Journal of Distance Education/
Revue de l'enseignement à distance
Spring/Printemps 1987
Vol. 2, No. 1, pp. 27-40

The Concept of Exchangeability in Designing Telecourse Evaluations

Richard J. Shavelson, Noreen M. Webb, and John Y. Hotta

Abstract

This paper examines designs for the evaluation of telecourses, that is college-level education television courses, from the perspective of exchangeability. The notion of exchangeability focuses evaluation on the level of knowledge attained at the end of tele- and traditional courses, not on the "growth" of knowledge from pretest to posttest. If tele- and traditional courses are exchangeable, students in both types of courses should attain the same level of knowledge, skills, and attitudes. Three attributes of exchangeability are: (1) telecourse treatments cannot be separated from their local implementation; (2) student characteristics and self-selection into tele- and traditional courses cannot be disentangled; and, (3) a balance between internal and external validity should be sought.

From the exchangeability perspective, randomized experiments for evaluating telecourses are usually inappropriate and uninterpretable. Telecourse populations typically differ from traditional course populations; "treatments" vary almost as much within telecourses as between tele- and traditional courses; and attrition is normal.

In view of these problems, four alternative evaluation designs are examined from the exchangeability perspective. These designs are able to handle population and treatment differences. Also, exchangeability with its emphasis on equivalence of outcomes, not "gains," avoids the problem of selection bias. We conclude that a combination of quasiexperimental and case study designs is most likely to provide the data that policymakers seek from telecourse evaluations.

Résumé

Cet article examine des modèles d'évaluation pour les cours télévisés du niveau collégial, du point de vue de leur interchangeabilité. La notion d'interchangeabilité concentre l'évaluation au niveau de la connaissance atteinte à la fin des cours télévisés ou traditionnels et non sur la "croissance" de la connaissance mesurée par un pré-test et un post-test. Si les cours télévisés et les cours traditionnels étaient interchangeables, les étudiants de ces deux types de cours devraient atteindre le même niveau de connaissance, d'habileté et d'attitude. Voici trois éléments distinctifs d'interchangeabilité: (1) le traitement des cours télévisés ne peut être séparé de leur mise en application locale; (2) les caractéristiques des étudiants et leurs motifs pour choisir des cours télévisés ou traditionnels ne

peuvent pas être démêlés; et (3) un équilibre entre la validité interne et externe devrait être recherché.

Selon la perspective d'interchangeabilité, les expériences non-contrôlées pour évaluer les cours télévisés ne sont habituellement ni appropriées ni interprétables. La population des cours télévisés est d'un type différent de celle des cours traditionnels; les "traitements" varient presqu'autant à l'intérieur des cours télévisés qu'entre ces derniers et les cours traditionnels et l'attrition est un phénomène normal.

En considération des ces problèmes, quatre modèles d'évaluation sont examinés dans la perpective d'interchangeabilité. Ces modèles sont capables de tenir compte des différences de population et de traitement. De plus, l'interchangeabilité, en mettant l'emphase sur l'équivalence des résultats et non des "gains," évite le problème des préjugés de sélection.

Nous tirons la conclusion qu'une combinaison de modèles d'études quasi-expérimentales et d'études de cas devrait fournir les données recherchées par les personnes chargées des décisions dans les évaluations de cours télévisés.

Introduction

American educational television has progressed a long way from when the medium was conceived mainly as a vehicle for transmitting classroom lectures or demonstrations to home audiences. From preschool to postgraduate levels, creative producers of educational television today utilize—often at great expense—the most sophisticated production technologies available. They aim not merely to replicate but to transcend traditional classroom teaching strategies. Capitalizing on distinctive attributes of the medium, they seek to exemplify verbal and written "lessons" with a visual immediacy never previously attainable. And they do so, or at least try to do so, at a technical level that attempts to satisfy the high expectations of audiences with considerable television viewing experience.

A series of television courses—or telecourses—recently created by the Annenberg School of Communications and the Corporation for Public Broadcasting (A/CPB) shows how far educational television has travelled since the era of the "talking head." Established in 1981 by a grant of 150 million dollars from the Annenberg School, the A/CPB Project's mission is to demonstrate the use of telecommunication systems to reach new audiences in higher education, and to develop innovative collections of high-quality, college-level courses to facilitate "distance teaching." At their best, the courses provide visually stimulating education that, in addition to being more intellectually rewarding, compares favourably as entertainment with commercial television.

