DOCUMENT RESUME

ED 038 807 EC 005 496

AUTHOR House, Ernest R.; And Others
TITLE The Visibility and Clarity of Demonstrations.
INSTITUTION Cooperative Educational Research Lab., Inc., Northfield, Ill.
SPONS AGENCY Illinois State Office of the Superintendent of Public Instruction, Springfield.
PUB DATE 69
NOTE 78p.; Appendices

EDRS PRICE EDRS Price MF-$0.50 HC-$4.00
DESCRIPTORS *Demonstration Centers, *Demonstrations (Educational), Educational Innovation, Educational Practice, Educational Programs, Evaluation Techniques, *Exceptional Child Research, *Gifted, Measurement Instruments, Observation, Program Evaluation, State Programs
IDENTIFIERS Illinois, Illinois Gifted Program

ABSTRACT

A state system of about 20 demonstration centers was developed in Illinois to exhibit a variety of model programs for gifted children, ranging from kindergarten to high school. Subjects ranged from foreign language to dance and dramatics. Evaluation indicated low quality in too many centers. The centers performed best on the awareness function, less well on the acceptance function. Demonstrations were found to lack intelligibility and to fail to illustrate both positive and negative features, thereby facilitating valid professional judgment. However, they ranked well for fidelity. Recommendations are made; a separate volume provides appendixes listing observed programs, describing a typical day in a center, and detailing procedures, the instrument used, and the obtrusiveness of measures. (Author/JD)

ED038807

THE VISIBILITY AND CLARITY OF DEMONSTRATIONS

BY

Ernest R. House

Thomas Kerins

Stephen Lapan

Joe M. Steele

MAY, 1969

Gifted Evaluation Project

COOPERATIVE EDUCATIONAL RESEARCH LABORATORY, INC.

Box 815, Northfield, Illinois 60093

ED038807

ED038807

THE VISIBILITY and CLARITY of DEMONSTRATIONS

Ernest R. House, Thomas Kerins, Stephen Lapan, Joe M. Steele

Assisted by

Sharon Gotch, Billie Hoyt, Judith Miley

Supported by

The Department of Program Development for Gifted Children

OFFICE of the SUPERINTENDENT of PUBLIC INSTRUCTION

Ray Page, Superintendent of Public Instruction

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

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Gifted Evaluation Project

COOPERATIVE EDUCATIONAL RESEARCH LABORATORY, INC.

Box 815, Northfield, Illinois 60093

MAY, 1969

OFFICE of the SUPERINTENDENT of PUBLIC INSTRUCTION

Ray Page - Superintendent of Public Instruction

Verne E. Crackel - Deputy Superintendent

John H. O'Neill - Associate Superintendent

Robert D. Cain - Assistant Superintendent
Division of Special Education Services

Herbert Baker - Director
Program Development for Gifted Children

MEMBERS of the STATE ADVISORY COUNCIL for the Gifted

Robert Bauernfeind
Professor of Education, Northern Illinois University

Mrs. Evelyn Carlson
Associate Superintendent of Schools, Chicago

Ray Caton
County (McLean) Superintendent of Schools, Bloomington

A. Gordon Dodds
Superintendent, Edwardsville Community School, District

David M. Jackson
Director, Cooperative Educational Research Laboratory, Inc.
Northfield, Illinois

Blair Plimpton - Chairman
Superintendent of Schools, District 44, Park Ridge

Richard Verduin
Coordinator of Professional Experiences
College of Education, Southern Illinois University

CONTENTS

APPENDICES (published in a separate volume)

I. WHAT IS A DEMONSTRATION?

The Illinois Gifted Program operates a system of approximately 20 demonstration centers intended to exhibit a variety of model programs for gifted children that range from kindergarten to high school. Subjects range from foreign language to dance and dramatics. (See Appendix A: "List of Centers and Programs Observed".) In all cases the centers are situated within school districts. They are located in different areas of the state, although most centers are in the Chicago Metropolitan area.

In order to visit a center, the visitor (usually a public school administrator or teacher) submits a formal request that the center acknowledges by specifying the day for the visit. After an orientation at the center, the visitor observes the demonstration classes. Often he also has an opportunity to talk with the teachers and students. After the visit, the demonstration director may offer to help the visitor with his own gifted program. (See Appendix B for a description of the typical day's visit.) The administrator or teacher may be reimbursed for his expenses from funds that his district receives from the Illinois Gifted Program.

The original rationale for the Illinois Demonstration Centers recognized three immediate operational goals for the centers[1]:

A. Awareness- Helping teachers and administrators become aware of innovations and ways to improve the quality of their programs.

B. Acceptance- Helping visitors decide whether the change or innovation is acceptable for him personally, to his district, and to his community.

C. Adoption - Helping schools adapt or adopt particular programs or procedures in which they are interested.

FIGURE 1[2] exemplifies how the demonstration centers might hope to accomplish these goals.

[1] William Rogge, "A Rationale for Demonstration Centers," Demonstration Director's Handbook, Mimeo., November 1965

[2] Ibid.

FIGURE 1: EXAMPLES OF PROCEDURES FOR EACH OBJECTIVE AND TWO KINDS OF INNOVATIONS OF DEMONSTRATION CENTERS

Innovation	Awareness (Information)	Acceptance (Attitudes)	Implementation (Behavior)
1. Acceleration of content	Brochures Speeches Articles Visits with administrators	Interviews with parents and students Summary of research findings	Consulting by director Exchange of Teacher
2. Modern Mathematics	Brochures Teach visitor a lesson Visit a class Talk to teachers	Interview students Talk to Teachers Read authoritative statements	Extension classes Joint in-service training programs Consulting by demonstration teachers

II. WHAT IS A "GOOD" DEMONSTRATION?

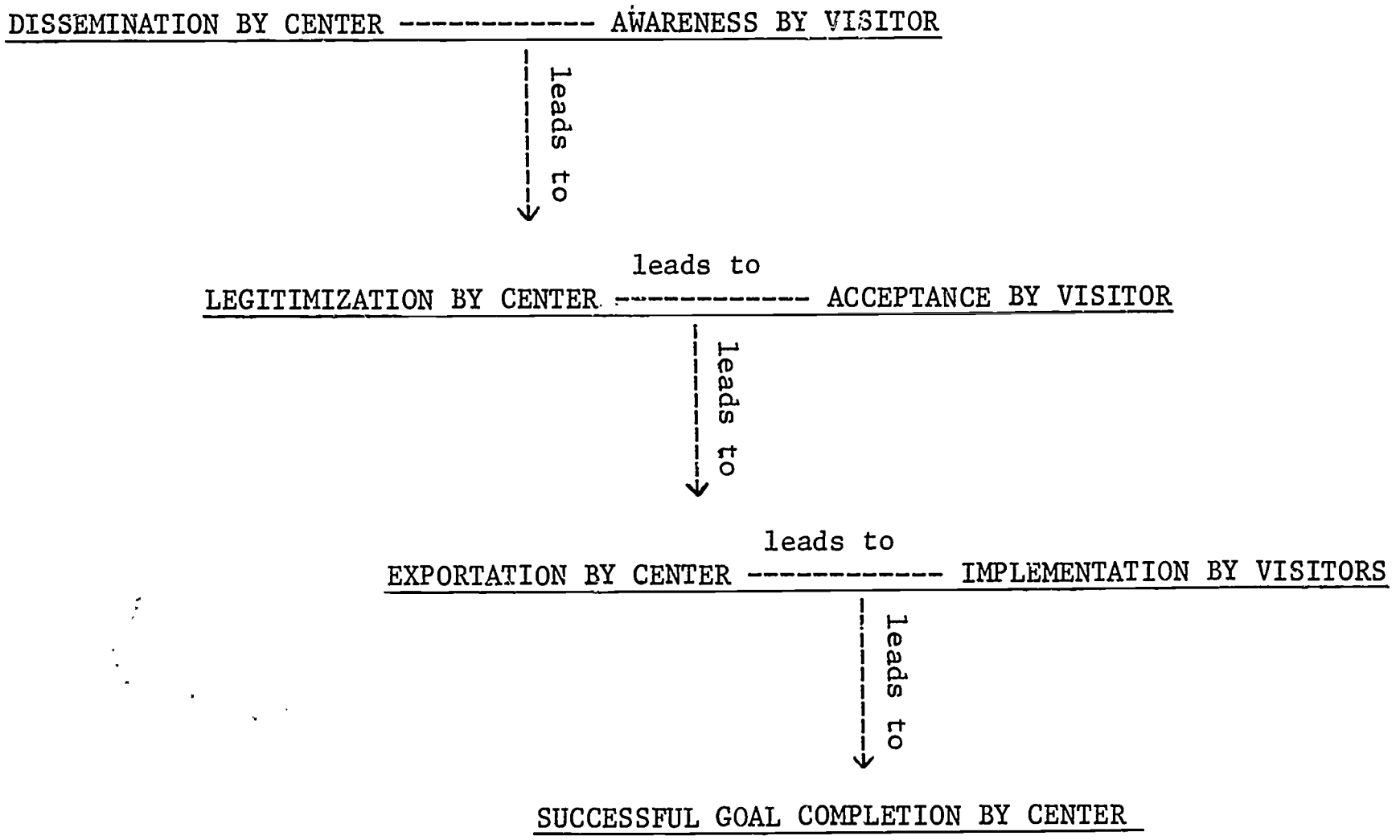
The success of the demonstration centers might be represented by Figure 2.

FIGURE 2: MODEL FOR DEMONSTRATION CENTER SUCCESS

IF THE VISITOR IS AWARE OF THE CENTER'S ACTIVITIES,
THE CENTER HAS ACCOMPLISHED ITS GOAL OF DISSEMINATION.

IF THE VISITOR ACCEPTS THE CENTER'S ACTIVITIES,
THE CENTER HAS ACCOMPLISHED ITS GOAL OF LEGITIMIZATION.

IF THE VISITOR IMPLEMENTS THE CENTER'S ACTIVITIES,
THE CENTER HAS ACCOMPLISHED ITS GOAL OF EXPORTATION.



The Illinois conceptualization of demonstration centers closely approximates the "diffusion" phase of the Clark-Guba model which divides this phase into a "dissemination" (awareness) stage and a "demonstration" (acceptance) stage.[3]

[3] David L. Clark and Egon G. Guba, "An examination of Potential Change Roles in Education," Seminar on Innovation in Planning School Curriculum, October 1965.

The purpose of the "dissemination" stage (Figure 3) is to inform about innovation:

> "It is the purpose of dissemination to create widespread awareness of the inventions among practitioners, that is, to inform or tell practitioners about the performance and process aspects of the invention. The criteria which are appropriate for the evaluation of dissemination activities include intelligibilty (is the message clear?), fidelity (does the message give a valid picture?), pervasiveness (does the message reach its intended audience?), and impact (does the message affect key targets?). The essential activities of dissemination are reporting and interpreting; these activities perform the function of informing about the innovation."[4]

The Clark-Guba model's (Figure 3) "demonstration" stage affords an opportunity for the target system to examine and assess the operating qualities of the invention, equivalent to what the Illinois Centers call "acceptance":

> "The criteria appropriate to an evaluation of demonstration functions thus seems to me to include credibility (is the demonstration convincing and does it build conviction?), convenience (is the demonstration accessible to those practitioners who ought to see it?), and evidential assessment (does the demonstration illustrate both positive and negative factors related to the invention so that the observer may reach a valid professional judgment about its utility?). The essential activities of demonstration are production and staging, and its purpose is to build well-founded professional conviction in relation to the innovation."[5]

As one of their main goals, the Illinois demonstration centers also have established "adoption" or getting the target population to try out the innovation. This formulation conforms to what Clark and Guba call the "trial" stage of adoption. In this phase, the appropriate criteria include:

> How "adaptable" is the innovation to the local scene?
>
> How "feasible" is it in the local setting?
>
> How does the innovation "act" in this setting?

Thus, the Illinois Demonstration Centers operate in the middle three stages of the Clark-Guba change model: dissemination, demonstration, and trial adoption.

[4]Egon Guba, "The Change Continuum and Its Relation to the Illinois Plan for Program Development for Gifted Children," presented to a conference on Educational Change, March 1966

[5]Ibid

FIGURE: 3 A CLASSIFICATION SCHEMA OF PROCESSES RELATED TO AND NECESSARY FOR CHANGE IN EDUCATION

	RESEARCH	DEVELOPMENT		DIFFUSION		ADOPTION		
		INVENTION	DESIGN	DISSEMINATION	DEMONSTRATION	TRIAL	INSTALLATION	INSTITUTIONALIZATION
OBJECTIVE	To advance knowledge	To formulate a new solution to an operating problem or to a class of operating problems, i.e., to innovate	To order and to systematize the components of the invented solution; to construct an innovation package for institutional use, i.e., to engineer	To create widespread awareness of the invention among practitioners, i.e., to inform	To afford an opportunity to examine and assess operating qualities of the invention, i.e., to build conviction	To build familiarity with the invention and provide a basis for assessing the quality, value, fit, and utility of the invention in a particular institution, i.e., to test	To fit the characteristics of the invention to the characteristics of the adopting institution, i.e., to operationalize	To assimilate the invention as an integral and accepted component of the system, i.e., to establish
CRITERIA	Validity (internal and external)	Face Validity (appropriateness) --- Estimated Viability --- Impact (relative contribution)	Institutional Feasibility --- Generalizability --- Performance	Intelligibility --- Fidelity --- Pervasiveness --- Impact (extent to which it affects key targets)	Credibility --- Convenience --- Evidential Assessment	Adaptability --- Feasibility --- Action	Effectiveness --- Efficiency ---	Continuity --- Valuation --- Support
RELATION TO CHANGE	Provides basis for invention	Produces the invention	Engineers and packages the invention	Informs about the invention	Builds conviction about the invention	Tries out the invention in the context of a particular situation	Operationalizes the invention for use in a specific institution	Establishes the invention as a part of an ongoing program; converts it to a "non-innovation"

Dealing only with dissemination and demonstration stages, this current report concentrates on the demonstration centers in their diffusion role. In this context, the centers may be diffusing an inferior program very well or an excellent program very ineptly.

In considering how these criteria might be applied to the centers, two techniques seemed feasible: one was to send observers into the centers and through direct observation to obtain information about what the centers were doing; the other was to collect the perceptions of regular visitors to the centers.

We decided to use both techniques. Direct observation would be particularly productive in focusing on such criteria as intelligibility, fidelity, and evidential assessment. Visitors' perceptions would focus on such criteria as convenience, credibility, and feasibility, which were more relative to the visitors' positions.

Considerable overlap was built into the two instruments for testing the observer's reliability and the visitor's reactions. This dual approach also fits one of the tenets of the total evaluation--that rather than relying exclusively on outcome measures considerable description of activities was highly desirable.[6]

Whenever possible, establishing the existence of a phenomenon (be it a demonstration or a program) before attributing causal effects to it seems worthwhile. In this respect, this report might be considered a description of the stimulus, i.e. the demonstration. Later reports will be studies of the response, i.e. visitor reaction.

