A RAND NOTE

PLAN FOR THE ANALYSIS OF DENTAL EXAMINATION DATA IN THE NATIONAL PREVENTIVE DENTISTRY DEMONSTRATION PROGRAM

Stephen P. Klein and Robert M. Bell

April 1981

N-1658-RWJ

Prepared For

The Robert Wood Johnson Foundation

Rand
SANTA MONICA, CA. 90406

This research is supported by Grant No. 4769 from the Robert Wood Johnson Foundation.

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

A RAND NOTE

PLAN FOR THE ANALYSIS OF DENTAL EXAMINATION DATA IN THE NATIONAL PREVENTIVE DENTISTRY DEMONSTRATION PROGRAM

Stephen P. Klein and Robert M. Bell

April 1981

N-1658-RWJ

Prepared For

The Robert Wood Johnson Foundation

Rand
SANTA MONICA, CA. 90406

PREFACE

The National Preventive Dentistry Demonstration Program is being conducted to determine the costs and benefits of various types and combinations of school-based preventive dental care procedures. This Program is supported by two separate grants from The Robert Wood Johnson Foundation of Princeton, New Jersey. One of these grants is to the American Fund for Dental Health (AFDH), a national, nonprofit organization dedicated to the support of dental education, service, and research. The other grant is to The Rand Corporation.

The initial concept for the Program was developed and presented to The Robert Wood Johnson Foundation by AFDH. Within the Program, AFDH is responsible for providing and supervising the preventive care, collecting the data, and conducting the annual dental examinations. Rand is responsible for monitoring these activities, developing the data collection forms and procedures, and conducting the data analyses. AFDH and Rand worked together in designing the Program, selecting the sites, and establishing the procedures for other areas of joint responsibility.

The Program began in 1976 and will run for a total of six and one-half years. The first year was devoted to planning, preparation, site selection, equipment development and procurement, and pilot testing. The sites were then activated on a phased schedule so that over the next four and one-half years each could provide preventive care for a full four years. The Program's final year will be spent analyzing data and reporting the results of these analyses. These data will include the results of dental examinations on approximately 25,000 children and information regarding the personnel, supplies, equipment, facilities, and time required to provide various kinds of school-based preventive dental care. The database also includes background information about the children at each site who did and did not participate in the Program.

The first chapter of this note provides an overview of the Program's design. The next chapter describes the variables that will be used to analyze the baseline and each succeeding year's dental examination data. The subsequent chapters describe the procedures that will be used to conduct these analyses, as well as those used to measure the degree to which the dental examiners agreed with themselves and each other in the application of the Program's diagnostic criteria.

The purpose of this note is to describe the factors that will have to be taken into consideration in analyzing the dental data and the methodology that will be employed in conducting these analyses. It is recognized, however, that as the analyses progress, additional factors may emerge that could lead to modifying the plans presented in this note.

A discussion of the procedures that will be used in analyzing the Program's cost data appears in A Cost Analysis Plan for the National Preventive Dentistry Demonstration Program, by Craig B. Foch, The Rand Corporation, N-1670-RWJ, 1981.

SUMMARY

The purpose of the National Preventive Dentistry Demonstration Program is to determine the costs and benefits of various types and combinations of school-based preventive dental care procedures. This note provides a description of the statistical methods that will be used to analyze the dental data collected in this Program.

The first chapter of the note provides an overview of the Program's design. It describes the sites, the longitudinal and cross-sectional samples of children involved, dental examination procedures, and treatment components. These components were organized into six treatment regimens. All six regimens were provided at each of the Program's ten sites. Children enrolled in the Program received a clinical and radiographic dental examination prior to the beginning of treatment. This was called the baseline examination. Children remaining in the Program to its completion will receive four annual examinations after baseline.

Chapter 2 describes the variables that will be used in analyzing the baseline and each succeeding year's dental examination data. The primary dependent variable for these analyses is the number of decayed, missing due to decay, and filled tooth surfaces per child (i.e., the child's DMFS score). Other dependent variables include the number of decayed, missing, and filled teeth (DMFT score); a measure of gingivitis (inflammation of the gums); and a general index of oral health. Independent and control variables include age, sex, race, socioeconomic status (SES), site, and treatment regimen.

Chapter 3 describes the procedures that will be used to analyze the baseline data. These analyses will focus on the relationships between DMFS scores and other variables. Most of this chapter is devoted to a discussion of the factors that have to be considered in analyzing the data and the methods for dealing with them. Some of these factors are the differences in sample sizes across sites, skewness of scores on the dependent variables, and nonlinearity of relationships. This chapter also includes a preliminary list of the tables that will be used to present the results of the baseline analyses.

Chapter 4 describes the procedures that will be used in analyzing the relative and absolute effectiveness of the treatment regimens and how this effectiveness is related to other variables, such as site, initial decay level, age, sex, and SES. The primary dependent variable for these analyses will be the increment in DMFS scores between examinations. This chapter also includes a discussion of the factors that have to be considered in analyzing increment scores, such as adjusting for differences in initial decay level and other imbalances between treatment regimens. The chapter concludes with a preliminary list of tables that will be used to present the results of the analyses of treatment effects.

The note's final chapter discusses the procedures that will be used to measure the degree to which the Program's 31 dental examiners agreed with themselves and each other in the application of the diagnostic criteria. Two general types of data will be used for this purpose: multiple examinations of the same child on the same day and the frequency of inconsistent diagnoses over time (e.g. a tooth being labeled as filled on one examination and free from decay on a subsequent examination).

ACKNOWLEDGMENTS

Many of our Rand and non-Rand colleagues contributed useful ideas and suggestions regarding the procedures that should be used in analyzing the Program's dental data. Particularly helpful comments were provided by Dr. Harry M. Bohannan, The American Fund for Dental Health; Craig B. Foch, The Rand Corporation; Prof. Richard C. Graves, College of Dentistry, University of Kentucky; Prof. Val Spolsky, School of Dentistry, University of California, Los Angeles; and Prof. John W. Stamm, School of Dentistry, McGill University. The authors also wish to thank John Rolph of The Rand Corporation for his thoughtful technical review.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS vii

TABLES x

Chapter
1. GENERAL PROGRAM DESIGN 1
Sites 1
Longitudinal and Cross-sectional Cohorts 1
Clinical and Radiographic Examinations 2
Treatment Components and Regimens 2
Assignment of Schools to Regimens 3
Site Activation and Termination 3

2. VARIABLES 6
Background Characteristics 6
Treatment Regimen 8
Clinical DMFS/DMFT Scores 9
Number of Teeth at Risk 10
Radiographic Examinations 10
Restorative Care Cost 11
Sealants and Fluorosis 12
DHC Index of Gingivitis 12
Referral Codes 13
Decision Rules for Retaining Children for Analysis . . . 13

3. ANALYSIS OF BASELINE DATA 14
Introduction 14
General Strategy 15
Analytic Considerations 16
Analysis of Covariance Model 17
List of Tables for Baseline Analysis 20

4. ANALYSIS OF TREATMENT EFFECTS 23
General Strategy 24
Analysis of Covariance Model 25
Estimation of a More Complete Model 30
List of Tables for Treatment Analysis 32

5. EXAMINER RELIABILITY 34
Introduction 34
Data Sources 35
Sources and Types of Unreliability 37
Clinical DMFS/DMFT Scores: Concurrent Indices of Reliability 38
Clinical DMFS/DMFT Scores: Longitudinal Index of Reliability 42

Radiograph Reader Reliability 42
DHC Scores 43
Miscellaneous Indices 43
Relationships Among Indices 44

Appendix
A. HANDLING THE LONG TAIL OF DMFS SCORES 45

B. EFFECT OF EXAMINER ERROR ON PRECISION OF THE ESTIMATES . . . 47

REFERENCES 51

TABLES

1.1 Sites Participating in the National Preventive Dentistry Demonstration Program 1

1.2 Description of Treatment Components 4

1.3 Organization of Treatment Components into Regimens Relative to Type of Water Supply 5

4.1 Information Obtained from Various Comparisons of Treatment Regimens 26

5.1 Data Sources for Examiner Reliability Analyses 35

5.2 Illustration of Procedures Used To Compute Reliability Indices 39

Chapter 1

GENERAL PROGRAM DESIGN

SITES

Two sites in each of five major regions of the United States were selected to participate in the Program. One of the two sites in each region had a nonfluoridated water supply (i.e., reported to be less than .2 ppm fluoride ion), while the water supply for the other site was considered optimal for that region (i.e., .8 to 1.0 ppm). Table 1.1 lists the two sites in each region.

Table 1.1

SITES PARTICIPATING IN THE NATIONAL PREVENTIVE DENTISTRY DEMONSTRATION PROGRAM

Region	Nonfluoridated Sites	Fluoridated Sites
Northeast	Billerica, MA	New York, NY
Southeast	Tallahassee, FL	Chattanooga, TN
Central	Wichita, KS	Minneapolis, MN
Southwest	Monroe, LA	El Paso, TX
Northwest	Pierce County, WA	Hayward, CA

Of the five fluoridated sites, only one (El Paso) has natural fluoridation. Analyses of water samples at each site have indicated that one site (Wichita) has a higher level of fluoride (.4 ppm) than was reported in its application to participate in the Program. There is also evidence that the amount of fluoride at the Hayward site has varied substantially over time.

LONGITUDINAL AND CROSS-SECTIONAL COHORTS

Enrollment in the Program was open to all children who, in the fall of 1977, were in grades 1, 2, and 5 at each site's participating schools. These children were designated as the Program's longitudinal cohorts in that they are scheduled to receive preventive care and/or a series of five annual dental examinations; i.e., one examination at the beginning of the Program (called the baseline examination or Exam #01) and one at the end of each of the following four years. About 82 percent of the eligible population received parental consent to enroll in the Program, resulting in a total of approximately 600 children per grade level per site.

Random samples of children who, in the fall of 1977, were attending grades 3, 4, 6, 7, and 8 in the same schools as the longitudinal cohorts were also examined during the baseline period. The primary purpose for examining these children was to develop benchmark data against which the longitudinal cohorts' data could be assessed in subsequent years. The children in these five grade levels were designated as cross-sectional cohorts. These children will not participate in any subsequent examination or preventive dental care activities. There were about 80 children per grade level per site in these cross-sectional cohorts.

CLINICAL AND RADIOGRAPHIC EXAMINATIONS

The children in the longitudinal and cross-sectional cohorts received clinical (i.e., visual-tactile) and radiographic (X-ray) dental examinations when the Program began at their site. The clinical examinations were conducted by six to eleven of the Program's cadre of 31 specially trained and calibrated dental examiners. These examinations were done to (1) assess the number of permanent tooth surfaces affected by decay (cavities) and (2) measure the amount of gingivitis (inflammation of the gums).

Approximately ten percent of the longitudinal cohort children received a second clinical examination during the baseline period. About one-half of these duplicate examinations were performed by the same dentist who had seen the child previously, while the other one-half were done by a different dentist. The purpose of these duplicate examinations was to determine examiner reliability, i.e., the degree to which the examiners agreed with themselves and with each other in the application of the Program's diagnostic criteria.

Radiographs of the children's posterior teeth (i.e., the ones most susceptible to tooth decay) were taken by trained radiographic technicians. A specially designed truck was used for this purpose in order to ensure adherence to strict safety requirements and to maintain standardization of procedures and equipment across sites.

Clinical examinations are scheduled to be repeated with the longitudinal cohort children at the end of each of the four years the Program operates at each site. Radiographic examinations are scheduled to be repeated for these same children only at the end of the fourth year. The one exception to this plan is at the New York site where the Program was discontinued after the children were examined (clinically and with radiographs) at the end of the third year.

