DOCUMENT RESUME

ED 041 401 24 EA 003 024

AUTHOR	Huberty, Carl J.
TITLE	Observational Procedures in Program Evaluation.
INSTITUTION	Georgia Univ., Athens. Research and Development Center in Educational Stimulation.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	RP-25
BUREAU NO	BR-5-0250
PUB DATE	Jul 70
CONTRACT	OEC-6-10-061
NOTE	17p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.95
DESCRIPTORS	*Classroom Observation Techniques, Educational Innovation, Language Programs, *Prereading Experience, Preschool Education, *Preschool Evaluation, *Program Development, *Program Evaluation, Verbal Communication

ABSTRACT

The use of standardized instruments has demonstrated the ineffectiveness of traditional educational programs in providing optimum learning situations for large numbers of children with diversified backgrounds. Consequently, many new innovative programs have been designed and implemented for preschool children. However, it has become apparent that a change in evaluation instruments and techniques is also needed. Observational techniques of evaluation over varying periods of time -- as opposed to formal or standardized testing -- are proving to be effective in program development and evaluation. Some evaluative implications were made from data gathered via a checklist of behavioral symptoms of young children. A prereading inventory of eight behavioral symptoms examined the readiness and progress of learning in a preprimary language arts program. Subjects used in this 2-year study were 3-to-6-year-old children in a southeastern suburban area representatively selected with respect to socioeconomic status and level of intelligence. (Author)

Research Paper No. 25

OBSERVATIONAL PROCEDURES IN PROGRAM EVALUATION

Carl J. Huberty

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

July, 1970

Research and Development Center in Educational Stimulation
University of Georgia
Athens, Georgia

This research reported in this publication was performed as part of the activities of the Research and Development Center in Educational Stimulation, University of Georgia, pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education, under Provisions of the Cooperative Research Program.

Center No. 5-0250 Contract No. OE 6-10-061

INTRODUCTION

In recent years great emphasis has been placed upon developing educational programs designed to meet the needs of all children more effectively than the traditional programs which, on the whole, seem to have favored middle class children. The standardized instruments designed to evaluate these programs likewise stressed the values taught in the traditional programs. Evaluation is an essential aspect of any program, since it provides feedback with which to judge students' performances and thus the effectiveness of the curriculum in meeting the students' needs and the program goals. Specifically, testing estimates the extent to which a student has developed a specific type of knowledge or skill (Bussis, 1965). The use of standardized instruments has demonstrated, to some extent, the ineffectiveness of the traditional program in providing optimal learning situations for large numbers of children. As a result, many innovative programs have been designed and implemented for preschool children to help them adapt to the new programs that are being started in schools.

However, it has become increasingly evident that a change in evaluation instruments and techniques is also needed. Since most of the current instruments were designed for, and standardized upon, middle class white children, they require verbal skills, knowledge, and experience present in the typical middle class environment. Thus, the instruments are not appropriate for use with children reared in environments differing from those of the middle class (Weick, 1954). The verbal orientation penalizes the very young child and especially the culturally deprived who come from an environment in which verbal communication is not greatly encouraged. Almost as debilitating is the

meaninglessness of the subject matter of standardized tests for those who live in different environments. Finally, the standardized instruments require sustained attention levels which are difficult, if not impossible, for young children.

In addition to the standardization subjects and materials causing limitations for the appropriate use of these instruments, they were often designed using different rationales than are now used in current theory and research. Knowledge concerning learning processes in children has greatly expanded since most standardized tests were developed. For example, the concept of intelligence itself is considered by some to have changed from an inherited ability to acquisition of skills (Dobbin, 1966). Therefore, these older instruments test from a different frame of reference than that which current programs emphasize (Bussis, 1965). As a result of these different bases for development of standardized instruments, their use in evaluating current programs often results in conflicting, uninterpretable findings. In addition, standardized tests constitute artificial testing in that they are often not directly related to the program. It is important to test for the effects of learning in the daily activities of the child (Wright, 1967); however, standardized instruments may interfere with the normal course of events (Wright, 1960; Caldwell, 1969), so that test results give no feedback of what the child and his activities are really like in order to improve the program.

