# MAKING INSTRUCTIONAL RESOURCE SENSE OUT OF GOVERNMENT POLICY DOLLARS*

Don Dorr-Bremme
Joan L. Herman
Edys S. Quellmalz

CSE Report No. 191

1982

CENTER FOR THE STUDY OF EVALUATION

Graduate School of Education

University of California, Los Angeles

*The papers in this report were originally presented in a symposium at the Annual Meeting of the American Psychological Association, Los Angeles, 1981.

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

# INTRODUCTION

The future of instructional research, at least in the present economic climate, is indistinct. The trends suggest a continuing reduction of support for basic research and a concomitant increase in competition for scarce resources. At the same time, evaluation or other policy directed studies may continue at their present level, if for no other reason than to provide rationales for budget reduction. This report considers the option of combining within a single study the needs of policy makers and the commitment to academic research. The decisions to be made involve real vs. laboratory settings, experimenter controlled vs. naturalistic designs, lean vs. thick data collection, and political reality vs. scholarly quality. The papers in this report, through illustration of research conducted within a policy framework, will identify problems and/or benefits of the forced marriage of knowledge production and decision-directed research. Methodologies for optimizing the match will also be explored. In each case example, the research focused on classroom behaviors and related instructional activities. Outcomes of interest included cognitive performance and affective responses from students and teachers.

The report considers future directions of research, not only as suggested by the specific fidings of theoretically derived inquiry, but also as such options may be influenced by the reality of political, administrative, and economic constraints. How can we serve self-interest, research, and policy interests? For example, the values of academic freedom come in direct conflict with centralized, e.g., policy, mandates.

In the report, Quellmalz identifies problems and limitations of current designs for serving instructional research needs, and suggests some alternative research strategies. Herman presents methodologies for combining research and policy needs, and suggests the advantages inherent in their merger. Finally, Dorr-Bremme highlights the advantages and problems involved in embedding a piece of instructional research in a larger policy study.

# ISSUES IN DESIGNING INSTRUCTIONAL RESEARCH: EXAMPLES FROM RESEARCH ON WRITING COMPETENCE

Edys S. Quellmalz

Instructional research ranges from broadly conceived national studies of schooling's effect on basic skills achievement to individual researcher's studies of specific variables promoting particular skills. Most of these studies tend to focus on features of the school, classroom, teacher, and curriculum to identify policies or actions facilitating learning. There are widely divergent perceptions, however, of the type and specificity of independent and dependent variables appropriate in large-scale (top down) and small-scale (bottom-up) studies of instruction conducted in the school context.

Much large-scale research is driven by evaluation methodology, while smaller-scale studies use paradigms from instructional technology and cognitive psychology. This paper describes two main categories of problems that seem to pervade school-based studies of instruction. The first set of problems relates to the design of outcome measures in terms of (1) the lack of sensitivity of many dependent measures used to document instructional effects, (2) the failure to collect corroborating measures of effect, and (3) the failure to match the content and processing requirements of alternative measures with each other. The second category of problems is in the design of context and process descriptions, including (1) the failure to describe contextual dimensions of the school and curricular systems that set the conditions within which instruction occurs,

(2) the failure to freely explore instruction as an interactive process; to relate instructional variables to logical or research bases to explain achievement results, and (3) the failure to compare the context and processing requirements of classroom tasks with test tasks.

The purpose of this paper is to describe features of instruction -- its context, processes, and outcomes -- whose relevance and utility for instructional improvement seem to have the strongest empirical support. The paper argues that researchers and evaluators studying instruction in schools should sharpen the focus of measures and better trace the interrelationships within and between independent and dependent variables.

## Problems in the Design of Outcome Measures

Lack of sensitivity. A prevalent problem in school-based instructional research is that test tasks are often insensitive to the logical and psychological aspects of tasks presented in instructional interventions. There is a gap between notions of the appropriate level of detail for describing and constructing dependent measures in laboratory-based instructional research designed by psychologists, on the one hand, and in school-based instructional research conducted by evaluators and psychometricians, on the other hand. For example, large-scale federal evaluations and many state and district evaluation studies still rely primarily or exclusively on norm-referenced tests to detect instructional effects. The many criticisms of norm-referenced tests for reflecting achievement of specific instructional goals have been described elsewhere (Glaser, 1963; Popham, 1978; Millman, 1974; Hambleton et al., 1978). The recognition of the need for a much closer match between testing and instruction first stimulated the call for criterion-referenced testing (Glaser, 1963). Furthermore, a growing body of learning research shows that student performance varies

significantly when task demands change. Classes of task or problem types require students to access different bodies of stored information and to activate different procedures, routines, or solution strategies.

For example, in math, the work of John Seely Brown and his associates demonstrates that different sets of subtraction problems elicit different solution schema (Brown et al., 1978). Thus descriptions of achievement at the molar level of "math achievement" or even of "computational skills" cannot sufficiently describe performance on homogeneous sets of skills, nor signal skill areas requiring attention at the program, classroom, or individual level. Similarly, reading research indicates that reading comprehension is not an undifferentiated construct; rather the type or discourse mode of reading material, such as narration and exposition, requires different schema for comprehension (Brown et al., 1978; Meyer, 1975). This research implies that, if tests are to be sensitive to different types of reading skills, they must be designed to provide subscore profiles on skills or inferencing required by different types of reading passages. They cannot merely report generalized scores for decoding and literal and inferential comprehension. Yet federal-level evaluations such as Follow-through and Cities in Schools (Murray et al., 1981) report global "reading achievement" scores.

Nowhere is the insensitivity of dependent measures more dramatically illustrated than in the recent surge of studies of writing. Like reading achievement, writing achievement must be decomposed into the level of skill demonstrated in relation to different types of writing tasks. The various controlling purposes of discourse modes or genre require students to use different kinds of topical information and different presentation strategies according to organizational schemes and development methods

conventionally expected in these various genres.

Large-scale evaluations of writing competence have too frequently used multiple choice items to measure writing achievement. Psychologists would deny the contruct validity of recognition tasks as anything other than enroute indicators of production capabilities. Research on the comparability of information derived from indirect (multiple choice) and direct (writing sample) tasks has primarily been conducted by psychometricians more schooled in metrics than psychology. Studies conducted within the psychometric framework have reported high correlations between total multiple choice test scores and holistic essay scores and cite these correlations as support for substituting multiple choice tests for essays. Recent research within a competency testing framework has investigated the comparability of information from these two measurement response forms. They have found lower total score correlations and, more importantly, much lower correlations between direct and indirect scores for subskills such as coherence, support, and mechanics (Moss, Cole & Khampalikit, 1982; Quellmalz & Capell, 1982; Quellmalz & Baker, 1981).

Studies of the effects of instructional interventions on writing achievement also demonstrate that holistic scores do not adequately describe how the varying skill levels in component features of the product contribute to the global quality score. For example, studies guiding students in writing strategies may find no significant differences in pre- and post-intervention judgments (e.g., Pearl, 1979), yet researchers discussing these inconclusive results cite observational information suggesting that student writing <u>really did</u> improve. At a conference of grantees of federally funded writing projects discussing their research progress, a dominant concern was the failure of holisitc

essay ratings to capture improvement or, at least, changes in student writing due to instructional treatments. The remedy, of course, is to design scoring schemes that include criteria, subskill ratings, and even detailed secondary discourse analyses detailing the features of interest.

Some writing researchers are doing this. Odell (1978) instructed students in Pike's (Young, Becher, & Pike, 1970) pre-writing discovery approach for selecting and organizing essay content. While judged essay quality as a whole was not affected, textual analyses showed that students' use of rhetorical devices such as temporal sequence and classification did increase. Similarly, Bracewell, Bereiter, and Scardamalia (1980) taught students rhetorical strategies for persuasion and found their use significantly increased, although overall ratings of essay quality did not. If holistic essay quality scores were the only ones reported, it might be concluded that the instructional interventions had no effect and should be dropped. But when more detailed analyses document that what was taught was used in students' writing, the implication may be that detection of overall quality effects requires

- more time and practice
- additional and different instruction on the subskill to help students use it more effectively
- instruction to help students integrate the strategy with other writing skills.

An analogy is seen in the case of a tennis instructor working with a student on his/her backhand. At the end of a series of lessons, two dependent measures seem appropriate: (1) is the backhand stroke and resulting ball placement better? (2) does the student win more games? If only the "games won" measure is used, it might be concluded that: (1) the backhand instruction had no effect and should not be used again,

(2) the student needs more practice time, (3) while concentrating on his backhand, the student's forehand went to pot, contributing to the "no increase in games won" score. This last phonomenon, all too familiar to athletes, implies the need for more practice and, most likely, for instruction on integrating use of the two strokes.

Therefore, many instructional studies would profit from more careful, detailed designs of dependent measures that document performance on sub-skills taught, as well as overall performance. Policy decisions at federal, state, district, and classroom levels which draw exclusively on the overall measure might conclude that the program or treatment had no effect at all. The treatment, however, might have been effective, but the measure was too gross to detect it.