TREATMENT COMPONENTS AND REGIMENS

Within each longitudinal cohort at each site, there are six groups of children. Five of these groups receive one or more preventive measures. All of these procedures have been proven safe, and clinical trials have shown that most are effective in reducing tooth decay. The children in the sixth group do not receive any of the preventive measures, but they are examined along with the other children in the Program.

Table 1.2 lists the various preventive techniques used in the Program and Table 1.3 indicates how these preventive measures were combined into treatment regimens. These combinations were chosen to provide information about the unique and combined effects of certain preventive measures as well as to test those combinations that were likely candidates for an operational school-based program. This latter consideration also led to the slight differences in the composition of the regimens in fluoridated and nonfluoridated sites.

ASSIGNMENT OF SCHOOLS TO REGIMENS

Schools, rather than individual children, were assigned to regimens. This procedure was adopted because certain preventive measures, such as bi-weekly toothbrushing, are most efficiently administered to children when they are in classroom groups. The assignments were made in a way that minimized differences in the number and characteristics of the children assigned to each regimen. Hand tallies of baseline dental examination results and the percent minority enrollment at each school were the major factors considered in the assignment process.

SITE ACTIVATION AND TERMINATION

Sites were phased into the Program on roughly a one-every-other-week basis (excluding vacation periods) starting in September 1977 and ending in February 1978. The first major activities at a site were those of securing parental consents and conducting examinations. Preventive care at a site was usually initiated about one month after that site's baseline examinations were completed.

One major exception to the foregoing procedures arose in connection with Minneapolis, where an unanticipated judicial desegregation order and a school board consolidation plan combined to threaten the loss of several schools from the Program. Moreover, these orders and plans were not made public until approximately three months after the baseline examinations were completed at this site. This situation led to the decision to add several more schools in Minneapolis, which in turn meant conducting a second wave of baseline examinations at this site during March of 1978.

The only other major exception to the procedures described above involved the New York site. Several factors, including the cost of operating this site, precluded extending its treatment regimens for a fourth year. Thus, the Program was discontinued at the New York site after its children were examined (clinically and with radiographs) at the end of its third full year of Program participation.

The Program will be terminated at each of the other nine sites on a schedule that is parallel to the sequence in which they were phased into the Program. This will permit each of these sites to provide its treatment regimens for four years, as well as maintain a 48 month interval between baseline and final examination cycles.

A complete description of the Program's procedures appears in The First Year of Field Activities in the National Preventive Dentistry Demonstration Program, by Stephen P. Klein and Harry M. Bohannan, The Rand Corporation, R-2536/1&2-RWJ, 1979.

Table 1.2

DESCRIPTION OF TREATMENT COMPONENTS

Treatment Component	Procedure and Rationale	Frequency of Treatment	Personnel Required
Fluoride paste prophylaxis	An acidulated phosphate fluoride (APF) paste is used in a professional cleaning of the child's teeth. This 30-minute procedure provides topical fluoride protection and prepares tooth surfaces for the acidulated phosphate fluoride gel treatment.	Twice per year.	Auxiliary personnel consistent with minimum state dental practice legislation.
Acidulated phosphate fluoride gel	Application of 1.23% fluoride ion gel in a styrofoam tray. This tray is kept in the mouth for a period of 4 minutes. Both arches are treated at same sitting. The gel treatment is applied immediately after the prophylaxis and provides topical protection to all tooth surfaces.	Twice per year in conjunction with the prophylaxis.	Auxiliary personnel consistent with minimum state dental practice legislation.
Sealant	A sealant is a plastic-like resin that adheres to the teeth. This transparent coating (Delton) is applied in about 30 minutes to the occlusal surfaces of posterior teeth of both arches that are not already carious or filled. The seal provides topical protection to the surfaces treated.	After initial application, seals are checked every 3 months.* If a seal is lost, it is reapplied a maximum of 3 times.	Dental or auxiliary personnel consistent with minimum state dental practice legislation.
Systemic fluoride tablet	One mg of fluoride by 2.2 mg of neutral sodium fluoride tablet is chewed and swished for one minute and then swallowed. This procedure provides both systemic and topical fluoride protection.	One tablet per day during the school year.	Classroom teacher or aide.
Fluoride mouthrinse	A 0.2% neutral sodium fluoride solution is served to the child in a paper cup. The child swishes the solution between the teeth for 60 seconds and then expectorates into the cup. The child does not swallow the solution. This procedure provides topical fluoride protection.	Once per week for each week of the school year.	The classroom teacher or aide supervises the administration of the mouthrinse.
Plaque control	Children in grades 1, 2, and 5 brush in school without a dentifrice. The grade 5 cohort also uses dental floss. These procedures are designed to remove plaque (and thereby the bacteria) that causes tooth decay and gingivitis. A supply of ADA approved fluoride dentifrice is provided for home use.	Twice weekly supervised exercises in the classroom. Disclosing solution is used at least once per month.	Dental auxiliary demonstrates appropriate procedures and periodically visits classrooms to ensure they are being followed. Teacher or aide provides routine supervision of classroom activities.
Education program	This program consists of a series of 10 units (about 50 minutes each) that were selected for the program from existing materials designed to promote appropriate oral hygiene and health. Different materials were prepared for each grade level.	Individual teacher decision whether to teach as a unit or spread over school year.	Classroom teacher after orientation by program education coordinator.
Diet regulation	This component consists of efforts to reduce the frequency of refined carbohydrates in school food programs and to encourage the elimination of cariogenic snacks.	Every school day throughout the year.	Education coordinator working with school dietician, administrators, parents, and children.

* After the first treatment year, the interval between sealant checks was increased to six months.

Table 1.3

ORGANIZATION OF TREATMENT COMPONENTS INTO REGIMENS
RELATIVE TO TYPE OF WATER SUPPLY*

		NONFLUORIDATED SITES		FLUORIDATED SITES	
	Regimen Type	Clinic	Classroom	Clinic	Classroom
1	Comprehensive	Examination Prophy/Gel Sealants	Mouthrinse Plaque Control Education-Diet Tablets	Examination Prophy/Gel Sealants	Mouthrinse Plaque Control Education-Diet
2	Modified Comprehensive	Examination Prophy/Gel	Mouthrinse Plaque Control Education-Diet Tablets	Examination Sealants	Mouthrinse Plaque Control Education-Diet
3	Clinic Care Only	Examination Prophy/Gel Sealants		Examination Prophy/Gel Sealants	
4	Classroom Care Only	Examination	Mouthrinse Plaque Control Education-Diet Tablets	Examination	Mouthrinse Plaque Control Education-Diet
5	Modified Classroom	Examination	Plaque Control Education-Diet	Examination	Plaque Control Education-Diet
6	Longitudinal Comparison	Examination		Examination	
X	Cross-Sectional Comparison	Examination at Baseline		Examination at Baseline	

*The only differences between fluoridated and nonfluoridated sites are: fluoride tablets are included in Regimens 1, 2, and 4 in nonfluoridated sites; and, Regimen 2 includes prophy/gel application at nonfluoridated sites but sealants at fluoridated sites. Children in the cross-sectional comparison groups were examined only at baseline.

Chapter 2

VARIABLES

Several data collection procedures were used to gather information about Program participants. This chapter describes the types of data collected and the general procedures used to gather this information.

BACKGROUND CHARACTERISTICS

Sex and Racial/Ethnic Group

When a child checked in to be examined, a desk clerk coded a box on the child's examination form to indicate the child's apparent sex and racial/ethnic group (Anglo, Black, Hispanic, and Other). Almost all the children categorized as "Other" were Asian Americans.

Age

Birthdates were obtained initially from parents at the time they enrolled their children in the Program. Many of the dates were subsequently checked against school records when it became evident that a small but significant portion of them were incorrect. Children who appeared to know their birthdate were asked to check if it was correct on their examination form. If it was not accurate, the form was revised to show the correct date.

Socioeconomic Status (SES)

A questionnaire was sent to all the parents whose children received a baseline clinical examination. This questionnaire inquired about family income and each parent's education level and occupation. Comparable data were obtained on most non-respondents via telephone or personal interviews.

The relationship between socioeconomic status and the various dental health measures will be studied in two ways. When possible, the effects of the three components--income, education, and occupation--will be measured separately, using multivariate analysis. When the data do not distinguish separate effects, the components will be combined into a composite variable. The construction of this composite variable will have to take into consideration two types of missing data:

- o No response to a particular question, such as might occur if the child came from a single parent household, if a parent decided not to answer a question, etc.

- o No response to the whole questionnaire; i.e., the child's parent(s) did not return the questionnaire and no personal or telephone interview was able to elicit the information.

In some instances, we will want to have an SES score for all children. When SES is used as a covariate, it would be unsatisfactory to delete all cases for which one or more components is missing. Thus, it is proposed that the first of the two types of missing data be estimated on the basis of the other questionnaire responses and the child's racial/ethnic group, school, and site. The second type of missing data will necessitate estimating a child's SES level solely on the basis of that child's racial/ethnic group, school, and site. It is anticipated that less than 15 percent of the sample will need to have their SES scores estimated. When the direct effects of SES, or individual components, are being estimated, only complete observations will be used.

Length of Residence

The mail/telephone questionnaire procedure also asked the respondent to indicate how long the child had lived in the site. The description of the site used for this purpose (e.g., "How long has your child lived within the city limits of ...") was tailored to the boundaries of the site's water supply. About five percent of the parents who returned the mail questionnaire did not answer this question (apparently as a result of its appearing in different type than the rest of the questionnaire). If a value for this variable is required for all children, it will be estimated on the basis of other data about the child that is correlated with length of residence. The estimation procedure would probably be site specific.

Previous Dental Care and Habits

The questionnaire process also elicited information about the child's dental habits (e.g., frequency of toothbrushing and type of dentifrice), whether the child received fluoride at home via vitamins or rinse, and how often the child and the child's mother had visited the dentist during the past year.

Grade Level

Grade level, rather than age, was used in designing the Program. This was necessary (even though age is more closely associated with dental decay) because of several administrative considerations associated with a school-based program, such as providing preventive care to classroom groups. However, the term "grade level" is not defined consistently across sites. The reason for this is that sites vary in their policies with respect to the minimum age required for entry into a given grade, the criteria for promoting a child to the next grade level, etc. Thus, the average age in a grade level at one site on a given date, such as September 15, could be quite different from the average age at that same grade level and date at another site. The range of age levels within a grade is also likely to vary across sites (and even across schools within sites).

Because of the inconsistencies in the definition of "grade level," this variable will not be included in the data analyses. If there is interest in what the average value of some variable (such as DMFS score) is at a grade level, then this could be computed for a defined population (where the definition would be in terms of the distribution of children in the grade level with respect to age, sex, SES, and racial/ethnic group). In other words, sufficient data will be provided in tabular and/or formula

form to permit the computation of a hypothetical grade level average for any major variable. The procedure for doing this will be discussed along with the presentation of the results.

It also should be noted that most of the dental literature in the area of prevention describes the sample in terms of age rather than grade level.

Cohort Type

The Program's design involved two types of cohorts: longitudinal and cross-sectional. The children in the longitudinal cohorts (i.e., those enrolled in grades 1, 2, and 5 in the fall of 1977) were assigned to regimens and were examined annually for four years after baseline. For the purposes of certain analyses, such as estimating the average SES level at a site, the Grade 1 and 2 children will be combined into one group.

The children in the cross-sectional cohorts (i.e., those enrolled in grades 3, 4, 6, 7, and 8 in the fall of 1977) only received a baseline examination. For the purposes of certain analyses, the grade 3 and 4 children will be combined into one group and the grade 6, 7, and 8 children will be combined into another group.

