EPSDT Impact on Health Status

by P.H. Irwin and Rosemary Conroy-Hughes

Early and Periodic Screening, Diagnosis and Treatment (EPSDT), a large-scale operational screening program which has generated a tremendous volume of data on the socio-demographic characteristics and health status of Medicaid-eligible children, seems to provide an ideal context within which to evaluate the effectiveness of preventive child health care.

Concerns about health care expenditures generally, and the effectiveness of preventive child health services specifically, lead to the question of whether the impact on the health status of the children served can be measured without significantly adding to the cost of these services with primary data collection. We employed a quasi-experimental research design using administratively-generated data from an operational EPSDT program to estimate program impact on the prevalence of serious abnormalities among the children served. We found that, compared either to themselves across time or to a control group, a representative sample of 1831 children had almost 30 percent fewer abnormalities requiring care on rescreening.

The ability to demonstrate the impact of EPSDT using these data suggests, among other things relevant to policy, that a national EPSDT monitoring system could be developed that would be cost-effective and could lead to program improvement.

Background

Researchers in the field of child health have been unable to agree on the value of preventive services for children. However, in a recent critique of thirty-eight controlled empirical studies of preventive-care for children (Shadish, 1980), one reviewer argues that the credibility of the evidence is significantly diminished by the poor methodological quality of much of the research. Although the studies are not totally inconclusive, Shadish reports that "preventive child health care has, in general, been given only the weakest of tests" (p. 27).

Evaluation of the long-term effectiveness of preventive care programs is a difficult task, involving the costly and data-intensive process of measuring change over time. The increasing burden of health care expenditures suggests that such evaluation is essential to the development of cost-effective services. The costs of mass screening programs and other early-intervention measures can only be justified if those programs produce clear and lasting benefits. It is therefore imperative that evaluative evidence on the effectiveness of preventive care be compiled and analyzed.

This project was funded under Grant Number 18-P-97115, Health Care Financing Administration.

Early and Periodic Screening, Diagnosis and Treatment (EPSDT) seems to provide an ideal context within which to evaluate the effectiveness of preventive child health care. This paper summarizes a methology for estimating the impact of EPSDT in southeastern Pennsylvania.

EPSDT in Southeastern Pennsylvania

The EPSDT program used in southeastern Pennsylvania is a multi-phase program of preventive health care for children of low income families. The immediate goal of the program is to detect potentially debilitating health problems, and, ultimately, to improve the health status and decrease the overall dependency of the target population. The screening procedure the EPSDT program employs is *not* intended to be a complete diagnostic examination; rather, the screen serves to identify the potential health problems and to distinguish those conditions that warrant immediate care or referral from those requiring no more than a simple followup. In Pennsylvania, all Medicaid-eligible children are entitled to be screened every three months for the first eighteen months of life and every other year thereafter, through age twenty.

The EPSDT program in the five-county Philadelphia area is administered by a private organization under contract to the State Department of Public Welfare and consists of five phases:

- Outreach or casefinding, to identify and contact eligibles, to inform them of the program and to assist them in gaining access to services;
- Administration and coordination of screening, to establish and direct a network of qualified screening units to serve the area's target population;
- The actual testing or screening by certified providers of medical care;
- Compilation and reporting of screening results; and
- Diagnosis and follow-through, to provide treatment when necessary.

Over the past seven years of the program's operation, a management information system has been developed and large quantities of managerial data have been produced. By 1980, over *700,000* screens had been performed in this system and the medical records stored in a large computer file.

However, these administratively-generated data are not ideally suited for research and evaluation, and suffer from serious problems of data quality and completeness. The eligibility status of the target population changes rapidly, and there is an unavoidable time lag before a change in eligibility is posted to the data system; the program's management information system (MIS) therefore does not provide an accurate estimate of the current number of eligibles at any point in time. Although screening data tend to be more complete than eligibility data, even this most reliable component of the MIS contains numerous instances of misidentification, duplicate records, and inconsistent data values. Finally, since a relatively small proportion of children are screened more than once, longitudinal data are available for only a very limited number of children.

Objectives of the Present Study

This study was undertaken as part of a larger effort to examine the process, effectiveness, and costs of the EPSDT program in southeastern Pennsylvania, and it represents an attempt to determine the extent to which EPSDT had improved the health status of children receiving periodic screens (that is, the program's effectiveness). The explicit demonstration component of the project attempts to show to what extent, and how, the administratively-generated data base can be used to estimate that effectiveness.

The project's initial attempts to measure the impact of the program on health status using only available data produced ambiguous results. In terms of change in risk status between screen and rescreen (that is, in the proportion of children having a serious problem on first screen compared to their second screen), no overall change is apparent in a sample of children who received two screens approximately two years apart. (See Table 1.)

The McNemar test results in a non-significant chi-square of 400/783 = 0.51 (d.f. = 1, p = 0.48) and indicates that no *net* change in risk status is evident in this sample. Similarly, there is no evidence of significant changes in risk status whether or not the child returns to the same provider (nearly 58 percent did) for rescreen. (See Table 2; chi-square is 0.037 and 1.786, respectively, with $p > 0.10$ for both.)

TABLE 1
Change in Risk Status from Initial Screen to Rescreen

		Time Two	
		No Problem (0)	Problem (1)
Time One	(0)	450	381
	(1)	402	590

TABLE 2
Change in Risk Status from Initial Screen to Rescreen for Same Provider and Different Providers

		Same Provider	
		Rescreen	
		No Problem (0)	Problem (1)
Initial Screen	(0)	255	219
	(1)	214	366
		Different Provider	
		Rescreen	
		No Problem (0)	Problem (1)
Initial Screen	(0)	203	162
	(1)	188	224

Tables 1 and 2 illustrate the impact of EPSDT by focusing on *turnover*—the clients who change risk status, from having a problem to not having one, and vice versa. In Table 1, (402 + 381 =) 783 children (42.7 percent of the sample) changed status, even though the observed *net effect* over time is only about a 2 percent decrease in those with problems. (Again, since it is not statistically significant, the change in the population *could* be zero or even in the opposite direction for the entire population.) We may therefore conclude that there is no evident impact on health status.

From another perspective, in terms of the persistence of specific problems, the curative impact of EPSDT can be quite clearly estimated with the same sample. Of the 992 children with problems that require care at Time One, 40.5 percent are problem-free at Time Two. This improvement is probably *under*estimated since we do not know from Table 1 whether the 590 children with problems at both times still suffered from the *same* abnormality identified at Time One (nor the same *number* of abnormalities).

Deriving a reliable estimate of risk reduction to determine if the *same* abnormality were present at both times requires considerable data manipulation and assump-tions about their meaning. We are able to determine whether the problems identified on rescreening were in the same general diagnostic area as those identified at the initial screen; if the subsequent problems are not in the same area, then

TABLE 3
Health Status at Time Two of Children
Who at Time One Had Had Problems Requiring Care
(By Diagnostic Area)

Diagnostic Area	Problem Resolved		Problem Persisted		Total Problems
	n	%	n	%	
Dental	308	70.3	130	29.7	438
Vision	189	53.5	164	46.5	353
Hearing/Speech	52	80.0	13	20.0	65
Develop/Behavior	41	85.4	7	14.6	48
Nutrition/Growth	24	85.7	4	14.3	28
Medical	332	69.7	144	30.3	476
Total Problems	946	67.2	462	32.3	1408[1]

[1]Sums to more than the total number of problem screens (992) because 19.3% of the children had multiple problems.

they are assumed not to be the same problem. Table 3 summarizes the results.