Failure to collect corroborating outcome measures. A second problem in the design of outcome measures is the failure to collect information about student performance on other facets of the skill. Too often, evaluation and instructional research studies report global performance on a single measure such as a math achievement or a reading achievement test. Again, writing assessments that collect only one sample dramatically illustrate the illogical and methodologically unsound nature of the "one-shot" performance index. Numerous studies of writing assessment demonstrate fluctuations in individual performance on different writing tasks (Crowhurst & Piche, 1979, Quellmalz & Capell, 1982). Certainly, we feel increasingly confident about students' competence when it is demonstrated repeatedly. Unfortunately, most commercially available tests present only one or two items per skill. While multiple performance indicators on other formal assessment devices are helpful, research suggests that progress on in-class work samples may provide better corroborating data. Studies of

test anxiety and contextual influence on performance, especially on writing performance, support the utility and validity of collecting classroom performance information, since the classroom is the more realistic and normal context for the student.

Failure to relate the context and processing requirements of alternative measures to each other. When corroborating data are collected, it seems reasonable that the data should be from performance on tasks similar in their processing requirements. Yet studies often fail to check or describe whether types of tasks on two formal tests or on classroom problems match. Whether direct records of classroom assignments and test performance are collected or teachers' indirect ratings of achievement progress are gathered, the comparability of tasks must be rigorously described.

## Problems in the Design of Descriptions of Instructional Contexts and Processes

Alternative research paradigms focus on the learning environment's features that differ substantially in specificity and in proximity to the learning event. The search for effective instruction in the complex formal school setting has, fortunately, grown from studies of teacher personality and vaguely defined teaching methods and now includes political and administrative contextual influences of the extended school system. Also, recent research is attempting to document the classroom's physical, social, and managerial context to explain factors influencing the interactive information processing of teachers and students. Few studies of instruction, large-scale or classroom level, trace the links between the conditions under which instruction occurs, features of the instructional process, and learning outcomes. While experimental researchers conducting laboratory studies are trained to describe the conditions in which treat-

ments occur, researchers studying the complex classroom environment may fail to describe extra-classroom conditions that constrain instruction. Similarly, large-scale evaluations may fail to describe alternative conditions of implementation that relate to program effects. Federally funded evaluation reports sometimes describe program implementation (process and context) in separate volumes. Policy makers thus find it difficult to trace cause/effect relationships between instructional implementation patterns and achievement data (e.g., Murray et al., 1981).

The context and process variables affecting learning outcomes are broad in scope and large in number. Figure 1 suggests categories of contextual and process variables with a research base supporting their influence on achievement. The contextual variables include constraints imposed by the existing organizational and curricular systems as well as by teachers' and students' entering perceptions and abilities. Variables involved in the course of the instructional process include the interactions of task features, teacher behaviors, and student behaviors in the hypothesized internal learning processes of students.

The system context: the school. Studies of effective schooling have identified policies enacted at higher levels of the educational and political system that profoundly affect the ultimate nature of classroom instruction. For example, legislative mandates affect the composition and stability of the school population. Funds allocated for the support of general education and special programs influence the range of available resources, usually personnel and materials. Resource allocation policy decisions seriously affect potential instructional quality (Harnishfeger & Wiley, 1976). Perhaps the most influential legislation affecting school instruction has been minimum competency testing re-

Figure 1:

Issues in Designing Instructional Research: Examples from Research on Writing Competence

FRAMEWORK OF VARIABLES FOR STUDYING INSTRUCTION

CONTEXT VARIABLES

PROCESS VARIABLES
INTERACTIVE INSTRUCTION

| System Context | Curricular Context | Teacher Characteristics | Student Characteristics | Task Features | Teaching Behaviors | Student Behaviors | Learning Process |
|---|---|---|---|---|---|---|---|
| Federal funding regulations | Mandated goals objectives syllabus | Experience in teaching in the subject | Ability level | Social context | Goal setting<br>Describe outcomes' content, form | Physical orientation<br>eyes, body | Attention |
| Laws on school student composition | Available test materials resources | Orientation teacher's role, preferred teaching methods | Achievement level documented<br>reading<br>writing<br>math | Functional purpose | Orient to relevant features | Ask questions | |
| Laws on minimum competency | Available curriculum consultants staff development | Concepts of subject matter | Language development (other dominant language) | Relationship to S's world knowledge | Presentation/Explanation<br>Present or elicit relevant content, strategies, rules | Rehearse<br>segment<br>label/categorize | Encode |
| District policy | | Information base about subject matter | Values and expectations about subject matter | Structural features<br>kind of task and interrelationship of information presented or required | Feedback | Elaborate/transform | |
| School administrative policy, emphasis, and style | | Judgments of student ability | Cultural background | Required processes<br>information component strategies<br>patterns to integrate above | Practice<br>Elicit response<br>Ask questions | Relate to other knowledge | |
| | | Expectations for student progress | | | Feedback<br>On appropriateness of details<br>codes<br>procedures<br>strategies<br>via<br>praise<br>confirm<br>correct<br>induce<br>tell<br>give explanation<br>rule<br>example | Use imagery | |
| | | | | | | Answer questions | Retrieve |
| | | | | | | Verbalize rules, strategies | |
| | | | | | | Solve problem<br>plan<br>write<br>revise<br>edit | |
| | | | | | Patterns of all above points of application orchestrated and integrated | Patterns of all above orchestrate and integrate | |

quirements and the accompanying scrutiny of the quality of instructional opportunities preparing students to pass those exams. Testing requirements influence curriculum emphasis and student achievement (Yeh, 1978; Schwille, Porter, & Garet, 1979; Floden, Porter, Schmidt, & Freeman, 1980). Also, administrative policy within the educational system, such as curriculum guidelines and state adopted texts, constrain classroom options. Edmonds (1979) cites several studies (e.g., Weber, 1971; New York, 1974; Brookover & Lezotte, 1977) showing that an active and supportive school administration leads to higher achievement levels in inner city schools. These studies point to the need to describe the systemic and curricular conditions limiting instructional options in the classroom.

Within the classroom itself, teacher effectiveness studies conducted through observations and naturalistic inquiry identify administrative policies that contrain instructional options. Number of pupils assigned to each class and range of pupil ability in a single classroom certainly bound teachers' planning (Dahllöf, 1971).

As Barr and Dreeben (1977) have noted, studies of classroom effects must refer to the broad social context in which the classroom functions. Their reviews of instruction in classrooms and Doyle's (1977) critique of paradigms for research on teacher effectiveness underscore the need for expanding the breadth and depth of variables considered when examining classroom ecology. They contend, and I agree, that most current research paradigms fail to consider the full range and the functional interdependence of contextual variables in instruction.

The system context: the curriculum. Classroom research across subject matter suggests that the basic unit of instruction is the assignment

or task. The task is described as a goal-directed set of activities presenting students with content of certain characteristics and required procedures for completion (Doyle, 1979; Mehan, 1974; Van Nostrand et al., 1980). For teachers, tasks involve content, materials, and activities (e.g., Morine-Dershimer, 1979; Shulman, 1980; Schutz, 1980). Research suggests that materials availability strongly influences the types of problems, practice, guidance, and feedback that students receive. Observations of classroom instruction for low achieving, low SES students in the elementary grades revealed that students spent as much as 70% of their time working alone with materials. Although studies at the secondary level indicate that activities are less materials-driven (Sirotnik, 1981; Applebee, 1981; Van Nostrand et al., 1980), it may be that appropriate materials are less abundant or unavailable. In any case, the instructional quality of commercially available materials has been criticized severly (Quellmalz et al., 1977; Van Nostrand et al., 1980).

In any study of effective instruction, then, one category of central questions should address the availability and quality of curriculum materials.

The interactive instructional process. When the teacher effectiveness literature is viewed from the perspective of theories of learning and instruction, many findings are rendered irrelevant or useless to the design of instructional research. Characteristics such as "businesslike" are too far removed from the refinements of student information processing. Medley's (1979) extrapolated effectiveness constructs certainly indicate that many descriptive studies are far removed from the learning act. For example, maintenance of a learning environment includes both "orderly" and "supportive" behavior. "Time on task" was reported effective in large group settings only. Methods of instruction generally thought to be important

were found to be ineffective with disadvantaged learners. Among these methods were high level questions, students asking many questions, providing more feedback, and increased teacher amplification.

It is clear that the analysis of teaching and learning must provide much more detailed descriptions of the conditions under which such findings prevail. Peterson (1979) notes the highly contingent interdependencies of instructional variables in her critique of Rosenshine's review of the effectiveness of the direct instructional model (Rosenshine, 1979). Rosenshine identified major components of this model as: (1) clear goals, (2) sufficient and continuous goals, (3) content coverage, (4) monitoring of performance, (5) low cognitive level questions, (6) immediate academically oriented feedback. Studies reviewed by Medley and Rosenshine focused on disadvantaged elementary age children. Thus, one might guess that low level questions were better predictors of performance because students were just learning skills and because low level items were on the test.