In summary, cohort type breaks down into the following four groups:

- o Longitudinal - grades 1 + 2
- o Longitudinal - grade 5
- o Cross-Sectional - grades 3 + 4
- o Cross-Sectional - grades 6 + 7 + 8

Fluoridation Status

Sites were categorized as being fluoridated or nonfluoridated based on the information they supplied at the time they applied to have the Program. However, data now indicate that Wichita has a medium level and Hayward has had sizable variations in level over the past 10 years.

TREATMENT REGIMEN

Each school (and thereby the children in it) was assigned to one of six treatment regimens at the beginning of the Program. If a child transferred schools during the Program, an effort was made to maintain the child in his/her regimen. If this was not possible, then the child will be dropped from the analysis of treatment effects. Similarly, children who were assigned to a regimen but who did not receive even a minimal portion of it also will be dropped. All other children with initial and final examinations will be included in the analysis, even if they only received a small portion of their designated treatment because of absenteeism and related factors. This policy reflects the overall purpose of the Program to provide information about costs and benefits under normal field conditions (i.e., where absenteeism would be expected to occur).

CLINICAL DMFS/DMFT SCORES

The first portion of a child's clinical examination involved determining the status of each permanent tooth, or more precisely, each permanent tooth space. This status was then communicated to a recorder on a tooth-by-tooth basis using the following code system:

S = Sound, permanent tooth.

D = Deciduous tooth regardless of whether sound, filled, or carious.

U = Unerupted permanent tooth.

E = Missing permanent tooth that was extracted because of decay.

Y = Tooth that could not be properly examined; i.e., lost because of trauma; extracted for orthodontic reasons; banded; fractured or restored as a result of trauma; hypoplastic; or affected by a calcification defect.

If the entire tooth did not fall into one of these five categories (i.e., it had one or more carious and/or filled surfaces), then each surface of that tooth was classified as being sound, carious, filled, or both carious and filled. In this way, each tooth and tooth surface was assigned to one of eight possible categories.

The diagnostic criteria for making these categorizations were essentially those adopted by the Conference on the Clinical Testing of Cariostatic Agents (Radike, 1972) and the same as those used by the National Institute for Dental Research (NIDR) in numerous studies since that time. The only major difference between the Program's and NIDR's criteria is that the latter agency records a tooth (for DMFT) or a surface (for DMFS) as "carious" if it is both carious and filled.

The following general rules will apply to the computation of scores:

- If a child has two complete concurrent examinations as a result of being in the sample of children used to assess examiner reliability, only the first examination will be used in computing that child's score. However, if the first examination is incomplete or otherwise unscorable, and the second one is complete, then the score on the latter examination will be used.

- If there is a discrepancy between two concurrent examinations of a child and if that discrepancy is resolved by a third concurrent examination, then the child's score will reflect the resolution. The procedure of using a third examination to resolve differences was initiated at baseline in Pierce County (the sixth site) and has been used in all subsequent sites and years when examination time and circumstances permitted.

- If a tooth (and thereby all of its surfaces) is classified as a Y on one examination, then it will be classified as Y on all subsequent examinations (i.e., even if on a subsequent examination it is judged to be something other than Y). Any carious surface in the examination immediately preceding the first Y call will be counted in that and subsequent DMFS scores.

- o Sometimes a child did not have every tooth classified and/or had an incompatible classifications on a tooth (e.g., it was marked as being both carious and sound). If there is a clear resolution to the inconsistency, that correction will be made. Otherwise no score will be computed for the child on that examination because there is no way of knowing where the error occurred in the sequence of calls.

- o No changes will be made as a result of a tooth (or surface) being classified as carious on one examination and sound on a subsequent examination; i.e., so called "reversals" will not be corrected.

A child's DMFS score will be computed by counting the number of surfaces on permanent teeth classified as decayed or as missing or filled because of decay. The following specific rules will be used in computing each child's DMFS score:

- o If a surface is classified as both carious and filled, it will be counted as only one affected surface in a child's total score.

- o If a tooth is classified as missing due to decay, then it will be given a score equivalent to the number of surfaces on that tooth; i.e., 4 for an anterior tooth and 5 for a posterior tooth.

For certain analyses, the overall DMFS score will be broken down into the following subscores.

- o Components (i.e., D, M, and F).

- o Type of surface (i.e., occlusal, interproximal, and buccal/lingual).

- o Ratios, namely: F/DMFS (as a measure of met needs) and D/DMFS (as a measure of unmet needs).

A child's DMFT score will be computed by counting the number of permanent teeth classified as decayed or as missing or filled because of decay. If a tooth is classified as both carious and filled, it will be counted as only one affected tooth in a child's total score.

NUMBER OF TEETH AT RISK

A count will be made of the number of permanent teeth at risk. This count (and/or the number of first and/or second molars) will be used as a covariate in certain analyses. It also is possible that the count will be weighted to reflect the probability of the tooth becoming decayed.

RADIOGRAPHIC EXAMINATIONS

AFDH's National Advisory Committee concluded that while radiographs may not be necessary in long-term studies of dental caries, they do reveal the presence of interproximal carious lesions earlier than can be observed by clinical examination alone. The use of radiographs, therefore, increases the accuracy of the measurement process in the sense that essentially all carious lesions are detected by the combination of clinical and radiographic examinations.

Rand supported the use of radiographs because of its concern with the possible interaction between a child's treatment regimen and the sensitivity of the examination process to detect the benefits derived from that regimen. For instance, the decrease in decay attributable to sealants is limited to the occlusal surfaces of a child's posterior teeth. These surfaces are diagnosed for decay by clinical means alone. A fluoride mouthrinse program, on the other hand, is presumably beneficial to all surfaces, but especially the hard to see inter-proximal surfaces on posterior teeth. Thus, a clinical examination might underestimate the benefits derived from a mouthrinse program but not underestimate those provided by sealants.

The foregoing rationale and other considerations led to the decision that all the children examined at baseline would be scheduled to receive both clinical and radiographic examinations. All the longitudinal cohort children are scheduled to receive a second radiographic examination at the end of the fourth treatment year at their site. The only exception to this plan occurs in New York where, because of the early phasing out of this site, the children received a radiograph examination at the end of the third treatment year.

Both longitudinal and cross-sectional children had bitewing radiographs taken at baseline. One radiograph was taken of each side of the mouth with each film designed to encompass eight teeth. Each pair of developed films was "read" (i.e., evaluated) by one of six dentists. The primary purpose of the readings was to detect possible carious lesions on interproximal surfaces that may have been missed by the clinical examination process.

The rule for combining concurrent clinical and radiograph data in the computation of DMFS and DMFT scores will be as follows: If a tooth (or surface) is classified as carious on either a clinical or a radiographic examination, then it will be classified as carious. The same rule applies to filled surfaces. If the two types of examinations disagree as to whether a surface is filled or carious, the radiograph call will be used. Since the radiograph readers only make carious or filled calls on interproximal surfaces, this is the only type of surface that can be affected by the radiograph examination.

RESTORATIVE CARE COST

An alternative method for assessing decay level involves estimating restorative care costs (RCC) from a child's examination form. Such costs could provide more useful information than DMFS or DMFT for estimating the benefits of various treatment components. For example, suppose that sealants prevent decay on about one surface per year and the mouthrinse component on about 0.5 surface per year. All other factors being equal, sealants would appear to be the better approach. However, sealants work solely on occlusal surfaces and mouthrinse primarily on interproximal surfaces. Since the cost of restoring an interproximal surface is almost twice that of a simple occlusal filling, the two procedures would produce similar savings in restorative care costs even though they had radically different DMFS score savings.

SEALANTS AND FLUOROSIS

After completing the DMFS portion of the clinical examination, the examiners indicated whether or not the child had sealants and whether or not there were any signs of fluorosis; i.e., a two point scale of Present versus Absent was used for each index. The fluorosis index was initiated at baseline in Wichita after the experience at the first four sites suggested that more than expected mild fluorosis was present. The criterion for fluorosis became "opacity" by Exam #03.

DHC INDEX OF GINGIVITIS

The gingivitis index was based on the one used by San Francisco Public Health Service Dental Health Center (DHC) (Suomi et al., 1969). This index involves evaluating the degree of inflammation of the gingival tissue along both the lingual and facial sides of several permanent teeth. For this study, six teeth--four first molars and two incisors--are examined.* Each of the 12 surfaces is then placed into one of the following four categories on the basis of color difference between the designated gingival area and the surrounding tissue:

0 = No inflammation--gingiva adjacent to the tooth surface being examined is pale pink in color and firm in texture. Swelling is not evident and strippling can usually be noted.

1 = Inflammation not encompassing all tissue adjacent to the tooth surface (including papillae)--gingiva is a definite red or magenta color; color change does not extend from mesial to distal.

2 = Inflammation encompassing all tissue adjacent to the tooth surface (including papillae); color change extends from mesial to distal.

NA = In this category are deciduous and missing teeth, and permanent teeth that are less than 50 percent erupted.

A child's DHC score is the average score assigned to all the surfaces that are examined; i.e., the possible score range is from 0.00 to 2.00. The following specific rules are used in computing each child's DHC score:

- If all of a child's teeth are classified as deciduous or not applicable, then that child does not receive a DHC score.
- If one surface of a tooth is classified as not applicable, then both surfaces of that tooth are eliminated from the calculation of that child's DHC score. Thus, there is some variation between children in the number of tooth surfaces that are used in computing their DHC scores.
- The rules for handling multiple DMFS examinations are applied to the scoring of the DHC.

---

*The surfaces chosen are the same as those used in the "Simplified Oral Hygiene Index," Greene and Vermillion (1964).

REFERRAL CODES

The final major step in the examination procedure involved noting on the form the general status of the child's oral condition. The following four referral codes were used for this purpose:

1 = No problems, only routine dental care indicated.

2 = One or more problems (or potential problems) that indicated that the child should see the family dentist in the near future.

3 = One or more serious problems requiring immediate attention.

4 = A complete examination could not be performed (e.g., more than 50 percent of the child's permanent teeth were banded).

DECISION RULES FOR RETAINING CHILDREN FOR ANALYSIS

Children will be deleted from the analysis of a given examination if one or more of the following conditions occurred:

- Fifty percent or more of the child's permanent teeth were "banded" (i.e., had some type of non-removable orthodontic appliance in which one-half of the erupted permanent teeth were banded).

- The child's baseline examination occurred more than 60 days after the regular examination week at the child's site. The only exception to this rule is the second round of baseline examinations at Minneapolis which were conducted in order to add more schools at this site.

- The child was given the baseline examination after receiving any type of one-on-one care in the Program and/or more than four weeks of a classroom component.

- No DMFS or DHC score could be computed for that child; i.e., both scores had to be missing in order for the child to be dropped from the data analysis.

- Technical problems or other considerations precluded scoring the child's examination form.

All between examination comparisons will be based on continuous residence samples; i.e., children who received both examinations.

Chapter 3

ANALYSIS OF BASELINE DATA

INTRODUCTION

The baseline data will be analyzed in order to answer the following questions:

- What were the background characteristics of the children participating in the Program and how did these characteristics vary across sites? For example, what minority groups were and were not represented at each site? How much variation was there within and between schools and sites on such variables as socio-economic status (SES)?

- What was the nature of the relationship between the dependent variables and the characteristics of the children participating in the Program? For example, how were DMFS scores related to age, sex, and other variables? Were these relationships the same across fluoridation levels?

- Within a given fluoridation status, did the children in one treatment regimen have essentially the same characteristics as the children in another treatment regimen? In other words, were the groups equivalent at the beginning of the Program with respect to such critical variables as DMFS scores, age, sex, racial/ethnic group, and SES?