This phase of the evaluation study sought to describe the treatment (demonstration) as fully as possible and to look at the variations in treatment among the demonstration centers. To that end, a 41-item observation schedule descriptive of a full day's activities at a demonstration center was constructed.* After obtaining reliability (See Appendix C: "Procedures"), two observers were sent simultaneously to each of 20 demonstration centers where they proceeded through the demonstration as though they were visitors and each marked his own observation schedule independently. The data in this report is based on a summary of these observations. Comparative data from regular visitors will be presented in a subsequent report.

[6]Ernest R. House, "Rationale For Evaluation of the Illinois Gifted Program" in Newsletter, Council of State Directors of Programs for the Gifted," May, 1968, Vol. 2, No.5 pp. 17-23 and reprinted in Illinois Journal of Education, October 1968, pp. 68-73

*The Evaluation Center at Ohio State helped construct the original instrument.

The major questions to be asked then were:

"Is the message clear?" (Intelligibility)
"Does the message give a valid picture?" (Fidelity)
"Is the demonstration convincing and does it build conviction?" (Credibility)
"Does the demonstration illustrate both positive and negative factors related to the invention so that the observer may reach a valid professional judgment about the ability?" (Evidential assessment)

In order to clarify the scheme, the 41 items were presented in five major sections based on components more relevant to the Illinois centers. Figure 4 enumerates these rating scale sections.

FIGURE 4: RATING SCALE SECTIONS

RATING SCALE SECTIONS	DIFFUSION: DISSEMINATION OBJECTIVES OF THE CENTER	DIFFUSION: LEGITIMIZATION OBJECTIVES OF THE CENTER
1) EXPLANATION OF PROGRAM	TO INFORM VISITORS BY... CREATING AN AWARENESS OF THE PROGRAM (INTELLIGIBILITY)	
2) EXPLANATION OF CLASS	TO INFORM VISITORS BY... CREATING AN AWARENESS OF THE CLASS (ES) TO BE OBSERVED (INTELLIGIBILITY)	
3) OBSERVATION OF CLASS		TO BUILD CONVICTION BY... AFFORDING VISITORS AN OPPORTUNITY TO EXAMINE AND ASSESS QUALITIES OF THE DEMONSTRATION ACTIVITIES (CREDIBILITY & FIDELITY)
4) EXPLANATION OF EVALUATION		TO BUILD CONVICTION BY... EXPRESSING EVIDENTIAL ASSESSMENT (EVALUATION) OF ITS OWN PROGRAM TO VISITORS (EVIDENTIAL ASSESSMENT OF PROGRAM RESULTS)
5) EXPLANATION OF FEASIBILITY		TO BUILD CONVICTION BY... ESTABLISHING ITS EASE OF ADOPTION BY VISITORS (EVIDENTIAL ASSESSMENT)

III. HOW DO THE ILLINOIS CENTERS RATE?

The items comprising the "Explanation of Program," how well the demonstration program was described, are given in Figure 5. Each "x" represents a demonstration center. For each item the further the "x" is toward the left, the fuller the explanation.

FIGURE 5: EXPLANATION OF PROGRAM (VERBAL ORIENTATION)

		Detailed (6)	(5)	General (4)	(3)	Little (2)	(1)	None (0)
1.	Were program objectives explained? (what, why how, when)		xxx x	xxx	xx	xxx xxx	x	xxx x
2.	Were program treatments explained? (e.g., methods, materials, management)	xxx	xxx	xxx	xxx x	xxx x	xx	x
3.	Was a description of school population given? (for example racial, socio-economic level, relation to program)		x			xxx	xxx xxx	xxx xxx xxx x
4.	Student selection procedures explained? (for example, tests used, who tested,cut-off pt(s)., weighting, relation to program, grouping arrangements, availability of test results)		xx	xxx xxx	xx	xxx xx	xxx x	x
5.	Historical explanation of program(s) given? (for example, date begun, who started, why, growth of program)			xxx	xxx	xxx xx	xxx	xxx xxx
6.	State plan described? (e.g., parts listed, explained, illus., related to visitors)	x	xx	xxx	xxx	xxx x	xxx	xxx x
7.	Teacher selection criteria explained? (e.g., who chose, minimums, recruitment)						xx	xxx xxx xxx xxx xxx xxx
8.	Teacher training for demonstration program(s) explained? (e.g., courses, internship, in-service)			x			x	xxx xxx xxx xxx xxx xxx

Maximum score possible: 48
Range: 0-29
Mean: 13.5
Highest score obtained: 29

The items on which the centers do best are (1) explaining program objectives, (2) explaining program treatments, (4) explaining student selection procedures, and (6) explaining the total state plan. Even on these items, however, a sizable number of centers give very little explanation. Notably lacking in the program explanations is how the demonstration teachers are selected and trained.

Figure 6 gives the items used in explaining the class that was demonstrated. As a group the centers did less well on these items.

FIGURE 6: EXPLANATION OF CLASS (VERBAL ORIENTATION)

	Detailed (6)	(5)	General (4)	(3)	Little (2)	(1)	None (0)
9. Today's class objectives explained? (e.g., were they related to overall program objectives)			xx	xx	xxx x	xx	xxx xxx xxx x
10. Today's class treatment explained? (e.g., were they related to overall program objectives)		x	xxx	xxx xx	xxx x	xxx x	xxx
11. Student Selection procedures for this class explained? (e.g., tests used, who tested, cut-off pts., weighting, relation to program, grouping arrangements for class, availability of tests, non-gifted)			xxx	x	xxx xx	xxx	xxx xxx xx
12. Intraclass academic progress (scores) explained? (e.g., speed, problems)			x		xxx	xxx x	xxx xxx xxx xxx
13. Intraclass characteristics explained? (e.g., social patterns, interests, study habits)			x		xxx xx	xxx xxx x	xxx xxx x

Maximum score possible: 30
Range: 1-15
Mean: 6.6
Highest Score obtained: 15

Figure 7 deals with what the visitor might see being demonstrated in the classroom. The items deal with how faithful the demonstration is to what it is supposed to be, whether the situation is natural or artificial, and other factors that might impair the visibility of the demonstrations. Items 17, 18, 19, 20, and 23 are scored positively where the answer is "no."

FIGURE 7: OBSERVATION OF DEMONSTRATION CLASS

		YES (4)	YES (3)	INCONCLUSIVE (2)	NO (1)	NO (0)
14.	Did the day's lesson reflect the overall program objectives?	xxx xxx xxx x.	xxx	xxx xxx	x	
15.	Did the day's lesson reflect the overall program treatment?	xxx xxx xxx xxx	xx	xxx x		xx
16.	Was competence of teacher adequate?	xxx xxx xxx xxx xxx	xx	xx	x	
21.	Were visitors able to see class proceedings clearly?	xxx xxx xxx xxx xxx xxx x		x		
22.	Were visitors able to hear class proceedings clearly?	xxx xxx xxx xxx xxx xxx		xx		
24.	Were visitors given a definite opportunity to talk to teachers?	xxx xxx xxx xxx xxx xx	x			xx
25.	Were visitors given a definite opportunity to talk to students?	xxx xxx xxx x	x	xx	x	xxx xxx

		NO (4)	NO (3)	INCONCLUSIVE (2)	YES (1)	YES (0)
17.	Was orientation, background or review given visitors as part of class sequence?	xxx xxx xxx xxx		xxx x		x
18.	Did total class sequence seem artificial?	xxx xxx xxx xxx xxx x		x		
19.	Were children continually distracted by the presence of visitors?	xxx xxx xxx xxx xxx x	xx	xx		
20.	Was visitor behavior excessively disruptive?	xxx xxx xxx xxx xxx xxx x		x		
23.	Were additional classroom materials needed to follow lesson?	xxx xxx xxx xxx		xxx xxx		xx

Maximum possible score: 48 Range: 28-48 Mean: 41.6
Highest score obtained: 48

As a whole, the centers did rather well on these items. There is very little in the way of artificiality or superfluous disruptions to distract the visitors. In most of the centers, the demonstration classes reflect the overall program. On the other hand, there is a sizable minority where this is doubtful (items 14 and 15). Giving visitors an opportunity to talk to teachers and students is considered to be particularly persuasive to visitors. While most centers give visitors a chance to talk to teachers (item 24), many centers do not provide an opportunity to talk with students (25).

Figure 8 deals with the information the center provides the visitor about the effect of the program on students, teachers, parents, etc. This information does not have to be formally collected and analyzed. Only one center discussed any kind of evaluation plan for assessing its program. The academic progress of the class was discussed by only a few. A few more centers discussed the effects of the program on student attitudes, the attitudes of the demonstration teachers, and the reactions of parents.

It is certainly no surprise that the demonstration centers have no evaluation going on, since few schools do. That this should be the case, however, is a sad commentary on education in general. Most programs proceed unassessed and unproved.

While the previous section dealt with evidential assessment of the effects of the program, "Explanation of Program Feasibility" deals with the problems of installing and maintaining it. (See Figure 9.)

It was deemed that discussions of the practical problems connected with the program would provide another opportunity for a different kind of evidential assessment, one that would enhance the feasibility of adopting the program in so far as the visitor was concerned.

As a group the centers do a very poor job of providing this type of explanation. A few discuss necessary equipment and materials slightly (items 36 and 37). Only one or two centers really discuss in any detail at all what is necessary for adopting their program. Item 34 is an item that in a general way incorporates all the others and gives a general picture of what the centers are doing with explanation of feasibility.

FIGURE 8: EXPLANATION OF DEMONSTRATION CENTER'S OWN EVALUATION

		Detailed (6)	(5)	General (4)	(3)	Little (2)	(1)	None (0)
26.	Was demonstration center's plan(s) for its own evaluation explained? (e.g., procedures, scheduling, rationale)			x			xxx	xxx xxx xxx xxx xxx x
27.	Interclass academic progress explained? (e.g., compared to last year or another group this year; compared to similar groups using local or national norms)				x	x	xxx xxx	xxx xxx xxx xxx
28.	Were effects of the demonstration program(s) on student attitudes explained?		x	x	x	xxx xxx x	xxx xx	xxx xx
29.	Were effects of the program on demonstration teachers' morale and attitudes given?			x	x	xxx x	xxx xxx xxx	xxx xx
30.	Were the reactions of the community to the project discussed?			x	x	xxx x	xxx xxx xxx	xxx xx
31.	Were the reactions of students' parents discussed?			x	xx	xxx	xxx x	xxx xxx xxx x
32.	Were effects of the demonstration program on non-program students explained?				x	xxx	xx x	xxx xxx xxx xxx x
33.	Were effects of the program on non-demonstration teachers discussed?				x	xxx x	xx	xxx xxx xxx xxx x

Maximum score possible: 48
Range: 0-13
Mean: 6.05
Highest score obtained: 13

FIGURE 9: EXPLANATION OF PROGRAM FEASIBILITY

		Detailed (6)	(5)	General (4)	(3)	Little (2)	(1)	None (0)
34.	Were possible problems of installation in other schools discussed?			x	xx	xx	xxx xxx	xxx xxx xxx
35.	Was an estimate of funds needed for installation of the program given?				x		x	xxx xxx xxx xxx xxx xxx
36.	Were necessary equipment and materials discussed?			x	xxx xx	xxx	xxx x	xxx xxx x
37.	Were the visitors told how to locate these materials and equipment?				xxx	xx	x	xxx xxx xxx xxx xx
38.	Were continuing costs of the program discussed? (e.g., maintenance)						xx	xxx xxx xxx xxx xxx xxx
39.	Was what you need to get in the way of training in order to start this program in another school explained?			x	x	xx	xxx	xxx xxx xxx xxx x
40.	Were weaknesses of the program explained?				x	xxx xx	xxx xxx	xxx xxx xx
41.	Were strengths of the program discussed?			x	xx	xxx xxx	xxx x	xxx xxx x

Maximum score possible: 48
Range: 0-17
Mean: 6.4
Highest score obtained: 17

IV. WHAT IS THE OVERALL PATTERN OF DEMONSTRATIONS?

Figure 10 presents information about the pattern of demonstrations overall--a very important question. The data for each item (weighted equally) was summarized in a total score for each section of the observation schedule. Each x represents each center's score on that section.

Relative to these scales, as a group the Illinois Centers did much better on "Observation of the Demonstration Class". This is probably because the demonstration class is the primary criterion for selecting centers and because that particular scale is somewhat easier than the others. The two scales on which performance was lowest deal with evidential assessment--the "Explanation of Evaluation" and "Explanation of Feasibility".

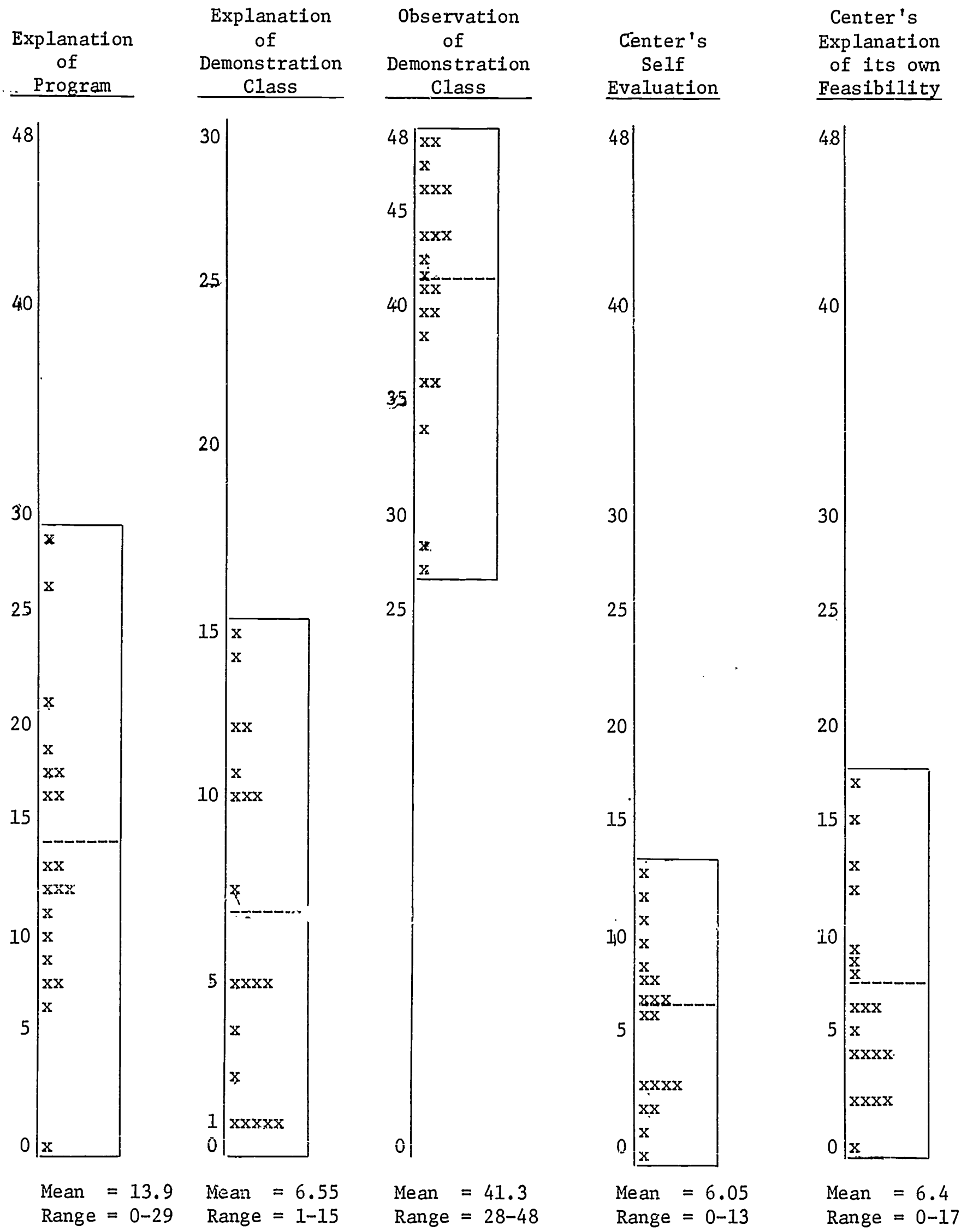
As a group then, relative to these scales, the Illinois demonstration centers are excellent in credibility and fidelity ("Observation of Class"), poor in intelligibility ("Explanation of Program" and "Explanation of Class") and very poor in evidential assessment ("Explanation of Evaluation" and "Explanation of Feasibility").

