THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Methods for Field Testing New Telephone Services

By D. J. EIGEN*

(Manuscript received December 29, 1981)

Telephone services are increasingly complex and diverse, and they require more human-machine interaction than ever before. A field test can help improve a new service by ensuring that it is easy to use with little chance for error. This paper discusses the methodology of field testing. Specially tailored telephone service evaluation methods, based on field test experience with the Calling Card Service, are presented in detail.

I. INTRODUCTION

New telephone services involve more customer-system interaction than ever before, and making the use of these services easy and error-free is a major goal of their development. Properly designed dialing plans, announcements, timings, tones, and instructions increase customer acceptance, minimize customer errors, and promote use of the service. The design of new services can be evaluated by coordinated studies that include:

- Analysis of present services,
- Interviews with customers,

* Bell Laboratories.

- Laboratory studies of proposed protocols for customer-system interactions,
- Field tests of services, and
- Product follow-up studies.

Field testing is the largest and most costly step in this coordinated set of studies. This paper discusses the methodology of field testing, using a field test of the Calling Card Service as an example.

II. CALLING CARD SERVICE FIELD TEST OVERVIEW

An analysis of operator-handled credit card service indicated that a reasonable proportion of credit card calls could be automated. Interviews with credit-card, bill-to-third-number, and collect customers verified their interest in and need for an automated Calling Card Service, and laboratory studies provided evidence that customers could use the Calling Card Service successfully.

The Calling Card Service field test was conducted in Milwaukee from November 1977, to June 1978.[1,2] Regular telephone credit card numbers could be dialed in order to place Calling Card calls from about 3000 noncoin phones in the Milwaukee area and from 70 coin phones at Milwaukee's airport, two downtown hotels, and a few local restaurants. Bright orange placards on the trial coin phones instructed customers on how to use their telephone credit card number. In addition, operators were trained to assist callers and answer questions. (Calls from unequipped stations were handled as usual.)

To use the trial Calling Card Service, customers first dialed zero plus the number they wished to call. Special programs in the Traffic Service Position System (TSPS) routed incoming "0+" calls from trial stations to a small team of specially trained operators who helped simulate Calling Card Service—in actual service no operators are used. Besides the TSPS console, these operators had a video display terminal linked to a minicomputer (see Fig. 1).

When a call arrived from a specially equipped station, the trial operator notified the minicomputer, which then delivered a tone to prompt the customer to dial a Calling Card number. Detectors received the dialed digits and sent them to the minicomputer over a data link for verification. Calls with valid Calling Card or credit card numbers were allowed to proceed and were billed appropriately.

Depending on the protocol being tested, the minicomputer displayed step-by-step instructions on a terminal screen to guide the operator in handling each call. For example, to encourage customers to redial after making errors, the minicomputer might display the instruction, "Please hang up and dial zero plus the number you are calling," which was to be read to the customer. By making simple changes in the minicomputer program, the operator's treatment of calls could be

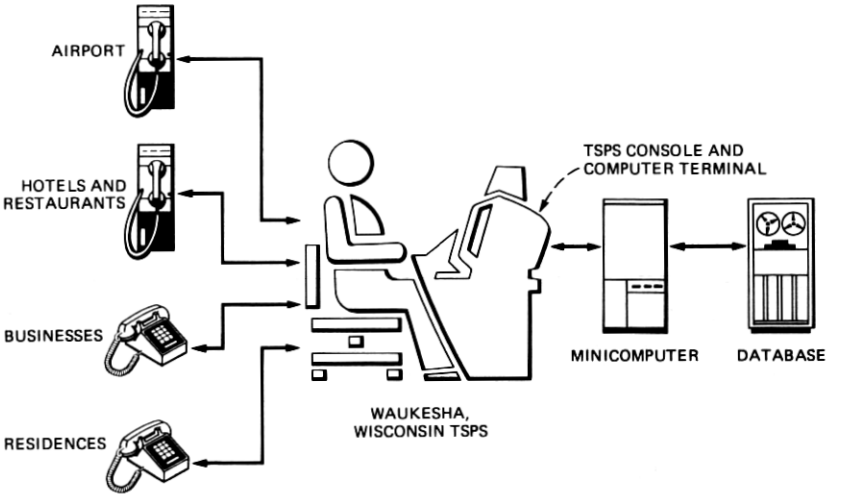


Fig. 1—Test setup.

altered, often without additional training. This flexible arrangement allowed easy testing of many different protocols and rapid changes among them.

The minicomputer recorded the time and details of each call. These records were analyzed rapidly to determine how the protocol could be improved. Throughout the trial, protocols were varied by changing announcements, timings, access to operators, error-correction procedures, and other aspects of the caller interface. In all, 24 variations of the protocol were tested at equipped coin phones, 14 at noncoin phones. Over 10,000 customers dialed more than 30,000 automated Calling Card calls during the trial and more than 5,000 customer interviews were obtained.

III. FIELD TEST ACTIVITIES

3.1 Field tests

In a field test such as that for Calling Card Service, a simulation of the proposed service is actually offered to a limited, but representative, set of customers on a trial basis. In some circumstances, when there is sufficient confidence in the form of the service, the actual product can be used as a test vehicle.

The field test can be used to adjust the technical and operational aspects of the service. It can be used to determine whether the service fits a customer need, and it can also be used to improve estimates of willingness to pay. And finally, it can be used to evaluate customer performance, satisfaction, and usage in the effort to provide a service

that optimizes the customer-system interaction on which new services rely.

The greater the fidelity of the test to the real service situation, the greater one's confidence in the final success of the service being tested. Field tests provide increased fidelity over more indirect techniques for evaluating services.

3.2 Study plan

Developing a study plan is the first task in planning a field test. A study plan must include the following five steps:

1. State the objectives of the study and delineate issues to be resolved by it. The primary objectives of the evaluation of the human-machine aspects of telephone services are to:
   a. Determine if usage, satisfaction, and performance are at acceptable levels.
   b. Predict the levels of usage, satisfaction, and performance in the final service.
   c. Refine the service to improve usage, satisfaction, and performance.

Many other detailed issues for a particular service may require resolution.

2. Determine constraints on study service and resources necessary to accomplish the test. Decisions on test methodology necessarily involve practical choices. For example, the marketplace often imposes serious time constraints on the development, deployment, and evaluation of a telephone service.

3. Design and refine the service and the human-machine interfaces involved. Some initial human-machine telephone service interface must be defined before the evaluation process can be initiated. Continued analyses, interviews, and laboratory studies are best used to generate and refine the service alternatives prior to the field test itself.

4. Delineate variables that may influence the results and hypothesize their interactions. Three categories of variables need to be specified for the field test:
   a. Independent variables—those variables that are to be deliberately manipulated or held constant. Among the possible variables of this type (with some examples) are the following:
      (1) Service protocol—announcements, tones, timings, error-handling strategies, and digit strings.
      (2) Capabilities of the service—billing, routing, and screening.
      (3) Availability—geographic constraints, time-of-day constraints, and station-type constraints.
      (4) Type of instruction for customers—media and format.
      (5) Marketing effort—promotion.
      (6) Rate—price structure.

b. Dependent variables—those variables whose values are affected by changes in the independent variables. Some examples are given below.
   (1) Subscription—initial interest and sign-up rate.
   (2) Usage—rate of first and repeated use.
   (3) Acceptance—judged worth and satisfaction.
   (4) Customer performance—speed, error, and abandonment rate.
c. Parameters—those identifiable variables that are free to vary. Among these are:
   (1) A priori condition—predisposition toward service, demographic mix of customer population, etc.
   (2) Internal characteristics—the test design or method used, intrinsic characteristics of the customer in the test, etc.
   (3) External characteristics—the geographic, environmental, and temporal setting of the test.