In order to provide accurate answers to these questions, the statistical analyses will need to take into account several factors that could bias outcomes and/or cause misleading findings. Some of these factors are as follows:

- There was almost a six month gap between the time the first child was examined in Chattanooga and the last child was examined in Minneapolis. Thus, on the average, the children at the latter site are likely to be older and therefore have more decay.

- There was considerable variation in the number of children participating at each site. Sites with large numbers of children will therefore carry more weight if all the children within a given category (e.g., seven-year-olds at nonfluoridated sites) are grouped prior to computing average scores.

- The distributions of the various types of DMFS and DMFT scores are extremely positively skewed. Thus, if a group is small, a few children with high scores could have an unduly large effect on the group's average.

- Many of the background variables are not independent; e.g., almost all of the black children in nonfluoridated sites are at Monroe. Thus, an observed difference in DMFS scores between racial/ethnic groups would be misleading, since it could be attributable to SES and/or site differences.

- For several children, data are missing from the parent questionnaire because parents refused to respond or moved soon after the baseline examination.

GENERAL STRATEGY

Three types of variables will be used: dependent, independent, and control. Dependent variables are those, such as DMFS or DHC score, that provide information about a child's dental health. The Program's only true independent variable is treatment regimen. However for the baseline analysis, several of the background characteristics such as site, age, sex, and SES will be treated as independent variables in certain analyses in order to study their relationships to the dependent variable. Those background characteristics not used as independent variables in a particular analysis may be used as control variables (or "covariates").

Two general types of analyses will be run. The first type will be used to describe the data base; e.g., numbers and characteristics of children at each site. These data will usually be presented in tabular form, such as the average DMFS score at each age level at each site.

The second type of analysis will be used to test for the statistical significance of observed differences between various groups and to assess the nature and strength of relationships between variables. For example, we will estimate the size of the differences in DMFS scores across sites for various age levels. For these tables, the values presented will be adjusted to take into account various factors that could otherwise result in misleading information. For example, the comparison of the average DMFS score at one site versus another site will control for differences that might be attributable solely to one of the sites having proportionately more girls than the others.*

These controls will be implemented by means of an analysis of covariance (ANOCOVA). The basic components of this analysis will be a dependent variable (such as DMFS score), one or more independent variables (such as site), and one or more control variables (such as sex). For example, an analysis of covariance will be used to determine whether there are any statistically significant differences in average DMFS scores between the five fluoridated sites at each of several age levels. In this analysis, the DMFS score is the dependent variable and site and age are independent variables. The analysis will indicate whether there are significant differences in average DMFS scores between (1) sites across age levels, (2) age levels across sites, and (3) whether the size of any differences between sites remains constant or varies with age level (i.e., is there an interaction between site and age level?). This analysis will control for such possible extraneous influences as differences in the proportion of

*Since girls tend to mature physically faster than boys, they will tend to get their permanent teeth sooner. Thus, it is likely that girls will tend to get decay on these teeth sooner than boys. If there is no control for sex, one site (or group) could have higher average decay scores than another site simply because it happens to have more girls in this sample.

each sex in each site-age level combination. It also will control for the chance that the children within a given category of age (e.g., eight-year-olds) are actually older (in terms of date of birth) than the children in that same age category at another site. Thus, "age" can be used as both an independent variable and as a covariate (i.e., control variable).

The decision whether to treat a variable as an independent variable and/or as a covariate rests on the purpose of the table. If the purpose is to explicitly show how the variables relates to the dependent variable, it will be an independent variable. If, instead, we only want to keep that variable from distorting the other results, it will be a covariate. With the exception of age, no variable will be used as both an independent variable and a covariate in the same ANOCOVA. The effect of age on DMFS score is large enough to require it as a covariate even within a two to three year age range.

ANALYTIC CONSIDERATIONS

Three potential conditions may complicate estimation and statistical inference in both the baseline and treatment effects analysis. These conditions are: (1) interactions among the predictor variables, (2) heterogeneity of variance, and (3) non-normality of the dependent variable. Classical analyses, like analysis of covariance and ordinary linear regression, are most suitable when none of these three conditions exist. When one or more of the conditions are present, such procedures may be inefficient and/or cause biased results. In this context, inefficiency means that the estimates will have larger than optimal standard errors. When this occurs, it is more difficult to detect treatment effects. When the estimates are biased, the significance levels of statistical tests will be different from their presumed levels. These problems could increase the chance of drawing false conclusions regarding whether or not a particular relationship exists.

Interactions

Interactions among the predictors are potentially the most troublesome of the three conditions noted above. These interactions occur if the effect of one predictor variable on the dependent variable changes with the level of other predictor variables. Interactions will be of special interest in the next chapter when we will want to know how treatment effects on the DMFS increment vary with control variables such as age, site, baseline DMFS score, and other background characteristics. We might also find some interactions among variables in the baseline analysis. For example, the effect of SES might be greater when there is a large amount of decay than when there is little decay. If this were so, then the effect of SES would be largest at sites that have relatively high levels of decay.

Two techniques are useful when interactions exist. The first adds predictor variables to the model to test for and fit interactions. Often, however, adequate accounting of all interactions requires a much more complex model than can be interpreted easily. Sometimes the second method, transforming the dependent variable, produces a model with few or no interaction terms. Both approaches will be investigated.

Heterogeneity of Variance

Classical statistical procedures usually assume that each observation of the dependent variable has the same variance. However, because the distribution of DMFS scores is highly skewed, the variance will almost certainly increase with the mean (as a function of treatment regimen and the control variables). When variances are heterogeneous, the proper weight to give each observation in the analysis depends on the variance of the dependent variable (given all the predictor variables). As the variance increases, the weights should decrease.

Ordinary least squares procedures give equal weight to all observations. While this misspecification fortunately does not bias parameter estimates, it can substantially increase their standard errors. It may also bias the presumed levels for significance tests. These problems can be handled by weighting each observation in the analysis by the inverse of the estimated variance of the dependent variable. In practice, the variance of an observation is modeled by a function of its predicted value. Then the variance derived from one fit of the ANOCOVA model is used to provide weights for a refit.

Transformation of the dependent variable provides a second possible solution to this problem. A transformation, which is found to reduce interactions, also might reduce, but not necessarily eliminate, heterogeneity of variance.

Non-Normality of DMFS Increments

As discussed in Appendix A, classical procedures (such as sample means and analysis of covariance) can be quite unstable when the dependent variable has an especially long tail, because a small number of outliers (extremely large or small observations) can unduly influence these estimates and statistical tests. As a result, these estimators may have much larger variances than other estimators which reject outliers. Classical tests also may have reduced power to detect significant relationships among variables. The degree of instability depends on the extent of the departure from normality in the tails and on the complexity of the model. In general, as one tries to fit intricate models with more estimated parameters, the data are spread more thinly. This increases the effect that a single observation can have on the analysis.

ANALYSIS OF COVARIANCE MODEL

An analysis of covariance (ANOCOVA) model will be used to estimate and test for relationships between dental health measures and background characteristics. Separate analyses will be performed for each fluoridation status by age cohorts (corresponding roughly to grades 1, 2, 3-4, 5, and 6-8). By analyzing these age cohorts separately, we expect to reduce substantially the extent of the problems mentioned above. The fixed effects ANOCOVA model provides a means for testing for differences across levels of the independent variable. It also provides estimates and confidence intervals for the mean of the dependent variable at each level of the independent variable.

Covariates

Covariates will serve two purposes: (1) to adjust estimated cell means for any imbalance of the values of control variables within the cells and (2) to reduce the standard errors of estimated parameters and thereby increase the power of statistical tests of hypotheses. Covariates will include various combinations of site, age, sex, socioeconomic variables, and measures of past dental care. By varying the set of covariates for a particular table, we can investigate the degree to which the relationship between two variables is a function of their relationships with other variables.

In the main ANOCOVAs, only age and sex will be used to adjust scores across sites. The phased-in examination schedule led to wide disparities in age across the sites. So that site differences will not be a byproduct of this phenomenon, age will be an across-site covariate. Sex will be used in the same way. In contrast, because other background characteristics are intrinsically related to certain sites (e.g., almost all the black, nonfluoridated children came from Monroe), these other characteristics will not be used as across-site covariates. Thus, if between-site differences in characteristics other than age and sex affect DMFS scores, the estimated site effects will reflect the differences.

A second set of ANOCOVAs will use the full set of covariates to adjust scores across sites. This analysis will only provide information about site differences that are not explained by background characteristics. By contrasting the two types of ANOCOVAs, we can estimate the degree to which site differences reflect differences in background characteristics.

Interactions

By fitting separate ANOCOVAs for each stratum of age by fluoridation status, we will allow certain interactions automatically. For example, separate SES coefficients will be estimated in each ANOCOVA. These coefficients will be compared across strata in an effort to detect any patterns (including whether the SES effect appears to be independent of age and fluoridation status). F-tests will be used to check for further interactions within the ANOCOVA tables. Aggregation of these tests across strata will help to indicate which interactions occur consistently.

Heterogeneity of Variance

For reasons discussed earlier, observations in different cells will probably have different variances. For example, DMFS scores in a site with a high average score might have a larger variance than scores in a site with less decay. The variance might also be a function of covariates such as age. By using a weighted analysis of covariance, with the weight of each observation depending on its predicted variance, we can obtain more efficient estimates.

Non-Normality of DMFS and DMFT Scores

Because DMFS, DMFT, and the various component scores all have very skewed distributions with a very long right tail, care must be taken to avoid estimates and statistical tests which are overly sensitive to extreme

observations. In order to see whether the results from the above analyses reflect general tendencies rather than a small fraction of the data, we propose to repeat certain analyses in a more robust form. Extreme values (those much larger or much smaller than predicted) will be downweighted in the analysis of covariance procedure. For nonsymmetric data, such as DMFS scores, this has the effect of slightly changing the interpretation of what is being measured. Thus, tables of (adjusted) means would not be comparable to the tables described earlier. However, tests for main effects and interactions would have almost the same interpretation as above.

Presentation of Results

The results of the analyses of covariance will be presented in two types of tables. One type of table will contain summary data of the statistical tests conducted (e.g., were the observed differences in DMFS scores across sites due to systematic versus chance differences in the decay levels?). This type of table will contain supporting documentation (such as the coefficients associated with the control variables so as to indicate how much impact they had and to provide a means for computing an average score for some hypothetical group of children, such as those in a given grade level). The second type of table will contain the adjusted mean score on the dependent variable for each level of each independent variable (e.g., each age level and each site) and each combination of levels (e.g., the average at each age level at each site). The adjusted mean score is the predicted mean score for a standardized population, such as an equal mixture of eight-year-old boys and girls with ages distributed uniformly over that age. Even if most of the eight-year-olds from a particular site had just turned that age, the adjusted mean for that site would represent the standardized population. Thus, results will be more interpretable and generalizable. Where appropriate, standard errors will be provided so that the reader can conduct more detailed tests. If the standard errors for a given cluster of groups are quite similar, then their average and range will be reported.

The lists of tables that follow indicate the specific analyses that will be conducted. In general, the first phrase of a table's title describes the dependent variable, such as mean DMFS score. This is followed by the independent variables that will be represented. The last part of the title describes the control variables. The reported figures are estimated means for the corresponding cell under the condition that the control variables have the same distribution across cells. In most cases, separate tables will be presented for each fluoridation status. The overall average for each fluoridation status also would be included in each table in which site is an independent variable. Thus, many of the tables, such as Table 14, will have subparts for each fluoridation status. These will be designated as A and B.