In this paper, we distinguish between *risk reduction* and *problem resolution.* A child is classified as "high risk" if at least one problem requiring care is identified on screening. It can be argued, however, that risk is reduced if at least one of those problems is resolved. Some children classified as high risk at both times will probably have enjoyed reduced risk before rescreening, but will have contracted another or similar problem. In Table 3, problems that "persist" (32.3 percent of the total) may in fact not be the same abnormality, but rather another problem in the same diagnostic area—another source of impact underestimation.

Undoubtedly some problems requiring care will persist because the condition is chronic. The remaining conditions that go untreated may be few indeed, which leads to a tentative conclusion that in terms of potential *curative* impact, the results are consistent with a significant improvement in health status among the EPSDT children in this sample.

While these conclusions from different approaches to the same data are not completely contradictory, they are ambiguous about the impact of EPSDT on health status. We may attempt to resolve this ambignity by developing a more complex research design, which incorporates longitudinal as well as cross-sectional comparisons, yet does not necessitate primary data collection.

Research Design

Conceptual Definition of Impact

Despite the limitations of the data system, it should be possible to determine, using only data generated by the program itself, the impact of EPSDT on the health status of the children it has served. Numerous confounding factors make this determination complex; nonetheless, it is important to know to what extent the impact can be isolated, without resorting to primary data collection or to experimentation with human subjects. "Impact" in this instance refers to a *change* in health status attributable to a given process (namely, EPSDT), and must be distinguished from those program performance measures which evaluate just cross-sectional differences between groups rather than longitudinal change.

Changes in health status can be determined only for a specified individual or group passing through a given process. *Differences,* on the other hand, can be observed between or among any units of comparison. The distinction between longitudinal changes and cross-sectional differences is important to maintain in this enterprise as elsewhere (see, for example, Lieberson and Hansen, 1974). The central methodological thesis of this study is that to analyze the impact of EPSDT on the health status of clients without collecting new data, it is necessary to compare both initial screen results to rescreen results *for the same children* and also to compare children who have participated in the program to those who have not.

A sample of children receiving two screens approximately two years apart is essential for this study. Even if a simple comparison of health status (to be operationally defined later) at both times were to reveal improvements for this group of children, the differences may be due to something other than the benefits of the EPSDT assessment and referral. Any rival explanations for observed differences in health status challenge our ability to attribute observed effects to the EPSDT program, but the challenge is a serious one *only if the alternative explanations are plausible.* The plausibility of competing explanations will be tested by comparing the two-screen sample to itself across time and to a control group of children just entering the program. To the extent that these comparisons tend to negate any rival explanations, inferences about program impact are valid.

Comparative Design

Of the many assumptions that are necessary in this analysis, the most impressive is that impact—any observed change in health status—is due to EPSDT alone. This admission is notably unsurprising. Outside (or inside) a laboratory, unequivocal attribution of cause and effect is not a simple task. Still, if EPSDT as a health services program has been ameliorative, then we can reasonably expect some evidence of the effect. It is also reasonable to specifically consider the possible sources of bias in available data.

In any research endeavor, success in attributing observed (and hypothesized) change to some experimental treatment (for example, the impact of a program) is primarily a question of the internal validity of the study. [The ensuing treatment of validity follows that of Campbell and Stanley, 1963.] Did the program produce the change, or is there some alternative explanation (referred to as a "threat to the internal validity")?

We propose to examine explicitly the threats to internal validity that challenge this impact analysis. We begin, therefore, by repeating the eight primary threats to internal validity identified by Campbell and Stanley (1963: p. 5):

- *History*—a specific event, in addition to the experimental variable, occuring between first and second measurement;
- *Maturation*—changes occurring in the respondents as a function of time;
- *Testing*—the effects of taking a test upon the scores of a second testing;
- *Instrumentation*—changes in the observers or scorers may produce changes in the obtained measurements;
- *Statistical regression to the mean*—operating where groups have been selected on the basis of their extreme scores;
- *Selection*—biases resulting from differential selection of respondents for experimental and comparison groups;
- *Experimental mortality*—differential loss of respondents for the comparison groups; and
- *Selection-maturation*—and similar interaction effects.

Each of these sources of bias seems more or less plausible in the context of EPSDT impact analysis, with one significant exception. "Testing" cannot be considered as a threat to the internal validity of a research design in which the measurement instrument (the EPSDT screen) *is* the experimental treatment.

Random assignment of subjects to experimental and control conditions may be the best design with which to assess the impact of social programs, but this decisive, single-experiment approach is not feasible in most field settings (Cook and Campbell, 1979). Without randomizing, quasi-experimental research designs can only partially resolve the issues of threats to internal validity. Fortunately, in contrast to conducting research in an experimental environment, conducting research in an operational environment offers a high potential for findings which can be readily applied or generalized to other operating environments (that is, it offers high *external* validity).

The potential for application across contexts is crucial in situations where challenges to internal validity will be made. If the same impact is found in several contexts or programs which are each subject to different uncontrolled variable effects, then the evidence for the single explanation of program impacts grows "just because the plausible rival hypothesis is different from study to study" (Campbell and Stanley, 1963: p. 36). *This inferential feature is deliberately introduced in the following design.* The design also attempts to provide not only a way around the ethical impasse to human experimentation by denial of indicated treatment, but also to demonstrate a cost-effective and unobtrusive mode of empirical research in preventive child health services.

Comparison Groups

Overall, the research design is comparative, with observations made both cross-sectionally (to measure *differences*) and longitudinally (to estimate *changes*). The groups to be compared are a two-screen sample (the experimental group), and two other samples of initial screens approximately two years apart.

Using the Campbell and Stanley diagrammatic convention, the research model may be formulated as follows:

Two-Screen Experimental Group	01	X	02
One-Screen Control Group (1977)	03		
One-Screen Control Group (1979)			04

The "0" stands for an observation at a given time; the "X" stands for "experimental treatment." Technically, the observation (assessment) is the EPSDT health screen, and the treatment is the referral. The EPSDT assessment and referral are only conceptually separable, however, since our concern is the outcome of exposure to an EPSDT screen. The comparison here is between those who have had this exposure previously and those who have not.

Sample of Rescreened Children

For a child to be included in the two-screen sample for this analysis, several criteria have to be met: (1) the child's first screen occurred between July 1, 1977, and December 31, 1977; (2) the child was at least 19 months of age at the time of the initial screen; and (3) the time interval from first to second screen is at least 18 (and up to 32) months. (After the restrictions to the *sampling frame* were imposed, only 1831 children qualified, so all were included in the two-screen group. Thus, these children do not constitute a simple random sample, although this virtual enumeration is no less representative than a random sample.)