Teacher effectiveness studies concentrating on "time on task" have primarily been large-scale (see Cooley & Leinhardt, 1980; Fisher et al., 1978). Findings about academic learning time were not startling; what was surprising was how little classroom time was provided for learning tasks. Clearly classroom time management is prerequisite to effective participation in instruction. In the Instructional Dimensions Study, within the instructional event, the techniques identified as related to the quality of instruction were (1) focusing attention on the task, (2) refferring to previously used material, (3) referring to earlier performance, and (4) effective classroom management. In the Beginning Teacher Evaluation Study, teaching methods associated with achievement and academic learning

time were (1) provision of tasks permitting a high success rate, (2) more presentation of information, (3) more monitoring of work, and (4) more feedback about academic performance. These findings coincide with those reported by Stallings (1980), who describes interactive on-task instruction time as characteristic of effective teachers. Effective instructional patterns were (1) more support and (2) positive corrective feedback. The nurturing environment is particularly important for secondary students with a history of failure. The need for positive, informative modes of feedback has also been reported. For example, Webb (1980) found that students working on cooperative tasks in groups participated more actively and achieved more when the group gave and received more explanation about how to solve problems.

While these teaching behaviors apparently can and do occur, some researchers using naturalistic inquiry methods report that teachers may leave performance expectations unsignaled, i.e., no clear goal setting, (Mehan, 1974) and provide inconsistant feedback. The question to be asked is how effective teachers plan and contruct instructional events to result in effective interactions. Borko et al. (1979) suggests that actual teacher planning is at odds with the idealized paradigm. As mentioned previously, teachers plan in terms of content, materials, and activities. The resultant instructional task or assignment becomes the basic unit of planning and action in the classroom (Doyle, 1977; Clark & Yinger, 1979). A major line of ethnomethodological and sociolinguistic inquiry focused on descriptions of teachers' decision-making for planning and aid during the interactional teaching-learning phase. One example of the detailed level of this research is reported by Dorr-Bremme. In a

untrained or unthinking.

Sensitive outcome measures can report students' performance on a reasonable number of items or tasks (not one or two) measuring subskills as well as total scores. Criterion-referenced testing programs are attempting this now. Data from a sensitive test can always be aggregated or disaggregated at a level appropriate to the policy decision (individual, class, school, district, state, or nation). Data from an insensitive test can never be disaggregated or decomposed. For example, policy makers are better served by data indicating the type of subskills on which students have difficulty, rather than a statement that they "can't read." Financial resources can then be focused on curricular and personnel selections relevant to areas of performance weakness.

Data collected through teacher records of performance on class assignments and tests can corroborate test information. By improving the designs of instructional research, projects' limited research funds should yield more valid, useful information for improving instruction.

## References

Applebee, A. M. Writing in the secondary school: English and the content areas. NCTE Research Report No. 21, 1981.

Barr, R., & Dreeben, R. Instruction in classrooms. In L. S. Shulman (Ed.), Review of Research in Education. Itasca, Ill.: F. E. Peacock, 1977. Pp. 89-162.

Borko, H., Cone, R., Russo, N., & Shavelson, R. J. Teachers' decision making. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and applications. Berkeley, CA: McCutchan Publishing Corporation, 1979. Pp. 231-263.

Bracewell, R. J., Bereiter, C., & Scardamalia, M. How beginning writers succeed and fail in making written arguments more convincing. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Brookover, W. B., & Lezotte, L. W. Changes in school characteristics coincident with changes in student achievement. East Lansing, MI: College of Urban Development, 1977.

Brown, J. S., & Burton, R. R. Diagnostic models for procedural lags in basic mathematics skills. Cognitive Science, 1978, 2, 155-192.

Brown, J. S., Stein, N. L., & Glenn, C. G. An analysis of story comprehension in elementary school children. In R. P. Freedle (Ed.), Discourse processing: Multi-disciplinary perspectives. Norwood, NJ: Able, 1978.

Clark, C. H., & Yinger, R. J. Teachers' thinking. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and applications. Berkeley, CA: McCutchan Publishing Corporation, 1979. Pp. 231-263.

Cooley, W. W., & Leinhardt, C. Instructional dimensions study. Educational Evaluation and Policy Analysis, 1980, 2(1), 7-25.

Crowhurst, M., & Piche, G. L. Audience and mode of discourse effects on syntactic complexity in writing at two age levels. Research in the Teaching of Writing, 1979, 13(2), 101-109.

Dahlöff, U. S. Ability grouping, content validity, and curriculum process analysis. New York: Teachers College Press, 1971.

Dorr-Bremme, D. Behavior and making sense: Creating social organizations in the classroom. Unpublished doctoral dissertation, Harvard Graduate School of Education, 1982.

Doyle, W. Learning the classroom environment: An ecological analysis. Journal of Teacher Education, 1977, 28.

Doyle, W. Classroom tasks and students' abilities. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and applications. Berkeley, CA: McCutchan Publishing Corporation, 1979.

Edmonds, R. Effective schools for the urban poor. Educational Leadership, October 1979, 15-24.

Fisher, C. W., Berliner, D. C., Filby, N. N., Marliave, R., Caben, L. S., Dishaw, M. M., & Moore, S. E. A summary of the Beginning Teacher Evaluation Study. BTES Technical Report VII-I. San Francisco: Far West Regional Laboratory for Research and Development, 1978.

Floden, R. E., Porter, A. C., Schmidt, W. H., & Freeman, D. J. Don't they all measure the same thing? Consequences of standardized test selection. In E. L. Baker & E. S. Quellmalz (Eds.), Educational testing and evaluation. Beverly Hills, CA: Sage Publications, 1980. Pp. 109-120.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Graves, D. Balance the basics: Let them write. New York: Ford Foundation, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of issues and developments. Review of Educational Research, 1978, 48, 1-48.

Harnishfeger, A., & Wiley, D. The marrow of achievement test score declines. Educational Technology, 1976, 16(6), 5-14.

Medley, D. M. The effectiveness of teachers. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and applications. Berkeley, CA: McCutchan Publishing Corporation, 1979. Pp. 11-27.

Mehan, H. Accomplishing classroom lessons. In A. V. Cicourel et al. (Eds.), Language use and school performance. New York: Academic Press, 1974.

Meyer, B. F. The organization of prose and its effect on memory. North Holland Studies in Theoretical Poetics (Vol. 1). Amsterdam: North Holland Publishing Company, 1975.

Millman, J. Sampling plan for domain-referenced tests. Educational Technology, 1974, 11, 17-21.

Morine-Dershimer, G. Teacher conceptions of pupils. East Lansing, MI: Michigan State University Institute for Research on Teaching, Research Series No. 59, 1979.

Moss, P., Cole, N., & Khampalikit, C. A comparison of procedures to assess written language skills of grades 4, 7, and 10. Journal of Educational Measurement, 1982, 19, 37-48.

Murray, L., et al. The national evaluation of the cities and schools program, Report No. 4. Final Report 1981.

New York State Office of Education Program Review. School factors influencing reading achievement: A case study of two inner city schools. March, 1974.

Odell, L. Measuring the effect of instruction in pre-writing. Research in the Teaching of English, 1978, 12, 228-240.

Pearl, S. The composing process of unskilled college writers. Research in the Teaching of English, 1979, 13, 317-336.

Peterson, P. L. Direct instruction reconsidered. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and applications. Berkeley, CA: McCutchan Publishing Corporation, 1979. Pp. 57-69.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education. Los Angeles, CA: Center for the Study of Evaluation, 1978.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice Hall, 1978.

Quellmalz, E. S., & Baker, E. L. Effects of alternative scoring options on the classification of entering freshman writing competencies. Report to the National Institute of Education. Los Angeles: UCLA Center for the Study of Evaluation, 1981.

Quellmalz, E., Baker, E., & Enright, G. Studies in test design: A comparison of modalities of writing prompts. Los Angeles: UCLA Center for the Study of Evaluation, 1980.

Quellmalz, E. S., Capell, F. J., & Chou, C. P. Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 1982, 19(4), 241-258.

Quellmalz, E. S., Snidman, N. S., & Herman, J. H. Toward competency-based reading systems. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Rosenshine, B. V. Content time and direct instruction. In P. L. Peterson & H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and applications. Berkeley, CA: McCutchan Publishing Corporation, 1979. Pp. 28-56.

Schutz, R. E. The design of measurement in instruction. In E. L. Baker & E. S. Quellmalz (Eds.), Educational testing and evaluation. Beverly Hills, CA: Sage Publications, 1980.

Schwille, J., Porter, A. C., & Garet, M. Content decision making and the politics of education. East Lansing, MI: Michigan State University Institute for Research on Teaching, Research Series No. 52, 1979.

Shulman, L. S. Test design: A view from practice. In E. L. Baker & E. S. Quellmalz (Eds.), Educational testing and evaluation. Beverly Hills, CA: Sage Publications, 1980.

Sirotnik, K. A. A contextual appraisal system for schools: Medicine or madness. Los Angeles: UCLA Center for the Study of Evaluation, Report No. 169, 1981.

Van Nostrand, A. D., Pettigrew, J., & Shaw, R. Writing instruction in the elementary grades: Deriving a model by collaborative research. Providence, R. I.: Center for Research in Writing, 1980.

Webb, N. M. A process-outcome analysis of learning in group and individual settings. Educational Psychologist, 1980, 15, 69-83.

Weber, G. Inner-city children can be taught to read: Four successful schools. Washington, D. C.: Council for Basic Education, 1971.

Yeh, J. P. Test use in schools. Los Angeles: UCLA Center for the Study of Evaluation, 1978.