Certain checks of the ANOCOVA model assumptions outlined above will also be made. Where these checks indicate a need, the analysis will be modified or supplemented.

LIST OF TABLES FOR BASELINE ANALYSIS

Number and Percentage of Children

Table 1: Number of children excluded from the analysis by age level and reason for the exclusion. These children will not appear in any of the other tables.

Table 2: Number and percentage of children with complete observations for each variable.

Table 3: Number of children with DMFS score by age, sex, and site.

Table 4: Percentage of children from each site for each combination of age and fluoridation status.

Table 5: Percentage of children from each site for each combination of racial/ethnic group and cohort type.

Table 6: Mean on each SES component and composite score for each combination of site and cohort type

Table 7: Percentage of children in each SES level at each site.

Table 8: Distribution of age on September 15, 1977 by grade and site.

DMFS and DMFT scores

Table 9: The cumulative percentage of children with each DMFS score within each combination of site and age level controlling on age and sex.

Table 10: The cumulative percentage of children with each DMFT score within each combination of site and age level controlling on age and sex.

Table 11: Ratio of mean DMFS to mean DMFT by age, sex, and fluoridation status controlling on age and site; a footnote will contain generalized equations for the prediction of DMFS from DMFT and vice versa.

Table 12: Mean DMFS by age, sex, and site controlling on age.

Table 13: Mean DMFS by age and fluoridation status controlling on age, sex, site, SES, length of residence, and teeth at risk.

Table 14: Mean DMFS by SES level, cohort type (4 levels), and site controlling on age, sex, length of residence, and teeth at risk.

Table 15: Mean DMFS by racial/ethnic group and age controlling on age and sex, at just those sites with sufficient minority population for analysis.

Table 16: Mean DMFS by site, length of residence at site, and age, controlling on sex and surfaces at risk.

Table 17: DMFS percentiles (e.g. 25, 50, 75, 90) by age and site controlling on age, sex, and teeth at risk.

Table 18: Percentage of total DMFS in group attributable to the highest 20 percent of the scores within the group when the scores are controlled for age and sex; and where the groups are each combination of age and site.

Table 19: Mean DMFS and DMFT scores with and without the inclusion of radiograph data by age and fluoridation status controlling on age, sex, and site.

DMF Component, Surface Type, and Ratio Scores

Table 20: Mean on each DMF component and surface type variable by age and fluoridation status controlling on age, sex, and site. Variables are: D, M, and F teeth; D, M, and F surfaces; occlusal, buccal-lingual, and interproximal surfaces.

Table 21: Mean on each DMF ratio score (FS/DMFS, DS/DMFS, FT/DMFT, and DT/DMFT) by age and site.

Table 22: Correlation coefficient between each DMF ratio score and SES within each cohort type and fluoridation status combination with the effect of site, sex, and teeth at risk partialed out.

Table 23: Mean SES score for each quartile level of each ratio score at each site.

Analysis of the Risk to Decay and Eruption Rate of Individual Teeth

Table 24. The percentage of children who have each tooth type at each age, sex, and fluoridation status combination controlling on age and site; tooth type is defined as one of 28 classifications (however, average of right and left sides presented in table).

Table 25: The percentage of children who have decay on each tooth type at each age and fluoridation status combination controlling on age, sex, and site.

Table 26: Same as table above with the exception that percentage is based only on those children who have the tooth at risk. Cells with less than adequate sample sizes will be left blank.

Analyses of Other Variables

Table 27: Mean DHC score by age, site, and sex.

Table 28: Mean DHC score by any other variables with which it is correlated (based on preliminary analysis).

Table 29: Percentage of children in each referral code category at each age and site combination.

Table 30: Percentage of children in each referral code category at each site controlling on age, sex, and teeth at risk.

Table 31: Mean SES score by referral code category for each fluoridation status controlling on site.

Table 32: Partial correlation of DMFS scores with the following variables: DHC score; number of times per day the child brushes teeth; use of fluoride toothpaste, tablets or vitamins; number of times visited dentist by child and by mother; and whether or not covered by dental health insurance. Partial correlations presented for each fluoridation status controlling on age and site.

Tables containing data on estimated restorative care costs will be presented if such data are collected, and if it appears that these costs can be estimated reasonably well from the clinical or from a combination of clinical and radiographic examinations.

Chapter 4

ANALYSIS OF TREATMENT EFFECTS

One of the two major purposes of the Program is to measure the effects of various types and combinations of school-based preventive dental care. One set of six treatment regimens was developed for nonfluoridated sites and a slightly different set of six regimens was developed for fluoridated sites to allow for a large number of potentially important contrasts to be made (see Table 1.3 for a description of the regimens in each fluoridation status). Treatment effectiveness will be defined in terms of the difference in dental outcomes between two or more treatment regimens--particularly between the control and any of the other five regimens. The most important dental measure for this study is the four year increment in DMFS scores from baseline to Exam #05. The discussion of treatment effects in this chapter will therefore concentrate on this variable.

The issues investigated in the analysis of treatment effects may be divided into two categories. One category deals with the relative effectiveness of the treatment regimens for various groups of children, such as all children of a certain age level. The second category focuses on the individual characteristics of children that may interact with treatment effectiveness, such as the relationship between initial decay level and the effectiveness of sealants.

The specific questions to be answered by the first set of analyses are as follows:

- o What is the relative effectiveness of the various treatment regimens within each fluoridation status?

- o How does the effectiveness of a treatment component vary with the amount of fluoride in the water? For example, is the component that is most effective in fluoridated sites also the one that is most effective in nonfluoridated sites?

- o How does the relative effectiveness of a treatment regimen vary with the general characteristics of the children receiving it? For example, does a given regimen work better for older children than younger ones, for children from one site versus another, etc.?

- o How long does it take for the effects of a given regimen to become apparent? Do some treatment components have delayed effects while others become apparent immediately? Is the speed of impact related to the type of fluoridation status?

- o What is the nature of a treatment effect; e.g., does one component tend to reduce decay on one type of tooth surface and another component prevent decay on another type? Which type of regimen is most effective in reducing the number of teeth lost due to decay?

The specific issues to be addressed in the second set of analyses are:

- o How is the preventive effect of a regimen distributed across children? Do all children within a given age level tend to benefit equally or do those who had initially high (or low) levels of disease tend to benefit more?

- o What are the specific background characteristics of children who tend to benefit versus not benefit from each type of treatment component and regimen? How accurately can these children be identified, such as for the purpose of targeting preventive care?

GENERAL STRATEGY

The main dependent variable for the study of treatment effects will be the four-year DMFS increment; i.e., the final DMFS score minus the baseline score. Because decay is presumably irreversible, the DMFS increment directly measures the decay which occurs during the treatment period. The alternative would be to use the final score as the dependent variable and the baseline score as a covariate. However, there is no major advantage to using this alternative approach in that it produces essentially the same results as the more familiar difference score method.*

Other dependent variables will include increments in DMFT; increments of the D, M, and F components of DMFS; and DMFS increments by surface type. To study the speed with which regimens take effect, we will also study increments during subperiods of the program, such as baseline to Exam #03 and Exam #03 to Exam #05.

In contrast to decay, gingivitis and the need for dental care are reversible. Thus increments of DHC and the referral code do not have as simple interpretations as the DMFS increment does. Instead we will present results in terms of the final DHC score or referral code, using the baseline value as a major covariate.

The independent variable for the analysis will be treatment regimen. It is important to note, however, that the treatment regimens used at fluoridated sites are somewhat different from those used at nonfluoridated sites. Thus, any analyses involving treatment regimen will be run separately for each fluoridation status.

As in the baseline analysis, background characteristics will be used as covariates, the choices being guided by the baseline analysis findings. For example, a regimen's average increment in DMFS score will be adjusted for the average age of the children in that regimen so that any comparisons among regimens will not be affected by one regimen having somewhat older or younger children than some other regimen. An additional covariate will be the baseline DMFS score. Background characteristics also will be used as a means to describe the children who are most and least likely to benefit from a given regimen.

---

*In general, using the baseline score as a covariate improves the precision of estimates compared with simply using increment scores without the baseline score as a covariate; c.f., Slakter and Juliano (1976).

Two sets of issues are presented in the introduction to this chapter. Because of differing levels of complexity of the questions being raised, we anticipate using distinct analyses to address each set. The first set, which will receive primary emphasis in the analysis, asks basic questions about the mean level of effectiveness for each site-cohort. These questions are very similar to those addressed in past clinical studies. Despite the data analytic complications raised in Chapter 3, we feel that most of these questions can be addressed adequately by standard analytic and reporting techniques common to dental studies. Because of the size and diversity of the sample for this study, these data also offer a unique opportunity to study how effectiveness varies with child characteristics. We anticipate that this individual-oriented study will require more complex statistical techniques.

The first type of analysis will be used to describe and test for differences between treatment regimens. Table 4.1 lists the major comparisons that will be made. The basic statistical technique used for this purpose will be an analysis of covariance in which the independent variables will be treatment regimen and site, and the covariates will be the children's background characteristics, such as age and sex. The choice of covariates will depend, in part, on the results of the baseline analyses because the variables that affect baseline scores are most likely to affect increment scores as well. Diagnostic tools, such as residual analysis, will be used to insure that the analysis of covariance model assumptions are adequately met.

The second type of analysis will be used to probe for and measure more specific relationships between increment scores and the children's background characteristics. For example, how does treatment effectiveness vary as a function of variables such as sex, SES, and baseline DMFS? These analyses will be more complex than those described above and will most likely involve transforming DMFS score increments to take into account both the skewness in these scores and the possibility of non-linear relationships with predictor variables. Alternative strategies for conducting this second type of analysis are discussed in the last section of this chapter.

ANALYSIS OF COVARIANCE MODEL

An analysis of covariance (ANOCOVA) model will be used to estimate and test for regimen and site effects on DMFS increments. Separate analyses will be performed for each fluoridation status by longitudinal cohort type (i.e., grades 1+2 and 5). The basic table will contain 30 cells in a two-way layout of 5 rows (sites) by 6 columns (treatments).* Covariates will be used to adjust for background characteristics across regimens within sites and to adjust for age and sex across sites. Each analysis of a given increment (such as between Exam #01 and #05) will be limited to just those children who (1) have that increment score (i.e., the continuous residence sample) and (2) have received some portion of their designated treatment regimen (i.e., as indicated by their treatment records).

*Because the project ended after three years in New York, the analysis of four year increments at fluoridated sites will be four sites by six treatment regimens.

Table 4.1

INFORMATION OBTAINED FROM VARIOUS
COMPARISONS OF TREATMENT REGIMENS

Treatment Regimens Being Compared	Information Obtained
1 vs. 6	Effect of full comprehensive program.
2 vs. 6	Effect of modified comprehensive program.
3 vs. 6	Effect of all 1-on-1 care.
4 vs. 6	Effect of all classroom components as a team.
5 vs. 6*	Effect of the dental health program (includes plaque control, dental education, and nutrition).
1 vs. 2	Effect of prophy/gel in fluoridated sites and sealants in nonfluoridated sites when used in conjunction with all other components.
2 vs. 4	Effect of sealants in fluoridated sites and prophy/gel in nonfluoridated sites when used in conjunction with all classroom components.
1 vs. 3	Unique effect of classroom components when used as a team in conjunction with 1-on-1 care.
1 vs. (3+4)	Effect of 1-on-1 and classroom components when used together versus separately.
3 vs. 4	Effect of 1-on-1 relative to classroom components.
4 vs. 5	Effect of mouthrinse component in fluoridated sites and mouthrinse plus fluoride tablets in nonfluoridated sites for children enrolled in a dental health program.