Due to the high cost, however, higher education policymakers are faced with justifying the use and development of educational programs using alternative instructional media. Policymakers need answers to such questions as: "Compared to alternatives available, is a telecourse (for example) cost efficient and effective?" (cf. Clark & Salomon, 1985); "Are student outcomes in telecourses exchangeable for student outcomes in traditional courses?" Yet media research, because it bears mostly on questions of media improvement, addresses the policymakers' concerns only indirectly. Clearly, a policy context demands summative evaluations as a basis for decision-making. This paper presents a conceptual framework, that of the *exchangeability* of telecourse outcomes with traditional course outcomes, and evaluates the ability of alternative designs to assess exchangeability.

Exchangeability: A Framework for Designing Summative Telecourse Evaluations

Comparative evaluations of tele- and traditional courses have used a measure of the amount of knowledge that students gain as a criterion of media effectiveness. With one or another version of the gain measure, studies comparing the effectiveness of television and classroom instruction overwhelmingly find no difference (see reviews by Schramm, 1962; Stickel, 1963; Chu & Schramm, 1967; Dubin & Hedley, 1969). Such research has been interpreted as showing that "media don't make a difference" compared to traditional courses (Schramm, 1977, p. 28; Clark & Salomon, 1985). That is, students will learn about as much from a telecourse as from a traditional course. If media "don't make a difference," there is good reason to believe that telecourse outcomes may be exchangeable with traditional courses.

Exchangeability refers to the extent to which the knowledge, skills, and attitudes acquired by students from a telecourse are interchangeable with the knowledge, skills, and attitudes that are: (a) valued by policymakers, faculty, and administrators, and (b) acquired by students enrolled in an equivalent course offered in the traditional curriculum. This concept translates into a set of questions that guide evaluation: (1) Does the telecourse content meet academic standards? If so (2) Do the students acquire the knowledge skills and attitudes imparted? and, (3) Do students in the telecourse acquire the knowledge, skills, and attitudes to the same level as do students in traditional courses that cover the same objectives, materials and so on?

Exchangeability focuses attention on the *level of knowledge and skills* attained at the end of the course, not the "growth" of knowledge and skills in the two courses from the beginning to end. The criterion that should be applied to comparative evaluations of tele- and traditional courses, then, is whether the two courses produce equivalent levels of knowledge, skills, and attitudes at the *end of instruction,* ignoring the starting place of students in the two courses at the beginning of instruction. Exchangeability sets forth this strong requirement because faculty, administrators, and policymakers share the responsibility for insuring the equivalence of tele- and traditional course outcomes so that academic and certification standards are met.

From the exchangeability perspective, then, demonstrating that, after covariance adjustment of one kind or another, tele- and traditional courses

produce equivalent outcomes, as the literature shows, is not adequate. Such adjustment, if applied correctly, shows that if the horses start at the same point, they will go (gain) about the same distance, on average. But, in absolute terms, students in one course may perform considerably less well than students in the other course, and this performance should not be considered equivalent from the perspective of academic standards or certification.

The exchangeability concept has an additional three noteworthy attributes. The first is that "treatment" and "control" groups in an evaluation—the telecourse and traditional course—are defined by their curriculum and their local implementation. By "local implementation" we mean instructor quality, course requirements and credit, and institutional facilities and support. No attempt is made to disentangle the complex that comprises the "course."

The second attribute is that student characteristics and self-selection into tele- and traditional courses cannot be disentangled; they are part of local implementation. Consequently, an evaluation should not attempt to "equate" or "match" students in telecourses and traditional courses, either by design (e.g., random assignment, matched assignment), or by statistical methods. Rather, the evaluator should recognize that student populations for telecourses and traditional courses are different and, consequently, should treat each student population as inseparable from course implementation.

The third attribute is that an evaluation should use designs that balance internal validity (correct attribution of the causes of observed tele- and traditional course differences) and external validity (inference to the larger set of telecourses and traditional courses not observed in the evaluation). For example, a randomized experiment might be preferred on internal validity grounds, but might not be preferred on external validity grounds.

In the remainder of this paper, we draw implications from the notion of exchangeability for the design and analysis of media evaluations. Where appropriate, we provide concrete examples from a recent evaluation of telecourses (Shavelson, Stasz, Schlossman, Webb, Hotta, & Goldstein, 1986) carried out for the Corporation for Public Broadcasting.