This is only part of the story, however. Within every section there is a very great difference between individual centers. For example, within "Explanation of Program" there is a great gap between the highest and lowest centers. Within "Explanation of Class", there are two distinct groups--a high group of centers and a low one. Even within the most narrowly prescribed range--"Explanation of Evaluation"--there is a sizable difference between the highest and lowest. The overall difference among centers is exemplified by the profiles of the centers highest and lowest in total scores. (See Figure 11.)

It is doubtful that a center with the lower profile should be demonstrating, however good its program may be. It is difficult to see how visitors can understand what is going on. Any operation as geographically decentralized as The Illinois Demonstration Centers is bound to have quality control problems. Inferences from the individual item data and from the overall profiles of centers indicate that it is indeed very serious with the Illinois Centers. The ability of centers to communicate their programs varies tremendously.

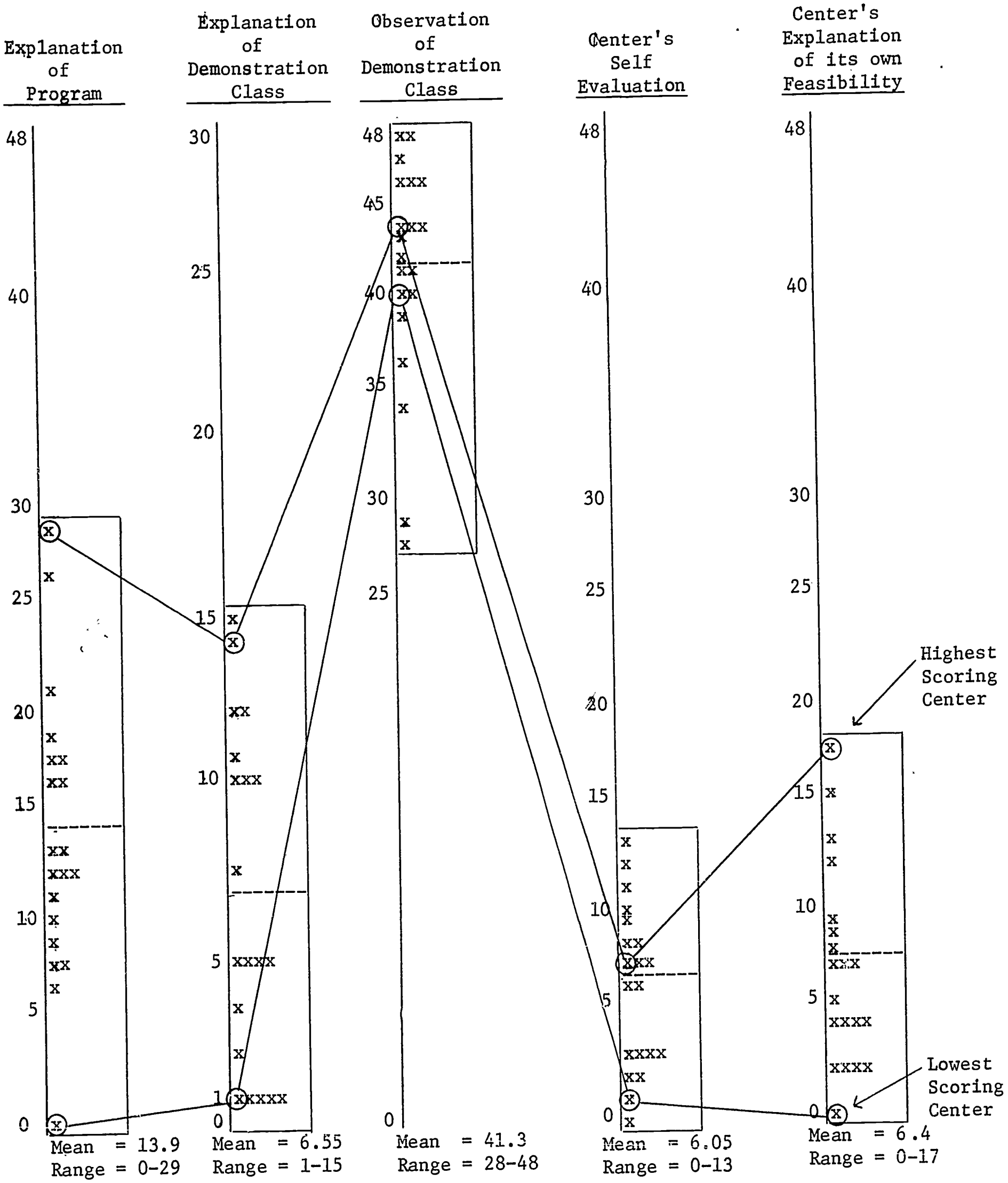
"Explanation of Program" and "Explanation of Class" deal more with how the visitor is persuaded or with implementation of the demonstrated programs. Hence, except for the class observation, the centers tend to do better at making visitors aware, rather than persuading or getting them to adopt a program.

FIGURE 10: PROFILE OF DEMONSTRATION RATING SCALE RESULTS



The dotted line represents the Mean or Average Score of the Centers for that particular section. The Range of scores is set off by the box.

FIGURE 11: PROFILE OF DEMONSTRATION RATING SCALE RESULTS



The dotted line represents the Mean or Average Score of the Centers for that particular section. The Range of scores is set off by the box.

V. WHAT DO DEMONSTRATION DIRECTORS THINK IS IMPORTANT?

All of the demonstration directors in the study were asked to designate how important each of the items was. The average (median) of all director responses was then compared with the average of how they actually rated as a group on each item. The comparative "ideal" and "real" scores give a discrepancy measure of how the centers are performing according to their own standard. Figure 12 gives the profile for "Explanation of Program."

The most important items according to the directors are explaining objectives, treatments, and student selection procedures (items 1, 2, 4). The state supervisor of demonstration centers thinks a description of the state plan (item 6) is most important. The greatest discrepancy between ideal and actual performance is in item 1 and 8. The best performance is on explanation of program treatment (item 2).

On "Explanation of Class" (Figure 13) the most important items are explaining objectives (item 9) and explaining treatment (item 10). The state supervisor also thinks these are important items. Significantly, the only items on the entire observation that are given the highest ranking of "6" are those dealing with objectives--items 1 and 9. Just as significantly, the centers perform poorest on these when compared to their ideal. Explaining class objectives reveals the greatest possible discrepancy. On the other hand, explaining class treatment is somewhat better.

On "Explanation of Evaluation" (Figure 14) the group ideal is considerably lower than for the other sections. Not only is the "ideal" for this section very low, the actual performance is even lower. The items felt to be most important are explaining the demonstration center's evaluation plan (item 26), on which the centers do extremely poorly, and explaining the effects of the demonstration on student attitudes (item 28). The state supervisor thinks explanation of interclass academic progress (item 27) is most important. Again, on this item, the centers scored zero on the scale. The best performance is on item 28.

On "Explanation of Feasibility" (Figure 15) the most important items are explaining necessary training (item 39) and discussing strengths of the program (item 41). On item 39, the centers scored zero on the scale. This item contains the second largest discrepancy in the entire analysis. To the state supervisor item 34 and 37 discussions of problems of installation and necessary equipment are most important. Again, the centers as a group scored zero on the scale.

On "Observation of Class" (Figure 16) the directors rated as most important giving visitors an opportunity to talk to teachers, giving visitors an orientation as part of the class (a negative item in our scoring), seeing and hearing class proceedings, whether the day's lesson reflected the program's objectives, and giving visitors an opportunity to talk to students (items 24, 17, 21, 22, 14, 25). The state supervisor thought items 14, 15, 16, 24, and 25 were most important. Interestingly enough, the top choices of the directors did not include whether the day's lesson reflected program treatment. In actual performance the centers did very well on items 24, 21, 22 and not so well on 14 and 15.