These variables are not necessarily mutually exclusive. Some parameters might be fixed or systematically varied, and, thus, be made independent variables.

It helps to have hypotheses about how these variables will interact. Figure 2 demonstrates one model of several possible models of the relationships among variables for telephone services. A different model might assume that instruction may affect acceptance and subscription more directly than is shown in Fig. 2.

5. Define methods for increasing confidence in the results of the study. Some of the most important aspects of a good study plan are motivated by the need to counteract rival explanations of the results

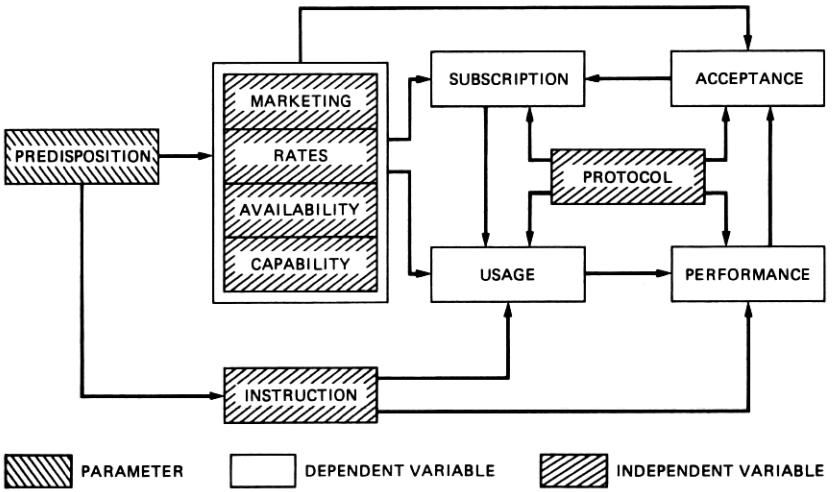


Fig. 2—Variable block model.

and, thereby, increase confidence in the validity of the field test as an indicator of the success of the service.[3–5] Section IV is devoted to this component of the study plan.

3.3 Test development

After the study is designed, the following tasks must be done.

1. Define data gathering tools.
2. Define data analysis techniques.
3. Define test service delivery vehicle.
   a. Define test service requirements.
   b. Design test service delivery vehicle.
   c. Implement service delivery vehicle (hardware/software development).
   d. Integrate trial system and test.
4. Acquire and prepare customers.
5. Acquire and prepare site(s).
6. Define test operations and procedures for utilizing results.

Figure 3 is an illustration of the interrelationships among these tasks.

3.3.1 Data collection

The following data collection methods were used in the Calling Card Service field trial.

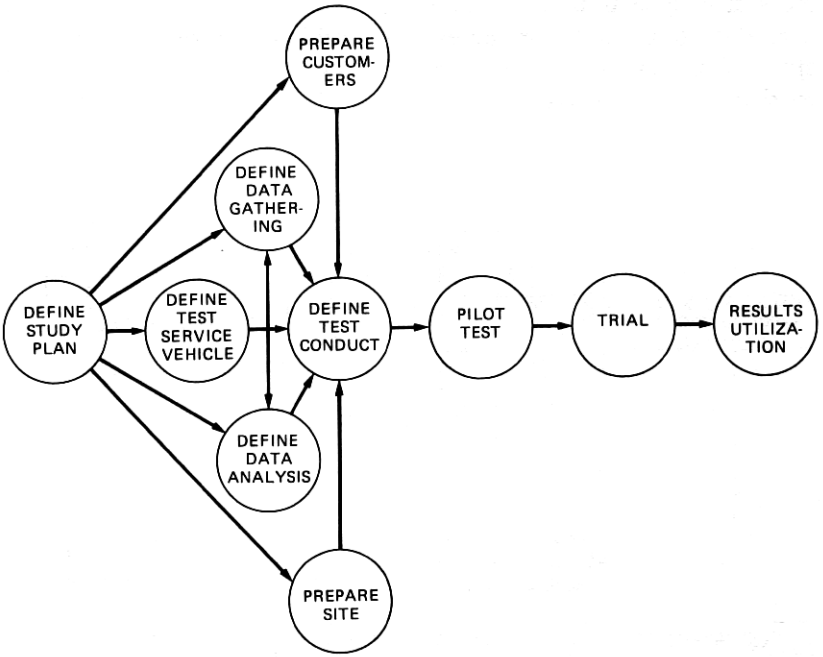


Fig. 3—Evaluation development paradigm.

3.3.1.1 Interviews/questionnaires. Interviews and questionnaires provide measures of attitude, as well as corroborative measures of usage, and can be given before, during, and after the field test. In-person, face-to-face interviews with customers who had used the Calling Card Service provided some of the more compelling evidence of the success of the service. The following spontaneous comments illustrate the range of such data.

Example of positive response:

1. "It's excellent. I compliment the phone company for coming through with this. Makes it a lot easier."

Example of negative response:

1. "In the past, you dialed Helen (operator) and said, 'I want to talk to Joe,' and right off Joe was on the line—it seems you are trying a lot of new services to get rid of Helen."

3.3.1.2 On-line measurements. These measurements are the machine recording of customer and system events, such as picking up a receiver or placing it back on the switchhook (hanging up), dialing digits, tones or announcements, error detection, and call progress states. By recording, time stamping, and storing these events, the sequence of customer-machine interactions can be reconstructed.*

3.3.1.3 Billing. Billing records routinely provide various data on a call, such as calling and called number, call duration, date, time of day, class of charge (e.g., collect, DDD, and so forth).

3.3.1.4 Observation. Observation provides a crude but valuable technique for gathering data that cannot be obtained by other means.

3.3.1.5 Customer feedback. Feedback mechanisms for customer complaints and suggestions are maintained by the telephone companies offering the service.

3.3.1.6 Records. Records and logs of equipment malfunction and events recorded in newspapers are important for detecting or counteracting erroneous or misleading results. Data in such records can be used to measure dependent variables.

These methods of collecting data are all relatively insensitive to one another in that a specific error in one measurement is very unlikely to affect another measurement. Thus, the methods can be used to corroborate each other.

---

* The Bell System as a common carrier is entitled to certain privileges to monitor the quality of its services. Bell Laboratories as an agent of the Bell System is extended these privileges. Only data necessary to ensure good service are gathered, and they are kept secure, in strict confidence, and statistically anonymous. To guarantee customer privacy, data are never gathered after a call is placed, i.e., after called party answers.

3.3.2 Analytic tools and techniques

The practice of feeding back the results as input to the experimental design process, while providing an efficient data gathering technique, places an additional burden on analysis. Results must be provided in a timely manner to be of use in formulating the next service improvements to be tested. Figure 4 illustrates a data processing stream used to field test the Calling Card Service.