Randomized Experiments and Telecourse Evaluation

In the "ideal" evaluation, subjects are assigned randomly to treatment and control conditions (Campbell & Stanley, 1963; Nunnally, 1975; Streuning & Guttenberg, 1975). This random assignment is supposed to ensure that subjects are equivalent in their characteristics across treatment and control conditions. The intent of the design is to be able to assert unambiguously that differences among treatment and control conditions, and not other factors, caused the observed differences in outcomes.

In a *telecourse evaluation*, the "ideal" experimental design would randomly assign students from a population to the treatment (the telecourse) and to a control

condition (traditional course). Comparisons of outcomes at the conclusion of the experiment would supposedly show whether the telecourse produced exchangeable knowledge and skills with those produced by the traditional course.

In the abstract, we can speak of "treatments," "controls," and "populations." However, telecourse populations are heterogeneous: they do not necessarily conform to the common stereotype of middle-aged students taking courses they would not otherwise have access to because of family and work obligations (e.g., Research Communications, 1985). Telecourses vary across sites depending on their implementation. And the choice of control groups is not always self-evident (Shavelson et al., 1986).

When implementing this experimental design the first problem lies in the definition of the treatments. The implementation of a telecourse can vary from semester to semester at one college, or from one college to another, on a number of dimensions (cf. Research Communications, 1985). *The New Literacy*, a telecourse aimed at developing "computer literacy," is a case in point (Shavelson et al., 1986). The implementations of the telecourse varied widely across the five sites studied (Table 1) and even though we can define dimensions along which

Table 1. Telecourse treatments across five implementation sites of *The New Literacy* (Shavelson et al., 1986).

Implementation Features	Site D1	Site E	Site F	Site G	Site H
Class meeting schedule:					
Weekly meetings					•
Orientation	•	•	•	•	
Study planning	•				
Midterm review	•	•		•	
Midterm examination	•	•	•	•	
Two field trips					•
Final review	•	•	•	•	
Final examination	•	•	•	•	•
Evening meeting				•	•
Weekend meeting	•	•	•		
Required attendance				•	
Course work:					
Midterm examination	•	•	•	•	•
Final examination	•	•	•	•	•
In-class quizzes					•
Take-home quizzes	•				
Extra credit questions and projects	•				
Computer programs			•		•
Instructor:					
Part-time	•	•	•		•
Full-time				•	
First-time telecourse instructor		•			
First-time New Literacy instructor		•			

telecourses may vary, telecourses exist, in reality, in specific combinations of dimensions. The same can be said for traditional courses. Hence, the *treatment cannot be defined independently* of the constellation of characteristics in the local implementation; multiple telecourse treatments abound (cf. White, 1980). Random assignment and strict control of treatment conditions, requirements of randomized experiments, may very well distort this "natural" phenomenon and drastically limit the generalizability of the evaluation findings.

As with treatments, telecourse populations are heterogeneous within and between sites. They vary from one implementation of the telecourse (e.g., site) to another. For example, students in *The New Literacy* telecourse varied widely in their characteristics, previous experiences with telecourses, and "studenting" while taking this telecourse (see Table 2). We believe that "treatments," be they

Table 2. Telecourse populations and participation rates across implementation sites of *The New Literacy* (Shavelson et al., 1986).

Population Characteristics	Site D1	Site E	Site F	Site G	Site H	
Sample Size [a]	29	29	32	16	15	
Populations						
Mean:						
Age	40.5	40.1	32.3	36.1	30.2	*
Verbal ability [b]	15.8	16.0	13.5	13.3	13.1	*
Computer knowledge pretest [c]	7.0	6.0	5.3	6.6	5.1	*
Computer knowledge posttest [c]	10.9	14.6	8.0	8.8	7.3	*
Percent:						
Female	69	61	59	31	33	*
White	93	75	63	75	93	*
Household income $30K	57	69	57	56	53	
Working full-time	66	72	84	75	80	
Enrolled for a degree	31	27	53	44	43	
Have B.A. or higher	17	43	13	25	7	*
B or better GPA	76	62	71	56	60	
Not taken telecourse	52	96	63	81	87	*
Not taken computer course	90	69	66	60	74	
Participation						
Mean:						
No. of on-campus meetings attended	2.0	2.6	2.5	3.3	15.5	
Study hours per week	4.4	3.6	3.0	2.6	1.9	
Percent:						
Reading 65% of textbook assignments	77	75	37	54	69	
Discussing TV program with others	40	69	51	67	56	