*If the dental health component has no effect, then Regimens #5 and #6 will be combined into a single longitudinal comparison group. This will permit more precise statements about certain treatment effects; e.g., the comparison of Regimen #1 vs. #3 would provide information about the unique effect of mouthrinse in fluoridated sites when used in conjunction with 1-on-1 care (i.e., because the dental health component was shown to have no effect).

The fixed effects ANOCOVA model provides a means for testing for differences between treatment regimens. It also provides confidence intervals and an overall test for treatment differences. Within each fluoridation status, we will also test for and estimate site effects on increments in the dependent variables. Finally, we will test for and estimate interactions between treatment and site effects.

Results will be presented in tabular form indicating the adjusted increment score at the end of one, two, and four years of program participation by treatment regimen, age level within cohort group, and site. Statistically significant differences will be identified, and standard errors will be provided.

Covariates

As in the baseline analysis, covariates will be used to adjust cell means for certain imbalances on background characteristics. As in that analysis, only age and sex will be used in the main ANOCOVA to adjust for between-site differences in background characteristics. Thus any differential treatment effectiveness caused by differing racial or socioeconomic mixes between sites will be reflected in the ANOCOVA. In contrast, all important background characteristics will be used to control for within-site, between-regimen differences. Because whole schools were assigned to regimens, it was impossible to completely balance treatment regimens within a site on the background variables. This task was especially difficult for such neighborhood related variables as race and SES. This necessitates using the background variables as covariates to adjust regimen means within sites.

A second set of ANOCOVAs will use the full set of covariates to adjust across both treatment regimens and sites. This analysis will only provide information about site differences that are not explained by background characteristics. By contrasting the two types of ANOCOVAs, we can estimate the degree to which site differences reflect differences in the baseline samples.

Interactions

An important question to study is: Within a fluoridation status, do treatment effects vary across sites? F-tests will be used to test for interactions in the ANOCOVA tables. If interactions are present in the first set of tables (without race, SES, etc., as across-site covariates), this indicates that certain treatments were more effective in some communities than in others. The same test for interactions in the tables with all the across-site covariates will indicate whether treatment effects still differ across sites after accounting for differences in the children's background characteristics.

One way for treatment and site to interact is that treatments might be most effective at sites with the largest increments. For example, fluoride mouthrinse might prevent a constant proportion of caries in nonfluoridated sites (e.g., 30 percent). If this were the case, mouthrinse would prevent 1.2 caries at a site where the average increment is four surfaces, but only 0.6 caries at a site with only one-half as much decay.

If interactions are found, various transformations of the adjusted means (e.g., logarithms) will be explored to determine the nature of the interactions. It may be possible to find a transformation which eliminates the site-regimen interaction, thereby simplifying the analysis. Because individual increments may be zero or negative, transformation of cell means is an easier first step than transformation of individual scores. Tukey's one degree of freedom test for nonadditivity provides help in selecting a proper transformation from the one dimensional class of transformations which take powers of the cell means. These transformations include the identity transformation (power = 1) and log transformation (power = 0) as special cases. These two transformations are of special importance because they are the most easily interpreted.

There also is the possibility that one or more site-regimen combinations will be substantially inconsistent with the typical pattern of results. If this occurs, even if there are no other interactions, the outlier site-regimens might make a test for interactions statistically significant. A striking example of what can happen is provided by regimen 4 in the fifth grade cohort in New York. The class comprising this entire site-regimen-cohort was taken to a public clinic where their cavities (and quite possibly some imagined ones) were filled. The result after two years was an average DMFS score about 3 surfaces-per-child higher than one would otherwise expect. This example is handled easily by using an unbalanced ANOCOVA which drops that site-regimen from the analysis. While other outlier site-regimens may be less blatant, they can be handled in the same manner. More sophisticated techniques also are available to do robust ANOCOVA.*

The analysis of covariance model proposed here is not designed to estimate treatment effects as functions of the control variables. Instead, it is designed to estimate the <u>average</u> treatment effect for a population, standardized on values of the control variables. If there are interactions between treatment and any of the control variables other than site (e.g., sealants working better for boys than girls), the simple ANOCOVAs described above might bias the estimated treatment effects. The size of the bias would depend on the extent to which the treatment regimens were unbalanced on the particular control variable (e.g., too large or too small a proportion of boys in the sealant regimen). We expect that any bias arising in this manner will be slight.

In contrast with the need to adjust the mean increments by using control variables as covariates, it seems unlikely that it will be necessary to adjust for treatment-control variable interactions. The reason is that these interactions are likely to be much smaller than the main effects of control variables on the dependent variable. For example, sex is likely to have a much larger effect on the size of DMFS score increments than it does on the size of the sealant effect. If the need arises to adjust for interactions, we will add special covariates to the ANOCOVAs that reflect these interactions (e.g., the product of sex by regimen).

---

* See, for example, Schrader and Hettmansperger (1980).

Heterogeneity of Variance

For reasons discussed earlier, observations in different cells will probably have different variances. For example, increments in an effective treatment regimen might have smaller variance than increments in the control regimen. By using a weighted analysis of covariance, with the weight of each observation depending on its predicted variance, we can obtain more efficient estimates.

Non-Normality of DMFS and DMFT Increments

By limiting the analysis of covariance to two independent variables, none of the site-regimen cells should include less than about forty children. Thus, the analysis will be less sensitive to outliers than a more ambitious model would be. However, the tail of the DMFS increments is long enough so that further precautions are advisable. Tests for treatment and site effects and for treatment-site interactions will be sensitive to the most extreme observations. In order to see whether the results from the above analyses reflect general tendencies rather than a small fraction of the data, we propose to repeat the analysis in a more robust form. We would replace the values of each of the two or three largest and smallest observations at each site-regimen (about 5 percent at each end) with the next most extreme values. For nonsymmetric data, such as DMFS increments, this has the effect of changing slightly the interpretation of what is being measured. Thus, tables of adjusted means would not be comparable to the tables described earlier. However, tests for main effects and interactions would have almost the same interpretation as above.

Examiner Reliability

The validity of ANOCOVA estimates, confidence intervals, and statistical tests is not affected by examiner inconsistencies. Such inconsistencies will, however, decrease the chances of detecting treatment effects. Systematic examiner biases, on the other hand, could result in misleading conclusions. If the examiner reliability analyses discussed in Chapter 5 uncover such systematic biases, the increment scores will be adjusted to remove their average effect.

Summary

The key to the ANOCOVA model described above is that only treatment and site effects are estimated. This eliminates most of the worry about interactions among predictor variables and the long tail of the DMFS increment score distribution. Moreover, this procedure is computationally simple, has built-in reliability estimates, and is straightforward to interpret and report.

ESTIMATION OF A MORE COMPLETE MODEL

While what the ANOCOVA model tells us is correct, it suffers from not telling the whole story. First, little effort is made to describe the effects that nontreatment variables have on DMFS increments. Second, no attempt is made to study how the treatment effects vary with the levels of other variables. Obviously, determining who benefits the most from treatment is an important concern.

To get a more complete picture of how different variables effect DMFS increments, we need to account for interactions in the model. Also, as the model becomes more complex, there is increasing disaggregation of the sample. This exacerbates the problem of non-normality to the point that the simple ANOCOVA model is no longer appropriate.

There are two analytic techniques that appear to offer a solution to this problem: (1) transformation of DMFS increments and (2) modeling DMFS increments. Each is designed to combat the three problems mentioned above: interactions, heterogeneity of variance, and non-normality. Unfortunately, there are also limitations with each method. Until some exploratory data analysis is done to check how the methods fit the data, we should not make a final decision about which technique to use. We outline both below.

Transformations

Transformations are a common method for treating non-normal data. This procedure involves changing each score in a systematic way (e.g., taking its square root) so that the distribution of scores is consistent with the assumptions of the statistical tests that will be run on the scores (e.g., making the distribution more normal in form). Standard analytic methods such as ANOCOVA or linear regression can then be used with the transformed scores. It is not unusual for one transformation simultaneously to (1) reduce interactions among the independent variables, (2) reduce heterogeneity of variance, and (3) produce a more normally distributed dependent variable. When the dependent variable is a count, it is common to use the square root transformation. Unfortunately, we cannot use that transformation for DMFS increments because they are sometimes negative. Slight modifications, such as adding a constant before taking the square root, are available.

A drawback of transformations is that they make it more difficult to interpret results. For example, if, on average, the square root of the DMFS increment is 0.2 lower in regimen 4 than in regimen 6, it would mean that as the DMFS increment in regimen 6 increased, the benefits from treatment regimen 4 would increase. A table showing how the benefit varies with the increment level would make this statement more precise. To the extent that a transformation uncovered any relationship between treatment effects and increment level, this method would be especially informative. Unfortunately, even if a "perfect" transformation exists, there will be some difficulty in discovering it empirically.

Modelling DMFS Increments

An alternative to transforming increment scores is to fit a probabilistic model to the raw scores. Such a model must specify both the distribution of increment scores and how this distribution varies as a function of treatment and control variables. A possible model for the distribution of DMFS scores would be the negative binomial distribution. Granger and Reid (1954) fit this distribution with some success to the number of smooth surface caries for eleven-year-old children. Specification of the rest of the model would require an equation relating the mean of the distribution to a function of the predictor variables. For example, the equation might be that the logarithm of the mean is a linear function of the predictor variables. Unfortunately, the negative binomial distribution, which only allows nonnegative integers, cannot be fit to DMFS increments because some increments are negative. A possible course would be to try to model the final DMFS score, using the baseline score as a predictor.

If a model can be found that fits the data well, it will solve the problems of heterogeneity of variance and non-normality. Furthermore, interactions can be handled much more easily than in a linear regression model. However, any model must be fit empirically. If the postulated model fits poorly, there are the same types of complications that arise from ill-fitting classical models. Unfortunately, even large data sets are often inadequate for testing goodness-of-fit of fairly complex models.

LIST OF TABLES FOR TREATMENT ANALYSIS

Two Year Continuous Residence Sample

Table 1. Number of children by fluoridation status, site, cohort, and regimen.

Table 2. Mean DMFS at baseline and Exam #03 by fluoridation status, cohort, and regimen.

Table 3. Mean DHC at baseline and Exam #03 by fluoridation status, cohort, and regimen.

Table 4. Mean age, SES, and percentage in each racial/ethnic group by fluoridation status, cohort, and regimen.

Table 5. Mean DMFS increment score by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 6. Mean DMFT increment score by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 7. Mean increment in DMFS (by surface type), DS, FS, DS/DMFS, and FS/DMFS by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 8. Mean increment in DT, MT, FT, DT/DMFT, and FT/DMFT by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 9. Mean change in DHC score by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and baseline DHC.

Table 10. Mean change in referral code by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Four Year Continuous Residence Samples

Tables 11-20 will be the same as 1-10. Tables 11-22 will be limited to children who received a clinical examination at both baseline and the end of the Program. Tables 24-27 will be limited to children who had readable radiograph examinations at the end of the Program. Tables 28-30 will be limited to children who received all three clinical examinations (i.e., Exam #01, #03, and #05).

Table 21. The mean number of times a child received each treatment component by fluoridation status, site, cohort, and regimen for children in the four year continuous residence sample.

Table 22. Distribution of four year DMFS increment scores in each regimen by fluoridation status and cohort controlling on age and site.

Table 23. Sample size attrition at the end of two and four years by fluoridation status, site, cohort, and regimen.

Table 24. Number of children with radiograph data at both Exams 01 and 05 by fluoridation status, site, cohort, and regimen.