Criterion 1 insures first that these children all represent first-to-second rescreens rather than second-to-third, and so forth. Using a more homogeneous sample should improve the precision of our estimates of risk reduction. The beginning of the six-month time frame selected for the first screen corresponds to an internal change in the data collection system (in July, 1977) that changed certain data items and codes.

Criterion 2 insures that we have *not* selected children who are 18 months of age or less, and consequently, are on a special infant rescreening schedule. These children receive a somewhat different series of tests, and are permitted to receive six screens from birth through 18 months. Our interest in this exercise is focused on children who are eligible for the two year schedule.

Criterion 3 insures that the two year schedule has been followed at least approximately. Again, this criterion should help provide a more homogeneous sample.

Control Groups

The *first* control group (the 03 sample of the preceding diagram) consists of a "sample" of all children receiving their first screen at about the same time as the two-screen sample in August, 1977, and who were also at least 19 months old (n = 2889). This group overlaps with children in the 01-02 sample—those of the two-screen group receiving their initial screen in August, 1977. Some of the comparisons we made assume independence of samples and require the exclusion of this overlap; we shall refer to the restricted 03 control group as 03', n = 2578. (Again, for both control groups, limitations to the sampling frame, cost, and convenience led us to include all who qualified in these nevertheless representative groups.)

The *second* control group (04) consists of children in the same age range who received their first screen at about the same time as the two-screen sample is being rescreened, between January and April, 1979. We tried to further restrict this group to those children from families who were new to the program as of January, 1979, to ensure sample independence (n = 1183).

Operational Definition of Impact

A complete EPSDT screen in Pennsylvania includes as many as 39 individual test areas (TAs), organized here into six, mutually exclusive summary Diagnostic Areas (DAs): medical, vision, hearing, dental, laboratory, and behavior. (This study does not explicitly consider assessment of the self-reported immunization status, which is also part of the screen.) Some of these TAs may not be assessed in every screen because they are discretionary or are not appropriate for a given child.

The response categories are not identical for all TAs, but the alternatives provided for the *medical* evaluations are typical:

Code	Response
1	Normal
2	Abnormal, treatment required
3	Abnormal, non-treatable
4	Not done

For the purpose of this analysis, only Code 2 (abnormal, treatment required) is taken as evidence of a "serious" problem. Whether the child is *referred* or not apparently indicates a provider's willingness, specialization, or experience in treating the condition rather than a medical judgment or assessment of the child's health status. Codes 3 (abnormal, non-treatable) and 4 (not assessed) are counted as missing values.

For the longitudinal and cross-sectional comparisons required in the present study, we operationally define Health Status (HS) in terms of the total *number of abnormalities* that require treatment. Again, abnormalities that are identified as not requiring care or as not treatable (a relatively rare occurrence) are excluded. To standardize HS for comparisons among different groups of children, the total number of Abnormal Test Areas (ATAs) found is divided by the *total number of test areas that are assessed* for a group of children. A *low* score on this index therefore indicates a *healthy* individual or group; a higher score indicates a relatively higher incidence of illness or injury.

The measure of HS is calculated according to the following formula and can be expressed either as a proportion or a percentage:

$$HS = \frac{ATA}{TTA - NA},$$

where, for a specified group of children,

- HS = health status index;
- ATA = total number of *abnormal* test areas where treatment is required;
- TTA = TAs x S, the number of test areas in which a treatable abnormality can be found (= 39) times the number of children screened;
- NA = total number of test areas not assessed, an adjustment for a given group of children to eliminate TAs not assessed.

With a slight modification of this formula, we can also confine the focus of the measure to a particular Diagnostic Area (DA); that is, any specified selection of TAs. For example, if we wish to consider only conditions within the dental DA, then the potential number of TAs is limited to three possible tests: caries, oral infections, and other dental problems. When a particular subset of TAs (<39) is selected for calculating the index, we shall refer to it as an *Abnormality Rate* (AR).

The HS index can be interpreted as the probability that the EPSDT screen will uncover a condition requiring treatment. Similarly, an AR for a specified DA can be interpreted as the probability that a specific *kind* of problem will be identified.

Adjustment Estimation Procedure

Before controlling for the relevant threats to internal validity, there seems to be virtually no change in HS for the longitutinal sample, between screen and rescreen, and no *difference* in HS for the cross-sectional samples, between rescreen and control group. For the unadjusted longitudinal sample, 2.61 percent of all TAs assessed on initial screening in 1977 were abnormal with a condition requiring treatment; that percentage appears to have "changed" by 1979 to 2.69 percent on rescreening. This latter health status level is also compared to 2.58 percent for the control group of children receiving their initial screen in 1979.

To conclude at this point that there is evidently no impact would be premature. Particularly in light of the need to control for alternative explanations, we maintain that these unadjusted comparisons are invalid: the groups are not yet comparable.

In the sections that follow, we describe in some detail the means of estimating the needed adjustments and then compare the groups again after adjustment; actual adjustments are calculated in the Findings and Conclusion Section. The procedures for estimating and adjusting may seem complex, but the conclusions about EPSDT impact on health status that we can make at that point will differ significantly from the preceding conclusions based on unadjusted HS. While not all sources of incomparability of groups can be controlled in each comparison made, by integrating two modes of comparative analysis—longitudinal comparison of change with cross-sectional comparison of difference—we can gain a new perspective that is essential for evaluating impact based on this administrative data base.

After estimating the magnitudes and directions of each of the adjustments needed, the adjustment procedure is: 1) ascertain the percent change or differences due to factors other than the program—the threats to internal validity; and 2) add or subtract this percentage to/from the percentage difference or change in the unadjusted HS rates. For example, if we found a 10 percent *increase* in abnormalities due to instrument changes, we would *subtract* this figure from the percent change in HS for the longitudinal comparison between screen and rescreen. The effects of this procedure will be tabulated both for the overall HS index and the selected DAs.

Instrument and History Effects

To gauge changes in health status over time, the first requisite is that the assessment (instrument) remain constant. A less rigorous procedure at rescreen than at initial screen, for example, would probably result in fewer abnormalities being found and could mislead us into accepting an apparent improved health status which is actually due to instrument change. When there are events extraneous to the EPSDT program, which clients commonly experience between screenings, these events could alter health status (history). The introduction between screens of a new child health initiative for Medicaid-eligible children (for example, an intensified nutrition or immunization program within the public schools) could mislead us into attributing improvements in health status to EPSDT rather than to history.

Instrument or history effects on health status assessment are comparable to the effect of inflation on income: health status may appear to improve across time only because its measurement is less rigorous. As with inflation, we can statistically control for differences due to changes in instrument or history, but only if they can be identified and isolated from changes attributable to the program.

A test of the instrument/history hypothesis can be made by comparing the two control samples of initial screen results—in terms of the diagrammatic convention, the 03-04 comparison. Since the 03 and 04 control groups are independent and representative groups of children being screened at two times, differences between them can be attributed to instrument or history effects. Those observed differences that are also *confirmed by program history* as resulting from protocol or programmatic changes, can be used to estimate the magnitude of the effect and to adjust apparent changes in the rescreened (that is, experimental) group.