Standard statistical routines must be used with caution since the assumptions on which they are based may be violated. For example, sampling is often nonrandom.

One heuristic is to leave the service unperturbed for some reasonable period of time (two or more weeks depending on the call volume). The extent of variations or noise in the data is measured and later used as a benchmark to assess meaningful variations potentially due to service manipulations.

3.3.3 Test service provisioning

A system to deliver a voice-prompted test service should include the ability to deliver tones and announcements, to route to an operator, and to receive dialed digits. Connection to the network, billing, and data processing and storage also need to be considered.

3.3.4 Acquire and prepare sample

Obtaining sufficient numbers of representative customers for the field test is important. The field test approach usually requires thou-

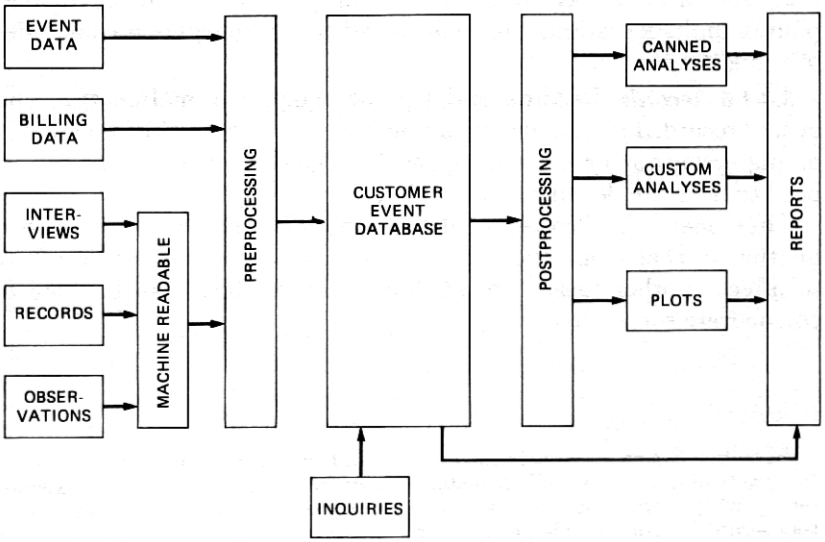


Fig. 4—Data processing stream.

sands of calls to provide adequate data. Inability to obtain sufficient customers, if not attributable to test limitations alone, may be a key indicator of the potential failure of a new service. The means for selecting and acquiring customers has to be designed with test validity in mind.

The process of obtaining participants and assigning them to groups can be the source of some of the strongest rival hypotheses. Random selection is the principal counteraction to ensure the representativeness of the population sample and group equivalence. But, randomization is not always possible. Moreover, even when it can be used, aspects of the customer acquisition and assignment process can still introduce biases that will limit the ability to generalize.

Selection processes are outlined below:

1. Determine target population using interviews and market studies.
2. Characterize the history, predisposition, environmental influences, etc., of target population.
3. Draw customers randomly from within the target population and from a similar group outside the target population. Sampling customers outside the target population will help to validate the use of the target groups for tests. If random selection is not possible, select customers who are matched to characterization of the target population.
4. Solicit participation for the test in a representative way. Preferably, contact customers in the same way they would be contacted for the final service.
5. Check participants for degree of similarity to target population characterization.
6. Interview random samples of those who agree and disagree to participate to determine the reasons for participation or nonparticipation. For example, the fact that it is a test rather than an actual service may have influenced either decision. Also, use a priori factors to account for any differences between those who agreed and those who refused.
7. Randomly assign participants to control groups and treatment groups. If random assignment is not possible, use self-selection of treatments.
8. Use measurements taken before the test to help determine the nature of group differences.

3.3.5 Results utilization

One of the largest pitfalls in the practice of implementing an evaluation of telephone service can be the lack of assurances and mechanisms for integrating evaluation results into the final product. Results from the test must be timed to allow development of changes

in the final product. A responsible organization must be named for coordinating the integration of results and the final product itself must be designed to allow changes. Appropriate flexibility in the initial design of the product can reduce delays in product introduction.

IV. DESIGNING TO COUNTERACT OTHER EXPLANATIONS

Steps taken in designing the test—which decrease confidence in rival hypotheses and increase confidence in the working hypotheses (the one we wish to prove)—are called counteractions. In laboratory research, some typical counteractions are:

- Orthogonality and counterbalancing of independent variables
- Control and preclusion of extraneous variables
- Random collection of subjects in the attempt to eliminate selection bias.

Field settings require adapting such measures to real-world constraints.

The counteractions useful in field tests can be divided into those required for evaluating a static service (i.e., determining if a service meets a priori acceptability standards) and those required for evaluating attempts to improve a service. Counteractions can be adapted to preclude, disconfirm, or control rival hypotheses.

4.1 Service evaluation

One of the simplest counteractions is to take multiple measurements of the same service, such as measurements of usage, satisfaction, and performance. Taking these measurements in more than one way (multiple methods) is also useful. Confidence in a particular result is increased if different sources point to the same conclusion.

One fear (rival hypothesis) that occasionally strikes is that the data are somehow mutilated to spuriously inflate results. This fear became real in the Calling Card Service field trial when the service began performing better than expected. Interviews, observation, and billing data corroborated the computer-collected data so strongly as to explode this rival hypothesis. Test calls removed any remaining shards of doubt.

Many rival hypotheses are based on test time. For example, the results may be due to some unusual circumstance or event, or the test may give rise to a trend in usage or performance that may be transitory (e.g., novelty effects or start-up effects) or the test may give rise to cyclic or delayed effects. One simple counteraction used in the Calling Card Service field trial to account for these problems was to repeat a test or take continued measurement of the service over time. If a unique circumstantial or historical event affected the outcome of a

test, then subsequent measures could provide evidence to support or negate this rival hypothesis.

Before, during, and after usage measurements of all coin stations (trial and nontrial) at the airport were inspected to determine if any events, trends, or cyclic phenomena confounded the data. Specifically, the Christmas holiday season and an airline strike had to be taken into account.

While multiple measurements and time-spaced testing counteract many problems, they give rise to problems of their own. One measurement may affect another. For example, repeated interviewing may cause customers to change their behavior more as a function of the interviewing process than the service itself.[6] Fortunately, the bulk of the measurements typically taken in a telephone service field trial are made on-line and are unknown to the telephone caller.* The customer is not usually aware in these tests that digits dialed, on-hooks and off-hooks are recorded and time stamped. These measures are presumably nonreactive, and thus there is no plausible explanation of measurement affecting customer behavior and attitude.

One could avoid the problem of reactivity by using only nonreactive measures. But interviews and questionnaires provide invaluable data. We handled this problem in the Calling Card Service field trial by interviewing some customers once, others twice, and still others not at all. Comparisons of these groups in terms of nonreactive measures and subsequent interviews, however, did not substantiate a reactivity effect.

Establishing a control group is another strategy that can be used to help determine if the results were due to historical events or arbitrary circumstances. The assumption is that those extraneous events or circumstances that affect the treatment group also affect the control group. To the extent that the precision of the measures allow, differences between the control group and the treatment group can, therefore, be attributed to the presence of the treatment, that is, the service.