Table 25. Mean DMFS increment with and without radiograph results included by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 26. Mean DS, FS, and interproximal increments with and without radiograph results included by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 27. Mean DMFT increment with and without radiograph results included by fluoridation status, site, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 28. Mean DMFS scores on Exams #01, #03, and #05 by fluoridation status, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 29. Mean DMFT scores on Exams #01, #03, and #05 by fluoridation status, cohort, and regimen controlling on age, sex, SES, and number of teeth at risk at baseline.

Table 30. Mean DMFS and DMFT two and four year increment scores by fluoridation status, age, and regimen controlling on age, sex, SES, and site.

Table 31. Mean DMFS increment by fluoridation status, cohort, regimen, and baseline DMFS (high, medium, and low) controlling on site. The level of baseline DMFS will be specified relative to age.

Chapter 5

EXAMINER RELIABILITY

INTRODUCTION

In the context of the National Preventive Dentistry Demonstration Program, the term "reliability" refers to the consistency (or agreement) between two or more measurements of the same oral condition. The examination process would therefore be classified as reliable if two examinations of a child on a given occasion led to the same assessment of the extent to which that child had tooth decay, gingivitis, or some other characteristic of interest. Reliability is, of course, a matter of degree. Thus, the focus of any measurement of examiner reliability is on the extent to which there is agreement as to whether a tooth surface is decayed, a gum area inflamed, etc.

The major reason for the Program's interest in examiner reliability is that any inconsistency in the measurement of a given level of decay (or some other oral condition) could affect the likelihood of detecting actual differences in the outcomes of the various treatment regimens. In other words, if there is any chance variation in the score assigned to a certain degree of decay, then this variation could reduce the likelihood of determining whether one regimen is better than another. For example, sealants may be an effective way of preventing tooth decay, but if decay scores were assigned randomly to children rather than in terms of how much decay they really had, then the measurement process would be completely "unreliable." Under these conditions, there would be no difference in the average decay scores of children who did receive sealants compared with those who not receive them even though the two groups might differ substantially in their average level of decay. Thus, unreliability in the data tends to mask (make it harder to detect) real differences between treatments.

This chapter describes the analyses that will be conducted to assess the degree to which examiners were consistent with themselves and with each other in the measurement of various oral conditions. Where appropriate, these analyses will also investigate the degree of examiner agreement within and between years. Some of the more important questions that will be answered by these analyses are:

- When examiners were inconsistent, what were the most common types of inconsistencies (e.g., sound vs. carious, sound vs. deciduous, etc.)?

- To what degree did examiner inconsistencies affect the chances of detecting actual differences between treatment regimens? And, was this influence the same across cohorts within each fluoridation status (i.e., fluoridated versus nonfluoridated)?

- Did the examiners have systematic biases; e.g., did some examiners have a greater tendency to call a tooth surface carious than did other examiners? If so, to what degree did these differences between examiners affect the measurement of treatment effects.

- Were some examiners significantly more reliable than others? If so, were differences between examiners in their respective levels of reliability consistent across dependent variables; e.g., were the examiners with the most reliable DMFS scores also the ones with the most reliable DHC scores? Was reliability related to examiner characteristics, such as the tendency to examine quickly or label borderline lesions as carious?

- How much would reliability have been increased if a child had been seen by the same examiner each year rather than by a different one? Were examiners sufficiently more consistent with themselves than they were with each other over time to argue for maintaining the same child/examiner pairings across examinations?

- Was examiner reliability a function of the child's age, sex, race, decay level, or other background characteristic?

- How reliable were the radiograph readers? Were some more reliable than others? How reliable were the combined clinical/radiograph DMFS and DMFT scores?

DATA SOURCES

The analyses of examiner reliability will use both cross-sectional and longitudinal data. The sources of these two kinds of data for assessing interexaminer and intraexaminer consistency are summarized in Table 5.1.

Table 5.1

DATA SOURCES FOR EXAMINER RELIABILITY ANALYSES

Focus	Concurrent	Longitudinal
Interexaminer Consistency	A child is seen by two examiners on the same day.	A child is seen by one examiner on one occasion and a different examiner on another occasion with a 12 month or longer interval between occasions.
Intraexaminer Consistency	A child is seen by the same examiner twice on the same day.	A child is seen by the same examiner on two or more occasions with a 12 month or longer interval between occasions.

The concurrent data were obtained by conducting two examinations per child on the same day for a random sample of approximately 10 percent of the children in the grade 1, 2, and 5 cohorts. About one-half of these children were seen by the same examiner twice (to assess intraexaminer agreement) and the remaining one-half were seen by two different examiners (to assess interexaminer agreement). Within a site, the children in each of these two subsets in each cohort were assigned randomly to examiners. Thus, the number of children in a subset who were seen by a given examiner was a function of the number of sites at which that examiner worked and the number of children at those sites. It was not a function of how quickly an examiner worked because all the examiners who visited a particular site participated in about the same number reliability examinations.

Concurrent interexaminer data will be gathered on all five examinations. Concurrent intraexaminer checks will be conducted on all but the last two examinations. This change was made for the following reasons: (1) with fewer children to examine per school, not as much time could be provided between examinations of the same child on a given day (i.e., to reduce the chances of the examiner remembering what was called previously); (2) intraexaminer checks are logistically more difficult to carry out; (3) substantial data on intraexaminer agreement had already been obtained on the first three examinations; and (4) the primary questions about reliability concern interexaminer agreement. Moreover, by dropping the intraexaminer checks on the last two examinations, it was possible to modify the examination procedures so that a child's second examination occurred almost immediately after the first one. This permitted the children who received duplicate examinations to stay with their classmates rather than having to come to the examination room twice (as was done with Exams #01, #02, and #03). This aspect of the procedures reduced the likelihood that the examiners would identify the children in the reliability analysis group.

Many of the examiners who participated at baseline did not participate during one or more of the subsequent years of the Program. This was done primarily to conserve funds because with the yearly reductions in the number of children to be examined at each site, a smaller team of dentists could still complete the examinations in the four to five days allocated for this activity. The smaller number of examiners also facilitated calibrating examiners with respect to the Program's diagnostic criteria. The decision of whether to retain an examiner was based on several factors, including availability, degree of agreement with other examiners in DMFS calls, and examination speed.

The examiners who were retained by the Program tended to be reassigned to the site(s) they had visited at baseline. An effort was also made to assign examiners within sites to the same children they had seen previously. However, intrasite school transfers, the practical constraint of completing the examinations on schedule, and similar considerations further limited the program's ability to maintain the same examiner/child pairings over time. Thus, a child who was seen by a given examiner at baseline may not have been seen by that same examiner at all the subsequent annual examinations.

SOURCES AND TYPES OF UNRELIABILITY

Unreliability can be due to chance and/or systematic factors. For example, distractions and fatigue may cause an examiner to make an inappropriate classification of the status of a tooth or a recorder to mark the child's examination form improperly. Systematic biases also may cause discrepancies, such as one examiner tending to classify borderline conditions as carious and another examiner tending to call that same condition sound.

Examiner inconsistencies are likely to be manifested in one of three ways. The first involves differences in how a given condition is classified, such as one examiner calling a tooth sound and another calling it deciduous, or one calling it sound and another calling it carious. Whether or not such misclassifications affect results is, of course, a function of the dependent variable under investigation. For example, a sound versus deciduous error would have no impact on DMFS score comparisons, but it would be important for studying the effect of fluoride on eruption patterns. The reverse would be true for a carious versus sound misclassification.

An analysis will be made of the extent to which various types of examiner inconsistencies occurred, such as a tooth being classified as sound on one examination and as decayed on another examination. The results of this analysis will be presented via a cross-tabulation of the frequency and percent of each of the 28 possible ways in which two calls on a tooth could differ (i.e., all possible pairs of S, D, U, E, Y, carious, filled, and both carious and filled). This analysis will be run separately for interexaminer and intraexaminer data collected each year. Results will be presented for cohorts 1 + 2 (as a group) and for cohort 5 separately by fluoridation status. A similar analysis will be run to determine the number of sound versus carious (and/or filled) disagreements by surface type (i.e., occlusal, buccal-lingual, and interproximal).

The second way in which different examinations of the same children may be inconsistent is in terms of how they rank order children on a particular dimension, such as disagreeing on which children have the most carious surfaces and which have the fewest. Finally, examiners may differ in score level. For example, one examiner may tend to call more surfaces carious than another examiner even though they both rank order the children in about the same way.

Recognition of the various sources and types of errors that could occur led the Program's staff to adopt a number of procedures in an effort to improve reliability. These procedures included the use of formal, written diagnostic criteria (with slides to illustrate different conditions), standardized examination equipment and methods, and extensive training and calibration sessions for examiners. A detailed description of these procedures appears in a previous report. The indices described below provide a means for determining just how well these procedures worked.

CLINICAL DMFS/DMFT SCORES: CONCURRENT INDICES OF RELIABILITY

Concurrent studies of examiner reliability have usually used correlational or categorical indices of examiner agreement. The correlational approach has involved computing Pearson Product Moment and intraclass correlation coefficients. The major reasons for using this approach are (1) these statistics are familiar to most researchers and (2) they focus on the dependent variable(s) used in the analysis of treatment effects (i.e., DMFS or DMFT scores).

One important assumption underlying correlational techniques is that the joint distribution of the scores being correlated does not differ markedly from a bivariate normal distribution. DMFS and DMFT scores do not satisfy this requirement because their distributions are highly skewed. Thus, correlational techniques are generally not appropriate with DMFS and DMFT scores (a discussion of this topic is presented in Appendix A).

Categorical approaches involve assessing how well the calls made on one examination of a child match (i.e., replicate) the calls made on another examination of that same child. For example, did both examinations classify a given surface on a particular tooth in the same way? The primary advantage of this approach is that it can be very sensitive to any differences in how teeth and surfaces are classified. Thus, it could be used to differentiate among examiners who were more or less reliable. An example of such an index is the Consistency Ratio (DePaola and Alman, 1972). In this ratio, cc = carious on both exams, cs = carious on first exam only, and sc = carious on second exam only.

$$\frac{cc}{cc + cs + sc}$$

Two major disadvantages of categorical indices are that they are not readily interpretable and there is no agreement in the field as to which one should be used. More importantly, they do not provide information with respect to the primary reason for being concerned about examiner reliability (i.e., assessing the degree to which examiner inconsistencies reduce the chances of detecting treatment effects). Thus, their value is limited to comparing examiners.

The foregoing considerations led us to use a set of three indices of examiner reliability that would circumvent the problems with categorical measures while still capitalizing on their respective advantages.* These three indices are as follows: the Average Squared Difference (ASD), the Average Absolute Difference (AAD), and the Average Observed Difference (AOD). The procedures used to compute each index are shown in Table 5.2.

*See Slakter et. al. (1976) or Haugejorden and Slack (1975) for further discussion of this point.

Table 5.2

ILLUSTRATION OF PROCEDURES USED TO COMPUTE RELIABILITY INDICES

	DATA		RELIABILITY INDICES		
Child	DMFS on Exam #1	DMFS on Exam #2	Observed Difference	Absolute Difference	Squared Difference
A	2	1	1	1	1
B	1	2	-1	1	1
C	2	0	2	2	4
D	3	3	0	0	0
Sum	8	6	2	4	6
Average	2.0	1.5	0.5	1.0	1.5

The average squared difference (ASD) provides the key to determining the extent to which examiner inconsistencies may have hindered the detection of treatment effects. In other words, it will be used to measure the extent to which examiner inconsistency increases the standard errors of the estimated treatment effects and decreases the probability of detecting effects of a given size (the power of statistical significance tests). The reason is that the average squared difference provides an unbiased estimate of the variance due to examiner error; i.e., (ASD) = (2)(variance due to examiner error). Appendix B contains a more complete discussion of this topic.