Of the 39 idividual comparisons included in the HS index, 9 revealed differences in ARs that are significant ($P<0.05$). In general, we found *higher* ARs in the later (04) group; this finding suggests that *the screening protocol may have become more rigorous over time.*

EPSDT program administrators have offered plausible explanations for virtually all of the observed differences. Administrative personnel attribute the differences in ARs to changes in EPSDT procedures, or to changes in the Medicaid service delivery system that have occurred in the two years between the screens. Thus, examination of independent groups from Time One and Time Two suggests that factors relating to *instrument* and *history* may indeed threaten the internal validity of our longitudinal comparisons. (See Table 4.)

TABLE 4
Selected First Screen Abnormality Rates (ARs) in 1977 (03) and 1979 (04)

	03		04		
DA (TAs included)	Total Tests Done	AR Percent	AR Percent	Total Tests Done	p-value
HS (39)	83608	2.13	2.58	34407	0.00
Medical (13)	37542	1.30	1.96	15375	0.00
Vision (2)	4508	9.05	9.96	1969	0.26
Hearing (2)	4502	1.38	1.99	1956	0.09
Dental (3)	8667	7.03	6.26	3549	0.12
Laboratory (10)	12635	0.87	1.25	5436	0.02
Behavoir (9)	15754	0.67	1.00	6122	0.02

Note: Indicated p-values are the two-tailed significance levels resulting from a difference-of-proportions test.

Table 4 presents the HS index and selected ARs for the 1977 and 1979 cross-sectional control groups (03 and 04, respectively). The overall HS in the 1979 group is approximately *21 percent higher* than the corresponding rate in the 1977 sample. This estimate of instrument effect will be used to adjust the percent change in HS between 01 and 02. (ARs are adjusted similarly according to their respective 03-04 differences.)

[Because the age distributions of the 03 and 04 groups are not exactly the same, AR differences between the groups might be attributable to age differences rather than instrument changes. As an additional check, we compared the ARs in two large cross-sectional groups which, while not carefully selected samples of first screens, are already age-comparable. Using administrative statistical reports for July through December, 1977 (covering 36,415 screens), and for October, 1979 (4,424 screens), we found differences similar to each of those reported here.]

The estimated difference may reflect increasingly rigorous quality control and standardization of tests (for example, "not assessed" vision TAs decreased from about 22 percent in 1977 to 17 percent in 1979). In addition to the HS adjustment, specific TAs vary in the amount and significance of difference. Therefore, any screen-rescreen change in HS or ARs (except Dental) toward *improvement* will be *under*estimated and therefore should be increased by the same amount. (Dental AR change will be overestimated because the probability of finding a dental abnormality *decreased*, although not significantly.)

Selection and Statistical Regression Effects

In gauging impact by cross-sectional comparison of rescreen results to the initial screen results of a control group, we realize that observed differences in health status could also be due to differential "selection" or "statistical regression." If those who received two screenings represent another, perhaps healthier, population, selection bias or statistical regression to the mean may explain any cross-sectional differences.

A test of this alternate hypothesis that we can do with available data involves comparison of the longitudinal sample *at initial screen* to others being screened initially at that time; that is, the 03' control group. If those children who eventually received rescreening are significantly different (for example, healthier) from others at the outset, a selection bias is indicated. The 01-03' differences can be used to estimate the magnitude of the bias and to adjust the cross-sectional comparison of the rescreen results (02) and control group results (04). If the rescreened group enjoys the same health status as others being initally screened at that time, we know that the rescreened children did not have a head start.

Unfortunately, since there could be no random assignment from all children who were screened once to the group of children who were rescreened, rescreening may itself be taken as evidence that the children are generally different (that is, extreme) in terms of health, specifically in tendency to improve, a selection-maturation interaction (Campbell and Stanley, 1963: p. 5).

Tendency to improve cannot be isolated easily using a longitudinal control group, observed at both Time One and Time Two but not treated, because the health assessment (or experimental *observation*) is the "treatment." Still, by extending the comparison at Time One of the inital screen children (01-03') to non-health factors (for example, demographic comparisons of available data) we can further support claims that the rescreen group is equivalent to the control group at initial observation (Campbell and Stanley, p. 59).

Receptiveness to intervention is not necessarily discernable from their initial HS; but some exogenous socio-demographic factors might present children with treatable abnormalities an environment that potentiates the effects of the EPSDT intervention. In other words, the chances that the indicated regimen will be followed may be higher in certain households and certain classes of children than in others.

Table 5 displays the summary DAs for 01 and 03', including general HS.

TABLE 5
Selected First Screen Abnormality Rates (ARs) in Two 1977 Samples

	01		03'		
DA (TAs included)	Total Tests Done	AR Percent	AR Percent	Total Tests Done	P-value
HS (39)	53,446	2.61	2.07	97,929	0.00
Medical (13)	23,801	1.58	1.27	33,500	0.00
Vision (2)	2,975	11.97	9.03	4,010	0.00
Hearing (2)	2,975	1.68	1.30	3,997	0.20
Dental (3)	5,493	8.47	6.66	7,734	0.00
Laboratory (10)	8,392	0.94	0.84	11,279	0.46
Behavior (9)	9,807	0.71	0.68	14,009	0.78

Note: 03' consists of all August, 1977, initial screens *excluding* those belonging to 01 sample. Indicated p-values are the two-tailed significance levels resulting from a difference-in-proportions test.

Generally, the rescreen group has a significantly ($p<0.0001$) higher abnormality rate at first screen (21.6 percent higher) than does the control group. The medical, vision, and dental DAs are the chief contributors to this difference.

Rather than enjoying a head start compared to others receiving their initial screening at the same time, the longitudinal sample appears to have had proportionately more serious problems. There are several possible reasons for this selection effect. One is that the identification of risk increases the probability that a child will be rescreened (that is, the presence of a treatable abnormality on initial screen co-occurs with an increased incidence of rescreening of almost 29 percent). While almost 60 percent of the 03′ group were known to be eligible for EPSDT through May, 1979, there is no record that any of them was recreened by March, 1980. Based on the entire 03 group of 2889 children who were screened for the first time in August, 1977, 1428 had problems and 1461 did not. Of those with problems, 173 or 12.1 percent were rescreened compared to 138 or 9.4 percent of those without problems. The association between whether or not a child has a problem and whether or not the child will be rescreened is significant (chi-square = 5.32, d.f. = 1, $p<0.05$).

Another possible reason for the higher 01 ARs is seasonality, or the fact that the longitudinal sample is drawn from a six-month sampling frame as compared to one month (August) for the comparison group. Other possible reasons for 01-03′ differences may lie in the differences in their age distributions (chi-square = 82.9, d.f. = 17, $p<0.05$) and in differences in the distributions of the providers who screened them (chi-square = 323.05, d.f. = 74, $p<0.05$). [Both chi-square statistics for age and provider differences are calculated after excluding categories with expected frequencies less than 5.]