FIGURE 12: EXPLANATION BY THE DEMONSTRATION CENTER OF ITS PROGRAM

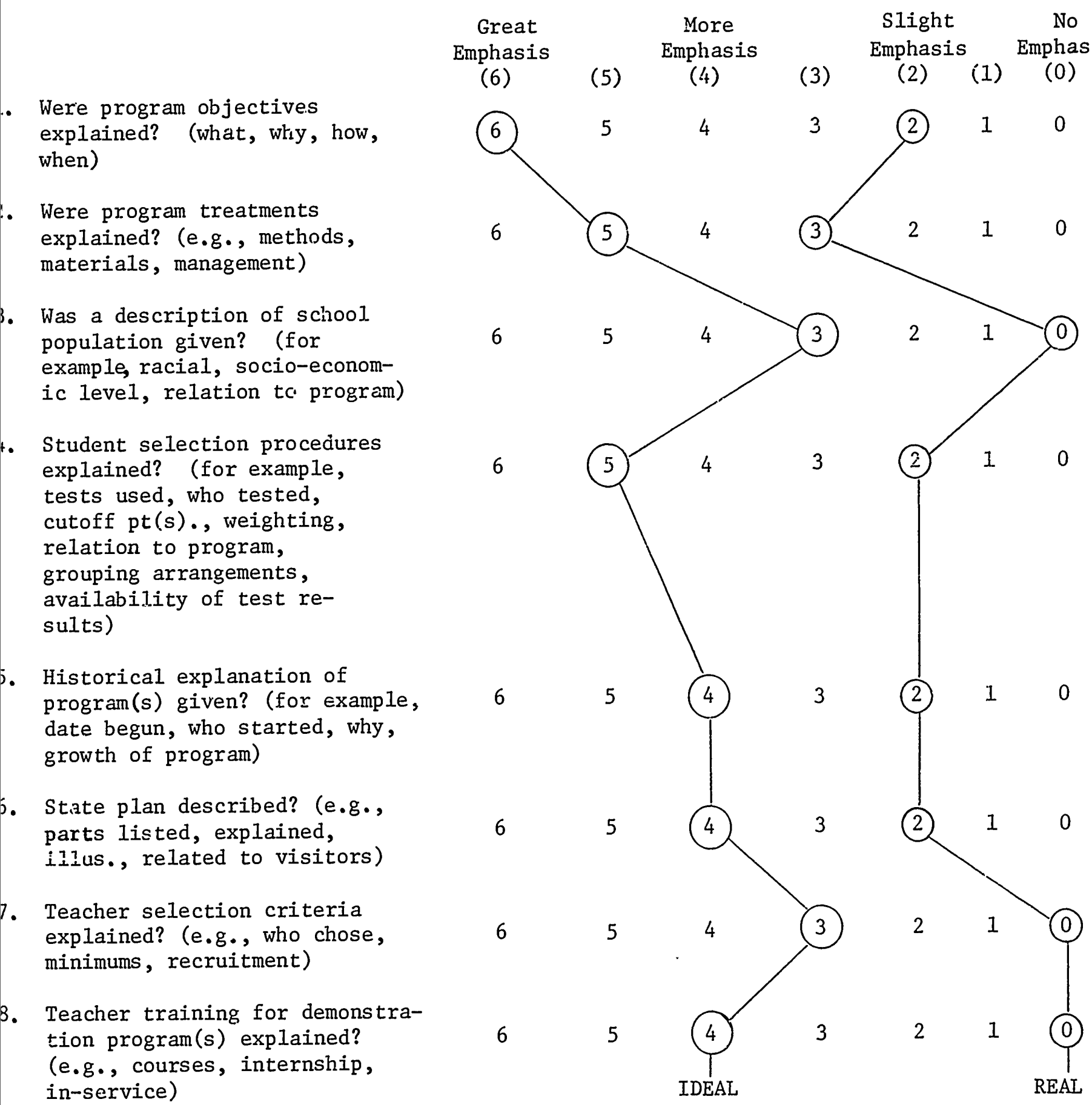


FIGURE 13: EXPLANATION BY THE DEMONSTRATION CENTER OF THE CLASS TO BE OBSERVED

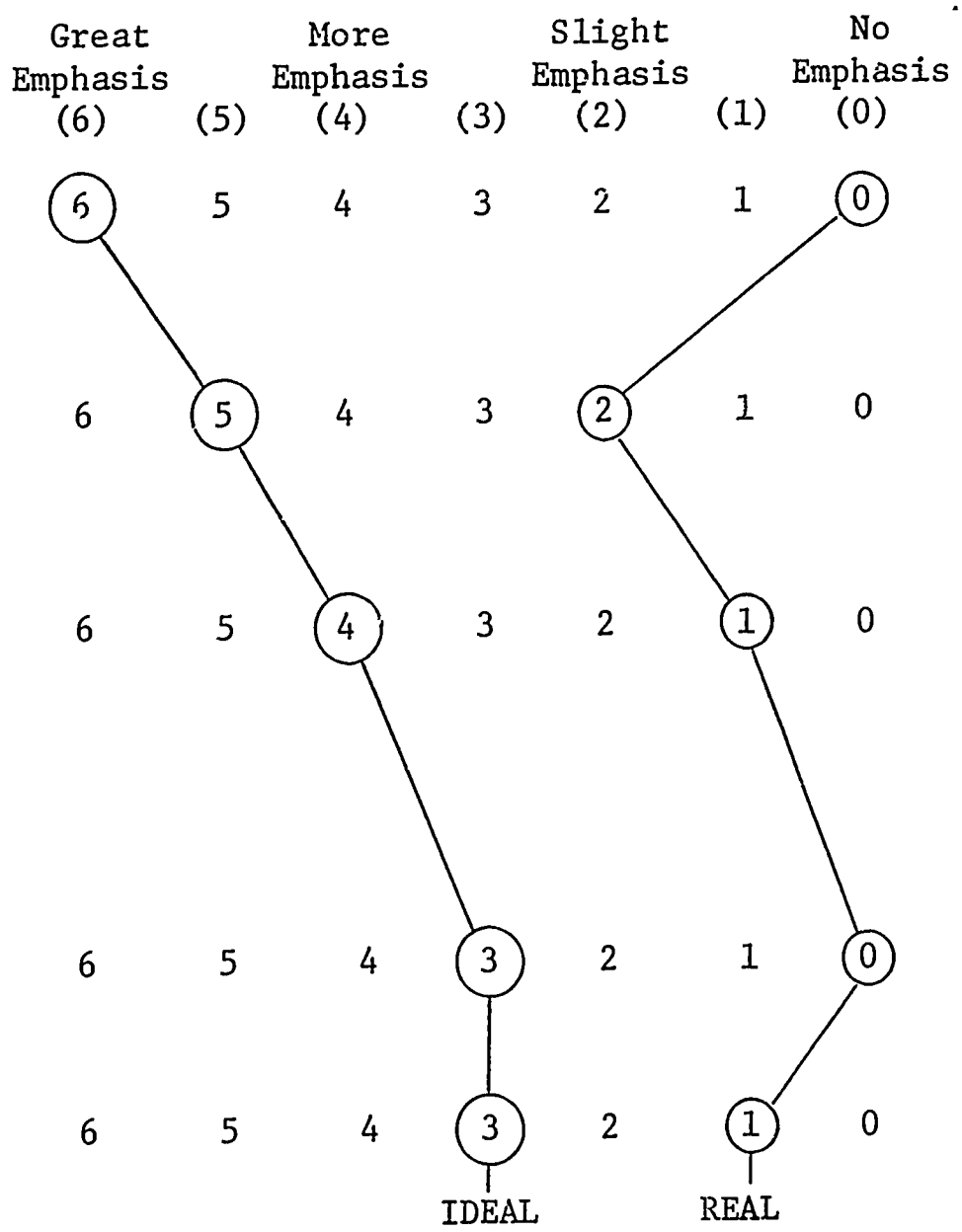


FIGURE 14: EXPLANATION BY THE DEMONSTRATION CENTER OF ITS OWN EVALUATION

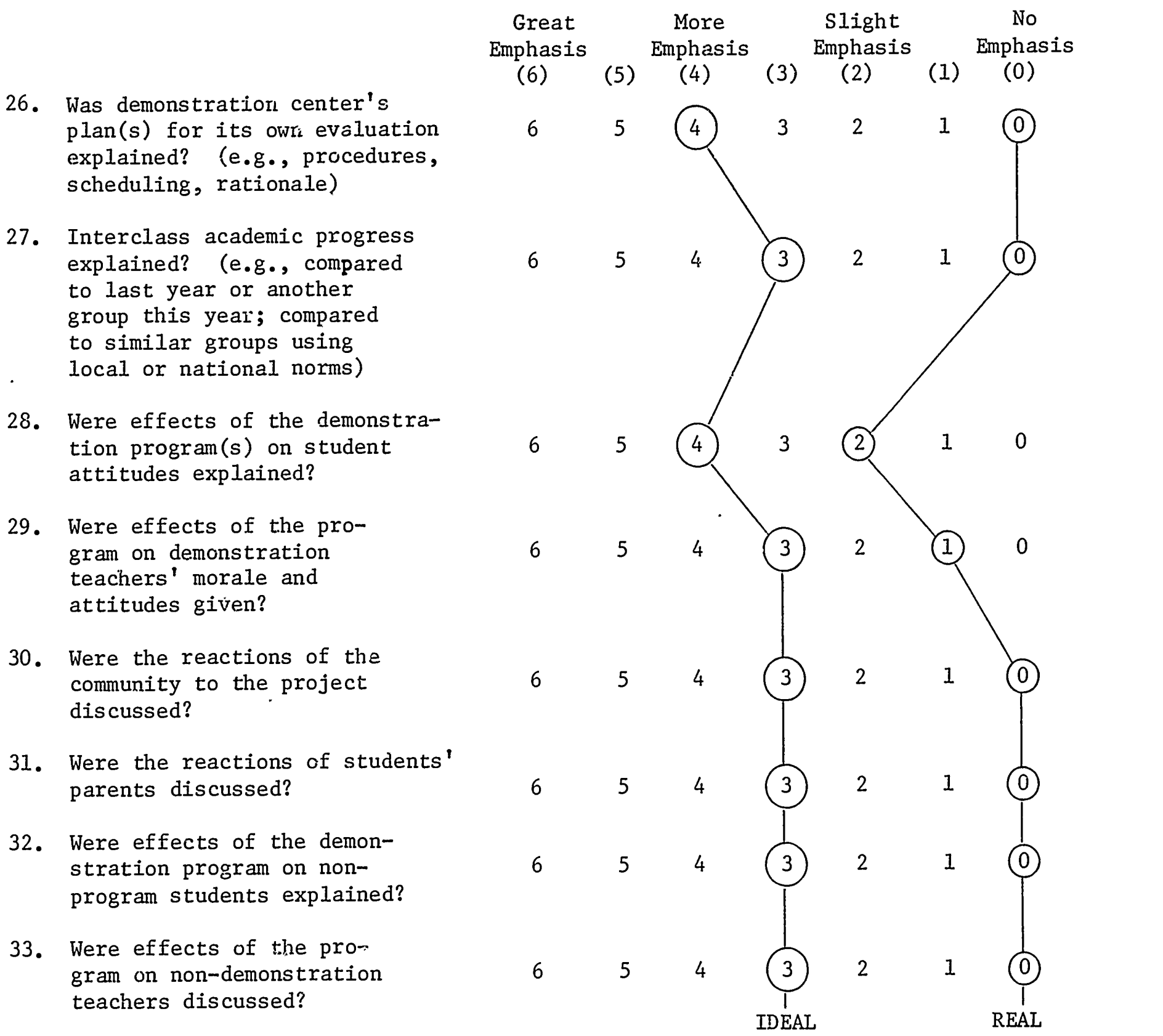


FIGURE 15: EXPLANATION BY THE DEMONSTRATION CENTER OF PROGRAM FEASIBILITY

FIGURE 16: OBSERVATION OF DEMONSTRATION CLASS

Due to differences in classroom observation a different scale was used for rating in this section. Note that "no" is a high score for items 17-23.

Item numbers circled indicate items directors felt were most important.

		YES (4)	(3)	INCONCLUSIVE (2)	NO (1)	(0)
(14.)	Did the day's lesson reflect the overall program objectives?	xxx xxx xxx x	xxx	xxx xxx	x	
15.	Did the day's lesson reflect the overall program treatment?	xxx xxx xxx xxx	xx	xxx x		xx
16.	Was competence of teacher adequate?	xxx xxx xxx xxx xxx	xx	xx	x	
(21.)	Were visitors able to see class proceedings clearly?	xxx xxx xxx xxx xxx xxx x		x		
(22.)	Were visitors able to hear class proceedings clearly?	xxx xxx xxx xxx xxx xxx		xx		
(24.)	Were visitors given a definite opportunity to talk to teachers?	xxx xxx xxx xxx xxx xx	x			xx
(25.)	Were visitors given a definite opportunity to talk to students?	xxx xxx xxx x	x	xx	x	xxx xxx

		NO (4)	(3)	INCONCLUSIVE (2)	YES (1)	(0)
(17.)	Was orientation, background or review given visitors as part of class sequence?	xxx xxx xxx xxx xxx		xxx x		x
18.	Did total class sequence seem artificial?	xxx xxx xxx xxx xxx x		x		
19.	Were children continually distracted by the presence of visitors?	xxx xxx xxx xxx xxx x	xx	xx		
20.	Was visitor behavior excessively disruptive?	xxx xxx xxx xxx xxx xxx x		x		
23.	Were additional classroom materials needed to follow lesson?	xxx xxx xxx xxx xxx		xxx xxx		xx

Maximum possible score: 48 Range: 28-48 Mean: 41 6
Highest score obtained: 48

When considering whole sections, the "ideals" and "reals" are indicated in Figure 17.

FIGURE 17: IDEALS AND PERFORMANCE ON WHOLE SECTIONS

Section	Demonstration Director Ranking	State Supervisor Ranking	Actual Performance
Explanation of Program	1	3	2
Explanation of Feasibility	2	2	4
Observation of Class	3	1	1
Explanation of Class	4	4	3
Explanation of Evaluation	5	5	5

The greatest discrepancy is obviously on "Explanation of Feasibility". The directors consider it very important but do rather poorly on it. The only disagreement between the demonstration directors and state supervisors is on the relative importance of "Explanation of Program" and "Observation of Class".

Several conclusions can be drawn from these comparisons:

1. When the Illinois demonstration center scores are averaged into one score, their performance looks considerably poorer than in a distribution of individual scores, as in Chapter III. This is because many centers are doing so poorly they pull the whole group down. Several centers are doing a good job, but many are doing a very poor job.

2. The directors' ideal indicates that three items in "Explanation of Program" are deserving of great emphasis, two items in "Explanation of Class", no items in "Explanation of Evaluation", and two items in "Explanation of Feasibility".

3. Although many commonalities exist in the priorities of the demonstration directors and the state supervisor, enough differences exist so that there are some chances for conflict over what a demonstration should do.

VI. WHAT CONDITIONS INFLUENCE THE DEMONSTRATION?

There are two sets of relationships among the sections of the observation schedule. If a center does well on "Explanation of Program", it also tends to do better on "Explanation of Evaluation" and "Explanation of Feasibility". These relationships indicate a concern with the overall program being demonstrated and with its acceptance by the visitor. On the other hand, if "Explanation of Class" is good,"Observation of Class"also tends to be good. This relationship indicates a concern for the particular activities of the classroom.

The centers that have been demonstrating the longest (six years as opposed to two) tend to have the poorest "Explanation of Class" and "Observation of Class". One contributing factor is that many of the older centers have hired new directors in the last year and many of the new directors are less familiar with specific classroom activities. The more experienced directors do better on "Explanation of Class" than the less experienced.

Furthermore, in the older centers, the director himself usually conducts the demonstration (rather than a teacher or assistant director). When the director does the demonstration, the "Explanation of Evaluation" is better.

Visitor behavior is also important in the demonstration. All questions asked by visitors were recorded by the observers. The more questions asked, the better "Explanation of Evaluation" and "Explanation of Feasibility" tended to be, indicating that many of the visitor questions were about evaluation and feasibility. This is noteworthy because it is precisely in evaluation and feasibility that the demonstration centers do the poorest job. Why visitors ask questions in one center and not in another is not known.

Which of these events lead to better visitor understanding, acceptance, and eventual implementation must await analysis of visitor responses. However, a suggestion may be found in the reactions of the observers who collected data. The better the center did on all the sections of the observation schedule, (except for "Explanation of Feasibility"*)the better our observers reported they understood the program. Also the more questions visitors asked, the better the reported understanding.

The observers were also asked how committed they were to the program as demonstrated. Again, the better the center performed on all sections except "Explanation of Feasibility", the higher the commitment. Once again the more questions asked by visitors, the higher the commitment. In addition, the observers were most committed to programs they felt they understood.

Finally our observers were asked what their ideal commitment was to each program, regardless of how well it was demonstrated. This time only good "Explanation of Class" and "Explanation of Evaluation" were associated with ideal commitment. The implication is that these two parts of the demonstration may play a significant role in the visitor's ultimate acceptance of a program. This is very provocative because the demonstration directors consider these two sections least important of all.

*"Explanation of Feasibility" was the one section on which we had reliability problems, thus reducing the chance of finding significant relationships.

VII. WHAT GENERAL CONCLUSIONS CAN BE DRAWN?

When all the sections of the observation schedule are combined to produce one total performance score for each center, the total score is very strongly related to both "understanding" and "commitment as demonstrated" (r=.8). However, the total score is not significantly related to "ideal commitment". One might speculate that in making an "ideal commitment" the values of the individual come into play. Confirmation of these trends must await analysis of regular visitor reactions.

A. In too many centers, the quality of the demonstrations is too low.

There are some very good demonstrations but more very bad ones. In any widely-dispersed, decentralized operation, it is very difficult to maintain quality of performance--in this case the quality of demonstration. Even on the observations where the centers as a group perform best, e.g. explanation of treatment, there are several centers that do very poorly. In fact, some centers should not be operating at all if they cannot do better.

RECOMMENDATION: A quality control system should be instituted by the state to insure that a minimum level of performance is maintained.

Each center should be allowed to operate its own unique form of demonstration. However,minimum requirements should be enforced if the whole program is to be effective. Whatever we continue to find out about the demonstration process, one thing is clear: if the most salient features are not communicated to the visitor, he cannot possibly understand the program.

One such system of quality control is to simplify the 41-item instrument we have used to rate centers by reducing the number of items to the 20 items demonstration directors, state staff, and our research indicate are most important. State staff members (or someone else) could be trained in using the instrument. These observers could then visit each center periodically, (perhaps three times a year) and record the center's performance. The demonstration director would know in advance what he was being rated on. At the end of the day, the director would be shown how well he had done. The report would then be filed in Springfield. Over a few years the progress and improvement of the center could easily be plotted. Hopefully, by the next funding period, all centers would be performing at an acceptable level and the next funding decision could be made on the basis of the program itself.

We repeat: The problem is serious enough that the entire demonstration project could be undermined by the poor performance of many of the centers.

B. The Illinois Centers are doing their best job on the "awareness" function of demonstration and rather less well on the "acceptance" function. (We momentarily have suspended judgment on the "implementation" function.) The demonstration centers do rather poorly on "Explanation of Program" and "Explanation of Class"--the criterion of intelligibility; they do excellently on "Observation of Class"--a mixture of intelligibility, fidelity, and creditability; and they do very poorly on "Explanation of Evaluation" and "Explanation of Feasibility"--the criterion of evidential assessment.

In so far as the Clark-Guba model is an accurate model of educational change, we would expect visitor acceptance of programs to suffer because of the poor handling of the latter two sections. Other evidence* indicates that the Illinois demonstration centers have traditionally emphasized "awareness" over "acceptance" and "implementation" as their operational goals. In fact, the more experienced the demonstration director becomes, the more important he thinks "awareness" is as opposed to the other goals. This has been interpreted as a function of career orientation and a distinct lack of diffusion technology. It should be noted that the Clark-Guba model calls for a demonstration to accomplish only "awareness" and "acceptance". The goal of implementation has been paramount with the state supervisory staff (not the directors) since the beginning of the Illinois Plan, however.

RECOMMENDATION: The demonstration centers should be more concerned with discussing evaluation and feasibility with visitors. The State Advisory Council and the State Staff should encourage centers that make explicit provisions for increasing visitor acceptance and implementation.

It is somewhat premature to predict exactly what components, if any, lead to visitor commitment. With our observers, "commitment as demonstrated" was associated with every section except "Explanation of Feasibility". "Ideal commitment" was associated with good "Explanation of Class" and "Explanation of Evaluation". However, our observers are not typical visitors. The best single strategy

*E. R. House "The Role of the Demonstration Director", unpublished doctoral dissertation, University of Illinois, 1967.

for a demonstration director to pursue is to improve performance on all sections. In promoting acceptance other activities outside the realm of this particular analysis, e.g. training institutes and conferences, are probably more effective than a one-day demonstration.