People at nontrial coin stations at the airport were interviewed as a control for interviews taken at trial coin stations. This procedure allowed us to determine if there were any changes in customer attitudes that might have been attributable to events alone and not to service. The interview control group thus served as a counteraction to the rival hypothesis of history. For example, those people interviewed some days were angry because of plane delays. If this irritation had an effect, it would show up in both treatment and control group scores.

---

* The customer generally is aware that a test is being conducted and that service is being measured; thus, the problem of test reactivity cannot be completely discounted.

Randomization in the selection process assures equivalence between the treatment and control groups. If randomization is not possible, a plausible rival hypothesis is that the people assigned to one group differ, on the average, from those in the others and that these intrinsic differences are the real cause of any observed difference among the groups. A method for determining whether such rival hypotheses are correct is to make measurements of the groups before the treatment (service) is administered. Comparing the pretreatment and post-treatment observations of the group given the trial service with simultaneous observations of the control group (not given the trial service) tests the rival hypotheses that these two groups are inherently different, or in other words, that results are due to the history of the customers or their maturation or learning during testing.

Customers were interviewed and billing data were tracked before and after the Calling Card Service field trial. These data were used to determine the differences among customers and between trial and nontrial stations.

4.2 Improving service

In the preceding discussion, counteractions for a static service evaluation were discussed. A new service can be said to consist of a set of variables that may be manipulated to possibly improve service.

The key question is: How can we be sure that changes in service outcome are due to our service manipulations? Establishing causal connections for what improves or degrades service increases understanding. Increased understanding will tend to increase confidence in predictions of how the final service will fare.

4.2.1 Varying the service

A simple model of service variation is depicted in Fig. 5. The knobs in the figure represent the independent variables; the meters, the dependent variables. The object of varying the service is to manipulate the knobs in such a way as to cause the meters to register a beneficial change in service. If the meters are considered to show customer satisfaction, usage, and performance, the object then is to find the combination of knob positions that maximizes the readings on each meter (metaphorically speaking).

With every knob manipulation, knowledge and understanding is increased, whether the meters change positively, negatively, or not at all. Knowledge gained from the previous adjustment permits efficient and effective design of subsequent service adjustments. However, such a test-adjust-test method requires a system that can provide very quick analysis of results.

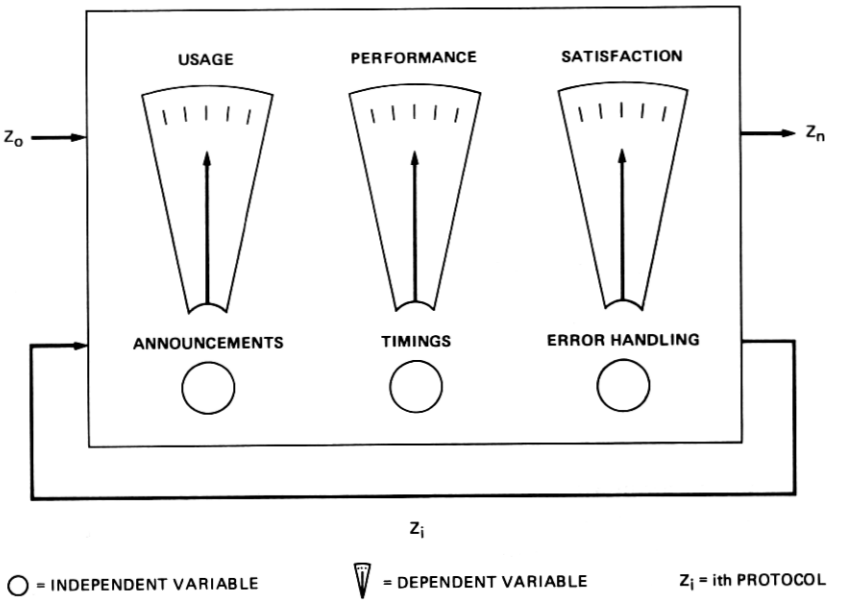


Fig. 5—Field trial metaphor—with results fed back as input into the field trial design.

4.2.2 Increasing confidence in service manipulations

Improving service can be thought of as repetitions of static service evaluation. The same counteractions discussed earlier can be used and augmented for service improvement. To account for the possibility of other events interposing effects on the dependent variables coincident with the adjustment of a service, the concept of the control group is again appropriate.

One new problem that is introduced in improving service is that multiple treatments (changes) may interact with each other. That is to say, customers' responses to a subsequent service may not be the same as those to the first service they experienced. Further, the very act of changing the service may impact attitudes and customer performance.

Adding a new group of customers for every new protocol iteration counteracts this threat to validity. (Appendix A contains notations for representing this design consideration and the others discussed here.)

The rudiments of these design principles were used in the Calling Card Service field trial. Service was manipulated in the trial in an attempt to improve service. A particular service configuration is here called a service protocol. Because many of the trial phones were in the airport, there were always a great number of new customers trying the service, more than enough to support the manipulations.

When customers dialed 0+NPA NXX XXXX at trial stations a tone prompt was given (sometimes followed by an announcement prompt) and the following things (in most protocols) could happen:

1. The customer could dial a credit card number.
2. The customer could time out without dialing and be connected to an operator.
3. The customer could dial 0 after the prompt and be connected to an operator.
4. The customer could abandon the call.

Ideally, all customers with credit card numbers would dial them. All others would dial 0 after the tone prompt to reach an operator. Practically speaking, this was not possible. Rather, different protocols were tested to attempt to increase the proportion of dialed credit card calls, increase the proportion of customers who needed an operator to dial 0, or decrease abandons.

Figure 6 summarizes the service manipulations for the placarded coin stations. Figure 7 shows the proportion of the four classes of call dispositions: (1) dial credit cards, (2) dial 0, (3) abandons, and (4) time-outs for each protocol for the placarded coin stations. The area under each curve represents the incremental proportion of 0+ calls.

While these curves are revealing, they cannot be used alone to determine the best protocol, i.e., the one with the highest satisfaction, performance, and usage. However, unacceptable protocols can readily be identified: they are those with abandonment rates in excess of 25 percent and those with dialed credit card call rates less than 25 percent.