* $p \leq .05$ between sites
[a] Students who completed the course
[b] 24 items—the Extended Range Vocabulary Test from ETS's Reference Tests for Cognitive Factors (French, Ekstrkm, & Price, 1963).
[c] 15 items—representing knowledge gained from *The New Literacy* course.

telecourse or traditional, not only cannot be defined independently of their implementation; they cannot be defined independently of the students served. On this score, randomized experiments are problematic. The very act of random assignment may very well undermine the purpose of the evaluation.

The problem of population differences and randomized experiments is reflected in *attrition* from telecourse and control conditions. Attrition is potentially the most devastating threat to the internal validity of a telecourse evaluation. Indeed, the literature on telecourses tells us that large dropout rates are to be expected (e.g., Research Communications, 1985; Rumble & Harry, 1982). In our evaluation of *The New Literacy* we considered those students who did not take the final examination as dropouts. The telecourse dropout rates ranged from 34 to 52%. The traditional course (the control condition) dropout rate was an enormous 88%.

Ironically, attrition is likely to be minimized when students are indifferent about which course, telecourse or traditional, they enroll in, or as is more common with telecourses, when they self-select into tele- and traditional courses. We suspect that even if the evaluator could randomly assign students to courses, attrition is not likely to be random or equal in the treatment or control conditions. Students who drop out of the traditional course may do so because attending regular class meetings is inconvenient or impossible (due to work schedule, distance to campus from where they live or work, and so forth). Students who drop out of the telecourse may do so because they are dissatisfied with, for example, the lack of instructor contact. These selective attrition factors will produce groups at the conclusion of the experiment that are neither equivalent nor representative of the population from which they were drawn.

If randomization is problematic, so is the alternative. Without randomization the issue of *selection bias* arises: perhaps the difference in outcomes between telecourse and traditional course was due to population differences, not to the instructional format of the course. The concept of exchangeability provides a solution to this conundrum. The focus of the evaluation is on the level of achievement at the end of the course, regardless of whether students entered the tele- and traditional courses on equal footing. Demonstrating growth in knowledge, skills, and activities is important. But if that growth still does not produce students with knowledge (etc.) expected for successful completior of the course, those students should not be certified as having adequately completed the course.

Solving Evaluation Design Problems: Some Alternatives

Here we describe four alternative evaluation designs that address the unique circumstances of telecourse evaluation. They can all be used to assess the exchangeability of tele- and traditional courses.

Uncontrolled Assignment to Form Non-Equivalent Groups

This design does not control assignment of students to courses, but instead compares intact courses. In effect, it allows students to select the course that they

normally would choose. The design would be used when the evaluator is willing to accept the fact that the student population cannot be separated from the course implementation. That is, the evaluator seeks to answer these specific questions: Do students learn more from the telecourse than students not taking the telecourse? Do students who enroll in telecourses learn what students who enroll in traditional courses with the same objectives, materials, and assignments learn? The advantage of this design is that the final samples will represent the self-selected populations.

Attrition is not a problem for this design as it is for the randomized experiment. Student drop-outs are considered part of the "real" situation. Students remaining in the courses at the conclusion of the evaluation are considered to be a representative sample of student populations that finish the courses. Because student populations differ across telecourses and traditional courses, this design has widespread applicability. The most obvious threat to the validity of interpretations from this design is selection bias: that the observed differences could be caused by prior existing differences between the populations of students taking the telecourse and traditional course. But, as explained above, the focus of the evaluation is on outcomes, not on adjustments for prior existing differences.

Patched-up Designs—The Recurrent Institutional Cycle Design

When institutions regularly cycle students through the same course, students from one cycle might be considered as a control group for students from another cycle (Campbell & Stanley, 1963). In the telecourse situation this design could be applied when the same telecourse is offered during fall and spring semesters in the same year (see Figure 1).