C. In summary, we asked the Illinois demonstration centers these questions: "Is the message clear?" (Intelligibility) and gave a qualified no. "Does the demonstration illustrate both positive and negative features so that an observer may reach a valid professional judgment?" (Evidential Assessment) and gave an unqualified no. "Does the message give a valid picture?" (Fidelity) and gave a yes. In making these judgments we have instituted an "absolute" set of standards. We were forced to do this since there is no comparison group. Admittedly our standards are tough. We think that regular visitors will be much less critical. However, we think that both sets of standards have merit. To re-emphasize, several districts do meet these standards, though the group as a whole does not. It is possible to conduct a good demonstration.

VIII. HOW CAN THIS DATA BE USED?

We have already suggested how a quality control system might be instituted to improve the visibility and clarity of the demonstrations. We have presented this information to the State Advisory Council, which oversees the Illinois Gifted Program, and to the State Staff, which supervises the program, prior to their refunding of the demonstration centers.

In addition, the data will be presented to the demonstration directors. Each director will receive a folder containing an "ideal" profile of what he has indicated he would like to achieve on each item and a "real" profile showing what he did achieve. He also will receive the group scores with his own scores circled for each item and each section in order that he may compare his performance to the entire group.

For our part, this is the first step in evaluating the Illinois centers. We will relate this information to how visitors actually reacted to the demonstrations and to what the visitors actually did as a result. In this way we hope not only to assess the effectiveness of the demonstration centers but also to ascertain empirically what a good demonstration is.

ED038807

The Visibility and Clarity of Demonstrations

APPENDICES

By

Ernest R. House

Thomas Kerins

Stephen Lapan

Joe M. Steele

JUNE, 1969

Gifted Evaluation Project

COOPERATIVE EDUCATIONAL RESEARCH LABORATORY

Box 815, Northfield, Illinois 60093

ED027962

ED038807

The Visibility and Clarity of Demonstrations

APPENDICES

Ernest R. House, Thomas Kerins, Stephen Lapan, Joe M. Steele

Assisted by

Sharon Gotch, Billie Hoyt, Judith Miley

Supported by

The Department of Program Development for Gifted Children

OFFICE of the SUPERINTENDENT of PUBLIC INSTRUCTION

Ray Page, Superintendent of Public Instruction

Gifted Evaluation Project

COOPERATIVE EDUCATIONAL RESEARCH LABORATORY, INC.

Box 815, Northfield, Illinois 60093

JUNE, 1969

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

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

OFFICE of the SUPERINTENDENT of PUBLIC INSTRUCTION

Ray Page - Superintendent of Public Instruction

Verne E. Crackel - Deputy Superintendent

John H. O'Neill - Associate Superintendent

Robert D. Cain - Assistant Superintendent
Division of Special Education Services

Herbert Baker - Director
Program Development for Gifted Children

MEMBERS of the STATE ADVISORY COUNCIL for the GIFTED

Robert Bauernfeind
Professor of Education, Northern Illinois University

Mrs. Evelyn Carlson
Associate Superintendent of Schools, Chicago

Ray Caton
County (McLean) Superintendent of Schools, Bloomington

A. Gordon Dodds
Superintendent, Edwardsville Community School, District 7

David M. Jackson
Director, Cooperative Educational Research Laboratory, Inc.
Northfield, Illinois

Blair Plimpton - Chairman
Superintendent of Schools, District 44, Park Ridge

Richard Verduin
Coordinator of Professional Experiences
College of Education, Southern Illinois University

APPENDICES

CONTENTS (published in a separate volume)

APPENDIX A

LIST OF DEMONSTRATION CENTERS AND PROGRAMS OBSERVED[1]

	Center	Grade Level	
1.	Belding	Ungraded Prim.	Special Curriculum--English
2.	Bowen	12	Special Curriculum; Cooperative/Team Teaching
3.	Bryn Mawr	5	Junior Great Books
4.	Carver	K	Culturally Disadvantaged
5.	Champaign	1	Special Curriculum--Math, Language Arts, Social Studies; Productive/Critical Thinking
6.	Charleston	12	Inductive Teaching
7.	Decatur	Sr. High	Small Group
8.	Edwardsville	6	Special Curriculum--Social Studies; Productive/Critical Thinking
9.	Elk Grove	Elem.	Individually Prescribed Instruction Learning; Resource Center
10.	Evanston	9	Fine Arts
11.	Evergreen Park	11	Special Curriculum--Creativity; Cooperative/Team Teaching
12.	Freeport	6	Cooperative/Team Teaching
13.	Lockport	4	Special Curriculum--Science, Inquiry
14.	Marion	6	Inductive Teaching in Language Arts
15.	Oaklawn	5-6	Special Curriculum--Reading
16.	Oak Park	5	Special Curriculum--Creativity
17.	Park Forest	Ungraded Prim.	Special Curriculum--Inquiry
18.	Signal Hill	2	Small Groups; Individualized Instruction--Reading
19.	Skokie	Primary	Music Instruction
20.	Urbana	Elem.	Individually Prescribed Instruction; Learning/Resource Center

[1]Many of these centers demonstrate other programs and include other grade levels in their demonstration. Those shown represent only programs and grade levels actually observed in the demonstration evaluation.

APPENDIX B

A TYPICAL VISIT TO A DEMONSTRATION CENTER

Early in the school year, I discovered that each teacher in the school system was allowed to take one day off school to visit another school district in the state. I believed that this would be a rewarding experience, so I had the district superintendent's office send me information on schools which encouraged visitors and which my district felt were conducting particuluarly interesting programs. I chose the district which was called a Demonstration Center and which demonstrated a program in Language Arts.

I received the pre-visit information which is sent to all the prospective visitors. It included a brochure explaining the basic concepts of the program, a schedule of the demonstration activities, and the days when visitation was possible. I also learned that the program I was to see was principally designed for the gifted students. It pleased me to have been provided with a map of the area to be visited. I requested a day to leave school.

The day of visitation arrived. I found a most welcome cup of coffee awaiting my arrival at the Demonstration Center office following the search down the unfamiliar halls. The Demonstration Center office is where the day's visitors were to assemble to begin the day's visit.

The orientation began at 9:20 rather than at 9:00, because three of the five visitors had had difficulty finding the Demonstration Center office. The director explained that he would have to omit some parts of the orientation due to the lack of time. The first observation class was scheduled for 10:30.

The director gave a presentation orienting us to the center's programs. During the orientation, the visitors were told that there seemed to be increased student and teacher interest since the program began. The director also explained the student selection procedures, grading practices, and parent attitudes toward the program. He said that when the program first began, the parents were skeptical about the added freedoms the students would be given. But after the parents found that the added freedoms were stimulating and permitted the students to learn more, the parents involved gave full backing. No negative aspects of the program were mentioned.

We were shown a series of slides as a partial explanation of the school district organization and program description. Toward the end of the orientation, the visitors were given a printed schedule of the day's planned observation activities. The morning allowed for the visitation of one of two classes. Following lunch, the visitors were to visit the Learning Resource Center to observe various programmed materials and to see students at work on independent study projects.

The center's plan called for the teacher of each class being visited to meet with her visitors before class and explain what they were to see. However, only the Language Arts teacher was free from 9:45-10:30, so she came to talk to all the visitors about the Language Arts program for the gifted and told us briefly what to expect in the class to be observed. Only five minutes were allowed for questions from the visitors, and I didn't get to find out how the teachers were chosen to take part in the program.

Each visitor was asked to indicate the activity which most interested him. I indicated my interest in the 7th grade Language Arts class. Other visitors were going to observe a 6th grade, self-contained class having a lesson in history.

One other teacher and I, who were interested in Language Arts, were taken to the Langugage Arts room, while the director took the other visitors to the 6th grade class.

We entered the Language Arts class and found two folding chairs placed at the back of the room. The class was already underway and the teacher acknowledged our presence with a nod. The students looked at us and then put their attention back to the teacher. The students seemed comfortable and unbothered by visitors. They were discussing among themselves, jotting notes on paper, and just watching what the teacher was writing on the board.

The teacher presented a lesson on modifiers and their purpose. I've taught adjectives and adverbs in many different ways, but I don't believe that my students ever seemed to learn it so easily. The lesson was taught inductively. The teacher wrote sentences on the board and underlined certain words. The students took over the entire discussion of analyzing the purpose of the underlined words in the sentences. From time to time the teacher would ask additional questions to bring out discussion. The students were very involved and interested.

When the students were asked to do a written assignment, the teacher came back and welcomed us to the class. She invited us to walk around the room and to talk with the students while they were

working on the assignment. Later, the teacher had the students arrange themselves in groups of 6 to discuss the paragraphs they had written using modifiers for description. She said we should feel free to sit in on any of the groups. "The students enjoy having you," she assured us.

Each of us joined a different group. I was very surprised and enthused to find how analytical the students were about their own work, even at the 7th grade level.

One student in my group wrote two paragraphs. The first paragraph told of a snake crawling across his body while he lay resting in the woods. In the second paragraph, the boy added many modifiers to add a more clear explanation of the senses he experienced. The students responded very positively to his effort, and his work was used as an example for the remaining discussion of the group's work.

I talked with the other visitor to the Language Arts class after the period was over. She too was amazed to find what the students could do when left on their own.

As the period ended, about 11:25, the teacher announced that as soon as the students arrived in class tomorrow they were to rewrite their "modifier paragraph" in its final form and turn it in to the teacher.

The other visitor and I talked with the teacher for a few minutes, and I did get to ask my earlier unanswered question concerning teacher selection. It seems that interested teachers sign up with the demonstration director and choices are made from the list of interested teachers. However, how they choose from this list was not made clear.

The visitors met back at the office following the morning class. Here, we were supplied with directions to the local restaurants.

At 1:00 we all met back at the Demonstration Center office. The afternoon observation was to be a Learning Center located in a nearby elementary school. As the K-5 school was only one block away from the Junior High and demonstration office, it was recommended that we walk.

Before going to the K-5 school, the Learning Center director came to the Demonstration Center office to explain her program. She first told us that the demonstration program funds provided her center with two part time "Teacher Aids." These aids allowed her to spend certain time with visitors and to give students more individualized attention while they were in the center. Some stu-

dents are programmed into the center for so many minutes and work while others are scheduled to use the center during the week as the teacher feels necessary.

Students milled freely about the center gathering materials and returning to a desk or study carrel.

The director mentioned that there were still some teachers who did not quite understand the concept of the center. "Some teachers think of it as a 'dumping' ground for students while the teacher goes to the lounge for a coffee break. They do not care to acquaint themselves with the wealth of supplementary resources which are available, thus many students are never programmed or scheduled into the center."

I wrote down several workbook titles and plan to write the companies as I feel these materials could also be used well in the regular classroom for those particularly slow or fast students.

At 2:15, we walked back to the Demonstration Center office and the director of the program gave us additional hand-out information and a follow-up evaluation questionnaire. The hand-out information listed the basic description of the program as it really existed. The questionnaire was two pages long and took about 15 minutes to fill out. I answered such questions as "What did the visitor like most about the day?" "What did the visitor like least?" and "Did the visitor plan to make any changes in his own classroom as a result of the visit?" The director expressed his willingness to assist any visitors with further implementation of programs in their home districts.

The director gave his thanks and goodbye about 3:00. I left, but two of the visitors stayed on to talk with him.

APPENDIX C

PROCEDURES

1. Rationale

There are several measuring instruments and combinations of instruments of potential value in approaching an evaluation of the demonstration centers. Through a visitor questionnaire it is possible to discover a visitor's immediate reactions to the center in terms of how well he is aware of the programs and if he is leaning toward acceptance. Through a post-visit questionnaire it is possible to find out whether or not the visitor has actually adapted a program or implemented observed demonstration center activities. However, if we only used these types of questionnaires we would not have a description of the treatment itself.

Therefore, in addition to these questionnaires it was decided that a rating scale should be developed and used to rate the centers' ability to make their demonstrations clear and visible to visitors. The first problem encountered in developing such a scale is the lack of uniformity in the objects to be measured.

In fact, if one word was used to describe the demonstration centers for gifted children in the Illinois Plan, that work would be "diversified." As is apparent from Appendix A, there is a wide variety of programs and activities at all grade levels available for demonstration. Due to these factors, along with each center's own methods of teacher and student selection, it is logical to expect that the demonstration process will vary from center to center.

The rationale behind this rating scale takes this situation into account but also assumes that there are basic elements necessary to the successful diffusion of a demonstration program. By using the Clark-Guba change model (page 5 of the text) as a starting point, a rating scale was developed which would measure these basic elements without penalizing the centers.

The section of the change model which is specifically correlated with the rating scale is the diffusion section and its subsections of dissemination and legitimization. Under this section we have hypothesized that the more visible and clear the demonstration process is to the visitor the more positive will be his later reaction to adopting a center's activities. Therefore, through the scale we are measuring the ability of a center to accomplish its dissemination and legitimization objectives with the change model as the standard. Through a comparison of our later data on visitor implementation with the centers' results on the scale we hope to

prove that the scale has predictive validity with the centers scoring the highest on the scale affecting significantly more visitors than the low scoring centers.

A center's score depends upon the verbal behavior of the demonstration center director and his staff as they attempt to accomplish their dissemination and legitimization objectives. All verbal statements throughout the day are written down by the raters who later classify the statements according to the items in each section.

The purpose of the first two sections is to rate the center's ability to disseminate its program by informing the visitor and creating an awareness about the center's program(s) and class(es) to be observed. (The intelligibility dimension of the Clark-Guba model.) The assumption is that the more the visitor knows about the center's relation to the Illinois Plan, its methods of student and teacher selection, its objectives and methods of treatments for the program and the particular class, the more likely will be his implementation of the center's activities. Or, as an alternate hypothesis, the more emphatic he will be in rejecting it.

In the last three sections of the rating scale, the rater looks at the center's ability to legitimize its program to visitors. First, the center must build the conviction of the observer by offering the visitor the opportunity to examine and evaluate at first hand the demonstration classes. Instead of measuring the verbal behavior of the teacher in the demonstration class as an interaction analysis would, the items in the third section rate the effectiveness of the class observation itself. If the days' lesson obviously reflected the overall program objectives, then the center received the maximum rating for that item. Since the opportunity for visitors to talk with the demonstration students and teachers may be quite necessary to build personal conviction, centers who did allow or encouraged this also received the maximum rating.

The fourth section of the rating scale measures the center's ability to build conviction by showing how they have informally and formally assessed or evaluated their program. This section is a good example of the fact that the centers were not expected to rate high on each item or even score on every item. It is very unlikely that a center could do or mention all the types of evaluation described by the items in this section. A center may be receiving some feedback from visitors and students in the program but none at all from the community or the demonstration teacher. However, some evidential assessments of a program and its results are expected since it would make the program credible to the visitor.

The last section on the scale is concerned with the center's ability to establish the program's ease of adoption. By rating the director's comments about the cost and location of materials, needed training, and the program's strengths and weaknesses, the breadth and depth of the center's verbal explanation of its own program(s) exportability can be determined.

