we must be able to say that a teacher is not following the model. We must also resolve the problem of comparing partially implemented classes. Within a single model, one class is, say, 60% implemented may not look like another which is 60% implemented because different parts of the model are present in each. We must be able to determine when these classes can be considered to be acceptable examples of a single treatment.

One approach to this question of limiting variation is to specify the key elements of a model, and require that those be present while letting the other classroom activity vary as it might. In our experience in trying to develop measures along these lines for a follow-up study, the approach had two complications. First, the nature of some models is to influence the entire pattern of interaction--these are the same ones that only specify principles--therefore, their key elements are of the same type: general principles which refer to all interaction, not just limited components. Second, some models do not really know themselves well enough to identify their most important parts. Their essential components therefore, tend to cover the same areas as the full model description, but are stated more generally.

We are still left then with the need to set standards for full implementation; to determine when variation is acceptable and when it is so great that a treatment is only partially implemented. Unless these limits can be set, there will be no acceptable operational definition of full implementation. As yet we have not satisfactorily resolved this issue within the framework of a systematic measure. This should not be interpreted as a statement that sponsors cannot personally distinguish between good and poor examples of their models. They hopefully can do that very well. We are only saying that thus far sponsors have not communicated their standards adequately enough to allow others to make systematic judgments.

Another issue in trying to measure the extent of treatment implementation is comparability. That is, whether one judgment about implementation can be compared with another. Comparability among raters is a concern which underlies the previous discussion about the importance of clearly defined treatments and explicit standards of implementation. Talking first about comparability within models, if judgments about the extent of implementation within a model were all made by one person, those concerns could be by-passed. If one person knew the model well and was familiar with all classes working under the model, he or she could probably tell us with a fair amount of reliability which classes were acceptable examples of the model or could roughly rank the classes in order of their performance without necessarily being able to articulate their criteria or definition of the model. The problem in HSPV was that for the most part, no single person, including the sponsor knew all the classes well enough to make good judgments.

We were faced then with different people making the ratings and in this situation external standards become important. Making comparisons without them is dangerous because judgments are based on the rater's personal standards and definition of implementation. As a result, different raters may be using the instrument differently such that one class would receive different ratings depending on who makes them. The relative rating of one site over another may have no correspondence with reality if the ratings are done by different people. This was the case with the sponsor ratings which had no explicit standards. Even with the consultant checklists where the components were listed, there were problems of comparability. In many cases the components were still general enough that ratings depended on observer judgments and inferences and a small reliability study showed that the inter-rater reliability varied widely.

Until now I have been talking only about comparability within models. Making comparisons among models is even more complicated, if not impossible. When treatments differ in their goals and levels of operationalization as well as in their methods, as was the case in HSPV, it is difficult to determine whether model A is implemented to the same extent as model B.

One might suppose that a common implementation measure for all models would solve this difficulty. Closer consideration shows, however, that a common measure will raise the same comparability issues that model-specific measures do because they still require judgments which are made on very different grounds from model to model. There is no basis for assessing, for example, whether the implementation of clearly specified prescriptions is equivalent to the implementation of general principles. There is no basis, therefore, for saying that 75% implementation in one model is comparable to 75% implementation in another.

From our work in HSPV, then, we are convinced of the importance of documenting the extent of treatment implementation. We learned, first, that the treatments did not meet the traditional laboratory standards of replication. There was a great deal of variation among classrooms both because some teachers only used parts of the model and because some models, as part of their philosophy, encourage variation among classes. We learned, second, how difficult it is to develop an adequate measure of implementation. Ideally, we would have liked to work with treatments which were clearly specified in operational terms and with an agreed-on standard of implementation so that the bases for judgments about implementation would be explicit and therefore comparable when done by a number of different raters. In HSPV, this ideal was not met. We also found, however, that some models came closer to the ideal than others. The behavorist models generally fit inot an experimental framework more easily than the other programs. Because they prescribe specific classroom activities more than other models,classes within those models tend to be more like one another on dimensions related to the models than do classes within other models. Similarly, the implementation of these models is more readily measured because the presence of their components can be more easily observed within a limited time. This is not to say that the behavior models are necessarily superior on theoretical grounds, only that they are more amenable to the experimental situation. The models based on general principles which encourage individual adaptation are more difficult to fit to the experimental mold and their assessment seems to suffer as a result.

More importantly, it appears that the reasons for our difficulty in HSPV, particularly with the latter group of models, are factors which will occur in other studies of this kind. Where the treatments are not independent variables in the traditional sense, but are educational and other social programs, treatment descriptions may always have elements of flux, variability, and generality. Not only will these treatments be difficult to observe, but it will be hard to specify what, short of exact replication, constitutes full implementation.

Thus, it may always be difficult to obtain clean quantitative measures of the extent to which educational programs are implemented. Two responses to that difficulty are possible.

First, we can limit our studies to those treatments which are conducive to experimentation. While this approach does reduce the measurement problems, it can also be argued that since practice refines theory and since participation in social experiments tends to be good publicity, the approach will artificially restrict the types of services offered in the country.

A second approach would be to continue to study all types of programs, and to rely on impressionistic judgments about implementation more heavily than quantitative measures. In this approach we acknowledge that we cannot develop the sound measurs we would like to have and bolster the measures we can develop by having the judgments made by people who know the model well and understand the evaluator's purpose in wanting the judgments and by having one person visit all the classes in a model.

Regardless of which approach is used, it is important to recognize that by some means we must determine what any program or treatment is in practice before we can justifiably draw conclusions about its effects.