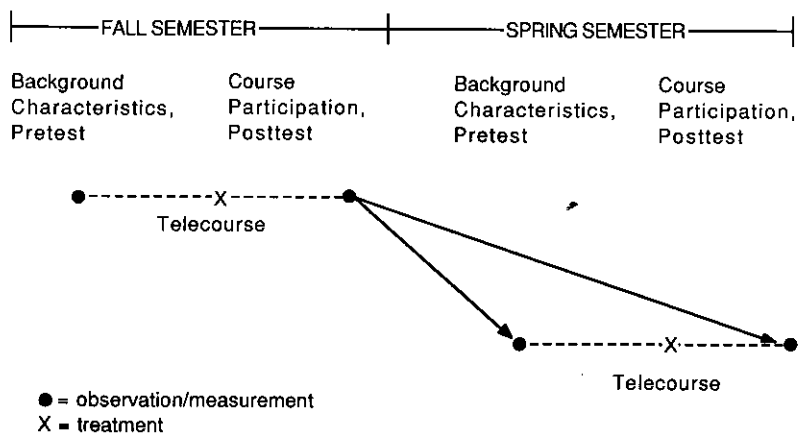

Figure 1. The patched-up design.

The minimum requirements for implementing this design are that the pretest achievement and other data are available from the spring semester students, and that pretest instrumentation overlaps posttest instrumentation significantly (e.g., the items on the pretest cover a substantial range of the final examination for both the fall and spring cohorts). Comparing fall posttest achievement with spring pretest achievement would help answer the question: "Do students learn more from the telecourse than they would have learned without it?" Comparing spring pretest achievement with spring posttest achievement provides an opportunity to replicate the "treatment effect" and increase confidence that instruction caused the change in performance. And finally fall posttest scores are compared with spring posttest scores to determine whether the outcomes in one semester are exchangeable with those in the other semester. The evaulator would not expect significant differences between the means of the fall and spring final examinations. Finally, the comparison between fall and spring pretest scores would show whether students enrolling in the two semesters are similar or different. If they are different, then the evaluator would focus on pretest-posttest information within a semester.

A major assumption underlying this design is that students enrolling during one semester are equivalent to (i.e., drawn from the same population as) students enrolling in another semester. Measurement of students' characteristics during both semesters (both at pretest and at posttest) can help address this assumption. This design has the advantage of not forcing assignment of students to the course, and not requiring special instrumentation outside the course implementation. Consequently, it has widespread practical applicability.

In our evaluation of *The New Literacy* (sites E, F, G, and H), we compared fall final examination scores with spring pretest scores, and spring pretest scores with spring final examination scores, to estimate mean gain. Then we compared fall posttest scores with spring posttest scores to estimate equivalence of outcomes. (Fall pretest scores were not available.)

The findings (Table 3) demonstrate remarkable consistency both within a site and across sites, with the exception of site E. That is, the effect sizes are equivalent across sites from fall final to spring pretest, and from spring pretest to spring final (an increase in average scores of about 1⅓ standard deviations). And the difference between fall final and spring final is not significant. The one exception is Site E where the instructor gave the students the answers to the questions on the final exam during the review session.

Table 3. Within- and between-site mean achievement and effect-size comparisons (Shavelson et al., 1986).

	Fall Final vs Spring Pretest		Spring Final vs Spring Pretest		Fall Final vs Spring Final	
Site	Signif. Level	Effect Size [a]	Signif. Level	Effect Size	Signif. Level	Effect Size
E	.01	1.29	.01	3.37	.01 [b]	2.82
F	.01	1.26	.01	1.42	n.s.	0.16
G	.01	1.07	.01	1.51	n.s.	0.64
H	.01	1.81	.01	1.10	n.s.	0.71

[a] The difference between means divided by the average standard deviation across sites.
[b] At this site the spring semester telecourse instructor gave students the answers to the examination questions, so scores were very high.

The Hybrid Design

The hybrid design is similar to the previous design with the important addition of a control group, the traditional course (see Figure 2). The hybrid design can be used when the same tele- and traditional courses are offered during fall and spring semesters of the same academic year. Students from the fall semester tele- and traditional courses would serve as control groups for the spring semester courses.