As the results illustrate, it is possible to do well on one or more sections of the rating scale and do very poorly on the remaining ones since all sections are scored independently. However, according to our model the centers should at least achieve a moderate score in each section since all five parts represent essential segments of the successful diffusion of a demonstration.

As the main text illustrates, the rating scale was constructed so that it would have many possibilities for analysis. However, the main outcome is the profile on page 15 which shows how the centers score on each of the five sections along with the distribution of scores on the individual items (pp. 8-15) which graphically illustrate what the staff of the demonstration centers are saying and what they are omitting--what they are stressing and what they are barely mentioning in their discussions with the visitors.

These results indicate to us the degree to which the demonstration centers are making their presentations intelligible and credible and simultaneously the degree to which they are accomplishing their dissemination and legitimization objectives. Therefore, the scale not only provides us with an overall picture of the performance of the demonstration centers in Illinois, but also which centers are strong and weak and the location of their strengths and weaknesses.

2. Instrument Construction and Field Testing

In developing the rating scale, which was titled the Demonstration Observation Schedule, large numbers of items (statements) about a day's activities at a center were pooled. The Ohio State Evaluation Center was contracted to study Illinois' Gifted Demonstration procedures, meet with people knowledgeable in the workings of the Illinois Plan, and finally construct an observation schedule representing many activities that the centers could be conducting. This original schedule was then tested for its appropriateness through discussions among a few demonstration directors, the evaluation staff, and the Ohio State group. The items were generated from a familiarity with both the Illinois Demonstration Centers and the Clark-Guba model.

After an initial draft was developed, assigned members of the evaluation staff began work on reorganizing the instrument giving particular attention to whether or not certain activities were observable. During this six weeks to two month period the overall structure and some key items were finalized keeping in mind that the applicability of the instrument needed to be tested using actual center visitations.

Before this time period one of the evaluation staff members was checking the feasibility of the untried instrument by visiting demonstration centers. This experience along with a visit to one center by another staff member contributed important data during the early stages of the instrument development. However, more formal field testing of the instrument and the observers was yet to come.

By the early fall it was necessary to bring together the entire evaluation staff to engage in discussion regarding the status of the observation schedule. During these preliminary exchanges members generated examples that were to exemplify each item. These ostensive definitions were then reworded until general agreement was reached on each of the over 50 items. (The schedule was eventually reduced to 41 items.) This long and tedious process included discussions over item meanings, item additions, and item deletions. (See Figure 1.)

FIGURE 1

EXAMPLES OF ITEM CHANGES

	Original Wording	Final Wording	Reason
Item #1	Were the demonstration center objectives explained to the visitors?	1 - Were program objectives explained?	1. A referent had to be specified for the term "objectives." Final wording refers to observed program objectives only.

Item #5 - Was the historical explanation of the demonstration center given?	5 - Historical explanation of programs given?	5.	History of center could refer to many different explanations. Final wording refers to programs demonstrated only.
Item #14 - Was an explanation given of the relationship between the objectives of the day's class to the overall demonstration program objectives?	14 - Did the day's lesson reflect the overall program objectives?	14.	This was a difficult relationship to draw. After some field tests, this was changed. It is better for observers to look for consistency.
Item #23 - Were visitors given textbooks, handouts, etc., necessary for following the lesson?	23 - Were additional classroom materials needed to follow lesson?	23.	The original wording presumed too much. Some programs use texts, others don't. Final wording provides for judgment based on program to be observed.

Other than the make-up of the group, the major differences between this process and the original instrument development tasks included continuous attempts to define each item in behavioral terms and to refine the individual item statements to achieve greater specificity.

At this juncture it became evident that real data from the actual treatment milieu (demonstration centers) needed to be obtained. For this reason, the staff began visiting centers, with the primary goal being that of seeing how well the instrument would work in the setting in which it was eventually to be used.

All four observers visited 3 centers together during the

of October, meeting for one to three days following these visitations. Since the ratings were independently recorded, wide disagreements were inevitable. The major purpose at this point was to find out if the observers could use the instrument.

By the end of the meeting following the third visitation, the decision was made to direct attention toward observer reliability.[1] The items, which by now were well refined, were to be changed as little as possible. On the other hand, each observer's perception had to be altered in relation to the other observers. For example, two observers considered naming the five parts of the Illinois Plan (Item #6) along with identifying in which parts the center was involved to be worth a "general" rating. The other two observers thought that at least one example, definition, or reason for existence should be given for each part of the Illinois Plan in order for the communication to earn a "general" rating. Such differences had to be solved by observer agreement and <u>not</u> by changing the item or the rating categories.

For the task of improving observer reliability four more centers were visited and one to three day meetings followed <u>each visit</u>. However, the content of the meetings was different in that the observers had to come to agreement about how they viewed centers as opposed to what the structure of an item should be. (Some item changes did occur, however.)

During the final field tests in preparation for the data collection phase, techniques (rules) for using the observation schedule were developed. One of the most important rules adhered to by the observers was that the observation schedule (rating scale) was a verbal analysis of the demonstration process. Thus, the observers were to record everything said at the center which was part of the formal order of things, i.e. anything presented by the center staff that was intended for the visitors. This also included any information that was given as a result of a visitor question (visitor questions were also recorded). The observers also noted which member of the center staff provided each kind of information. All of this verbal behavior was recorded on note pads by the observers. At the end of the day each observer would then independently spend from two to four hours going through the notes categorizing every piece of information according to which item in the schedule it corresponded. They then would rate how well each item was communicated.

[1]See the discussion of Reliability, section <u>4</u> in this appendix.

In the final data collection, each of the two observers visiting a given center would categorize and rate according to items on the observation schedule just as they had done in the field test. Also, as in the field test, each observer carried out these tasks independently, after the day's observation, without any benefit of knowing how the other observer was rating the center.

3. Data Collection

The observers were confined, then, to recording verbal behavior at the center. Thus, pre-visit information and hand-outs at the center were not counted unless they were verbally referred to during the demonstration day. It was also decided that both observers would visit the same class (whichever class most of the visitors first visited), and this would be the only class responded to on the observation. (This was a practical decision based on the difficulty of scheduling that would occur if both observers tried to stay together and also remain with the same group of visitors.) This also meant that the overall observation would include only the program that was represented by the class. Thus, if the first class visited was in independent reading, the observers would record only those communications during the day which dealt with the independent reading program.

The directors were told on the phone that the team would want to visit in this manner, and they were also told to choose their best demonstration for this particular time period.

There were other rules that were to be followed by the observers which pertained to their behavior at the center. The observers were to act as normal visitors never purposely indicating their reason for visiting. The one exception to this rule was that the observers were not to ask questions or act in any manner which would affect the demonstration procedure. (See Appendix E, Obtrusiveness of Measures.)

The directors were sent a communication outlining the observers' behavior during the visitation. They were asked not to single out the observers in any fashion other than by name and city. The directors were also told about the administration of the Visitor Questionnaire and to withhold their own instruments on that day. These directions were then restated during the telephone scheduling.

In late October and early November the scheduling for visiting demonstration centers began. Each director was telephoned and asked to pick from dates available to the team chosen to visit his center.

This telephoning continued through the month of November with a few centers still not scheduled.

Fifty-three telephone calls were necessary in order to schedule the twenty-one centers to be visited with one of the centers never settling on an open date. The only requirement for the centers to meet on the scheduled day was that at least two (normal)[2] visitors had to have been scheduled to visit other than the two observers. This was done so that the director and his staff could operate normally, expecting questions and any other behaviors that typical visitors would exhibit. As indicated before, the observers could not ask questions. Also, since the observers would be busy recording verbal behavior, they were not to fill out any forms that the center might ask the visitor to complete.

During the data collection phase of the study the observers had certain other tasks to perform, the Demonstration Visitor Questionnaire was to be administered to the visitors. The same individual in each pair of observers was responsible for this administration each time. The other observer had the responsibility of interviewing the director.

The director interview was conducted for the purpose of getting written and verbal information about the program not given during the day. This information was not considered in rating the center, but will be analyzed separately along with other data gathered about the centers and their programs. The most important question asked the director with regard to the day's data collection was "Would you say that today was a typical demonstration day?" One director out of the twenty answered "No" to this question giving a specific reason for this answer. The evaluation staff arranged to revisit this center at a later date.

Two other centers were revisited also--one because the director requested it some weeks after the first visit and the other because of a scheduling confusion which afforded the observers a distorted view of the center.

On January 22, 1969 the last center observation was completed with the exception of the one center which did not settle on a visiting date.[3]

[2]Normal visitors by our definition were any public school professional personnel, i.e. teachers, administrators.

[3]Given three months (November, December and January), it is not explicitly clear why this center could not schedule one day for the observers.

4. Reliability

A great deal of confusion exists in the reporting of reliability estimates for observation and rating instruments. Some investigators report correlations between raters, some report correlations between observations by the same raters, some report correlations of different raters observing at different times.

Other studies, such as the CUE Evaluation of NYC Title I, report percentages of agreement among raters. Even here there are variations. The New York study reported the percent of the time raters assigned ratings which were the same or within one scale point. This degree of agreement would be markedly higher than percentages based on identical assignment of ratings.

Other data used to estimate reliability include analysis of variance, and Scott's pi coefficient, which is an adaptation of Chi-square. The latter has been used in observational systems such as Flanders.

Because of the wide variation in the meaning of information reported as "reliability" data, the recommendations of experts were sought. Here too there exists a great deal of ambiguous and contradictory information. Many standard texts on educational statistics (such as Cronbach) make reference only to the estimation of test reliability. This kind of analysis, as Remmers points out, is not appropriate for many kinds of rating scales and observational systems. Perhaps this lack of discussion by statisticians has given rise to the variety of approaches in use.

Kerlinger recognizes the many forms of reliability reported, and states that "reliability is usually defined as the agreement among observers... Practically speaking, then, the reliability of observations can be estimated by correlating the observations of two or more observers. When assessing the reliability of the assignment of behaviors to categories, percentage of agreement between judges is often used. But, as with all kinds of measurement, there are other ways to estimate reliability, for example, repeat reliability and reliability estimated through analysis of variance."[4]

Medley and Mitzel[5] define the reliability coefficient to be the correlation between scores based on observations made by dif-

[4]Fred N. Kerlinger, Foundations of Behavioral Research, New York; Holt, Rinehart and Winston, 1964, p. 507.

[5]Donald M. Medley and Harold E. Mitzel, "Measuring Classroom Behavior by Systematic Observation," in N. L. Gage (ed.) Handbook of Research on Teaching.

ferent observers at different times. They give the name coefficient of observer agreement to the correlation between scores based on observations made by different observers at the same time. This, they say, tells something about the objectivity of an observational technique.

A third coefficient identified by Medley and Mitzel is the stability coefficient, which is the correlation between scores based on observations made by the same observer at different times. This coefficient tells something about the consistency of the behavior observed from time to time. They suggest that unreliability comes about most commonly when two measures of the same class tend to differ too much. However, as Remmers[6] points out, if the interval between observations is long, there may be real changes which lower such coefficients. "If such fluctuations do occur, a low "reliability" coefficient would be more desirable than a high one."

Remmers lists five criteria on which to judge rating scales as measuring devices. Two of these are relevant to the discussion of reliability.[7] "1. Objectivity: Use of the instrument should yield verifiable, reproducible data not a function of the peculiar characteristics of the rater. 2. Reliability: The instrument should yield the same values, within the limits of allowable error, under the same set of conditions. Since basically, in ratings, the rater and not the record of his response is the instrument, this criterion boils down to the accuracy of observations by the rater."[8]

The criterion of objectivity would seem to be similar to what Medley and Mitzel are defining as reliability. The criterion of Reliability appears to be similar to what Medley and Mitzel call the Coefficient of Observer Agreement. Remmers seems to be in agreement with Kerlinger that the estimate of reliability refers to the agreement among observers.

Perhaps the crux of the differences among these experts lies in what they regard as the instrument. Both Remmers and Kerlinger stress the fact that when a rating scale is used, the person doing the rating is the instrument. It is the observer's inferences based on what he sees that are recorded as values on the rating

[6]H. H. Remmers, "Rating Methods in Research on Teaching," in N.L. Gage, ibid.

[7]The other three criteria--Sensitivity, Validity, and Utility--while affected by reliability, are not directly pertinent to this discussion.

[8]Remmers, op. cit., p. 330.

scale. From this point of view reliability does have to do with the degree of agreement among judges. Even Medley and Mitzel note that "So crucial is the observer's judgment in coding behavior that the major effort in instrument construction is usually devoted to the task of defining categories as unambiguously as possible to make the judgments as easy as possible."[9] However, the assumption seems to be made by the latter that once such problems of interjudge agreement have been minimized, a reliability coefficient can be derived for the written scale itself administered by a number of raters. They state that "A measure is reliable to the extent that the average difference between two measurements independently obtained in the same classroom is smaller than the average difference between two measures obtained in different classrooms."[10] They develop a general design for reliability estimation based on four-way analysis of variance. This definition of reliability is an extremely rigorous one which requires a major investment of time and resources independent of any use to which the rating instrument might eventually be applied. While this approach appears to be eminently respectable, its use in the early stages of instrument development is simply not feasible.

It should be noted that Medley and Mitzel go so far as to report for some studies that "Information is not yet available regarding the reliabilities of these measures, but a number of statistically significant findings are reported, indicating that they were reliable."[11] In other words, they feel that the obtaining of statistically significant findings is de facto evidence of reliability.

The conclusion to be drawn is that there are increasing refinements that can be considered in estimating reliability. A judgment must be made as to the time and expense that can be invested at a particular stage in the development and use of an instrument. In any case, care should be taken in reporting the exact circumstances from which a particular coefficient is derived so that it may be correctly interpreted. The contribution of whatever results that are reported should be made quite clear.

For this study a detailed report of estimated reliabilities appears later in this section. What follows is a summary of reliability data for all observation combined. The coefficient of observer agreement for all ratings combined is .75. This, in Kerlinger's or Remmer's terms, represents the reliability. The percentage of observer agreement for identical assignment of ratings is 73 %. The observers attained 93.4% agreement on assigned ratings which were identical or within on scale point. Only incomplete data is available on reliability as defined by Medley and Mitzel. Such comparisons as are available indicate a reliability of .80.

[9] Medley and Mitzel, op. cit., p. 253.

[10] Medley and Mitzel, op. cit. p. 250.

[11] Medley and Mitzel, op. cit., pp. 274 and 283.

To adequately validate and establish the reliability of a new instrument for assessing behavior is extremely time consuming. The task usually consumes from three to six years of extensive refinement, field testing, training, and data collection. This time period does not include the use of the instrument in actual data collection for research purposes.

Some of the stages of development through which a new instrument, such as a rating scale, moves are:

(1) Development of a theoretical rationale for item selection.
(2) Selection and screening of items.
(3) Clarification and definition of items.
(4) Field testing and redefinition of items.
(5) Studies of stability of rating by the same rater over time.
(6) Studies of interrater agreement based on simultaneous observation.
(7) Studies of rater interpretation of items used.
(8) Studies of reliability of the instrument based on use by different raters observing the same activities at different times.