To test the possibility of a selection-maturation interaction or *tendency* to improve, we also compared the two groups (01-03′) on a series of demographic variables. In terms of sex, race, and whether the child is from Philadelphia or the suburban counties, there are no significant differences. In terms of payment category ("categorically needy" or "medically needy" versus "general assistance"), household size, and age distributions, the differences are statistically significant ($p<0.05$) but otherwise very small. (See Figures 1A and 1B).) [Other socioeconomic comparisons can be made of measures derived from 1970 Census tract-level data for most of the children in the samples. Given the nature of the data, these comparisons are extremely tentative, but they may provide a useful sense of environmental factos. The average (mean) median family income differs significantly between groups ($6337 for 01 and $6606 for 03′, $p<0.05$); while the mean percent living in crowded housing does not differ significantly (18.1 percent and 19.8 percent, respectively).]

Generally, the demographic differences described are interesting and suggestive but, in terms of practical significance in support of the selection-maturation interaction hypothesis or as a source for an adjustment to the longitudinal comparison, they are too tentative. Moreover, we note that the coeval cross-sectional comparison does not suffer from interaction threat to validity, as does the longitudinal comparison. If the results of both comparisons are the same, after other threats are controlled, tendency to improve would lack support as an alternate hypothesis.

Maturation Effects

Another of the plausible alternative explanations for observed longitudinal changes is maturation. Campbell and Stanley (1963: p. 5) define maturation as "processes within the respondents operating as a function of time." As such, maturation is a form of *change*, and can affect only the longitudinal (01-02) comparison. Inferences drawn from cross-sectional comparisons (02-04), on the other hand, may be invalidated by age *differences* between the experimental and control groups. An age-adjustment procedure must therefore be used to control for *both* maturation and age differences.

The mechanics of age adjusting involve a procedure similar to deriving standardized specific rates (for example, age-specific birth or death rate). The age standardization recalculates HS (or specific ARs) as if the two groups were identically composed regarding age. The formula is simply a weighted mean that expresses the age-specific HS of the two first screen groups (01 and 04) weighted by the age distribution of the rescreened children (02) (Mueller, Schuessler, and Costner, 1977: p. 134). If there are no age differences between any two comparison groups, then standardization will not change the HS, although the converse is not necessarily true.

Since no child in the rescreened group is under three years of age when rescreened, and none is over 18 in the first screen groups, it is important to note that standardization will limit the comparisons to only those children 3 through 18 years of age (Figures 2A-C).

FIGURE 1A
Demographic Comparisons at First Screen Between Those Who Received a Rescreen (01) and Those Who Did Not (03').

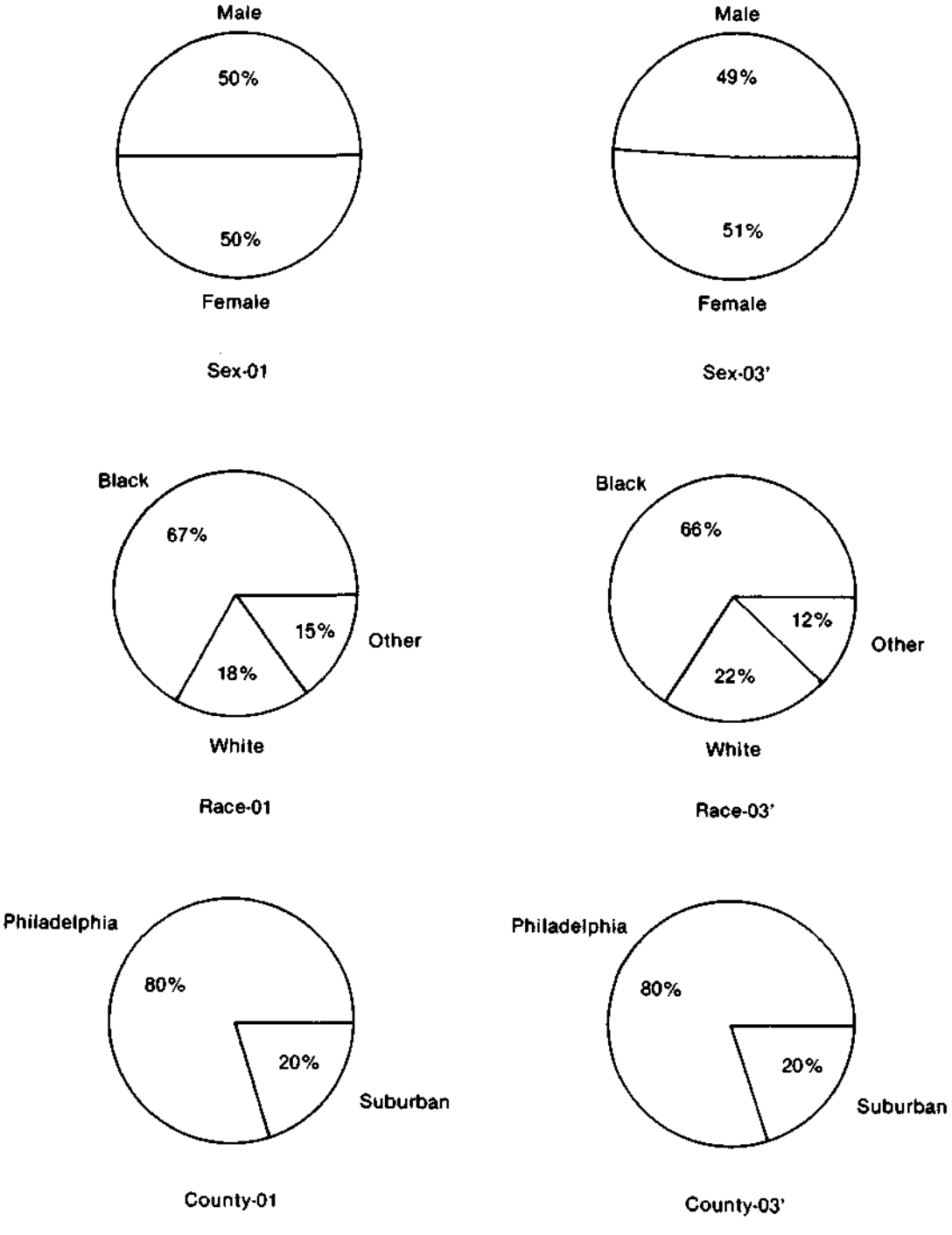


FIGURE 1B

Demographic Comparisons at First Screen Between Those Who Received a Rescreen (01) and Those Who Did Not (03').