Young, R., Becher, A., & Pike, K. Rhetoric: Discovery and change. New York: Harcourt Brace & World, Inc., 1970.

# MERGING POLICY AND RESEARCH INTERESTS: A CASE FOR MUTUAL NEEDS

Joan L. Herman

## Introduction

Social science research dollars are dwindling and the outlook for sponsored research grows dimmer on a daily basis. But while the picture is bleak, their may be a faint light on the horizon. The continuing need for and commitment to evaluation research may brighten some of our futures.

Funds allocated for evaluation and policy studies have increased dramatically in the last decade, and while such escalation is unlikely to continue, available funds may hold their own--a marked contrast to the outlook for other social science research. The evaluation funds currently tied to bloc grants, for instance, are hardly insignificant, and the emphasis on local rather than federal program evaluation increases their appeal. Can educational research find a home, health, and happiness with these available dollars? Perhaps. Certainly some compromises will have to be made, but evaluation studies can serve some mutual needs of instructional researchers and of policy makers, and the merger can benefit both parties.

Evaluations, after all, can be conceived as hypothesis-testing ventures. That is, consider the proposition that many special programs, especially school reform efforts, are social experiments which, among other things, attempt to translate research ideas into practice to achieve particular outcomes. For example, California's School Improvement Program and its predecessor, Early Childhood Education, as well as many federal educational programs, are based on

a number of premises about what factors contribute to and foster school effectiveness and student achievement; e.g., the efficacy of parent involvement, systematic planning and evaluation, lower adult-student ratios, individualized instruction, etc. More straightforward examples are the Followthrough programs, which are based on fairly specific models of how instruction ought to occur.

Given the perspective that educational programs embody, or at least imply, particular treatments, then the task of evaluation is to test the hypothesis that the specified treatment is, in fact, associated with the desired outcomes. The applications of research methodologies and notions of operationalizing and measuring the independent variables as well as the dependent variables are obvious here, as are the potential relationships between legitimate evaluation questions and research questions. While evaluation conducted in a real-world setting may be sloppier than work conducted in more controlled research environments, the principles are largely the same.

Obviously, if you want to know whether a treatment works, or whether an independent variable has particular effects, sound research design suggests that you first define the treatment, and then make sure that it in fact occurs. You can't evaluate the effects of an empty set nor draw inferences about the results of an absent treatment. We know this--it's obvious--but evaluations often miss this essential point. Too many evaluations try to answer the question "Does the program work?" without first asking "Was there a program?" This practice may occur because of client unsophistication in research and evaluation design and lack of interest in program processes. Program managers and operators ask simple questions and want simple

answers. But they can be convinced of the need for more. For instance, in our early experiences in evaluating California's Early Childhood Education Program, the funders initially were interested only in outcome data. However, we at CSE held firm on the needs for process data and were permitted to proceed as we desired.

So, evaluation is, and ought to be, concerned with school process as well as outcome, and for programs aimed at student achievement, it's not hard to bring evaluation studies into the classroom. That is, if you agree that student achievement is principally a function of what teachers and students do in the classroom, it is easy to build the case for why evaluation ought to look at instructional practice.

Where does this take us? In place of the single basic question "Is the program effective?" sound evaluation will ask:

1. What is the treatment implied by the program?
2. To what extent is the treatment implemented?
3. What are the outcomes of the program?

If the answer to question two is positive and there is a demonstrable treatment, then the outcome data give a valid answer to the initial basic question--i.e., does the program work? But aligning answers to questions two and three provides food for instructional research and asks "what are the effects of the treatment and to what extent do the independent variables affect the dependent variables of interest?" Such process-product research has been with us for some time, with a somewhat checkered history, but prior specification and newer causal modeling techniques can increase its power. Let me provide an example of how evaluation provides opportunities for instructional research. The example is imperfect, but does demonstrate how educational

evaluation can contribute to our knowledge base.

An Example: A Study of Individualized Instruction

The Center for the Study of Evaluation conducted a study of California's Early Childhood Education (ECE) Program (Baker, 1976). This program, according to the then current section 6445 of the California Education Code, provided that early childhood education be designed, among other things, to assure:

(a) a comprehensive restructuring of primary education in California, kindergarten through third grade, to more fully meet the unique needs, talents, interests, and abilities of each child.

(b) the cooperation and participation of parents in the educational program to the end that the total community is involved in the development of the program.

(c) that pupils participating will develop an increased competency in the skills necessary to the successful achievement in later school subjects such as reading, language, and mathematics.

Thus, one could reasonably infer that the ECE program was intended to foster student achievement through, among other things, more individualized instruction and community involvement, i.e., program processes were means to an end. Alternatively one might take a more coordinate view that higher student achievement and more individualized instruction for students were equally valued program outcomes. In any case, it was clear that individualized instruction was an important component of ECE and, consequently, CSE's study collected a range of questionnaire and observation data about how teachers implement individualized programs. In addition, because ECE claimed an interest in the "whole child," criterion-referenced tests of reading and mathematics as well as measures of students' attitudes were also collected. The data set allowed us to look at how individualized

instruction operates in classroom practice, and to examine its affect on student outcomes.

The secondary analyses posited a model to explain expected interrelationships between attributes of individualized instruction and their direct and indirect effects on second grade students' achievement and attitudes. The underlying assumption of the hypothesized model was that classrooms which were more individualized in terms of instructional decisionmaking, activities, and teacher-student interactions would provide more appropriate instruction for students and thus result in improved student achievement and attitudes. It was also assumed that if an individualized program was implemented systematically, the degree of individualization in decisionmaking, activities, and interaction with the teacher would be interrelated. That is, individualized decisionmaking would lead to different prescriptions and different kinds of instructional activities for different students, based on assessment of need. Further, having more activities going on in the classroom should allow the teacher to interact on a more individualized basis with students working on any single activity. Aides and volunteers were conceived as serving a support function in the classroom, i.e., their presence allowed teachers to manage the individualization effort. Socio-economic status was also included as a control in the model as well as to examine its effects. Path analysis was used to test the direct and indirect effects predicted by the model.

## The Data Set

The data used for the analyses were a subset of those used for the main ECE evaluation. A stratified random sample of 256 schools

were selected for participation in the main study to represent three levels of ECE status (0, 2, and 3 years) and four levels of compensatory education funding (receipt and non-receipt of federal and/or state level funding). From within these 256 schools, 72 were selected for more intensive study. Two second-grade and two third-grade classrooms within the 72 schools were randomly chosen for data collection. The study of individualized instruction was limited to data collected in second-grade classrooms (n = 90).

Multiple data sources were available for composing the independent variables, including teacher questionnaire and interview responses and brief (20 minutes) classroom observations during both reading and mathematics instruction.

Degree of individualization in decisionmaking. Three variables were included to operationalize the degree of individualization in decisionmaking: sources used for placement, frequency of progress monitoring, and frequency of remediation and/or corrective actions derived from progress-monitoring.

Degree of individualization by activity. During classroom observations in both reading and mathematics, observers recorded the number of different activities occurring in the classroom. An activity was defined as a unique student assignment, often related to materials in use. For example, if some students were working on one workbook assignment while others were reading a text this would reflect two activities. However, if all students were working in the same workbook, but on three different assignments within the workbook, then this occurance would be recorded as three activities.

Degree of individualization in teacher-student interactions. Teachers responded to questionnaire items asking what percentage of instructional time they typically spent in whole class, large group, small group, and individual instruction during reading and mathematics.

Number of aides and volunteers. Teachers indicated, during interviews, how many aides and/or volunteers assisted them during reading and during mathematics instruction; classroom observations also recorded the presence of aides and volunteers.

Socio-economic status. SES was a school level index provided by the California State Department of Education. This three-point index was based on parent's occupation; three was the highest rating.

Achievement measures. Criterion-referenced tests of reading and mathematics were constructed specifically for the main study. Objectives were those agreed upon as central in the primary grade curriculum; their importance was verified by teacher questionnaire responses. Because individualized instruction is supposed to permit all students to learn basic objectives, both level of achievement and classroom variation in achievement were included as variables of interest.

Student attitudes. Items dealing with students' attitudes toward reading and toward mathematics were adapted from the School Sentiment Index (IOX, 1972). Three items were included in the reading scale, and four items were included on the mathematics scale.

## Results

Path analysis was used to examine the significance of the hypothesized relations for both reading and mathematics. To examine

whether the patterns of relationships were the same for higher and lower SES groups, interaction terms were added to the model.

In reading, as predicted, socio-economic status was positively related to achievement, and whole class instruction, for higher SES groups, was negatively related. However, teacher consulting with students and providing one measure of corrective action were negatively related to achievement, and whole class instruction was associated with greater achievement for lower SES classrooms; these latter findings are in direct contradiction to the concept of individualized instruction. With respect to attitudes toward reading, consulting with students was a negative predictor, and presence of more adults in lower SES classrooms was associated with more positive student attitudes.

As expected, SES also was positively related to student performance in mathematics and whole class instruction was negatively related to achievement. An unexpected finding was that the number of adults was negatively related to achievement in lower SES classrooms. Grouping was found to contribute both to more variation in classroom achievement and to less positive attitudes toward mathematics. For lower SES classrooms, more activities and a teacher's use of corrective action were positively related to attitudes towards mathematics.