An accurate estimate of reliability requires a balanced research design utilizing four-way analysis of variance. It is obviously out of the question to develop fully refined instruments for use in evaluation studies. The limited resources in time, funds, and properly trained personnel prohibit such refinement. Even if these were available, the essence of evaluation is its timeliness in reporting findings for use in decision-making. The delays necessary for extensive instrument development would render the evaluation findings worthless.

The more we are engaged in evaluation activities, the larger time looms as a primary enemy. To spend three years developing this observation schedule would make the data totally irrelevant to the people to whom it is directed. Even as it was, there was a year and a half gap between the original conceptualization of the instrument and presentation of the data--much too long a time. During this time period the evaluation project also had many other evaluation activities and kinds of data to collect.

Exclusive devotion to the observation schedule would have shortened the time gap, but the result would have been the small amount of data contained in this report compiled at great expense of time and money--a bad bargain from the consumer's viewpoint. Other instruments could not have been substituted since none appropriate existed. So we traded off a certain amount of reliability for time and for

other overlapping information, e.g. questionnaire data, etc., that we could buy with that time. We think we did not pay too high a price. One of the other instruments that we subsequently developed, an attitude inventory, we later abandoned completely because we were not satisfied with its reliability by the time we were to use it. The name of the evaluation game then is not primarily instrument development, but rather providing pertinent data to those who need it. While such trade-offs may be odious in research, in evaluation they are mandatory.

A primary concern during the development of the Demonstration Observation Schedule was to obtain stability for each of the items on the scale. In a very real sense when a rating instrument is used, the person doing the rating is the instrument. The items need to be clarified to the degree that the rater is consistent in his rating: he should consider the same kinds of things each time he uses the scale. Also, when a number of raters are using the same rating scales, there not only needs to be consistence in each of their performances from one time to the next, but congruence among their performances at any one time. That is, when all raters rate the same behavior there should be agreement on what kinds of things are considered for each item and the judgment that is made when these things are considered.

One solution to this dilemma of obtaining commonly understood and stable items would be to have each item refer to one specific feature or behavior to be observed. The instrument then becomes a very limited set of scales, but a highly precise one. It also requires an extremely refined theoretical model to determine all the important behaviors to look for. This approach would require such a large number of items to actually describe the activity that it would be almost impossible to use if it could be developed.

An alternative approach, which is the one used, is to have each item refer to a group of behaviors which could be expected to occur. No list is made of all the behaviors that is included in the item description. Instead the general description is written as clearly as possible and then defined operationally by the raters using the item. Over a period of trial ratings by the observers a number of examples of behavior appropriate for each item are accumulated. This use of ostensive definitions serves to clarify the items and orient the rater to the appropriate categories of behavior to observe.

One of the problems early in the study was to develop a sampling plan for observing demonstrations. Several alternatives were considered. One approach called for sending observers to each center separately, so that more than one sample of behavior could be ob-

tained. This plan was rejected because it was felt that the objectivity of the rating scale was not sufficiently developed to attribute all differences of rating to real differences in behavior rather than differences resulting from rater bias or variability in interpreting the scales. The plan would also have involved extensive travel by single observers (three of them women) over wintry roads. In addition there would have been extensive scheduling complications due to the necessity to observe the centers when regular visitors were present.

The sampling plan that was adopted was to send two observers to each center at the same time. They each independently rated the center's demonstration. In obtaining a single score for the center, the two ratings would be combined by deriving a mean for any items where different values were assigned. The rating instrument contained a four point scale (none, little, general, detailed) which were assigned the numerical values 0, 2,4, and 6. In deriving an average, some items would receive the intermediate values of 1, 3, or 5. This results in a scoring system providing a range of seven values. (See Section III of the report.)

This plan provided a means of checking the objectivity of the raters. Both percentages of agreement and coefficients of observer agreement are reported. The latter is what is generally reported as the reliability of the instrument. (Even with two teams of observers, this plan required two and one-half months to implement, due to the scheduling difficulties noted above.)

This sampling plan was felt to provide an accurate indication of the demonstration activities for all of the Illinois Demonstration Centers considered as a group. The major intent of the evaluation was to determine the variation in behavior across centers rather than for each center. While the rating of any one center based on one visit might not truly represent the activities of that center, errors in rating would tend to cancel themselves out when all ratings were combined. Thus the results would be highly representative of the kinds of demonstration activities engaged in by the Illinois Centers as a whole.

Medley and Mitzel report that sending two observers into a classroom at the same time is more wasteful than sending them in at different times. When the number of visits is increased, the errors due to instability of observed behavior as well as observer errors tend to cancel out and reliability is markedly increased. It is unfortunate that the objectivity of the rating scale was not sufficiently established to have made use of the first plan. A great deal more weight could then have been given to findings for indi-

vidual centers.

The detailed report of estimated reliabilities is discussed in the following paragraphs.

As can be seen from Table 1, the per cent of complete agreement of observers ranged from 56% to 95% with an average per cent of agreement across all centers of 73%. When agreement is defined as assigning values within one scale point, the per cent of agreement ranged from 85% to 100% with an average across all centers of 93%.

Table 1 also shows the coefficient of observer agreement derived by correlating the scores of the two observers at each center. This is the coefficient usually reported as representing the reliability of the rating instrument. There was a range of observer agreement for Team I from .67 to .97. The mean correlation[12] for the twelve centers observed was .81. For Team II, the range was from .47 to .85 with a mean correlation for the eight centers of .65. Considering both teams, a mean correlation of observer agreement for the twenty centers was .75.

As has been noted, extensive field testing was conducted prior to actual data collection. For the final field test, all four observers visited the same center and independently completed the Demonstration Observation Schedule. Table 2 presents data on the agreement of ratings among all four observers. Two-way analysis of variance was used to calculate the reliability (coefficient of observer agreement) based on Guilford's formulation for reliability of ratings.[13] The reliability of ratings for the four observers combined was .92. This figure indicates that an extremely high degree of objectivity and agreement of ratings can be obtained by trained observers using the Demonstration Observation Schedule.

Table 2 also shows correlations of observer agreement for every possible combination of pairs of observers. The obtained coefficients of .78 and .77 for observers 1 and 2, and 3 and 4, were judged quite adequate for this combination of observers to collect the actual data.

[12] A mean correlation is estimated by converting the individual correlations to Z scores, computing an average and then converting to the equivalent correlation coefficient.

[13] J.P. Guilford, Psychometric Methods, 2nd Edition, N.Y.:McGraw-Hill Book Co., 1954, pages 395-397.

The information presented in Table 2 provides an indication that both teams of observers were interpreting and using the rating scale in the same way. Thus results obtained by the two teams are judged to be comparable.

An estimate of the degree to which the rating scale is assessing dimensions of demonstration activities that remain relatively stable from one presentation to another can be obtained with the Stability Coefficient. Table 3 shows stability coefficients based on ratings by the same observer visiting the same center at two different times. Data on only four centers are available and for two of the centers (F and M) one of the observations occurred during field testing. Considered by center, the combined Stability Coefficients range from .66 to .98. This indicates that relatively little change occurred in the way a particular center demonstrates its program. The Stability Coefficient for all four centers combined was .85. Thus the Demonstration Observation Schedule appears to be tapping dimensions of demonstration that are relatively stable.

Table 4 presents estimates of reliability based on the rigorous definition of Medley and Mitzel. Reliability coefficients are based on ratings by different observers observing the same center at different times. Data is available for only three centers. The information for center F is based in part on ratings made during field tests.

As the table shows, a combined reliability coefficient of .53 was obtained for center F. There was a six week time interval between observations. This coefficient is no doubt lower due to changes resulting from the field tests.

The reliability coefficient obtained for center L is .62. This data was collected during December and January. Again there was a six week interval between observations.

Partial data is available for a third center (J) indicating a reliability of .96. Only one team of observers visited this center and the time interval between ratings is seven weeks. It is felt that reliabilities of .62 and .96 are quite satisfactory for this stage of instrument development.

TABLE 1

DEGREE OF OBSERVER AGREEMENT FOR EACH OBSERVATION TEAM
FOR EACH OF THE TWENTY DEMONSTRATION CENTERS OBSERVED

TEAM I (OBSERVERS 1 AND 2)

CENTER	DATE VISITED	PER CENTER OF AGREEMENT		COEFFICIENT OF OBSERVER AGREEMENT
		Identical ratings	Within one scale unit	
A	11/13/68	78	95	.80
B	11/14/68	83	98	.90
C	11/21/68	56	93	.67
D	11/26/68	78	95	.72
E	12/4/68	73	93	.77
F	12/5/68	71	95	.78
G	12/6/68	71	90	.70
H	12/10/68	63	93	.69
I	12/12/68	73	95	.81
J	1/13/69	95	100	.97
K	1/17/69	78	98	.77
L	1/22/69	78	98	.77

TEAM II (OBSERVERS 3 AND 4)

CENTER	DATE VISITED	PER CENTER OF AGREEMENT		COEFFICIENT OF OBSERVER AGREEMENT
		Identical ratings	Within one scale unit	
M	11/13/68	78	93	.72
N	11/21/68	71	88	.59
O	11/26/68	73	100	.85
P	12/3/68	71	90	.61
Q	12/4/68	76	93	.73
R	12/5/68	59	85	.47
S	12/10/68	66	88	.55
T	12/18/68	71	90	.50
Mean % of Agreement: (Teams I and II combined)		73%	93%	

Mean Coefficient of Observer Agreement (Both Teams): .75

TABLE 2

AGREEMENT AMONG ALL FOUR OBSERVERS RATING
THE SAME CENTER AT THE SAME TIME

(Based on the final field test of the Demonstration Observation Schedule.)

Two-way Analysis of Variance

Source		Sum of Squares	Degrees of Freedom	Variance
From items	(i)	137.96	41	3.364
From raters	(r)	.495	3	.165
From remainder	(rm)	32.255	123	.262

$$r_{\text{all raters}} = \frac{Vi-Vrm}{Vi} = \frac{3.364-.262}{3.364} = .92$$

Matrix of Intercorrelations Among All Observers

	1	2	3	4
1				
2	(.78)			
3	.62	.78		
4	.82	.74	(.77)	

The circled coefficients represent the correlations of the two teams of observers who worked together during actual data collection.

TABLE 3

RATING SCALE STABILITY COEFFICIENTS BASED ON RATINGS

BY THE SAME OBSERVERS VISITING THE SAME CENTER AT TWO DIFFERENT TIMES.

Center	Dates Visited	Observer 1	Observer 2	Combined
F	(10/24/68 and 12/5/68)	.74	.55	.66
K	(12/13/68 and 1/17/69)	.63	.76	.70
J	(11/25/68 and 1/13/69)	.98	.98	.98

Center	Dates Visited	Observer 3	Observer 4	Combined
M	(10/30/68 and 11/13/68)	.65	.86	.78

Stability for all four centers combined = .85

TABLE 4

RATING SCALE RELIABILITY COEFFICIENTS BASED ON RATINGS
BY DIFFERENT OBSERVERS OBSERVING THE SAME CENTER AT DIFFERENT TIMES

Center F (Visited 10-24-68 by Observers 3 and 4 during field tests; visited 12-5-68 by Observers 1 and 2)

Observers	r
1,3	.73
1,4	.45
2,3	.46
2,4	.43
Combined r* =	.53

Center L (Visited 12-11-68 by Observers 3 and 4, and 1-22-69 by Observers 1 and 2)

Observers	r
1,3	.70
1,4	.55
2,3	.54
2,4	.67
Combined r* =	.62

Center J (Visited 11-25-68 and 1-13-69 by Observers 1 and 2)

Observers	r
1,2	.95
2,1	.97
Combined r* =	.96

*Combined r is based on z score mean.

APPENDIX D

THE INSTRUMENT

1. How the Items Were Rated

The rating scale consists of forty-one items divided into five sections. Four of these sections utilize a four position rating scale: Detailed, General, Little, None. The twelve items in the other section, "Observation of Demonstration Class," utilize a three position scale: Yes, Inconclusive, No. (Due to differences in the nature of the items on which classroom groups were rated, the four position scale was felt to be inappropriate.)[1]

In considering the section, "Observation of Demonstration Class" positive responses to all of the 12 items were numerous. In general the visitors were not disruptive (Item #20) and the visitors could see and hear (Items 21 & 22). The positive and negative reactions by the observers were based on whether or not they experienced these reactions. Thus, either they could see or they could not see; either they could hear or they could not hear.

The observers would assign the Inconclusive scale position when they found it impossible to make a clear choice. For example, there were cases when it was not clear whether it was permissable to talk with students. In one case, although the observers were near the students during and immediately after the observation, the center personnel had not given any indication that talking with students was acceptable. On the other hand, if any one of the visitors would have chosen to talk with a student, surely he could have done so.

Unlike the other sections of the observation schedule, the observation section was marked at the time of observation. However, as in the case of the other sections, these items were rated independently by the two observers.

The definitions of scale positions varied with each item on the observation schedule. Each of the scale positions above None (or No) were defined operationally by the raters. The field tests provided the examples for these ostensive definitions.

[1]For further discussion see Appendix C, especially Section 1.

The None category, however, was determined in generally the same manner for each item. In the case of rating an item as None, the raters would do so if they heard nothing verbally stated regarding the particular item being rated. This also included any references made to the content of an item without the content being identified. This rating of None was exemplified by such statements as "We have objectives that fit our program" and "this works according to our objectives."[2] Thus, in each case the term "objectives" was used, but an identification (naming) of those objectives was not given.

Other examples of where None ratings were given included the following statements made by center personnel:

For Item #6 "You, of course, know about the Illinois Gifted Program."

For Item #7 "Then the teachers were selected for the program."

For Item #8 "Our demonstration teachers have been specially trained."

For Item #11 "We have homogenous grouping here."

For Item #26 "We are planning to evaluate our program."

For Item #35 "This isn't an expensive program."

For the other categories used in rating the quality of communication (Little, General, Detailed), a comprehensive account is provided regarding how these ratings were assigned for Item 1.[3]

[2]Referring to Item #1 on the schedule.

[3]When the reader notes the nature and amount of information needed to earn a "detailed" rating for item #1, he will realize how difficult it is to attain such a rating. It would be impossible for any center to score perfectly on every item on this observation schedule. As a matter of fact, there would not be enough time during a visiting day to cover all of these items in a detailed fashion. Thus, the purpose of the observation was not to see if every center could get a perfect score, but rather to see what centers were emphasizing.

Following this account, selected examples are provided for each category for items 2, 10, 28, and 34. This should give the reader a representation of how the items were rated. The examples found herein are verbal statements recorded during the data collection visits to the centers.

How Item #1 Was Rated

#1 Were the program Objectives explained?

Little: This rating was given when the objectives were listed, named, or in some way stated, but no other information was provided. Such was the case when the following two statements were made at one of the center's visited:

Example: "We believe in individualization starting at kindergarten. We want to maximize the intellectual potential of youngsters."

Comment: The goals here are individualization and maximizing intellectual potential. There was the naming of the goals, but no communication regarding what these goals meant either by definition or example. Also no reasoning or justification was provided as to why these particular goals were chosen.