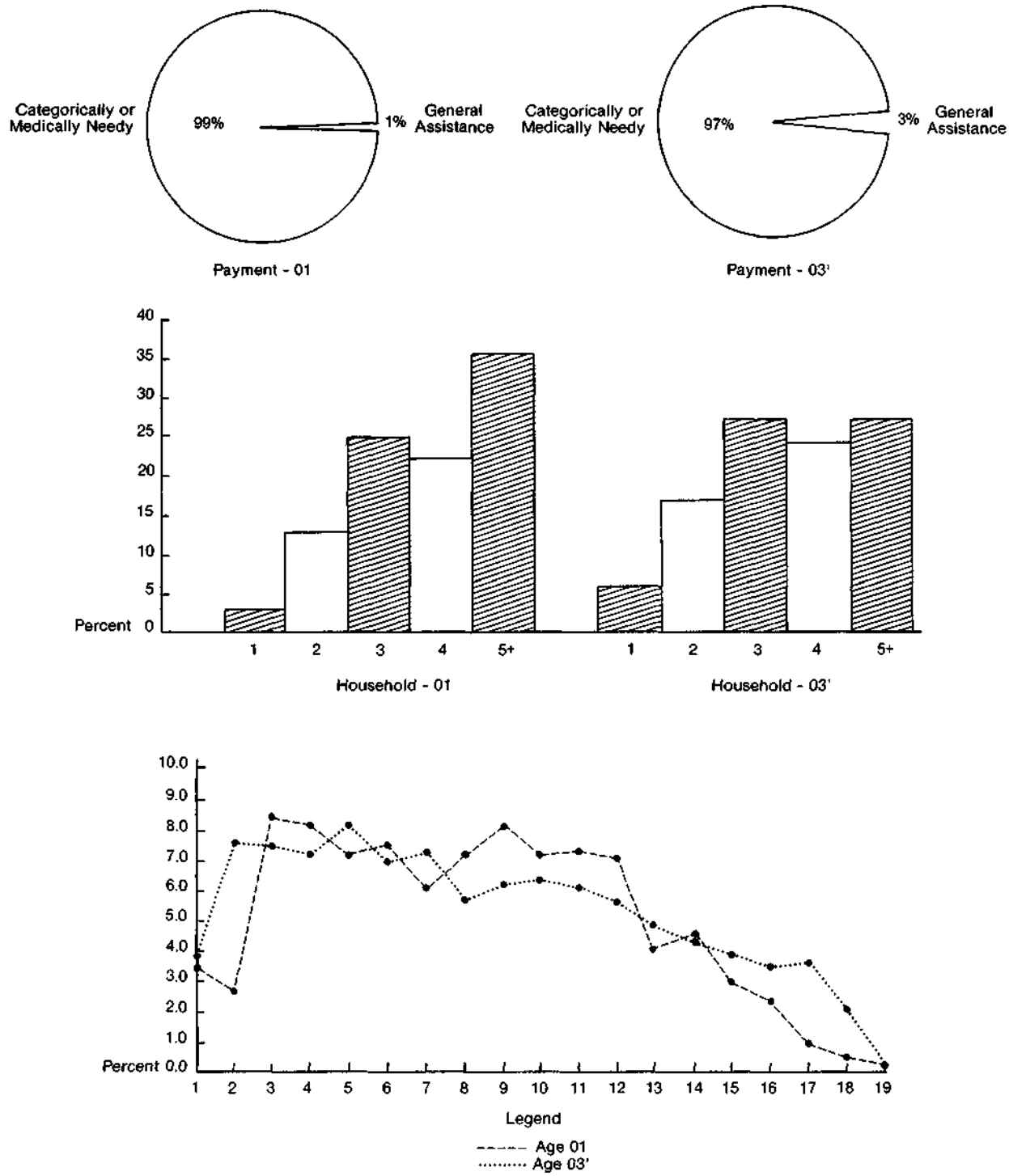
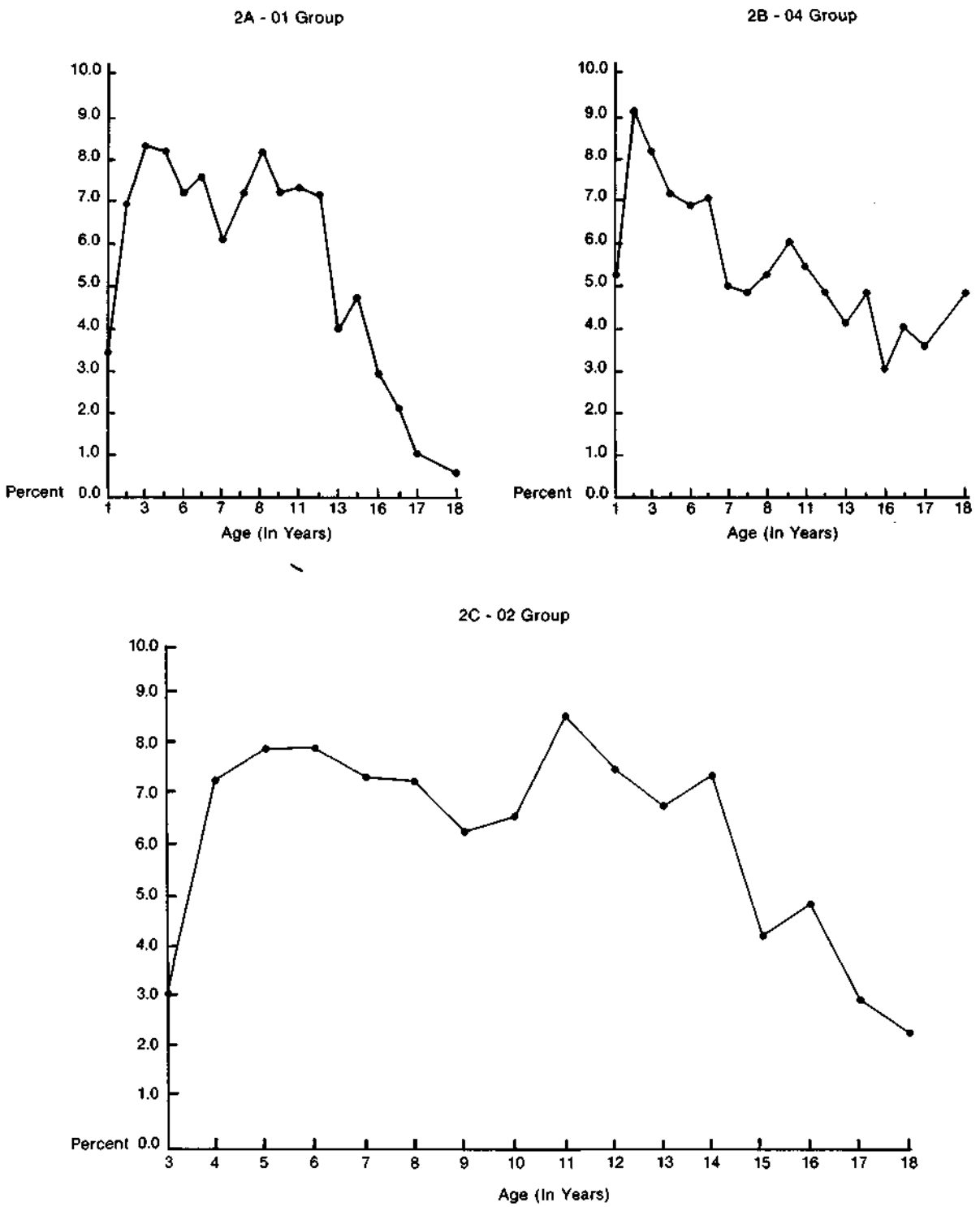


FIGURE 2
Age Distributions Among the Comparison Groups



Experimental Mortality

The validity of comparing rescreened children's health status to that of children just receiving their initial screens (02-04) is also threatened by the fact that, if the program's future emulates its past, most of the children in the comparison group (04) will not be rescreened. Since there are differences at first screen between those children who are eventually rescreened and those who are not (the selection effect), we may therefore suspect that, had all first screen children received a second health assessment, our estimation of EPSDT impact would be different. The differential loss of children from the time series implies that the observed cross-sectional 02-04 differences may be due to "experimental mortality" rather than to EPSDT.