These results showed some support for individualized instruction. The negative effect of whole class instruction on mathematics achievement, and on reading achievment for students from higher socio-economic status backgrounds, supported one of the major premises of individualized instruction, i.e., providing only one

instructional treatment is inappropriate to the needs of many students within a class. The relationship between whole class instruction and within-class variation in reading achievement was similarly supportive of the premise.

However, these data suggest, as we might guess, that providing alternative materials and more individualized instructional settings does not solve the problem of student learning. The relationship between grouping and within-class variation contradicts the theoretical rhetoric, i.e., certain strategies associated with individualized instruction may magnify differences among learners. The relationship between individualized instruction variables and the reading achievement for students from lower socio-economic backgrounds is particularly discouraging. The results may imply that, in current practice, providing more individualized strategies may not be appropriate for these students, a conclusion supported by research in teacher behavior and direct instruction models (Rosenshine, 1977; Bennett, 1976).

The relationship between process variables and student attitudes was similarly contradictory. The results in mathematics suggest that while grouping practices associated with an individualized approach may be detrimental to student attitudes for lower SES classrooms, other processes which facilitate instructional responsiveness and variety appear to enhance their attitudes.

Despite the contradictions in the data and lack of relationship between most of the process variables and student outcomes, one conclusion is clear. One cannot assume that classrooms which appear more individualized are, in fact, more facilitative environments for

students than are classrooms which appear less individualized. Processes underlying students' learning are more complicated than surface appearances regarding teachers' use of assessment or provisions for instructional alternatives.

This obvious conclusion has serious implications for evaluation policies at various levels. For example, some SEAs and LEAs evaluate schools on the extent to which they provide individualized instruction, based on brief classroom observations and interviews (Herman & Hanelin, 1977). Certainly the validity of using such ratings to help assess school quality is suspect, a serious concern given the potential impact of such practices on funding allocations.

The mixed findings regarding the effects of individualized instruction may be a function of the fact that classroom practice does not mirror the theory espoused by advocates of classroom individualization; that is, teachers do not truly implement individualized programs. Although teachers may look like they are individualizing instruction in terms of assessing students' progress and providing instructional alternatives, etc., the results suggest that these actions are unrelated and that the link between diagnosis and prescription is missing--a finding that again points to a need for looking below the surface before passing evaluative judgment.

What insights does the example elaborated in this paper provide --beyond the not so astounding conclusion that molar variables often leave more questions than answers? I think it supports a few conclusions.

1. Evaluation of school programs, and particularly those programs that focus on student achievement, benefit greatly from an instructional research perspective. Evaluations often ask questions that are too simple; and simple answers that ignore school and classroom processes are likely to be invalid. Good information requires a knowledge of what goes on in classrooms and schools.

2. Evaluation can provide good data for instructional research and it contributes to our knowledge base. The example reported here perhaps does not reveal too much about individualized instruction, but the findings are consistent with other findings in the field: for example, the results with regard to lower SES classrooms support much of the work of the direct instruction advocates (see, for example, Stallings et al, 1977; Soar, 1973; Rosenshine, 1977); that individualized instruction, in the example, tended to magnify differences between learners is consistent with some of the research from Wisconsin (personal correspondence, 1978). Convergence of data from several studies certainly adds strength to the knowledge base.

3. Evaluation studies which may necessarily have to look at more molar variables can nonetheless support the need for more fine-grained research. Similar findings across studies, in particular, provide a good rationale for why deeper understandings are necessary--and for the compelling need for instructional research.

Is there a case for mutual needs? I think so. Evaluation certainly requires the research perspective, and we can benefit from the need as well as contribute to informed, rather than simplistic, public policy.

REFERENCES

Baker, E.L. The evaluation of the California early education program. Los Angeles, California: Center for the Study of Evaluation, 1976.

Bennett, N. Teaching styles and pupil progress. London: Open Books, 1976.

Herman, J., & Hanelin, S. Audit of the monitor and review process. Volume II in E. Baker, The evaluation of the California early childhood education program. Los Angeles, California: Center for the Study of Evaluation, 1977.

IOX. School sentiment index. Los Angeles, California, 1972.

Personal correspondence, 1978.

Rosenshine, B. Academic engaged time, content covered, and direct instruction. Paper presented at the annual meeting of the American Educational Research Association, New York: April, 1977.

Soar, R.S. Follow through classroom process measurement and pupil growth (1970-71): Final report. Gainesville: College of Education, University of Florida, 1973.

Stallings, J., Gory, R., Fairweather, J., & Needles, M. Early childhood education classroom evaluation. Menlo Park, CA: SRI International, 1977.

# HITCHHIKING ON FAST-MOVING POLICY RESEARCH: A CRITIQUE

Don Dorr-Bremme

## Introduction

This paper addresses two premises: (1) in a time of fiscal restraint, government funds are likely to be more available for policy studies and program evaluations while grants for basic research on teaching and learning become less available; (2) researchers may try to "hitchhike" on government-sponsored policy studies--to use them as vehicles for doing research on instruction. These two premises raise the question of whether research on teaching and learning can be built into government-funded program evaluations and policy studies and, if so, under what circumstances?

This paper addresses these questions through a case study. It tells the story of how I tried to hitch a ride with some research of my own on a policy study that happened to be passing by, and elaborates the circumstances under which one can make good instructional research sense out of government policy dollars.

First I'll provide a description of the vehicle--a federally funded policy study--and the questions which drove it. Then I'll talk about the hitchhiker: me with my small piece of research that seemed to be going in exactly the same direction as the vehicle. Finally, I'll report how the ride seems to be going and why.

## The Vehicle: A Federally Funded Study of Testing and Test Use

The vehicle passing by was a piece of policy research: a national study to inform federal, and especially state- and local-level policy, on achievement testing.

Student achievement testing in the nation's schools has become a vast enterprise, and both the amount and variety of testing continue to grow. Across the country, more than 40 states have now mandated tests of minimum competency for school children. (Some states require the tests for promotion and graduation; others, merely to check students' basic educational needs at milestones in their school careers.) The testing of student achievement remains a primary way of meeting the evaluation requirements that federal and state programs include. School districts have expanded their testing programs: many have developed or purchased assessment tools to monitor student's progress along district-mandated continua of skills or objectives. Teachers, meanwhile, develop and administer their own tests as well as other tests that come with curriculum materials they use. All in all, hundreds of millions of dollars in public monies are expended annually on testing.* Amidst this testing "boom" various types of tests, and testing in general, have become controversial. (The National Education Association and the American Federation of Teachers, for example, have taken official and somewhat opposing positions on testing.)

However, there has been very little research to inform debate or decisions about testing. Just how much testing is going on, how much it costs, and what specific benefits derive from particular types of

---

* It has been estimated, for example, that in 1976 standardized testing in the elementary grades alone cost well over a quarter of a billion dollars (EDC News, 1977). A study done by Lyon (1978) found that budgets in school districts' evaluation and testing units range between $2,000 and $4,000,000 annually. These estimates, however, omit substantial indirect costs: e.g., teachers' and administrators' time spent in preparation for testing, test administration, etc.

tests and testing programs under what circumstances, all remain largely unknown. The policy study on which I attempted to hitch a ride, then, since it would focus on test-policy issues, had to address several broad questions:

1. With what frequency and distribution are particular types of tests given in the nation's schools?

2. In what ways do particular types of tests and testing programs impact upon schools and those within them?

   a) through their very presence, required or recommended?

   b) through educators' utilization of their results?

3. What factors influence

   a) where and how much particular types of testing are done?

   b) the ways that tests and their scores impact on schools and those within them (students, teachers, etc.)?

4. What are the costs--direct and indirect dollar costs; opportunity/ educational and psychological costs--of different types of tests and testing programs?

Of course, these research questions generate information to address policy issues such as: (1) What do we get and what do we trade off when we invest our testing dollars in this, that, or some other test or assessment program? or (2) If we want to accomplish "X," what's the best investment of our testing dollars?

These concerns and questions drove the policy study, which took shape as a three-year effort. The first year would entail planning the design of a national survey of teachers and principals (and some district officials). Exploratory fieldwork was included, along with a literature review and reanalysis of some test-use data CSE had previously collected. The second year would see instrumentation and

fielding of the survey in a national sample of districts (over 100) and schools (ideally two elementary and two high schools per district), for a total of roughly 2100 respondents answering questions about testing and test use in the basic skills areas (reading/English and math). While data from this survey were being analyzed in the second year, planning for year three would begin, again including a good deal of on-site fieldwork. Finally, in the third year, ethnographic studies in three or four schools, as well as less intensive fieldwork in other sites, would be carried out to follow up on the survey and, especially, to get a close-up look at testing costs.

This, then, was the vehicle--the policy study of testing and test use--and the questions which drove it. One more point is worth noting about this before moving on. Although the words test and testing recur above, the study was equally concerned from the outset with other, less formal means of assessment: teachers' observations and daily interactions and the information they yield, routine classwork and homework, etc.