General: This rating was given when the objectives were stated and one of the following units of information was also given: Examples of each goal or reasons for choosing each goal or definitions of each goal. Such was the case when the following statements were recorded at a center:

Example: "In the student program we want to develop responsibility, self-direction, and decision making skills... (For this reason we) give the student the opportunity to evaluate his own work by establishing his own criterion and selecting judges... Students have the actual experience of making decisions."

Comment: The goals listed are responsibility, self-direction, and decision making skills. These are named along with some examples provided to clarify what is meant. One named goal is decision-making skills. The examples are having experience in making decisions and the student deciding for himself how he will be evaluated. Information not provided here included reasons for choosing these goals and specific definitions for each goal.

Detailed: This rating was given when the objectives were stated specifically and one of the following units of information was given; an example for each goal along with reasons for choosing each goal or an example for each goal along with a definition for each goal or reasons for choosing each goal along with a definition for each goal. Such was the case when the following statements were recorded at a center visited:

Example: "The program is designed to develop higher level thought processes...according to the Guilford Model...(The teachers) teach for higher level thinking...(Such thought processes would include) the practice of divergent thinking which is like creative thinking--students coming up with new solutions to old problems...(This is what the) leaders of tomorrow should be able to do."

Comment: There is one major goal, that of getting the students to think at higher levels according to the Guilford Model. Examples of this include creative thinking. A definition of creative thinking was new solutions to old problems. Further, the reason for choosing this goal was that the center personnel believe this is the type of leader that should be prepared. The amount of information was in large quantity, which also accounts for the "detailed" rating.

Selected Examples for Some Representative Items

#2 Were Program treatments explained?

Little: We have "concept instruction in language arts and math at this grade level."

General: We do "individualizing, testing, and setting of definite goals...divide into ability areas...stress games with a purpose...start children where they are at each grade level."

Detailed: We have "team teaching, large group instruction, small group work, modular scheduling...individualized instruction using the instructional material center...Room 115 is used for seminars and the cafeteria for informal sessions...once the students choose one of these modes of instruction they must stay there for at least one module...there are teacher assigned tasks along with project work and contract study..."

#10 Today's class treatment explained?

Little: Today you will"...see the student's working with programmed materials in math along with their folders, and some working on tapes...teachers and aids will be wandering around helping each as they need it."

General: "Today the students are working on the renaissance man, listing traits of such a person, then look for patterns in the listings...There will be no rejection of answers by the teacher, so it is the students responsibility for the answers...the kinds of questions asked are related to the higher thought processes according to the product and operation dimensions...Evaluative questions will be emphasized."

Detailed: You will see "the teacher emphasizing creative thinking skills that generate fluent, flexible, and original responses from the students...it will be a student centered atmosphere making provision for choice...the students will show acceptance of rules if they are explained...they will be working in groups part of the time...The teacher will accept all student responses, never giving verbal or non-verbal rejection to any student statements...all of these activities relate to idea of creative thinking skills which is one main...objective of our program."

#28 Were effects of the demonstration program on student attitudes explained?

Little: "Last year's students were a problem...not courteous... most of the students this year like the class...can do more things."

General:* "Really like the program that we are in...you get a chance to think for yourself...not like the class last year...lot of memory of facts...the other kids like it too because there is less homework and class is more fun."

*These comments were recorded from student statements about the program.

Detailed:* "Most of us don't feel grades are important...we like the program, but it is harder...at first I goofed around a lot, but my conscience caught up with me... the ground rules are O.K., you can spend the whole year on one project if we want to...like it better than the regular class because you are your own master...feel better prepared for college...feel better to have teachers as equals..."

*These comments were recorded from student statements about the program.

Concluding Comments

Two factors influenced the level of rating for each of the items. One of these factors was the kind of information that was presented. Specific topics were expected to occur in order for the center to receive a high rating for a particular item. The other factor was that of a time/quantity measure. If a lot was said and a good deal of time was taken to express it, the item was considered to have been emphasized relative to that length of time and amount of information.

Hand-outs were also counted as part of the rated content as long as they were in some way referred to verbally. The system for rating the hand-outs was as follows:

None: Here is a packet of information about our program (no identification of what the packet specifically contained).

Little: Here is a packet of materials and in here you will find a list of our programs, the objectives of those programs, and a schedule for you to follow today (the packet was given and the contents identified).

General: Here are some materials, they may help you understand our programs a little better. You will find our program objectives and program descriptions enclosed. Please take 5 minutes to read this over and I'll be glad to answer any questions you might have (the packet was given, identified, and visitors were given a definite time to read it).

Detailed: Here are some materials for you to look over regarding our program here. I would like for you to read it right now...You will find a description of each program...Now, I would like you to pay particular attention to what the program objectives are on page 3 of your materials. As you can see there are three and I would like to tell you something about each, etc...(the packet was given, identified, and time for reading was provided, along with some explanation of the printed matter).

Finally, in rating each item greater credit was given where information about any given item was made relevant to an individual visitor's needs. Such was the case when one director said, "...now at Cahokia your program director has been at the junior high, mainly...how do you get your top groups...your program director has been in the content area of social studies...in your situation you might consider using the Guilford structure...have you had any background in it?

2. Copy of Instrument

DEMONSTRATION OBSERVATION SCHEDULE

Center _____ Rater _____ Date _____

SECTION 1: EXPLANATION OF PROGRAM (VERBAL ORIENTATION)

	Detailed	General	Little	None
1. Were program objectives explained? (what, why, how, when)				
2. Were program treatments explained? (e.g., methods, materials, management)				
3. Was a description of school population given? (for example racial, socio-economic level, relation to program)				
4. Student selection procedures explained? (for example, tests used, who tested, cut-off pt(s)., weighting, relation to program, grouping arrangements, availability of test results)				
5. Historical explanation of program(s) given? (for example, date begun, who started, why, growth of program)				
6. State plan described? (e.g., parts listed, explained, illus., related to visitors)				
7. Teacher selection criteria explained? (e.g., who chose, minimums, recruitment)				
8. Teacher training for demonstration program(s) explained? (e.g., courses, internship, in-service)				

SECTION 2: EXPLANATION OF CLASS (VERBAL ORIENTATION)

		Detailed	General	Little	None
9.	Today's class objectives explained? (e.g., were they related to overall program objectives)				
10.	Today's class treatment explained? (e.g., were they related to overall program objectives)				
11.	Student Selection procedures for this class explained? (e.g., tests used, who tested, cut-off pts., weighting, relation to program, grouping arrangements for class, availability of tests, non-gifted)				
12.	Intraclass academic progress (scores) explained? (e.g., speed, problems)				
13.	Intraclass characteristics explained? (e.g., social patterns, interests, study habits)				

SECTION 3: OBSERVATION OF DEMONSTRATION CLASS

	YES	INCONCLUSIVE	NO
14. Did the day's lesson reflect the overall program objectives?	___	_____	___
15. Did the day's lesson reflect the overall program treatment?	___	_____	___
16. Was competence of teacher adequate?	___	_____	___
17. Was orientation, background or review given visitors as part of class sequence?	___	_____	___
18. Did total class sequence seem artificial?	___	_____	___
19. Were children continually distracted by the presence of visitors?	___	_____	___
20. Was visitor behavior excessively disruptive?	___	_____	___
21. Were visitors able to <u>see</u> class proceedings clearly?	___	_____	___
22. Were visitors able to <u>hear</u> class proceedings clearly?	___	_____	___
23. Were additional classroom materials needed to follow lesson?	___	_____	___
24. Were visitors given a definite opportunity to talk to teachers?	___	_____	___
25. Were visitors given a definite opportunity to talk to students?	___	_____	___

SECTION 4: EXPLANATION OF DEMONSTRATION CENTER'S OWN EVALUATION

	Detailed	General	Little	None
26. Was demonstration center's plan(s) for its own evaluation explained? (e.g., procedures, scheduling, rationale)	____	____	____	____
27. Interclass academic progress explained? (e.g., compared to last year or another group this year; compared to similar groups using local or national norms)	____	____	____	____
28. Were effects of the demonstration program(s) on student attitudes explained?	____	____	____	____
29. Were effects of the program on demonstration teachers' morale and attitudes given?	____	____	____	____
30. Were the reactions of the community to the project discussed?	____	____	____	____
31. Were the reactions of the students' parents discussed?	____	____	____	____
32. Were effects of the demonstration program on non-program students explained?	____	____	____	____
33. Were effects of the program on non-demonstration teachers discussed?	____	____	____	____

SECTION 5: EXPLANATION OF PROGRAM FEASIBILITY

	Detailed	General	Little	None
34. Were possible problems of installation in other schools discussed?	_____	_____	_____	____
35. Was an estimate of funds needed for installation of the program given?	_____	_____	_____	____
36. Were necessary equipment and materials discussed?	_____	_____	_____	____
37. Were the visitors told how to locate these materials and equipment?	_____	_____	_____	____
38. Were continuing costs of the program discussed? (e.g., maintenance)	_____	_____	_____	____
39. Was what you need to get in the way of training in order to start this program in another school explained?	_____	_____	_____	____
40. Were weaknesses of the program explained?	_____	_____	_____	____
41. Were strengths of the program discussed?	_____	_____	_____	____

Appendix E

THE OBTRUSIVENESS OF MEASURES

On February 4, 1969 the following reaction form was sent to the directors of the gifted demonstration centers providing them with the opportunity to indicate the reactive influence The Gifted Evaluation Project had created (See Figure 1).

FIGURE 1:

1. Make three short statements about your negative reaction to the presence and behavior of the data collectors.

   a. _____

   b. _____

   c. _____

2. Make three short statements about your negative reaction to the overall Gifted Evaluation Project.

   a. _____

   b. _____

   c. _____

General Comments

The information (reactions) provided by the directors was not given to the evaluation staff in any form which would identify its source.

Thirteen of the twenty-one directors responded to the questions. These reactions were mixed.

The reactions to the presence and behavior of the data collectors ranged from statements such as "they fit right in" to "they refused to act normal." The major concern about the data collectors had to do with their abnormal behavior--"Their methods alienated our visitors."

Regarding the reactions to the overall evaluation project, some directors continued to aim their comments at the data collection teams with such statements as "one visit is a short-sighted view" and "the collectors did not enter into the spirit of a demonstration center visitation." Other directors indicated major concerns in schedule shifts, length of forms, and "The futility of collecting data which cannot be used in decision making...for coming biennium."

In general then, the directors indicated an uneasiness about the data collectors' presence, and noted a concern about to what use the data should and/or would be put. A statement by statement account of the directors' reactions follows in Figure 2.

Figure 2: EVALUATING THE EVALUATORS DEMONSTRATION DIRECTORS' REACTIONS

(Number of responses 13)

1. MAKE THREE SHORT STATEMENTS ABOUT YOUR NEGATIVE REACTION TO THE PRESENCE AND BEHAVIOR OF THE DATA COLLECTORS.

I have none whatsoever

None - they fit right in - perhaps it was because we were familiar with the people and therefore, did not feel threatened by them.

They created an abnormal orientation & visitation. They refused to act as normal visitors. Their methods alienated our visitors.

One of the data collectors showed disinterest in what was going on.

Their presence made me somewhat self-conscious. I kept wondering about their "objectivity." I felt they should have been around more than one time.

Figure 2 (Continued)

One data collector did not express interest in the demonstration program. This was made evident to a school secretary and a demonstration teacher. One principal commented that the state should not send people to evaluate unless they were truly interested in the demonstration.

Furious note taking and recording was distracting. Non-participation in discussion activities. Unwillingness to share findings.

Note taking was very disturbing to demonstration teachers. Data collectors should have paid attention to the demonstration rather than continuously taking notes. Should of at least looked interested instead of laughing among selves during some parts of presentation.

No negative feelings about data collectors or techniques of data collection.

Anonymity of evaluators to other visitors. Partial observation of the demonstration program.

2. MAKE THREE SHORT STATEMENTS ABOUT YOUR NEGATIVE REACTION TO THE OVERALL GIFTED EVALUATION PROJECT.

More specifics on how and by whom the data is to be used. Would have enjoyed immediate feedback but we understand your position.

Too slow in coming--available information was/is too technical. No one will pay much attention to it.

The flurry of rumors concerning the use of the evaluation data in regard to selecting centers which should be refunded.

I wondered, if during one visit, if enough information could be gained to really make a good evaluation of a demonstration center.

It made us do considerable shifting (the Feb. feedback sheet did). Some how I felt out of touch with the ongoing evaluation process. I wanted to know more about what was happening. I don't think the Springfield office will give the CERLI information enough attention.

The forms were too lengthy. The host center should not have to give up their own evaluation program.

Figure 2 (Continued)

Not built in from the beginning. Too much potential impact on people who don't know what gifted program is and does (legislature).

One visit is a short-sighted view since follow-up with teachers is usually necessary to know if anything was really carried away by visitors.

Team was too conspicuous for the collectors did not enter into the spirit of a demonstration center visitation. The questionnaire for the month of February is very unrealistic for visitors to complete - too complicated after a long day.

The futility of collecting data which cannot be used in decision making about direction of program for coming biennium.

Lack of immediate feedback to centers--perhaps some separate short form could be used for this purpose.)

Greater communication with directors regarding degree of involvement in the evaluation (administration of forms, etc.)

GENERAL COMMENTS

Found the CERLI evaluators fitted in very well with the other visitors and on occasion we forgot who they were. If their presence helps in an objective evaluation of the functions of a demonstration center, I have no qualms about their returning as often as necessary. I do not think the other visitors were aware of the CERLI people as being different.

My only comments would be in relationship to the questionnaire we are using during Feb. as ordered by you, I have not read it, but the majority using it the first week found it hard to determine what was meant by many of the questions.

We have felt the evaluation has caused us no problems. In each case we attempted to make the evaluators work as easy as possible. We felt that it was part of our responsibility to make the evaluation project as valuable and accurate as possible.

The only negative comment from one of our Feb. visitors was that he felt that some of the questions did not leave him enough flexibility. Since I have not looked at the questionnaire to avoid reacting to it, I can't even say for sure that this comment is valid.

Figure 2 (Continued)

The Gifted Evaluation Project came at a time when we were in the midst of changing personnel director (director had been in charge of center for only one month.) Also, there has been a change in personnel of teachers within the five-year period. Only one teacher has remained in center since its inception 6 years ago.

I enjoyed the visiting teams. I'm certain they were trying to do a good job. I appreciate the CERLI'S team. They tried to cooperate with us in every way even though it was difficult to make arrangements for a visit.

None worth sharing.

I am pleased by the efforts put forth---hope that some of the findings can eventually be used to upgrade program planning.

The above comments express minimal negative reactions from our center; those directly involved in verbal or written interviews were satisfied with the techniques and processes used.

At the close of a very full day with visitors, directors are somewhat hard pressed to respond with a relatively high degree of effectiveness to the many and varied questions posed them. These questions call for information regarding program development, research data, evaluation procedures, dissemination effects of center, etc., etc.

Timing of interviews as comprehensive as the above needs to be reconsidered.