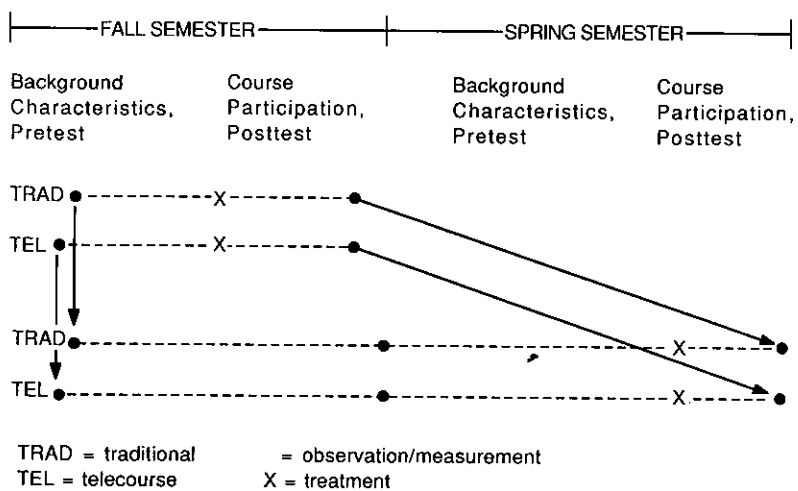

Figure 2. The hybrid design.

The hybrid design would help answer these questions: Do students learn more from the telecourse or the traditional course than they would have learned without it? Do students who enroll in telecourses learn what students who enroll in traditional courses learn? As before, the minimum requirements of this design are

that pretest achievement and other data are available from spring semester students and that pretest instrumentation overlaps posttest instrumentation significantly (e.g., the items on the pretest cover a substantial range of the final examination for both the fall and spring cohorts).

Case Study Methods

Bates (1981) has questioned the value of experimental methods in providing timely and relevant information for educational media decisions. The lack of significant differences between television and traditional instruction, according to Bates, has arisen because in many laboratory-controlled experiments: (a) important variables were ignored or not recognized (e.g., quality of program production); (b) organizational/contextual variables (e.g, class scheduling or viewing times) were often ignored; (c) differences between the quality of the treatment and control conditions were not accounted for; and (d) individual differences in responses to tele- and traditional courses were not examined. These problems are similar to those we have identified. Such observations have led evaluators to consider case study methods as potentially more appropriate and more responsive to decision-makers' choices about future uses for, and improvements in, educational media (Bates, 1981; Prosser, 1984).

By "case study," we mean a narrative account of an object of social inquiry—such as a classroom, a college, school system, or any other bounded system (cf. Stake, 1978)—in its cultural context; a case study is usually more descriptive than theoretical. Case study research rests upon the assumption that "inner understanding enables the comprehension of human behaviour in greater depth than is possible from the study of surface behaviour, from paper and pencil tests, and from standardized interviews" (Rist, 1979, p. 20).

Given the problems that regularly beset experimental and quasi-experimental telecourse evaluations, Bates (1981) and Prosser (1984), for example, have recommended that media evaluations be eclectic and include a strong case study component (see also Harris & Bailey, 1982). We concur. The context in which the telecourse is implemented cannot be separated from the treatment itself and, for reasons of ecological validity, should not be separated. By bringing both contextual and quantitative data to bear on the evaluation, the evaluator may: (a) document and completely describe the implementation process and potential threats to validity (e.g., document and describe reasons for attrition); (b) generate a wider range of alternative explanations for the observed findings than would be possible in an experiment (e.g., the instructor at Site E gave precise answers to examination items at the review meeting); and (c) examine a wider range of data than would be possible with an experiment alone (e.g, examine instructional methods and interaction patterns during class meetings and telecourse viewing).

Case studies alone, however, are unlikely to resolve questions of exchangeability raised by policymakers. Policymakers ultimately demand quan-

titative achievement treatment effects in different, representative contexts as an essential part of the policy evaluation.

Conclusions: Evaluating Student Outcomes From Telecourses

It is NOT feasible or desirable, except in unique circumstances, to use randomized experiments to evaluate student outcomes from telecourses. This evaluation design imposes too many unrealistic requirements on sites and students. Moreover, it usually addresses the wrong evaluation question—namely, are tele- and traditional courses exchangeable for the same (rather than self-selected) student population? Further, it assumes random attrition, which is untenable when students desiring telecourses for scheduling reasons are randomly assigned to a traditional class.

We conclude that by using quasi-experimental designs (Campbell & Stanley, 1963), evaluation of student outcomes is feasible given constraints imposed by evaluation sites and students. However, these constraints are considerable, even with the "patch-up" and "nonequivalent-control group" designs.