Fortunately, the first coin placarded service protocol had low abandonment rates and high credit card dialing rates. The subsequent

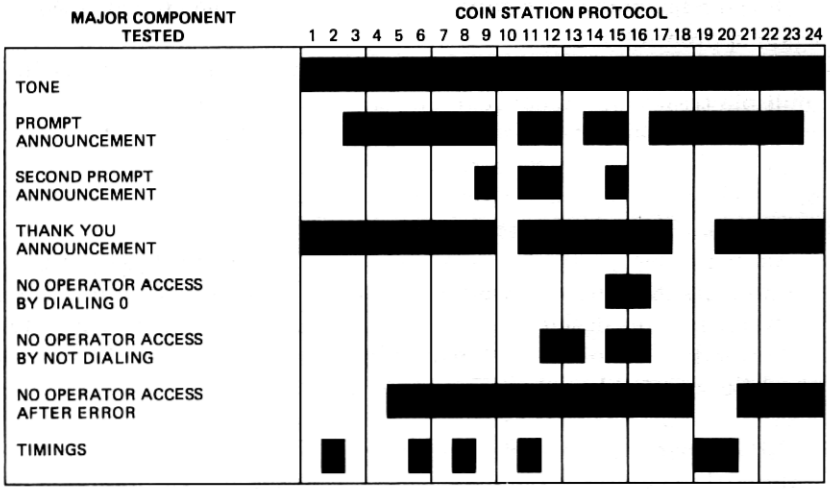
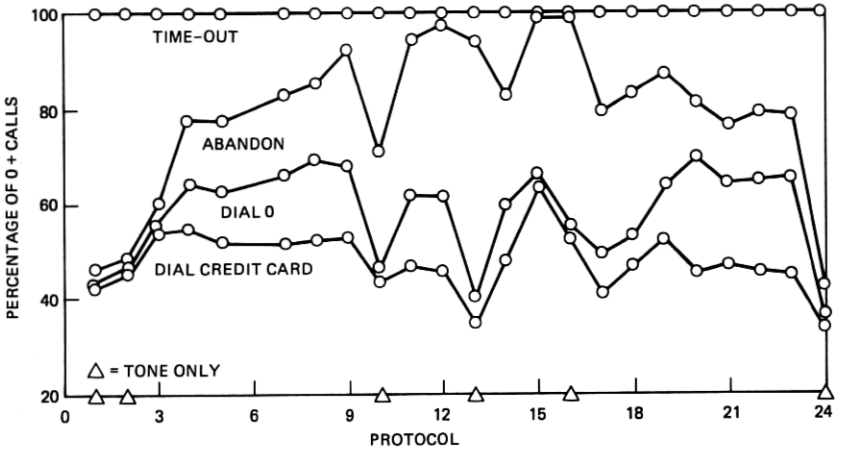


Fig. 6—Service manipulations.



NOTE: PROTOCOLS WITHOUT A TONE-ONLY INDICATOR HAVE A TONE AND AN ANNOUNCEMENT. PROTOCOL 17 INCLUDES A SPECIAL ANNOUNCEMENT DESIGNED TO DISCOURAGE CALLS FROM UNENABLED ROTARY STATIONS.

NOTE: THESE CURVES REPRESENT INCREMENTAL PERCENTAGES.

Fig. 7—Dialing performance at placarded coin stations by protocol.

manipulations were primarily aimed at improving service, although clearly the opposite effect was sometimes achieved.

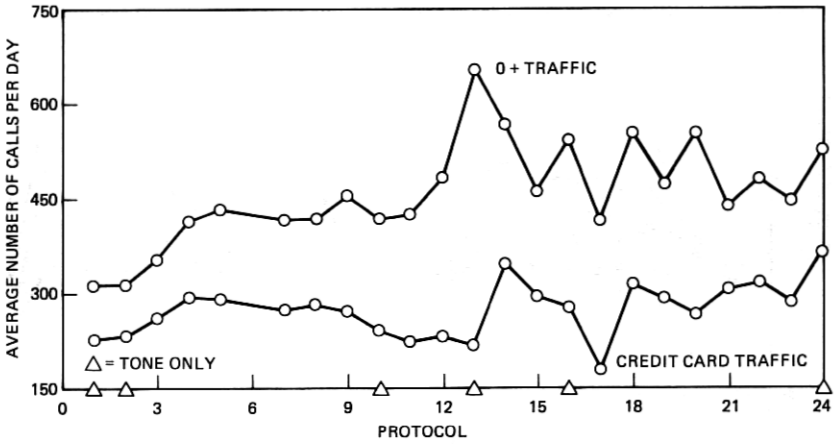
A changing volume of 0+ calls and a changing call mix were plausible explanations for one service protocol doing marginally better than another. As Figure 8 shows this was sometimes a consideration.

Another rival hypothesis was related to assumptions about those who abandoned. If all or a large number of those who abandoned were credit card customers, estimates of the proportion of credit card dialers could vary significantly. Figure 9 shows the percentage of dialed credit card calls (of all credit card calls) with and without abandons counted as credit card calls. Including and excluding abandons in this way provides lower and upper bounds, respectively, of the percent of credit card calls dialed.

Still another rival hypothesis was that usage (or attempts) may be higher, but performance lower when comparing one protocol to another. Figure 10 illustrates success rates by protocol.

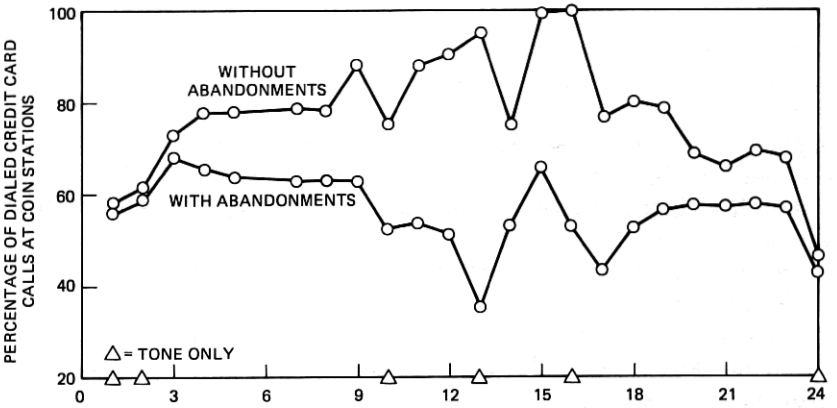
Rival hypotheses due to repeated usage (familiarity and learning) were accounted for by sorting out and examining results from repeat users.

Service protocols were compared to determine the effect of a particular manipulation. Figures 11 and 12 compare two protocols in terms of abandonments and percent of 0+ dialing. These figures lead to the conclusion that repeating a prompt announcement in this service does not increase dialing but does increase abandons. Thus, repeating announcements is a less desirable alternative.



NOTE: PROTOCOLS WITHOUT A TONE-ONLY INDICATOR HAVE A TONE AND AN ANNOUNCEMENT. PROTOCAL 17 INCLUDES A SPECIAL ANNOUNCEMENT DESIGNED TO DISCOURAGE CALLS FROM UNENABLED ROTARY STATIONS.

Fig. 8—Volume of 0+ traffic at placarded coin stations.



NOTE: PROTOCOLS 12, 13, 15 AND 16 MANIPULATIONS INVOLVED LIMITING OPERATOR ACCESS WHICH REDUCES 0 VOLUME AND CAUSED 0 + CUSTOMERS TO EITHER ABANDON OR DIAL A CREDIT CARD NUMBER.

NOTE: PROTOCOLS WITHOUT A TONE-ONLY INDICATOR HAVE A TONE AND AN ANNOUNCEMENT. PROTOCOL 17 INCLUDES A SPECIAL ANNOUNCEMENT DESIGNED TO DISCOURAGE CALLS FROM UNENABLED ROTARY STATIONS.

Fig. 9—Dialed credit card calls with and without abandonments by protocol.