Observational techniques have been suggested as a solution to some of these problems in the evaluation of programs for young children. The technique of observation may be defined as a systematic recording in objective terms of behavior in process of occurring, in a manner that will yield quantitative, individual scores (Gellert, 1955, p. 179). Gellert (1955) and

Wright (1967) elaborated upon the features that make observational methods useful in solving the discussed problems related to research with children:

(1) Observation better suits young children with little verbal facility, since observers can see whether a child has developed an understanding that he is unable to express verbally.

(2) Children are more natural in the presence of observers than they are in a formal testing situation.

(3) Observation, in contrast to standardized tests which cause manipulation within the environment, does not interfere with the stream of events, thereby letting things happen as they may.

(4) Since behavior is recorded as it occurs, the ambiguities of projective tests are avoided and the omissions and distortions obtained from later recall of events are minimized.

(5) Observation instruments can be tailored to meet specific needs and goals of a program, thus staying relevant for the new theories and research on programs.

The advantages of the use of observational techniques with young children are being increasingly recognized. A group in Santa Barbara, California, developed a new approach to prediction of school success, based on learning in kindergarten, called the Kindergarten Evaluation of Learning Potential (KELP) (Wilson and Robeck, 1966). The evaluation takes place as a continuing part of the learning situation in that the child is given the opportunity to learn to do the things that measure his potential, thus fusing testing and teaching. This procedure is also valuable in that it extends the

observation skills of the kindergarten teacher. The rationale behind the program assumes three levels of learning: (1) making appropriate associations; (2) grasping whole ideas, concepts; and (3) developing creative self-direction.

Recently, the New York City Board of Education (ETS, 1965) began a study of their problem of educating children having very diversified backgrounds; standardized tests were judged as being inappropriate for their purposes, especially for first graders. In addition, the currently available tests did not tell the teacher (a) how children learn, (b) how their intellect develops, (c) where they are in respect to some cycle of development, and (d) what the teacher may do to further development along the scale. They began with teacher observations to get actual, natural samples of the behavior of children. These samples were incorporated into a working model based upon Piaget's theory of development (Dobbin, 1966). A curriculum, along with observational scales, was developed, using as guides the achievement of certain specific skills by the child (this was similarly done with KELP). This program is still in the experimental development stage but is another example of the use of observational techniques in curriculum evaluation.

The use of observational techniques with young children is appropriate for three means of evaluation:

(1) The formative evaluation procedure is, usually, a short term evaluation (e.g., unit check list at the end of every small unit of learning), or an evaluation after every three months. Here, periodic feedback is provided for interpretive and evaluative judgments. This procedure may suffer from the shortcoming of delayed acquisition; e.g., a skill taught in November may not become apparent until April.

(2) The summative evaluation procedure appraises the program over a usually longer period of time (e.g., at the end of a school year) so as to help those concerned know when and to what extent the program has been effective.

(3) Through prediction, observational techniques can be used as a diagnostic tool to vary the program to meet the individual needs of the children. An instrument capable of predicting where a child might later have difficulty could conceivably provide the pertinent information needed in order to avert that difficulty. Early recognition of a deficit (in skills leading to reading, for example) which can be remedied easily will prevent the confounding effects of inability to read in other subject areas as a child progresses.

In developing an observational instrument for any of the preceding three uses, the main concern is that of validity--in this case, content validity. Of course, the third use, that of prediction, implies a concern for predictive validity. The problem of reliability (here the concern would be that of observer agreement, both among different observers and at different times) is a more difficult one, in terms of practicability, to solve. Several good references are available pertaining to the reliability and the construction and use of observation instruments (e.g., see Medley and Mitzel, 1963).

PURPOSE

The purpose of this study was to point out the usefulness of observational techniques in program development and evaluation. As an illustration,

some evaluative implications from data gathered via a check list of behavioral symptoms of young children were made.