## The Hitchhiker and His Study of Teachers' Thinking and Decision Making

I turn now to the hitchhiker--myself--and the small piece of research I carried in my pack. First, it is important to know that I have an interest in what can be construed broadly as social cognition. More specifically, I'm concerned with understanding the everyday knowledge (c.f., Sudnow, 1968), the "background understandings" (Garfinkel, 1967), and practical reasoning (c.f., Cook-Gumperz, 1975) which are presumed to underlie the practical affairs of members of particular social groups. Put another way, in

my work I attempt to describe the "system of standards for perceiving, believing, acting, and evaluating" (Goodenough, 1970, 1971) or the cognitive "rules, maps, and plans" (Spradley,, 1972) evident in participants' routine actions and talk, and how these are functionally relevant (Erickson, 1978; Erickson & Schultz, 1977) to the performance of particular educational events (e.g., lessons, morning circle time in elementary classes, etc.)*

These interests are basically psychological in nature, but as the language and citations may indicate, interests that I pursue through the adjacent and sometimes complementary theories of cognitive anthropology, (Goodenough, 1964, 1971, 1975; Tyler, 1969; Wallace, 1970), ethnomethodology (Cicourel, 1974; Garfinkel, 1967; Mehan & Wood, 1975; Turner, 1974), and sociolinguistics (Hymes, 1972, 1974) are merged in what Hugh Mehan has called "constitutive ethnography" and Fred Erickson has termed "microethnography."

It may by how be evident why the vehicle stopped to let me on as a hitchhiker. I was not the designer of the policy study, but I was clearly interested in how things get done in schools and classrooms, in how people routinely think and act, and what influences how they think and act. Furthermore, I had the fieldwork training and experience that would be needed recurrently throughout the policy study.

I, in turn, was interested in climbing aboard, for it seemed to me that my research interests were virtually congruent with the policy study's questions. And there was plenty of fieldwork in the project

---

* See for example, Bremme and Erickson, 1977; Dorr-Bremme, 1981(b).

through which to pursue those interests following the methodological canons of my field. More specifically, the policy study would seek to determine what types of tests and other means of assessment educators in schools use in making particular instructional decisions and how each "counts" as a given decision is made. That is essentially a cognitive issue, which I could restate as: "What knowledge and processes of reasoning do teachers routinely employ as they make particular instructional decisions?" Or again, "What system of standards for believing, perceiving, evaluating, and acting are routinely in use as teachers make particular instructional decisions?" I would summarize those standards or cognitive "rules" in a flow chart such as the figure:

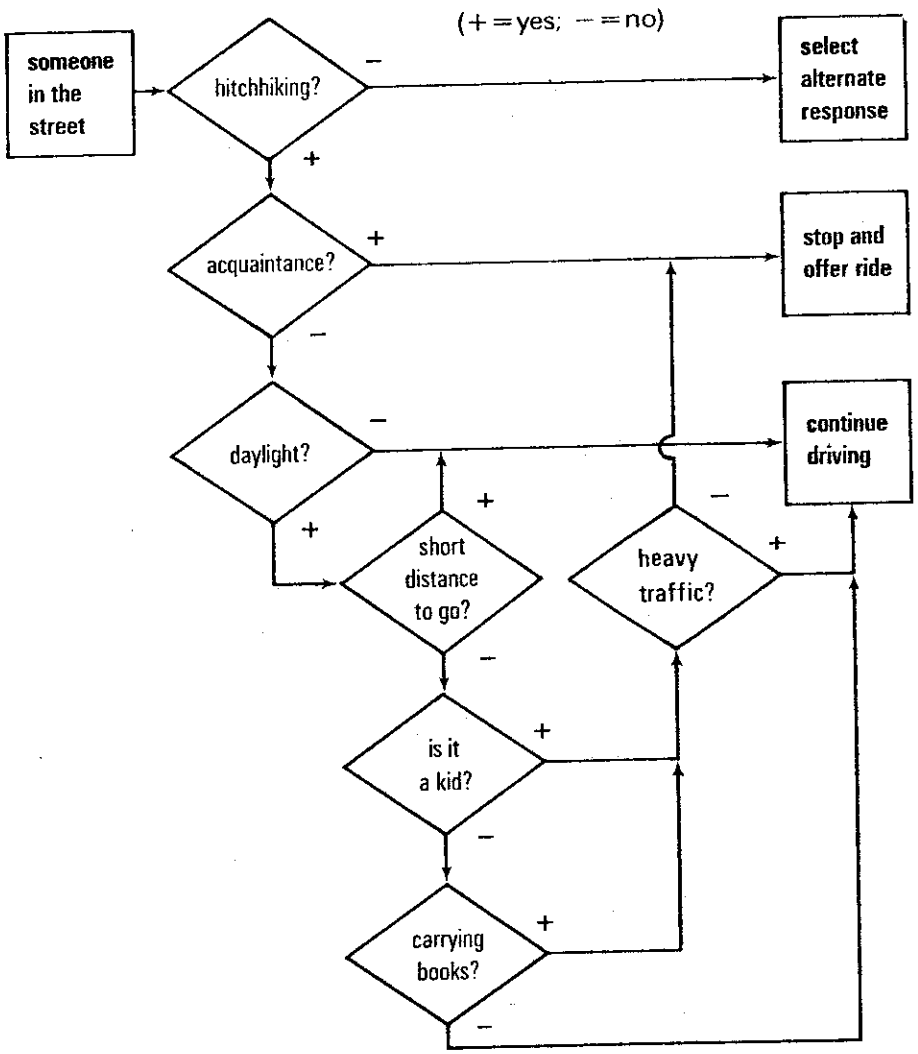
---

Insert Figure 1 About Here

---

The flow chart here is, of course, more appropriate to the metaphor in this paper than to the study I wanted to do. But in that study, times of day, kinds of people, driving conditions, etc., that appear in this chart would be replaced with kinds of tests and other assessment means, types of students, and instructional options that inform school decisions. Specifically, ethnographic work would reveal how these elements figured in a particular instructional decision--where they fit in the decision-making process, how much each "counted" in the decision, etc. The study that led to these findings would also tap teachers' "background understandings"--their opinions of testing in general and of the worth of different types of tests and other assessment results. These issues which would provide context for the

Figure 1




Reproduced from J.P. Spradley (Ed.), Culture and cognition: Rules, maps, and plans, 1972, p. 32. By permission of the author.

findings summarized in my flow chart were also concerns of the policy study.

The data needs of my piece of research, then, would be merely a subset of the policy study. It would only be a matter of asking an extra question or two here and there in the fieldwork, looking a bit more closely at certain events we'd be looking at anyway,, and I'd have done a reasonable piece of research on an aspect of teacher thinking and decision making that would be relevant to educational practice. This work wouldn't be as fine-grained as usual constitutive or micro-ethnography, but it would use the same principles and it would be supplemented with a great deal of information from the larger policy study in which it was embedded.

I didn't decide, as I climbed aboard the policy vehicle, just what instructional decision I would focus my research upon. I had in mind looking at the ways in which classroom teachers decide a student needs extra help. But I could hold that as a tentative choice pending the results of the first exploratory fieldwork.

In general, then, this hitchhiking would proceed as follows: (1) during the first-year exploratory fieldwork, the goal would be to get a general map of the kinds of ways teachers had of knowing about students' performance and progress: the kinds of decisions they made (as they saw them), and how the two seemed to relate. I would be sure to include questions on "special help" decisions in this work which would serve my substudy as background, and which we had to do anyway for the policy study. (All this almost happened. Some 80 interviews were conducted with teachers, specialists, principals, and counselors, department chairpeople, and others. Documents were collected as well

as copies of tests, and we tried to "look ethnographically" in the limited time we had in three schools in each of three districts.) (2) In the survey, I wanted to center inquiry on the functional relevance of particular types of assessment results--tests and others--for particular kinds of decisions. This would be extremely important for the policy study, as I saw it,, and it would (a) help me decide what sort of decision to focus my hitchhiking effort upon, as well as (b) provide some broad data on that decision to contextualize the fieldwork. (3) Then, in the fieldwork planning for the third year, and in the third year ethnographies themselves, I would begin in earnest to do the extra little pieces of work which would give me my hitchhiking research on teacher thinking and decision making.

To this point, I have reviewed the policy study (the vehicle), how I hitchhiked on it and where I was sitting, and what I was carrying in my valise (the decision-making study.)

The Ride: Results of the Study and Some Caveats for Hitchhikers

In a short description of the ride I'd have to say that the questions and concerns driving the policy study keep sneaking things from the hitchhikers--time, money, and other resources--that they need to be methodologically robust. Both are still on the way to their destination. (Planning for the third year is concluding; survey results have had preliminary analyses.) They may well make it to where they are going, but the driver is getting awfully large, there in the front seat: room for the hitchhiker seems to be shrinking.

Less whimsically, data to set up the substudy as I intended it--data which ultimately will be a part of it if it does not fall off the vehicle--has begun to come in. Intuitively, the data seem solid

to me. But the documentation is weaker than I had hoped for, less systematic than I feel good about. This has resulted, I think, from certain generic features of policy studies and circumstances in this particular study. I will describe these with examples, and underscore the lessons I think they teach. But first let me say something about the "findings," hypotheses really at this point, derived from the year one fieldwork, year two fieldwork to date, and some preliminary examination of the survey questionnaire data.