4.2.3 Comparison groups

There are two ways of making service comparisons in the test design just discussed, which consisted of staggered subjects and measurements of responses to service refinements. First, customer comparisons can be made across service changes. This consists of multiple measurements or time-spaced measurements of the same customer. The

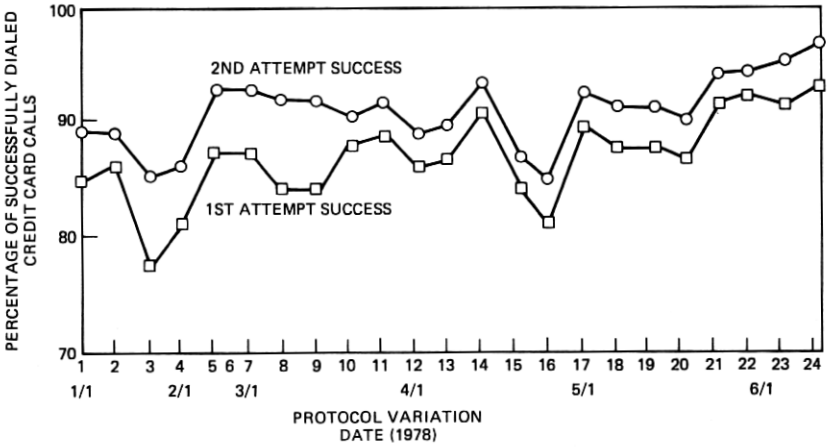


Fig. 10—Customer dialing success rate.

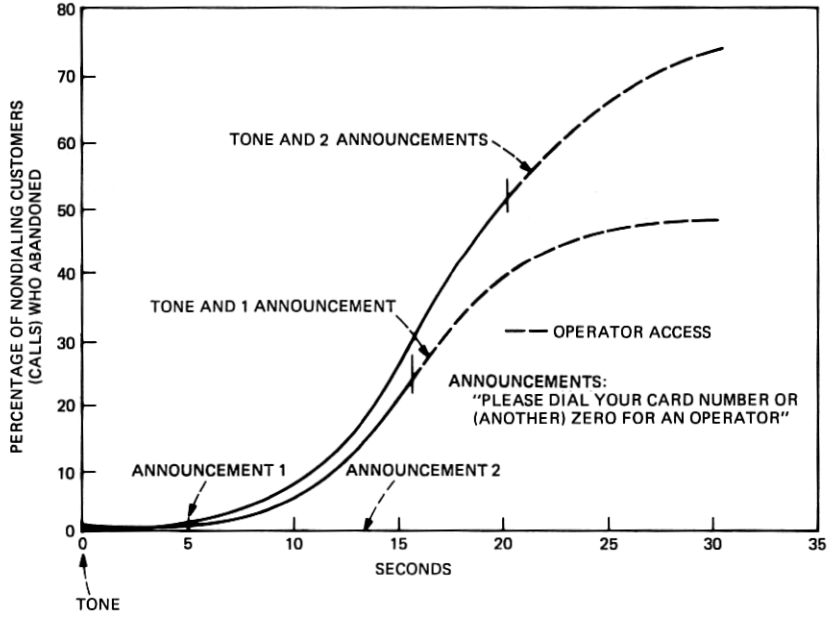


Fig. 11—Abandonments as a function of repeated announcements at coin placarded stations.

error variance is small because there are no intersubject differences, but there is multiple treatment interference. Second, service comparisons can be made across new subject groups, but treatment effects are confounded with history and by any bias introduced by the selection of groups.

Comparison groups can be arranged to preclude history as a con-

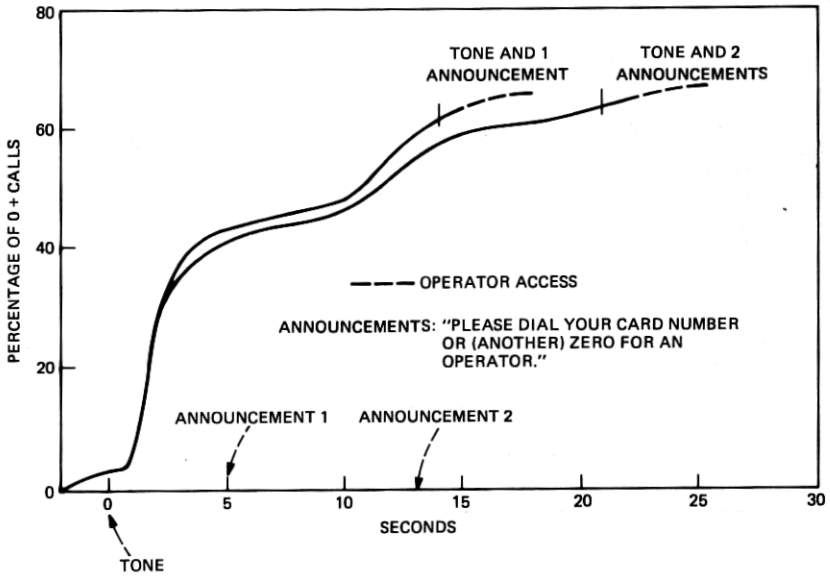


Fig. 12—Dialing as a function of repeated announcements at coin placarded stations.

founding variable by offering different service arrangements simultaneously (at different telephones, for example). For best results, each of these comparison groups should be as similar as possible except for any planned differences in treatment (service).

In the Calling Card Service field trial, we devised three comparison groups: noncoin stations, placarded coin stations (with special bright orange instruction cards), and placardless coin stations. These comparison groups were not mutually exclusive. Crossovers occurred, especially between placarded coin and placardless coin groups. But for every comparison group, and, as discussed, for every manipulation, there were new customers. Service changes were replicated across these comparison groups further increasing confidence and providing information about the effects of the station type.

4.2.4 Refinement

Because the number of variables and variable states in a field test of this kind is usually large, only a small subset of the number of possible treatment combinations can be used. Moreover, the protocol is evolving and a delay is often necessary in the testing of different combinations. Unfortunately, this delay means the manipulation is confounded with the history of the customer population. However, prudent selection of changes and repetitions within a treatment group and across comparison groups can yield a more complete understanding of the effects of the service components and their interactions.

In the Calling Card Service, the primary objective of the service manipulations was to make reasonable changes which were hypothesized to increase usage, satisfaction, and performance. This objective was sometimes subjugated to the purpose of increasing confidence in the belief that the changes in the independent variables (and not the parameters) caused the observed changes in the dependent variables. One such manipulation was to deliberately make it progressively more difficult for the customer to get to an operator, i.e., protocols 12, 13, 15, and 16. It was hypothesized that more customers would use their credit card if it were made relatively more difficult to obtain an operator. However, this hypothesis was not borne out; the proportion of credit card calls dialed was either unchanging or decreasing, and abandons increased as one would suspect. Clearly, other manipulations were more successful.

As the treatment change chart shows, changes in the state of more than one independent variable were made more often than not. The large number of variables prohibited single sequential changes in each variable. For example, the error-leg protocol was manipulated independently and simultaneously with the access-leg protocol many times. Evidence gathered from protocols prior to this practice showed little or no interaction from changes in the access leg and the error leg, except that the number of people making errors fluctuates.

Inferences made about the protocol change were made in terms of the multiple change because the independent variables were confounded. Inferences based on effect separation and independence assumptions were made cautiously and verified, when possible, in later manipulations.