SUBJECTS

The experimental subjects used in this study were three-, four-, five-, and six-year-old children in a southeastern suburban area. For the first year (1969) there were 78 three-year-olds, 63 four-year-olds, and 55 five-year-olds; in the second year (1970) there were 60 three-year-olds, 75 four-year-olds, 60 five-year-olds, and 50 six-year-olds. (The four-, five-, and six-year-olds of the second year were the three-, four-, and five-year-old children of the first year.) As can be noted, the attrition rate regarding the size of each group from 1969 to 1970 was relatively small. Subjects in the experimental school were representatively selected with respect to socio-economic status and level of intelligence. For further discussion of the selection of subjects and other organismic data, see the report by Huberty (1969) and another Research and Development Center publication* for a description of the curricular treatment to which the children were subjected.

INSTRUMENTATION

The Evaluation Division of the University of Georgia R & D Center in Educational Stimulation developed a "prereading" inventory based upon the procedure used by the New York City Board of Education. This observational technique was constructed in order to examine the readiness and progress of learning in the preprimary Language Arts program which was implemented at the experimental field center for the Research and Development Center.

---

*This publication will be released July 31, 1970. Copies can be obtained from Mrs. Gretchen McCann, Research and Development Program, U.S. Office of Education, Department of Health, Education, and Welfare, Room 3139, 400 Maryland Avenue, Washington, D.C. 20202.

In 1968-69, teachers of the ten preprimary groups of children ages three, four, and five listed various symptoms that they had observed in the classroom which they felt showed developmental progress and which they considered important behaviors to be demonstrated before the introduction of specific reading instruction. These lists were revised by teachers and evaluators. Symptoms were then categorized and arranged in sequential order.

Main categories were labeled in the following way: (1) Directions, (2) Dramatizing, (3) Being Read To, (4) Bookhandling, (5) Persons and Names, (6) Visual Discrimination, (7) Auditory Discrimination, and (8) Attempts to Read. Under each of these principal categories the individual symptoms which were judged pertinent are listed. Symptoms to be observed positively are, for example: "Orients book correctly," "Turns pages correctly," "Recognizes written names," and "Sees simple likenesses and differences." These are a few of the symptoms from the various categories.

On this inventory, teachers attempted to record the observed symptoms as they were exhibited by each child, noting the date when they observed a positive demonstration of the symptom. Thus a profile of an individual child's development was revealed as relevant symptoms became evident and were noted. Emphasis on the positive identification of evidence of progress to the exclusion of negative reports is a special feature of this approach; teachers report only what a child can do.

DATA COLLECTION

The check list was accessible to each teacher and the two teacher aides for each class within each age group for the month of May, 1969, (during 1970 the check lists were available for approximately six months). Each teacher or aide checked those symptoms as they were observed; then, in late

May of each year the recordings were completed via discussion among the teacher and two aides until concensus had been attained. Percentages of children exhibiting each symptom were calculated for each age group. In October, 1969, the director of the reading curriculum program* estimated separately the proportion of children in the three- through six-year age groups that he predicted would exhibit the symptoms, thus supplying expected percentages. (Up until this time, and after the task of predicting the percentages, this director was not involved in the compilation of this inventory and subsequent data analysis and interpretation, due to the fact that it was a separate and distinct project of the Evaluation Division of the Research and Development Center.) For example, (see Table 1, p. 14 - 16), he predicted that 80% of the four-year-olds would exhibit the symptom "Letter order" under the major heading of Visual Discrimination, whereas, only 32% of the 1969 four-year-olds actually exhibited this symptom, but 93% of the 1970 group displayed it. On the other hand, the director predicted that 50% of the three-year-olds would display the symptom, "Recognizes written names (others, some)" under the major heading of Persons and Names, whereas 76% of the 1969 three-year-olds and 62% of the 1970 three-year-olds actually exhibited this symptom.

It should be noted that in the present study the check list or inventory was employed for purposes of "summative" rather than "formative" program evaluation or for establishing potent predictors of success in reading.