The picture that is emerging of how teachers routinely think about and handle student assessment is quite consonant with the picture Eliot Friedsen paints of the "clinical mentality" in his sociology of applied knowledtge, Profession of Medicine, (1973). In unfortunately loaded language, Friedsen says,

> The practitioner is a fairly crude pragmatist ... prone in time to trust his own accumulation of personal first-hand experience in preference to abstract principles or "book knowledge", particularly in assessing and managing those aspects of his work that cannot be treated routinely. As Sharif and Levinson noted in the case of psychiatrists in training, "The dangers of intellectualizing and "book learning" are stressed. The highest value is placed on emotional experience, on widening the range of "gut response" as a means of understanding what is going on in oneself and in the patient". This represents a certain subjectivism in his approach.

Further on, Friedsen adds: "Thus, a rather thoroughgoing particularism, a kind of ontological and epistemological individualism is characteristic of the clinician." Shed of its pejorative language, Friedsen's description aptly describes elements of structure in the teacher's thinking about assessment.*

---

* The data which support the generalizations offered here may be found in Dorr-Bremme, et. al., 1980; Dorr-Bremme, 1981.

1. Teachers' Thinking is Pragmatic and Experience Oriented.

Our preliminary results show that the tests teachers give most often, devote most class time to, and rely upon most heavily have three qualities:

- face validity--in the teacher's eyes, they match with what was actually taught
- immediacy--they are immediately available, may be given discretionarily, and the results are immediately available
- consonance with teacher's routine practical tasks--placement test for placement; unit tests for unit grades; tests labeled diagnostic for diagnosis, etc.

Furthermore, clinical experience overrides, in many cases, test results. Almost invariably, the teachers we spoke with said (without explicit elicitation from us) that they might use a placement test, e.g., to group children for reading; but whether the placement was correct was determined on the basis of the teachers' judgement of the child's work. Similarly, according to teachers, some children are "good test-takers;" others choke or may just not try; they may be having a bad day when the test is given, and so on.

2. Teachers' Reasoning About Test Results and Students' Performance in Particular.

The last point above illustrates one form of particularism. Another emerged in the regularity of teachers reponding "it depends..." when we asked them how a decision would be made. Family circumstances, classroom social behavior, other teachers' remarks and opinions, oral performance, patterns in routine classwork, the appearance of interest or motivation, together with a wide variety of types of test scores are available to teachers and figure in most decisions they make about their students. Similar evidence, albeit organized

differently, also figured in teachers' assessments of their own performance--in their judgments of their effectiveness. It appeared--though it was difficult to tell certainly in interviews--that this information would be weighted differently, in making the same type of decision on different occasions or with different students.

3. Teachers' Reasoning Processes Appear to be Rational when Viewed within a Clinical Framework.

Studies of teacher decision making offer conflicting results reagarding how "rational" or valid teachers' clinical decisions are. Vinsonhaler (1980, 1981), for instance, has demonstrated that the same reading specialist often diagnoses an individual student's "case" differently on two different occasions separated in time; from specialist to specialist, there also seems to be little reliability in the diagnosis of a case. On the other hand, similar "policy capturing" studies using case simulations reported by Shavelson and colleagues (e.g., Borko, Cone, Russo, & Shavelson, 1979) indicate that teachers can readily recognize and usually tend to employ information from more reliable sources. Other work (Pedulla, Airasian, & Madaus, 1980) shows that teachers typically predict students' scores on standardized tests quite accurately. Most research which has attended to the practical circumstances teachers confront as they make instructional decisions tends to depict them as fairly reasonable. (For a comprehensive review, see Shavelson & Sterns, in press).

Work on teacher thinking and decision making conducted thus far within the test use policy study tends to support and extend this view of the teacher. Exploratory findings, for example, suggest that classroom practitioners rarely rely on one source of information in

making a given instructional decision (Dorr-Bremme, et al, 1980; Dorr-Bremme, 1981). Like the scientist, the teacher looks for replication of results generated through different types of measures:tests of different types, class and homework assignments that embody different performance conditions, etc. This makes good sense in light of recent work in human cognition (e.g., Griffin, Cole & Newman, in press), and language (Bloom & Gumperz 1972; Phillips 1972; Gumperz & Hernandez-Chavez, 1972) which shows that the demonstration of competence in performance varies with context.

This is only a brief overview of findings to date, but it should indicate that instructional research embedded within a policy study can contribute to our collective understanding of teachers' thinking and decision making. As things stand, however, our findings are not based on evidence as solid and as systematically obtained as most researchers--even ethnographers, who are wrongly reputed to be less concerned with "hard" data--would like. In the end, it is quite possible that our "findings" will remain provocative impressions. They may not attain the status of research results in the study I had hoped to do. This, as I noted earlier, is largely because the policy study has consumed more resources I thought would be readily available for the hitchhiking research.

I do not think that this is a peculiarity of the policy study in which I am engaged. Policy studies are, I believe, a ravenous species. They are (to switch metaphors) generally subject to centrifugal forces. In the remainder of this paper, I want to indicate where these centrifugal forces come from and then illustrate with some examples how they work to the disadvantage of a "hitchhiker." I will also

offer one or two caveats, drawn from my experience, for researchers who are considering hitchhiking on policy studies with some research of their own.

The Nature of Policy Studies

For a number of reasons, policy studies have a tendency to "want to be larger" than one originally anticipates they will be. For example:

1. Funding is most often offered for research which is germane to national or statewide policy. The results, then, must be generalizable to the nation or to a state or to other units which embody diverse program settings across a very large number of potentially program-relevant (and policy-relevant) variables. Specifying a small number of these, a priori, as sampling variables, is often difficult. This has at least two implications for research time and other resources:

   (a) There is a tendency to be inclusive rather than exclusive: to sample along more rather than fewer variables and thus to expand a sample which was rather large to begin with in order to obtain a sufficient n in each of many sampling cells. This tendency has ramifications throughout the study. A larger sample demands more time for actually drawing the sample, for contacting the sites to be surveyed and gaining their cooperation, for conducting the survey and managing the data, etc.

   (b) There is rarely one single best way to draw a national or statewide sample for a given policy study. Alternative sampling plans offer the possibility of different--but equally policy-relevant--analyses. Examining these alternatives requires time and review that may exceed original projections, especially when project staff, representatives of the funding agency, reviewers, and consultants disagree on the merits of the different sampling plans and the analyses they facilitate.

2. Policy research usually requires that data be collected on a wide range of dependent variables. Previous studies have identified a large set of generic factors that can influence any program's outcomes. (A partial list includes leadership, participants' feelings of program "ownership," the nature of the informal social structure of the implementing institution, participants "sense of efficacy," the number of other programs extant simultatneously at a site, participants' "angle of vision" or perspective on the program, and the frequency and quality of staff development and other support services.) Particular programs, of course, are susceptible to the influence of other variables in addition to such generic ones. Thus, the number of domains relevant for inclusion in study instruments is large. Furthermore, a nationwide or (to a lesser extent) a statewide sample entails consider-

able diversity in local conditions and practices, as well as in local terminology for describing conditions and practices. To take this diversity into account, questions in research instruments must often be long and complex, rather than simple and succinct.

3. Policy research usually has clearly evident political implications. As a consequence, policy makers and stakeholders in programs want to be sure their interests and perspectives are reflected in the research design and represented in the research instruments. The funding agency may have one or multiple agendas for the study, some of which are politically motivated. Responding to the concerns and wiches of various interest groups is often unavoidable. In many cases, their involvement in research planning is critical to the success of the study. (For instance, the support and/or endorsements of certain groups may facilitate local agencies' participation in the study, promote higher rates of return of survey questionnaires, etc. Support and endorsements may only come in exchange for a voice in research planning.) Involving various interested groups in planning and/or reviewing the research design and instruments consumes further time, energy, and research funds. Including questions that interest groups suggest adds to the length of research instruments.

4. Policy research usually must address multiple audiences. Political considerations aside, these audiences do not always share the same concerns and questions. They may need and expect very different kinds of information. (The study of testing and test use, for example, is expected to provide information to policy makers in federal, state, and local education agencies. Their interests and information needs are not identical, however.) Balancing these competing information needs is another centrifugal force with which policy research must contend. Again, there is a tendency to resolve the problem of multiple information needs by making the research more inclusive.

5. When policy research is undertaken at a large scale (which it most often is when government-funded), a team effort is most often required. Members of the research team may not agree on the best resolution of research issues. Compromising among researchers often results in expanding the scope of inquiry.

6. Today, requests for policy research frequently call for inclusion of fieldwork of some kind--ethnograpy, case studies, etc. On-site work generates centrifugal forces: the closer one looks and the longer one looks on site, the more issues seem to deserve investigation. To accomodate these issues, research instruments and research time call out for expansion.

For a variety of reasons, then, there is a tendency for policy research to spin out beyond its projected boundaries. Even under the best of circumstances, coping with these centrifugal tendencies--

intelligently restraining the expansion of a policy study--requires time, staff energy, and research dollars. And it seems that when a substudy is along for the ride, it is that substudy that suffers first and most from the centrifugal forces.