4.2.5 Replication

Any changes in the dependent variables caused by a change in the protocol (or service) and not by history should be present in successive identical tests (within the bounds of measurement error). Replication of effects increases confidence in the hypothesis that a change in protocol actually causes a change in the dependent variable being measured. As a practical matter, the utility of repeated replications of a single treatment is sharply reduced by the need to manipulate as many variables as possible in the time allowed. However, intermittent replications can serve to benchmark drifting or trending data.

On several occasions, we repeated protocols within and across comparison groups in the Calling Card Service field trial. Replications afforded an increase in confidence by allowing an assessment of changes in dependent variables due to confounding of external and internal parameters with the independent variables. The replications, in effect, provided a basis for comparison. And the replications in-

creased confidence in the assertion that the observed changes in the dependent variable were caused by the deliberate change in the independent variables. For example, protocol 14 amidst protocols 12, 13, 15, and 16 (coin placarded stations) was a replication of an earlier protocol and supported the hypothesis that the poor performance of the other protocols was real and not an artifact of internal or external conditions.

Some replications were designed to provide more information about the contributing effects of certain service components. Most notably, major service changes were tested with both the tone-only and tone-and-announcement conditions. For example, protocol 13 (coin placard station) was the same as protocol 12, but without the prompt announcement. This is similarly true for protocols 15 and 16.

4.2.6 Partial counterbalancing

Arranging the treatment order in special ways—e.g., the Latin Square design—across comparison groups provides for pairwise comparisons of different treatment orderings. Such arrangements, however, are often not practical in field trials of telephone services because of the large number of variables and variable states (treatment conditions) that should be tested. The large number of treatment combinations required of typical counterbalancing techniques preclude their use. Manipulation of protocols on the basis of data gathered from prior manipulations also precludes planned counterbalancing.

But it may be possible partially to counterbalance treatment pairs across comparison groups in order to assess some multiple treatment order effects. If the manipulations evolve, the application of counterbalancing schemes have to be delayed by at least one time frame to have knowledge of the treatment pair.

4.2.7 Post-manipulation static study

After service manipulations are finished, it is useful to study the final service at rest. In this period in the Calling Card Service field trial (protocol 24), transient effects due to changes in the service were not present, thus improving predictions of the final service.

V. CONCLUSION

Many service issues were resolved by the trial of Calling Card Service. Most important, the customers who tried the service continued to use it because they felt it was faster and more convenient than operator-assisted credit card calling. Moreover, the design of the service was critically dependent on the results of the field trial manipulations. Both announcements and instruction placards were found to be very effective in stimulating customers to dial. The field trial data

provided objective measurements of optimum customer usage and acceptance and lower rates of error and abandonment in the service offered. Table I lists the criteria found most useful in selecting the attributes of the final service.

When Calling Card Service was first offered in Buffalo in July 1980, a product follow-up evaluation study was conducted to monitor customer use, performance, and acceptance levels. The measurements of actual service conformed closely to estimates made from field trial data and, thus, support the utility of test methodology in guiding service design choices. The methodology developed for evaluating Calling Card Service is now being applied to the next generation of services, such as teleconferencing.

VI. METHODOLOGY SUMMARY

The pre-field trial activities of analysis, interviews, and laboratory studies are undertaken to refine the service definition and assess the potential of its success. A study plan is constructed which defines objectives, resources and constraints, study variables, acceptance criteria, and study design. Field test requirements are developed that specify the data collection, analysis, and customer acquisition and preparation procedures as well as specify the hardware, software, and network aspects of a telephone service test vehicle.

Given adequate requirements the test is developed and implemented. During the test, the service is manipulated and data are collected and analyzed to optimize criteria of usage, performance, and satisfaction. Trade-offs are made relative to available resources, data validity, confidence in results, and time constraints. If success is predicted for the service being tested, results are then integrated into the design and the service is offered and monitored. Figure 13 summarizes the total process of conducting a field test of a telephone service.

Table I—Criteria for selecting service attributes

Independent Variables	Dependent Variables
1. Announcement Presence	a. Percentage Dialed b. Abandons
2. Announcement Wording	a. Percentage Dialed b. Confusion (interviews) c. Abandons
3. Number of Attempts Allowed	a. Success Rate b. Percentage of Subsequent Attempts
4. Operator Access	a. Percentage Dialed b. Abandons c. Satisfaction (interviews)
5. Interevent Timing	a. Distribution of Times to First Digit
6. Interdigit Timing	a. Distribution of Pauses Between Digits
7. Protocol (service in general)	a. Percentage Dialed b. Satisfaction (interviews) c. Success Rate
8. Delays	a. Abandons

Fig. 13—Design and evaluation process for human-machine telephone service protocols.

VII. ACKNOWLEDGMENTS

I would like to thank E. C. T. Walker and C. A. Riley for their valuable comments on this manuscript. K. R. Hickey, R. J. Jaeger, N. S. Pearson, and J. L. Santee are thanked for their input as well. The Calling Card Service test depended heavily on the contributions of T. M. Bauer and E. A. Youngs.

REFERENCES

1. M. R. Allyn, T. M. Bauer, and D. J. Eigen, "Planning for People: Human Factors in the Design of a New Service," Bell Lab. Rec., *58*, No. 5 (May 1980), pp. 55–161.
2. D. J. Eigen and E. A. Youngs, "Calling Card Service—Human Factors Studies," B.S.T.J., *61*, No. 7 (September 1982), pp. 1715–35.
3. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Design for Research*, Chicago: Rand McNally, 1966.
4. D. T. Cook and D. T. Campbell, *Quasi-Experimentation Design and Analysis Issues for Field Settings*, Chicago: Rand McNally, 1979.
5. H. M. Parsons, *Man-Machine System Experiments*, Baltimore: Johns Hopkins, 1972.
6. E. J. Webb et al., *Unobtrusive Measures: Nonreactive Research in the Social Sciences*, Chicago: Rand McNally, 1966.

APPENDIX A

Field Evaluation Design Notation

To provide another perspective on the field design methods, the following notation is provided.

A.1 Service

A service (or treatment) can be defined by set of dependent variables with specific values, represented here as a vector $\dot{Z}$.

$$\dot{Z} = \begin{vmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{vmatrix}$$

A.2 Measurements

Let 0 stand for observation. A matrix of n kinds of measurement taken in m different ways is represented by $\ddot{0}$.

$$\ddot{0} = \begin{vmatrix} 0_{11} & 0_{12} & \cdots & 0_{1m} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 0_{n1} & 0_{n2} & \cdots & 0_{nm} \end{vmatrix}.$$

When multiple measures (taken multiple ways) are repeated, the notation is:

$$\ddot{0}^m = \ddot{0}_1 \cdots \ddot{0}_m .$$

A.3 Improving service

Two aspects of the field study, static service evaluation and improving service, could be represented as:*

Static Service Evaluation		Improving Service
$\dot{0}_0^m \dot{Z}_0 \ddot{0}_1^m$	$\dot{Z}_1$	$\ddot{0}_2^m \cdots \ddot{0}_n^m \dot{Z}_n \ddot{0}_{n+1}^m$

This design effectively constitutes repetitions of the static service evaluation.