Longitudinal comparisons (01-02), on the other hand, do not suffer from the effects of differential selection, since the same group is compared across time. Therefore, although this research design requires both cross-sectional and longitudinal comparisons, only the inferences from cross-sectional data are subject to the challenge of experimental mortality.

The plausibility of experimental mortality as an alternate explanation for observed cross-sectional differences might be tested by locating and screening those children who were not yet rescreened. If this new estimate of health status for rescreened children does *not* differ from that of the previous rescreened group, the hypothesis of experimental mortality could not be supported. Even if the test were feasible, however, the resources required to locate and rescreen these children would be prohibitive.

Given the practical difficulties of eliminating experimental mortality as a rival explanation for cross-sectional differences, it is appropriate to ask what, if any, effects this factor could possibly have on the estimation of impact in this study? The evidence suggests that most of the arguments that come to mind are in the wrong direction. For example, we might assume that those who drop out will be less healthy because they are likely to be less conscientious about their health and medical care than are those who stay. In fact, as Table 5 illustrates, the opposite is true—the "experimental group" is *less* healthy at initial screen than is the control group.

There does not seem to be a feasible way to control for experimental mortality in the cross-sectional (02-04) comparison. We can, however, again use the inferential feature of this research design to combine the results of cross-sectional and longitudinal comparisons. If the longitudinal (01-02) comparison, which is unaffected by experimental mortality, discloses the same impact after all other threats have been controlled, then there would be no empirical evidence that the EPSDT program impact hypothesis is less plausible. To the extent that they differ, experimental mortality may be the source.

Findings

Age Adjustment

Table 6 shows fairly consistent increases in 01 and 04 ARs when we make the age adjustment. The difference between each group's crude and standardized health status must be attributed to the differences between age distributions compared to the two-screen group (02) at rescreening: for the 01 group, there is about 8 percent difference between its unstandardized and standardized HS (2.61 percent versus 2.83 percent) and 6 percent difference for the 04 group (2.58 percent versus 2.74 percent). In both comparisons, *the effect of age standardization is to improve the experimental* (02) *group's health status relative to those of the comparison groups.* In other words, compared to children at rescreening (02), their own health status at first screen (01) and that of a control group being screened for the first time (04) appear to be slightly worse due (first) to the favorable age distributions of the first screen groups.

Longitudinal Adjustment

After the age distributions of the control groups have been standardized (indicated by an asterisk), further adjustments are needed: observed *changes* (01*-02) must be adjusted to control for the effects of increased screening rigor, and cross-sectional *differences* (02-04*) must be adjusted to control for the effects of self-selection for rescreening. Differences that are observed in both the longitudinal and cross-sectional comparisons of adjusted abnormality rates may then be interpreted as evidence of program impact.

The adjustment procedure applied to instrument changes is simply to calculate the percent change from 01* to 02 and then to subtract from that the percent "change" from 03 to 04, due to instrument/history change. The adjustment factors are based on estimates from Table 4 and are signed—positive if more abnormalities are being found in 1979 than in 1977, and negative if fewer abnormalities are found (Table 7). Similarly, the signs of the unadjusted and adjusted percent changes indicate the direction of change in ARs before and after adjusting for instrument effect—a negative change indicates reduced abnormalities.

Before the adjustment for instrument effect, there is evidence (Table 6) to support claims of some improvement in health status between age-comparable screen and rescreen results across time. When we adjusted for the increased rigor of the 1979 screening protocol, compared to that in 1977, we find the abnormality rate in 1979 to be about 27 percent lower. Thus, in terms of the original HS index for the two-screen group, in 1977, on an average, *26.1* abnormalities were found per 1000 tests; for the same children in 1979 *19.0* abnormalities per 1000 tests were found.

Table 7 also allows us to estimate relative impact in different diagnostic areas, although these estimates are based on differing numbers of tests and are therefore of variable reliability. Even after adjusting for age and instrument changes, the greatest improvement appears in the behavoral area, with an almost 60 percent lower AR.

TABLE 6
Comparison of Abnormality Rates for 02 with 01 and 04 Before and After Age-Adjustment

DA	AR Percent				
	02	01	01*	04	04*
HS	2.66	2.61	→ 2.83	→ 2.58	2.74
Medical	1.98	1.58	→ 1.55	→ 1.96	1.82
Vision	11.95	12.00	→ 13.13	→ 9.96	11.01
Hearing	2.15	1.68	→ 1.79	→ 1.99	2.09
Dental	7.13	7.57	→ 9.79	→ 6.26	7.58
Laboratory	1.06	0.94	→ 0.85	→ 1.25	1.08
Behavior	0.63	0.71	→ 0.70	→ 1.00	1.05

Note: Asterisk indicates an age-adjusted rate.

TABLE 7
Longitudinal *Changes* Adjusted for Instrument/History Effects

DA	Percent			
	(01* — 02 Change)	—	(03 — 04 Adjustment Factor)	= (Adjusted 01* — 02 Change)
HS	- 6.01		+21.13**	-27.14
Medical	+27.74		+50.77**	-23.03
Vision	- 8.99		+10.06	-19.05
Hearing	+20.11		+44.20	-24.09
Dental	-27.17		-10.95	-16.22
Laboratory	+24.71		+43.68**	-18.97
Behavior	-10.00		+49.25**	-59.25

Note: Asterisk indicates an age-adjusted rate. Two asterisks indicate adjustment factors based on significant 03—04 differences ($p<0.05$). (See Table 4.)

Cross-Sectional Adjustments

Since cross-sectional comparisons of rescreen results (02) to those of an age-adjusted control group (04*) are subject to selection/regression effects, the percent differences in rates between 02 and 04* are adjusted in Table 8 based on estimates from the 01-03' comparisons in Table 5. The sign of the adjustment factor is positive for all DAs, indicating that the two-screen sample had, on the average, higher abnormality rates on first screen than others receiving first screens at the same time. We must recognize that having a significantly sicker group of children for the two-screen group (at Time One) could mean that the observed differences between 02 and 04* are at least partly due to statistical regression, in spite of the control. Campbell and Stanley suggest, however, that if the experimental group is self-selected—such as the two-screen group—the problem of uniform regression effects between experimental and control group becomes less likely, but interaction effects are more likely (1963: p. 50).