Within the context of exchangeability, we consider evaluation of student outcomes feasible provided that it is possible to:

1. use evaluation designs that permit student self-selection into tele- and traditional courses;
2. oversample, to alleviate problems of attrition, small enrollments, and course cancellations caused by underenrollments;
3. establish tele- and traditional course comparisons;
4. require students to complete pretests by:
   a. mandating attendance at telecourse orientation,
   b. requiring completion of pretests for course credit, and
   c. using other mechanisms (e.g., inducements); and to
5. carry out a case study at each site to:
   a. document and completely describe the implementation process and potential threats to validity,
   b. generate a wider range of alternative explanations for the observed findings than would be the case in an experiment, and
   c. examine a wider range of data than would be the case with an experiment alone.

References

Bates, A.J. (1981). Towards a better research framework for evaluating the effectiveness of educational media. *British Journal of Educational Technology, 12,* 215–233.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychology Bulletin, 54,* 297–312.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

Chu, G.C., & Schramm, W. (1967). Learning from television: What the research says. Stanford, California: Institute for Communication Research.

Clark, R.E., & Salomon, G. (1985). Media in teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: McMillan.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally College Publishing Co.

Dubin, R., & Hedley, R.A. (1969). *The medium may be related to the message: College instruction by TV.* Eugene, Oregon: University of Oregon Press.

French, J.W., Ekstrom, R.B., & Price, L.A. (1963). *Manual for kit of reference tests for cognitive factors.* Princeton, N.J.: Educational Testing Service.

Harris, N.D.C., & Bailey, J.G. (1982). Conceptual problems associated with the evaluation of educational technology. *British Journal of Educational Technology, 12,* 4–14.

Nunnally, J.C. (1975). The study of change in evaluation research: Principles concerning measurement, experimental design, and analysis. In E.L. Streuning and M. Guttentag (Eds.), *Handbook of evaluation research.* Beverly Hills, California: Sage Publications.

Prosser, M.T. (1984). Towards more effective evaluation studies of educational media. *British Journal of Educational Technology, 15,* 33–42.

Research Communications. (1985). Executive summary for research on student uses of the Annenberg/CPB Telecourse in the Fall of 1984. Chestnut Hill, Massachusetts: Research Communications, Ltd.

Rist, R.C. (1979). On the means of knowing: Qualitative research in education. *New York University Education Quarterly, 10,* 17–21.

Rumble, G., & Harry, K. (Eds.). (1982). *The distance teaching university.* New York: St. Martin's Press.

Schramm, W. (1977). *Big media little media: Tools and technologies for instruction.* Beverly Hills, California: Sage Publications.

Schramm, W. (1962). Learning from instructional televison. *Review of Educational Research, 32,* 156–167.

Shavelson, R.J., Stasz, C., Schlossman, S., Webb, N., Hotta, J., & Goldstein, S. (1986). Exchangeability of student outcomes from regular and telecourse instruction: A feasibility study. Santa Monica, California: The Rand Corporation.

Stake, R.E. (1978). Case study evaluations. *Educational Researcher, 7,* 5–8.

Stickell, D.W. (1963). *A critical review of the methodology and results of research comparing televised and face-to-face instruction.* Unpublished doctoral dissertation, Pennsylvania State University, University Park, Pa.

White, P.B. (1980). Educational technology research: Towards the development of a new agenda. *British Journal of Educational Technology, 11,* 170–177.

RICHARD J. SHAVELSON is Professor of Research Methods and Inquiry, Graduate School of Education, U.C.L.A.; consultant to the RAND Corporation; and president, American Educational Research Association. At the time of this study, he directed RAND's Education and Human Resources Program. He currently conducts research on teaching mathematical and scientific problem solving; on measurement theory including generalizability of performance measurements, and modeling graduate school academic programs; and on policy issues including national indicator systems for monitoring mathematics and science education, and military manpower, especially accession policy.

NOREEN M. WEBB is Associate Professor of Research Methods and Inquiry, Graduate School of Education, U.C.L.A. She has done research and published papers on classroom processes and learning, and measurement theory and applications, especially generalizability theory.

JOHN Y. HOTTA is a research consultant at the The RAND Corporation in Santa Monica, California, and a graduate student in the Graduate School of Education, U.C.L.A. He is currently studying advanced educational technologies, including telecourses and intelligent tutoring systems.