IMPLICATIONS

Implications drawn from the agreement and disagreement between the estimated and actual figures between age groups and between data collection years

---

*Special thanks are due Dr. George Mason who was kind enough to perform this task for the Evaluation Division of the Research and Development Center.

were formulated. It should be noted that the symptoms listed were assumed by the teachers, in and by themselves, to be important considerations in reading program development. It must be realized, of course, that added experience with preschool children and changes in the reading program may produce a change in the list of symptoms. In fact, by the fall of 1969 new knowledge about the experimental reading program dictated necessary changes in the current instrument (some symptoms then appearing inappropriate). The original check list was retained, however, so as to gain information with respect to year-to-year change in observations.

Several possible percentage combinations may result which give rise to potential questions relevant to program revision. Comparisons of percentages may be made in either of two manners: 1) between age groups within data collection years, and 2) within age groups between data collection years. An inspection of Table 1 facilitates these comparisons and reveals specific (though in some cases, isolated) examples of the possible implications to be touched upon.

If expected percentages of children exhibiting the behavioral symptoms are not, in fact, obtained, [e.g., as in the case of the symptom, "Asks to read from certain book (even if not able)" for the age groups four and five in Table 1] perhaps such an outcome would call for a reevaluation of the program goals or another look at the capability of the children. The group may have been poorly evaluated in terms of readiness or IQ or in terms of appropriateness of the program. If such is the case, a reevaluation through a closer look at the objectives, materials, or instruction may be necessary. If, on the other hand, expectations are surpassed (with respect to a given age group) [e.g., such was the case for age groups four and five regarding the symptom, "Retains delayed directions"], the explanation may be that (1) some phase of

the program has been overemphasized at the expense of some other phase of the program, or (2) the group may have been poorly evaluated from the beginning.

In either case (differences between the expected and observed percentages as stated in the preceding two paragraphs), discrepancies can occur, of course, between both high and low percentages. For instance, in the age five group the expected percentage for the symptom, "Composes original story," was a high 100%, while observed percentages for both the 1969 and 1970 groups were considerably lower (74% and 40%, respectively). On the other hand, for the age three group, the expected percentage for the symptom, "Writes names (others, some)" was 5%, a low estimate with somewhat lower observed percentages for both the 1969 and 1970 groups (1% and 3%, respectively). Other examples in Table 1 reveal similar tendencies for other symptoms, but in the opposite direction (i.e., higher observed percentages than expected percentages in both a high and low range). For example, the observed percentages for the symptom, "Follows simple directions (group)," were greater for the age three group (1969 group--95%, 1970 group--98%) than the expected percentage, which was 75%. The symptom, "Knows where ending of book is," revealed an observed percentage for the age three, 1969 group, of 76% and for the 1970 group of the same age the percentage was 87%, while the expected percentage was 30%.

If the same instrument is used for evaluation across different age groups, some idea of retention may be obtained. For example, at age four there may be a certain objective of the program which at age five is no longer a specific objective of the program. If the same instrument is utilized, some idea of how that particular skill is retained may become apparent.

In summary, if expected percentages are not attained, or those attained, not expected, at any given age, the direction of the program and/or the expectations may need to be changed. Also, those symptoms that younger children exhibit and

older ones do not, suggest that, (a) changes need to be made in the inventory since different behavioral signs need to be considered; (b) retention of the skills over age should be questioned; or, (c) a decrease in emphasis on the symptom in the program writing could be the explanation.

One example of use of data such as these which have been discussed can be found in a report by Mason (1970).

DISCUSSION

Inventories developed which are relevant to curricular programs provide the opportunity to check on the appropriateness of the specific objectives of a given program as well as general objectives that would be considered important by substantive experts. Such inventories assessing developmental programs may point out any need for revision in order to more closely meet the needs of the program. More generalized objectives, or those objectives which are goals of any program, may, on the other hand, be more invariant. Thus, with any given program, it can be seen that there may be specific program objectives as well as general objectives, and inventories can be constructed to meet a variety of needs, depending on whatever the program coordinator feels is appropriate.

As with many instruments in the developmental stage, changes in the inventory are necessary to meet the evaluation needs of changing curricula. Objectives may change, or different instruments for different age groups may be necessary. The changes and revisions will depend, of course, upon the goals of the program and the objectives of the evaluation. Hence, some of the selected items on an instrument may not be appropriate for programs in succeeding years, or for programs based on different theories of learning and instructions, or for use in prediction.