The adjustment procedure summarized in Table 8 parallels the one used for the longitudinal comparison. In this procedure, the percent AR *difference* between 01 and 03' is subtracted from the percent difference between 04* and 02. As in the longitudinal comparison, the negative signs of several 04*-02 differences, including the overall HS, indicate that the rescreened children are slightly healthier than those of similar age just entering the system in 1979. The estimated adjustment for selection/regression results in the rescreened group showed about 29 percent fewer abnormalities than the age-adjusted comparison group at first screen.

Conclusions

The similarity of results of the two modes of comparison is more than accidental—the rescreened group has between a 27 and 29 percent lower overall abnormality rate, whether compared to itself over time or to a control group.

It is still possible for either one of these differences, in isolation, to be explained by rival hypotheses. The longitudinal difference could be due to uncontrolled statistical regression or selection-maturation (and other) interaction effects; the cross-sectional difference may be attributable to experimental mortality. But the plausible rival hypotheses are different in the two comparisons. The research design included this inferential feature of changing contexts deliberately to enable us to isolate and guage the impact of the EPSDT program based on secondary data analysis.

TABLE 8
Cross-Sectional *Differences* Adjusted for Selection/Regression Effects

DA	Percent (04* — 02 Difference)	— (01 — 03 Adjustment Factor)	= (Adjusted 04* — 03 Difference)
HS	- 2.92	+26.09**	-29.01
Medical	+ 8.79	+24.41**	-15.62
Vision	+ 8.54	+32.56**	-24.02
Hearing	+ 2.87	+29.23	-26.36
Dental	- 5.94	+27.18**	-33.12
Laboratory	1.85	+11.90	-13.75
Behavior	-40.00	+ 4.41	-44.41

Note: Asterisk indicates an age-adjusted rate.
Two asterisks indicate adjustment factors based on significant 01—03' differences ($p<0.50$). (See Table 5.)

Summary

The findings presented here attest to the beneficial effect of EPSDT on the health status of the children served. Although narrowly focused, our conclusion is definitive and unequivocal—periodic screening is associated with a decrease in the incidence of abnormalities requiring care.

The results of this study do not permit a definitive statement on the ultimate impact of EPSDT, and certainly cannot be construed as a blanket endorsement of screening programs in general. We have, on the contrary, deliberately restricted the operational definition of "impact." Our concern is with health status impact, and not with such equally important outcome measures as equity, cost-benefit, or client satisfaction. Furthermore, the term "health status" is used exclusively in terms of abnormality rates, and does not measure level of functioning, risk status, or any of the other significant dimensions of well-being. Health status (and, ultimately, impact) can be measured only by isolating a set of very specific variables on which data are available and from which valid conclusions may be drawn. Consequently, the results of this analysis cannot be interpreted as an estimate of the only, or even of the most important, impact of the EPSDT program. Also, these results do not permit an evaluation of the extent to which EPSDT has achieved its ultimate goal of decreasing overall dependency in the target population. In fact, such an evaluation may never be possible for a program with ultimate objectives as complex and comprehensive as those of EPSDT. Recognition of these limitations in no way diminishes the significance of specific impact that has been identified and documented.

The credibility of any impact estimate depends, to a great extent, on the power of the research design to eliminate competing explanations for the observed phenomena. Thus, the distinctive methodology employed in this study suggests policy implications which are no less important than those suggested by the substantive conclusions. Specifically, the value and feasibility of secondary data analysis as a method for evaluating program impact has been convincingly demonstrated.

However, it is important to recognize that such secondary analysis is not easily accomplished. Our initial attempts to measure impact produced ambiguous results—no significant change in risk status occurred between first screen and rescreen, despite clear evidence of problem resolution. The actual changes in health status were revealed only after a complex series of mathematical adjustments.

In spite of its complexity, the research design employed here offers several distinct advantages. Using a quasi-experimental design, we can estimate impact without incurring the cost of primary data collection; in addition, such a design avoids the ethical difficulties of selective denial of treatment. Because it is both cost-effective and unobtrusive, the methodology used in this study may be applicable to other operating environments. Through the use of a similar secondary data analysis, a national EPSDT monitoring system could be developed, and the impact of screening on the health status of children in a variety of environments could be estimated.

EPSDT, like other public services programs, presently is in an environment of both scarce and shrinking resources and a constant or increasing demand. Estimating the impact of such a program on those it serves is therefore an essential first step to improve the efficiency of that program by maintaining or increasing the level of services, but at a lower cost.

References

Campbell, D.T., and J.C. Stanley. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand-McNally, 1963.

Clearinghouse on Corporate Social Responsibility, American Council of Life Insurance and the Health Insurance Association of America. "Fact Sheet: Lifecycle Preventive Health Study," Washington, D.C., July, 1980.

Consumers Union. "Those Costly Annual Physicals: Are they Really Worthwhile?" *Consumer Reports* 45:601-607, October, 1980.

Cook, T.D., and D.T. Campbell. *Quasi-Experimental: Design and Analysis Issues for Field Settings.* Chicago: Rand-McNally, 1979.

Eisenberg, L. "The Perils of Prevention, A Cautionary Note," *New England Journal of Medicine* 297(22):1230-1232, December, 1977.

Foltz, A.M. "The Development of Ambiguous Federal Policy: EPSDT." *Milbank Memorial Fund Quarterly* 53:35-74, 1975.

Frankenberg, W.K. "Pediatric Screening," *Advances in Pediatrics* 20:149-175, 1973.

Gibson, Robert M. "National Health Expenditures, 1978," *Health Care Financing Review* 1:1-36, Summer, 1979.

Gibson, Robert M. "National Health Expenditures, 1979," *Health Care Financing Review* 2:1-36, Summer, 1980.

Kristin, M.M., and C.B. Arnold. "A Mini Cost-Effectiveness Analysis of Mamography as a Procedure for the Mass Screening of Asymptomatic Populations." Presented at the 54th Annual Western Economics Association Conference. San Diego, CA, June, 1980.

Lieberson, S., and L.K. Hansen. "National Development, Mother Tongue Diversity, and the Comparative Study of Nations." *American Sociological Review* 39:523-541, 1974.

Mechanic, D. "Approaches to Controlling the Cost of Medical Care: Short-Range and Long-Range Alternatives," *New England Journal of Medicine* 298:249-254, 1978.

Miller, John C. "Super Salesmen Make EPSDT Work in Pennsylvania," Health Care Financing Administration *Forum* 2:15-18, 1978.

Mueller, J.H., K.F. Schuessler, and H.L. Costner. *Statistical Reasoning in Sociology.* Boston: Houghton-Mifflin Company, 1977.

Public Health Service, United States Department of Health, Education and Welfare. *Health United States 1979.* Washington, D.C.: USGPO.DHEW Publication No. (PHS)80-1232.

Shadish, W.R. *Effectiveness of Preventive Child Health Care,* Working Paper #38, Department of Psychology, Evaluation and Methodology and Center for Health Services and Policy Research. Northwestern University, Evanston, IL, 1980.

Spitzer, Walter O., and Bruce P. Brown. "Unanswered Questions about the Periodic Health Examination," *Annuals of Internal Medicine* 83:257-263, 1975.