A.4 Control group

${}_{A'}\ddot{0}_0^m \dot{Z}_0 {}_A\ddot{0}_1^m \dot{Z}_1 {}_A\ddot{0}_2^m \cdots {}_A\ddot{0}_n^m \dot{Z}_n {}_A\ddot{0}_{n+1}^m$	Treatment Group
${}_{A'}\ddot{0}_0^m {}_{A'}\ddot{0}_1^m {}_{A'}\ddot{0}_2^m \cdots {}_{A'}\ddot{0}_n^m {}_{A'}\ddot{0}_{n+1}^m$	Control Group

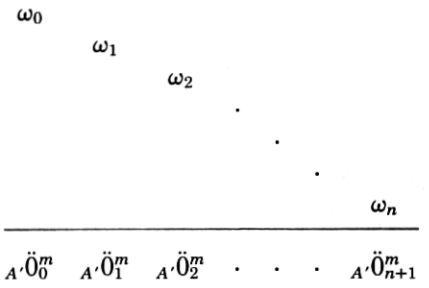
A.5 Staggered treatment groups

New subjects are added (or are available) at each treatment change. Limited resources usually necessitate the continued manipulation of the variables within treatment groups. The line offsets the control group from the treatment groups and denotes unequivalence of the groups.

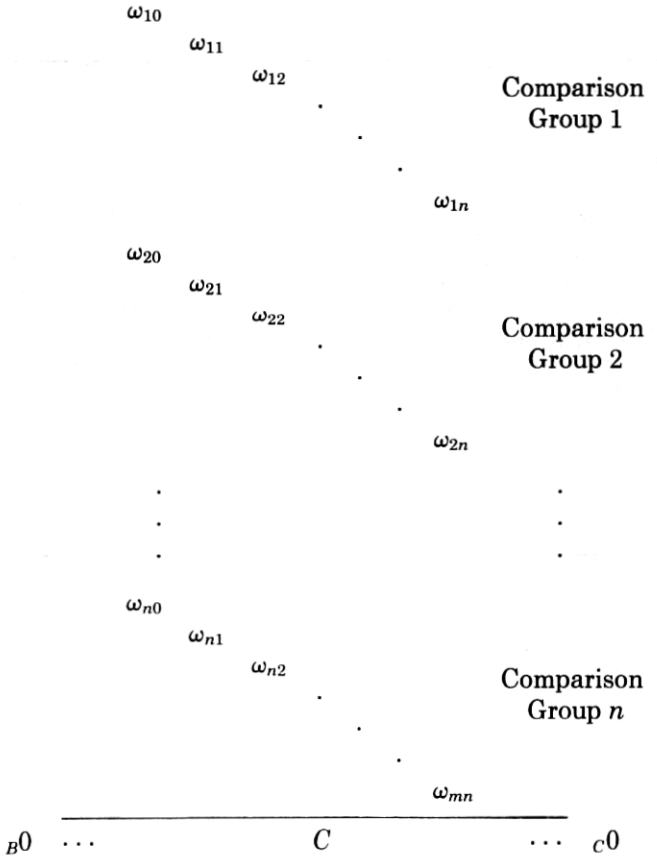
${}_{A'}\ddot{0}_{10}^m \dot{Z}_0$	${}_{A'}\ddot{0}_{11}^m \dot{Z}_1$	${}_A\ddot{0}_{12}^m$	$\cdots$	${}_A\ddot{0}_{1n}^m \dot{Z}_n$	${}_A\ddot{0}_{1n+1}^m$	Treatment Group 1
	${}_{A'}\ddot{0}_{21}^m \dot{Z}_1$	${}_A\ddot{0}_{22}^m$	$\cdots$	${}_A\ddot{0}_{2n}^m \dot{Z}_n$	${}_A\ddot{0}_{2n+1}^m$	Treatment Group 2
				${}_{A'}\ddot{0}_{nn}^m \dot{Z}_n$	${}_A\ddot{0}_{nn+1}^m$	Treatment Group n
${}_{A'}\ddot{0}_{n+10}^m$	${}_{A'}\ddot{0}_{n+11}^m$	${}_{A'}\ddot{0}_{n+12}$	$\cdots$		${}_A\ddot{0}_{n+1n+1}^m$	Control Group

* $\dot{Z}_i$ denotes the vector of service-independent values for the ith service change (treatment). ${}_A\ddot{0}_i^m$ denotes the matrix of multiple measurements by multiple methods measured m times for the ith service (treatment) change. The left-hand A' subscript (versus A subscript) denotes that some measurements present in the trial may not be available in the pretrial measures (or in the control group).

To reduce the notational burden, let ω_i denote the ith staggered treatment group.



A.6 Staggered treatment groups with comparison groups



Note that the Control Group, C, is augmented with historical and follow-up measurements.

A.7 Replications

Replications are denoted by the appropriate repetition of protocol time frame subscripts on the independent variable vector:

$$ {}_{A'}\ddot{O}_0^m \boxed{\dot{Z}_0} {}_{A'}\ddot{O}_1^m \dot{Z}_1 {}_{A}\ddot{O}_2^m \boxed{\dot{Z}_0} {}_{A}\ddot{O}_3^m \dot{Z}_2 {}_{A}\ddot{O}_4^m \cdots \dot{Z}_n {}_{A}\ddot{O}_p^m. $$

A.8 Partial counterbalancing

$$ \begin{array}{llll} \boxed{{}_{A'}\ddot{O}_{101}^m \dot{Z}_0 {}_{A}\ddot{O}_{111}^m \dot{Z}_1} & \ddot{O}_{121}^m \dot{Z}_2 & \cdots & \left.\begin{array}{l}\\ \\ \\ \end{array}\right\} \text{Comparison Group 1} \\ {}_{A'}\ddot{O}_{211}^m \dot{Z}_1 & {}_{A}\ddot{O}_{221}^m \dot{Z}_2 & \cdots & \\ & {}_{A'}\ddot{O}_{321}^m \dot{Z}_2 & \cdots & \\ {}_{A'}\ddot{O}_{102}^m \dot{Z}_0 \quad {}_{A'}\ddot{O}_{112}^m \dot{Z}_1 & {}_{A}\ddot{O}_{122}^m \dot{Z}_0 & \cdots & \left.\begin{array}{l}\\ \\ \\ \end{array}\right\} \text{Comparison Group 2} \\ \boxed{{}_{A'}\ddot{O}_{212}^m \dot{Z}_1} & {}_{A}\ddot{O}_{222}^m \dot{Z}_0 & \cdots & \\ & {}_{A'}\ddot{O}_{322}^m \dot{Z}_0 & \cdots & \end{array} $$

The two boxed groups are the time-delayed counterbalanced treatment pairs $\dot{Z}_0$ and $\dot{Z}_1$.

AUTHOR

Daryl J. Eigen, B.A. (Psychology), 1972, M.S. (Electrical Engineering), 1973, University of Wisconsin; Ph.D. (Industrial Engineering), 1981, Northwestern University; Bell Laboratories, 1973—. Mr. Eigen initially worked in the Human Performance Technology Center. He then was involved in feature and service planning for the Traffic Service Position System and, later, the No. 4 ESS. He is currently Supervisor of the System Analysis and Human Factors Group for No. 4 ESS. Member, IEEE, APA, Human Factors Society, and Tau Beta Pi.