Another possible outgrowth of the development of such an observational inventory or check list is that of specifying content areas that will define items for an objective test. That is, the instrument may serve as a means to an end as well as an end in itself.

The purpose of this report was to point out the need and usefulness of observational techniques in program evaluation. Although data presented here were considered summative data, probably a more sound idea would be to consider such an instrument an integral part in a continuous, on-going evaluative process. One such approach to overall program evaluation that may be followed is the CIPP (context, input, process, and product evaluation) model (Stufflebeam, et al. [in press]).

Table 1

Actual and Expected Percentages of Experimental Children Exhibiting Prereading Skills Symptons

Symptom	Age 3			Age 4			Age 5			Age 6	
	Exp. %	1969	1970	Exp. %	1969	1970	Exp. %	1969	1970	Exp. %	1970
Directions											
A. Follows simple directions--one step											
1. Individual	90	100	100	95	100	100	100	100	100	100	100
2. Group	75	95	98	85	95	100	95	100	100	100	100
B. Follows multiple directions--two or three steps											
1. Individual	70	87	85	90	94	100	100	100	100	100	100
2. Group	50	87	80	80	90	99	90	100	100	95	90
C. Retains delayed directions	50	35	48	75	84	95	80	100	87	85	92
Dramatizing											
A. Spontaneous reacting to pictures or stories (observable reacting)	100	87	53	100	79	100	100	78	63	100	74
B. Finger play-attending	95	100	100	100	100	100	100	100	100	100	68
C. Finger play-participating--acting out	90	100	93	100	100	100	100	100	100	100	68
D. Role-playing											
1. Spontaneous	80	87	77	85	81	95	90	89	43	95	92
2. Directed	90	100	97	100	94	51	100	100	100	100	66
E. Composes original story	75	53	50	85	71	43	100	74	40	100	66
Being Read To											
A. Comes readily to reading	85	88	98	85	86	100	90	100	55	90	92
B. Listens to stories											
1. With pictures	90	97	100	95	98	100	100	100	100	100	100
2. Without pictures	80	0	38	95	37	60	100	100	100	100	66
C. Asks to have reading done	80	29	50	90	57	80	95	100	62	95	100
D. Listens to stories											
1. Short	90	100	100	100	100	100	100	100	38	100	100
2. Medium	85	100	93	90	89	100	95	100	98	100	100
3. Long	75	74	88	80	40	72	85	100	58	85	92
E. Gets book to retell story	50	26	63	75	24	32	75	85	28	75	66

Bookhandling											
A. Book is to read	100	92	100	100	100	100	100	100	100	100	100
B. Orients book correctly	50	88	100	100	100	100	100	100	100	100	100
C. Knows where beginning is	30	79	75	100	100	100	100	96	100	100	100
D. Knows where ending is	30	76	87	100	100	100	100	95	100	100	100
E. Knows where title of book is	0	90	100	90	97	100	100	85	67	100	100
F. Knows where front of book is	30	90	98	95	100	100	100	100	100	100	100
G. Knows where back of book is	30	92	98	95	100	100	100	100	98	100	100
H. Turns pages correctly (R-L)	30	90	95	100	100	100	100	100	100	100	100
I. Knows where top of book is	30	94	87	100	100	100	100	100	100	100	100
J. Knows where bottom of book is	30	95	87	100	100	100	100	100	100	100	100
K. Left to right reading	0	51	88	100	62	92	100	100	95	100	100
L. Rows-return sweep	0	41	80	100	62	91	100	100	92	100	100
M. Observes relationship of text and pictures	30	86	82	100	90	84	100	100	97	100	100
Persons and Names											
A. Recognizes oral names											
1. Self	100	100	98	100	100	100	100	100	100	100	100
2. Others	95	100	98	100	100	100	100	100	100	100	88
3. Absentees	80	82	93	95	92	100	100	100	100	100	94
B. Recognizes written names											
1. Self	100	99	100	100	100	100	100	100	100	100	100
2. Others (some)	50	76	62	100	60	93	100	93	73	100	98
C. Writes names											
1. Self	50	41	37	100	81	77	100	100	98	100	100
2. Others (some)	5	1	3	60	33	51	90	78	30	100	96
D. Spells names (orally)											
1. Self	25	51	70	100	73	69	100	87	73	100	100
2. Others	5	1	20	60	17	13	80	28	0	90	76
Visual Discrimination											
A. Seeing simple likenesses and differences											
1. Knows likenesses	90	87	98	100	100	100	100	100	100	100	100
2. Knows differences	90	81	90	100	97	100	100	100	100	100	100
3. Discriminates between likenesses and differences	90	82	90	100	97	100	100	100	100	100	100
B. Distinguishes letters from other forms (any)	75	79	90	100	73	80	100	100	100	100	100
C. Distinguishing words											
1. First letters	75	47	82	100	70	99	100	96	85	100	96
2. Letter order	50	54	72	80	32	93	90	84	85	100	96
3. Whole words	60	65	73	90	62	100	100	84	85	100	100