Some Examples

My own experience provides several examples of how this happens. Earlier, I explained that my hitchhiking plan included adding a question or two to the exploratory fieldwork in the test use study's first year. I had also planned to use that initial fieldwork as background for my teacher decision-making study. To assure adequate time for this, I had hoped to conduct the fieldwork in states which had different testing requirements (a necessity for the policy study), and which were geographically close to California. Money saved in travel expenses could then go toward making the exploratory work more systematic and rigorous. The funding agency, however, urged from the outset that the study be national in scope. Responding to that suggestion was in the best interest of good and continued relation between the agency and our research center. As a consequence, dollars were consumed in travel; time on site was reduced to the minimum.

Moreover, long distance negotiations abut our site visits sometimes resulted in unavoidable deviations from our research plan. We spoke on several occasions by phone with key personnel in each district we planned to visit. We also exchanged several letters with them. During these contacts, we stressed the importance of our speaking with each interviewee for forty to forty-five minutes, and suggested a number of measures we were willing to take in order to arrange for that. Each district acknowledged our request and assured us that

they would respond to it. Nevertheless, once we arrived on site, we found in several schools that interviews were scheduled for shorter periods than we had requested. Thus, we had to cut back our interview questions: those critical to the policy study remained; those additional one or two questions most germane to the decision- making study had to be cut. And with time on site already at a minimum, there was no possibility of returning later to gather information lost. The exploratory fieldwork yielded a wealth of information that proved extremely useful for designing the survey research which followed. (That was its primary function.) But with abbreviated interviews in some schools and time on site focused on policy study issues, the results were too asymetrically gathered to count, analyze, and include as background in the substudy on teacher decisionmaking.

Later on in the project, other centrifugal forces applied pressure on the substudy. In preparing for the national survey, two complete sampling plans were developed and discarded before arriving at a third and final one. This was not due to any incompetence of the plans' designers. (All of them were highly skilled with considerable experience in sampling for large-scale survey research.) The difficulty was simply that each plan posited a slightly different set of variables (and survey analyses) as most important. In successive reviews, the advocates of each plan and the consulting reviewers argued effectively for the value of different sampling approaches. Each of these different points of view had merits and drawbacks within the context of the study. Resolving these and maximizing the analyses the sample would permit required a good deal of time and research dollars. All of this, of course, strengthened the policy study. But

it delayed the start of the next phase of fieldwork, to be undertaken once sampling for the survey was underway. Once again, reduced time meant restricting the focus of fieldwork to issues most germane to the policy study, and passing over the extra little bit of work which would have fleshed out the research on teacher decision-making.

In the survey itself, I had hoped to focus the inquiry on the functional relevance of different types of assessment results--test scores and other student products--on teachers' decisions. A number of pressures came to bear, however, which minimized the attention which could be given to this domain of inquiry. First, a large number of domains had to be covered on the questionnaire, and questions on each domain grew longer in order to take into account the diversity of practices across the nation. Project officers in the funding agency emphasized that certain areas were of critical importance, given agency information needs. Representatives of teachers' organizations, commercial test publishers, and others involved (for political reasons) in the review process called for inclusion of certain types of questions. Project staff members argued effectively for different emphases. To avoid an extensive burden on questionnaire respondents, difficult choices were necessary. In the end, collective thinking and conflicting demands assured comprehensiveness in the survey. Questions on how teachers used particular types of assessment results were included, but only as one of several domains of inquiry important to the policy study.

These were only some of the ways in which the centrifugal forces inherent in policy research compressed the decision-making study I planned to conduct. But these examples should be sufficient to illustrate the general process.

I should add that this was not my first experience with policy research: I was aware that each phase and task of the project would tend to expand in scope and complexity, and had anticipated that tendency as I planned my hitchhiking. But there seemed to be so much overlap between the issues of the policy study and the issues of my decision-making study that it seemed I would be able to address both even though the former were likely to grow as work continued.

Caveats

The observations I have made here about the nature of policy research are unlikely to be new to those experienced in such work; they routinely experience the centrifugal forces I have described. But as grants for research become scarce, scholars new to the policy research road will be stepping onto it in greater numbers in search of vehicles on which they can hitchhike with their own instructionally relevant studies. For them, I offer the following words of advice. It is probably best not to hitchhike with strangers. That is:

1. Your research on instruction is likely to lose weight to the extent that its questions are not exactly congruent both with the research questions of the policy study and with the actual items in research instruments.

2. Get a ride for your study on the research methods that are absolutely central to the policy study, no matter how extensive some other methods may appear to be in the research design.

The substudy research discussed in this paper was accomplished effectively largely because it followed these caveats. Large-scale policy research, when it is done reasonably, provides very little room for naive hitchhikers.

BIBLIOGRAPHY

Blom, J.P., & Gumperz, J.J. Social meaning in linguistic structures: Code-switching n Norway. In J.J. Gumperz and D. Hymes (Eds.), Directions in sociolinguistics: The ethnography of communication. New York: Holt, Rinehart, & Winston, 1972.

Borko, H., Cone, R., Russo, N.A., & Shavelosn, R.J. Teachers' decision making. In P.L. Peterson & H.J Wahlberg (Eds.), Research on teaching: Concepts, findings, and implications. Berkeley, CA: McCutchan, 1979.

Bremme, D.W., & Erickson, F. Relations among verbal and nonverbal classoom behaviors. Theory into Practice, 1977, 16, 153-161.

Cicourel, A.V. Cognitive sociology: Language and meaning in social interaction. New York: Free Press, 1975.

Cook-Gumperz, J. The child as practica reasoner. In M. Sanches and B.G. Blount (Eds.), Sociocultural dimensions of language use. New York: Academic Press, 1975.

Dorr-Bremme, D.W., Burry, J., Lazar-Morrison, C., Moy, R., Polin, L., & Yeh, J. Test use project annual report to the National Institute of Education, (Vols, I & II). Los Angeles: Center for the Study of Evaluation, 1980.

Dorr-Bremme, D.W. Test use project progress report in Phase II planning to the National Institute of Education. Los Angeles: Center for the Study of Evaluation, 1981 (a).

Dorr-Bremme, D.W. Behaving and making sense: Creating social organization in the classroom. Unpublished doctoral dissertation. Harvard Graduate School of Education, 1981 (b).

Erickson, F. On standards of descriptive validity in studies of classroom activity. Paper preented to the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March, 1978.

Erickson, F., & Shultz, J. When is a context? Some issues and methods in the analysis of social competence. Quarterly Newsletter of the Insitute for Comparative Human Development, 1977, 1, 5-10.

Friedson, E. Profession of medicine: A study of the sociology of applied knowledge. New York: Dodd, Mead, 1973.

Garfinkel, H. Studies in ethnomethodology. Enlgewood Cliffs, N.J.: Prentice-Hall, 1967.

Goodenough, W.H. Description and comparison in cultural anthropology. Cambridge, England: Cambridge University, 1970.

Goodenough, W.H. Culture, language and society. (Addison-Wesley Modular Publications Number 7). Reading, MA.: Addision-Wesley, 1971.

Goodenough, W.H. Cultural anthropology and inguistics. In D. Hymes (Ed.), Lanugage in culture and society: A reader in linguistics and anthropology. New York: Harper & Row, 1964.

Goodenough, W.H. Multiculturalism as the normal human experience. Paper presented at the annual meeting of the American Anthropological Association, San Francisco, December, 1975.

Guffin, P., Cole, M., & Newman, D. Locating tasks in psychology and education. In L. Cherry-Wilkerson (Ed.), Discourse processes, (in press).

Gumperz, J.J., & Hernandez-Chavez, E. Bilingualism, bidialectism and classroom interaction. In C.B. Cazden, U.P. John, & D. Hymes (Eds.), Functions of language in the classroom. New York: Teachers College Press, 1972.

Hymes, D. Introduction. In C.B. Cazden, U.P. John, & D. Hymes (Eds.), Functions of lanuguage in the classroom. New York: Teachers College Press, 1972.

Hymes, D. Foundations in sociolinguistics: An ethnographic approach. Philadelphia, PA.: University of Pennsylvania, 1974.

Lyon, C. Evaluation and school districts. Preliminary results reported to the NAtinal Institute of Education. Los Angeles, Ca.: Center for the Study of Evaluation, 1978.

Pedulla, J.J., Airasian, P.W., & Madaus, G.F. Do teachers' ratings and standardized tests results of students yield the same information? American Educational Research Journal, 1980, 17, 303-307.

Phillips, S. Participant structures and communicative competence: Warm Springs children in community and classroom. In C.B. Cazden, U.P. John, & D. Hymes (Eds.), Functions of language in the classroom. New Yrok: Teachers College Press, 1972.

Spradley, J.P. (Ed.) Culture and cognition: Rules, maps, and plans. San Francisco, CA.: Chandler, 1972.

Shavelson, R.S , & Stern, S. Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. (To appear in Review od Educational Research, in press.)

Sudnow, D. Remarks. In R.C. Hill & K.S. Crittenden (Eds.), The Purdue symposium on ethnomethodology. (Monograph Number 1, Institute for the Study of Social Change). Purdue University, 1968.

Turner, R. (Ed.) Ethnomethodology. Harmondsworth, Midddlesex, England: Penguin Education, 1974.

Tyler, S.A. Cognitive anthropology. New York: Holt, Rinehart, & Winston, 1969.