Auditory Discrimination											
A. Distinguishing familiar sounds	90	100	100	100	98	100	100	100	100	100	100
B. Rhyming	50	33	0	90	94	100	95	73	100	100	100
C. Letter sounds											
1. Beginning sounds	10	0	0	85	44	96	100	87	95	100	96
2. End sounds	25	0	0	85	14	64	95	69	42	95	84
Attempts to Read											
A. Relates to book--responds to pictures in books	95	9	65	100	94	100	100	100	100	100	98
B. Attaches name he originates for book	80	5	30	90	62	100	30	56	48	20	58
C. Reads (tells) stories from pictures in unfamiliar book	75	10	27	85	57	0	90	76	93	80	100
D. Asks to read from certain book (even if not able)	10	0	25	60	30	0	80	40	65	100	100
E. Learns words	15	72	0	100	71	97	100	100	100	100	100
F. Spontaneously picks out words	15	10	0	100	30	97	100	89	83	100	98
G. Asks help with reading	5	12	0	100	51	59	100	76	63	100	86
H. Reads whenever asked to	0	10	0	70	73	97	100	100	100	100	100
I. Resists interruptions of reading	0	0	0	50	38	32	75	55	33	80	36
J. Helps others read	0	0	0	50	29	75	60	24	23	70	70
K. Brings library or other books to read (himself)	0	1	0	20	38	0	30	47	40	50	100

REFERENCES

Bussis, A. M. From theory to the classroom. New York: Board of Education of the City of New York, 1965.

Caldwell, B. M.; Fannenbaum, J. A.; & Honig, A. S. Patterns of information processing used by and with young children in a nursery school setting. Mimeographed paper of University of Arkansas, College of Education, Center for Early Development and Education, 1969.

Dobbin, J. E. Some observational techniques for appraising development. Paper presented at the University of Georgia Teacher Education Conference, Athens, Georgia, January 20, 1966.

Educational Testing Service. Let's look at first graders. New York: Board of Education of the City of New York, 1965.

Gellert, E. Systematic observation: A method in child study. Harvard Education Review, 1955, 25, 179-195.

Huberty, C. J. Basic assessment of subjects in the Clayton County model. Research Paper No. 23. Athens, Georgia: Research and Development Center in Educational Stimulation, University of Georgia, November, 1969.

Mason, G. E. The language arts program in the Clayton County project: The third year. Technical Paper No. 6. Athens, Georgia: Research and Development Center in Educational Stimulation, University of Georgia, January, 1970.

Medley, D. M. & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. E. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Stufflebeam, D. L., et al. Educational evaluation and decision-making. Itasca, Illinois: Peacock Publishing Company, in press.

Weick, K. E. Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology, (2nd ed.), Vol. 2, Reading, Mass.: Addison-Wesley, 1954.

Wilson, J. A. R. & Robeck, M. C. Kindergarten evaluation of learning potential. New York: McGraw-Hill, 1966.

Wright, H. F. Observational child study. In P. H. Mussen (Ed.), Handbook of research methods in child development. New York: John Wiley & Sons, 1960.

Wright, H. F. Recording and analyzing child behavior. New York: Harper & Row, 1967.