The results of the computations of statistical power, using the ASD, will be presented in tabular form. These tables will indicate the probability of detecting a score difference of a given size between the beginning and end of the Program when examiner inconsistencies are and are not present. The calculations for this table will consider the sample size for the comparison, the correlation between a child's score on the first and last examination, and the variance of the scores on each examination.

Because the ASD is very sensitive to any large disagreements, it has limited value for assessing individual examiner reliability. The ASD for an otherwise consistent examiner could balloon with a single large disagreement. A more stable estimate of an examiner's reliability is provided by the average absolute difference (AAD) because it is more sensitive to the how much the examiner usually disagrees than it is to the size of the largest disagreements. The AAD is also directly interpretable; e.g., an AAD of 1.0 in DMFS scores means that on the average, the examiner disagrees with another examiner on one surface per child.

The AAD will be used to assess the degree to which examiners differed from one another in their respective levels of interexaminers and intraexaminer reliability. This will be done by means of one set of linear regressions for interexaminer absolute differences and another set of linear regressions for intraexaminer differences. In both sets of regressions, (1) the dependent variable is the absolute difference, (2) there is one observation for every child reexamined, (3) there is one predictor for every examiner, and (4) each of the following child characteristics are included as additional predictors: age, sex, site, teeth at risk, and the average D component on the two examinations. These child characteristics are included to adjust for the likelihood of making an error when examining the child. On all annual examinations after baseline, there will also be one predictor for each treatment regimen because the number of borderline calls may vary systematically with regimen. If regression diagnostics indicate that the variance of the absolute difference increases with its predicted value, then weighted rather unweighted least squares will be used.

Each examiner's coefficient in the equations above provides an AAD score for that examiner that is adjusted for both interexaminer pairings and for the likelihood of making an error relative to the characteristics of the children examined. The coefficients (i.e., AAD scores) for each examiner will be presented in a table along with an indication as to whether a given examiner's AAD score deviated significantly from the AAD scores of the other examiners. The average AAD score by fluoridation status and cohort controlling on age, site, and sex also will be presented.

The average observed difference (AOD) provides an index of the extent to which there were systematic biases in classifying borderline conditions as carious versus sound. For example, an examiner with an AOD of +1.0 would have a tendency to call an average of one more surface per child than would other examiners. Systematic biases of this kind could affect the measurement of treatment effects because it was not feasible to have all examiners see an equal number of children at every Program school (and thereby counterbalance any biases that might exist). Moreover, schools rather than children were assigned to regimens, and these assignments were made after the baseline data had been collected. Thus, the primary purpose of computing AOD's for each examiner is to determine the degree to which the pattern of examiner assignments to schools indirectly led to a regimen having a higher or lower average score than another regimen. If it appears that the assignment pattern led to a regimen having a higher or lower average score than it deserved (i.e., because of the systematic biases of the examiners), then its average score would be adjusted accordingly in the analysis of treatment effects.

The adjusted value of an examiner's AOD will be used to investigate whether that examiner tended to call (or not call) borderline surfaces as carious. This score is the ratio of the the sum of the observed differences (adjusted for interexaminer pairings) to the sum of the estimated AAD scores for the children examined (where the estimate is based on the child's age, regimen, sex, and teeth at risk) after this ratio has been scaled by the overall mean AAD score. The formula for computing an examiner's adjusted AOD score is given by the following equation:

$$\text{Examiner's Adjusted AOD Score} = \left[\text{Overall Mean AAD Score}\right] \left[\frac{\text{Sum of the ODs Adjusted for Examiner Pairings}}{\text{Sum of the Estimated AAD Scores}}\right]$$

The procedure for adjusting the observed differences for interexaminer pairings will be the same as that used with the AAD scores, except that the child characteristics will not be included because the likelihood of an error should not be systematically related to its direction. No adjustment is necessary for intraexaminer data because such data only involve examiners being paired with themselves.

The examiner's estimated AAD score is used to adjust the observed differences for the number of times the examiner is likely to make an error (i.e., relative to the characteristics of the children examined).

The numerator of the ratio indicates the direction of the bias (i.e., + = calls more than average and - = calls less than average). The ratio itself provides an indication of the average size of an error when an error occurs. This ratio will therefore tend to range between -1.00 (i.e., all the errors were due to calling too few) to +1.00 (i.e., all the errors were due to calling too many). The examiner's ratio is then multiplied by overall AAD score to reflect the magnitude of the bias (i.e., after adjusting for both examiner pairings and the characteristics of the children seen by the examiner).

An F-statistic will test the hypothesis that all the examiners are unbiased. If this hypothesis is rejected (indicating that some examiners call systematically higher than others), then there may be a need to adjust each child's DMFS score to correct for who conducted his/her examinations. Fortunately, that adjustment would only be necessary if those dentists with systematic bias examined more children from certain regimens than from others. If the imbalance of examiner/regimen assignment is large enough (in conjunction with the size of systematic bias), then the adjustments will be made.

The AOD also will be used to investigate whether there were any systematic differences in DMFS scores between the first and second concurrent examinations of a child; i.e., did one of these scores tend to be higher or lower than the other? This set of analyses will explore this issue in terms of cohort and fluoridation status. If systematic differences are obtained, then it may be inferred that some of the discrepancies between examinations were due to the procedure of conducting concurrent examinations (such that the condition of one or more teeth was affected by the examination process) rather than to differences in the application of the Program's diagnostic criteria. In other words, the call on a tooth when it is examined for a second time may be different from the call the first time because the condition of the tooth was changed by the first examination. If that has happened, then concurrent indices of examiner consistency will underestimate the true level of agreement by the size of the AOD between the two examinations.*

To contrast the AAD with reliability indices that have been used in other studies, the consistency ratio (CR) described previously will be computed on the interexaminer and intraexaminer data collected each year. The CR scores will be adjusted in the same way as were the AAD scores.

---

*The calculation of AOD and the other indices discussed in this section may exclude children with all deciduous teeth at baseline.

CLINICAL DMFS/DMFT SCORES: LONGITUDINAL INDEX OF RELIABILITY

The standard longitudinal index of examiner reliability is the reversal. A reversal occurs when a tooth or surface is classified as decayed, missing, or filled on one examination and as sound (or unerupted) on a subsequent examination. The rationale for studying reversals is that the decay process is irreversible. Thus, examiner and/or recorder error are considered to be the only explanations for a tooth (or surface) progressing from carious to sound. The reversal rate between two examinations is the number of reversals from the first to the second examination divided by the the number of decayed, missing, and filled calls on the first examination (i.e., the number of possible reversals). It is generally further assumed that the actual number of misclassifications is two times the reversal rate, since a tooth could be classified as sound on the first examination when in fact it was carious (however, more true decay is likely to occur the longer a child's teeth are at risk).

Reversal rates will be calculated for each examiner who participated in more than one annual examination cycle. These data will be subjected to an analysis of covariance where the dependent variable is reversal rate (or some transformation of rate to adjust for possible skewness in this variable), the independent variable is examiner, and the covariates are age, sex, site, treatment regimen, and the average of the DMFS scores across the two examinations. Reversal rates also will be calculated on a Program-wide basis broken down by regimen and fluoridation status (while controlling on age, sex, and teeth at risk).

A separate analysis will be run to determine the effect on reversal rate of maintaining versus not maintaining the same child/examiner pairings over time. This will be done by dividing the children into two groups, namely those who were and those who were not seen by the same examiner across examinations. The reversal rate in each group will then be contrasted, using an analysis of covariance where the dependent variable is the number of reversals per child examined; the factors are type of child/examiner pairing (same versus different across examinations), fluoridation status, and cohort (1 + 2 versus 5); and the covariates are age, sex, site, average DMFS across the two examinations, and treatment regimen. This analysis will be run on Exam #01 to #03 data and on the Exam #01 to #05 data.

RADIOGRAPH READER RELIABILITY

A stratified random sample of children with readable radiographs at baseline was selected for the reader reliability analyses. Only children whose films were evaluated by one of the six major readers were included in this sample. The stratification variables used in selecting the sample were site, school, cohort (1, 2, and 5), and initial reader. About 50 percent of the children selected had their films read again by the same dentist who evaluated their films initially. The remaining 50 percent had their films read again by one of the other five dentists.

A count will be made of films classified as readable on the initial evaluation but categorized as nonreadable on the second evaluation. These films will be excluded from subsequent reliability analyses.

A cross-tabulation (controlling on fluoridation status and cohort) will be run on the interreader and intrareader subsamples to determine the number and percentage of each of the 21 possible discrepancies that could have occurred on a tooth (i.e., a reader could classify a tooth as sound, unerupted, missing, nonreadable, filled, carious, or both filled and carious). If the reader classified either the distal or mesial surface of a tooth as carious, then the tooth is classified as carious. The same rule applies to fillings. If one surface is carious and the other is filled, then the tooth is classified as both. The both category is also used if a surface is classified as being both carious and filled.

Observed and absolute differences in radiograph DMFS and DMFT scores between the two readings of a set of films will be analyzed in the same way as the clinical DMFS and DMFT scores. A decision rule will have to be developed for handling surfaces (or teeth) that were classified as non-readable on one reading and as carious and/or filled on the other reading. As in the clinical reliability analyses of covariance, an indicator variable for each examiner will be used to adjust for any imbalance in the examiner pairings. The results of the AOD analyses will be inspected to determine what adjustments, if any, have to be made in a child's combined baseline score (i.e., clinical plus radiograph) to control for systematic biases in who read the child's films.

DHC SCORES

The second portion of the clinical examination process involved an assessment of the status of a child's gingival tissues. The procedures and editing rules used in computing this index for each child have been discussed previously. The children and examiners included in the analysis of DHC score reliability will be the same as those used in the clinical DMFS/DMFT analyses. The one exception to this rule will be the exclusion of children who had all deciduous teeth (since no DHC score was computed for these children).

Unlike tooth decay, gingivitis is reversible. Thus, examiner reliability on the DHC index can only be measured by using concurrent data. This will be done by using the same procedures as those employed for the clinical DMFS data, i.e., analyses involving the calculation of ASD, AAD, and AOD scores. In other words, separate analyses will be run for interexaminer and intraexaminer consistency within each fluoridation status for each cohort type. There will also be a check on whether there is a systematic difference between the first versus second of two concurrent examinations of a child. Results will be presented in tabular form by examiner. They also will be presented for each cohort by fluoridation status combination across examiners (controlling, of course, on age, sex, and site).

MISCELLANEOUS INDICES

The last major phase of the clinical examination of a child involved the examiner indicating (1) whether there were any sealants present, (2) whether the teeth showed signs of fluorosis, and (3) the type of letter that should be sent to the child's parents to give them appropriate information about their child's oral status. The first two variables are dichotomies (i.e., present versus absent), while the last variable has four categories (no problem, minor problem, serious problem, and incomplete examination).

Concurrent interexaminer and intraexaminer agreement on all three indices will be assessed in the same way. This will involve computing the percentage of agreement for various subgroups, where the grouping variables include cohort type, sex, fluoridation status, SES, racial/ethnic group, and treatment regimen. A parallel set of analyses for individual examiners may also be carried out; however, these will have to use a reduced number of variables and/or a collapsed set of categories within variables to handle the problem of small cell frequencies. Where appropriate, phi coefficients will be presented.

RELATIONSHIPS AMONG INDICES

A correlation matrix will be computed for each examination where the unit of analysis is examiner and the variables are all the indices of examiner reliability (e.g., interexaminer and intraexaminer AOD and AAD scores). When scores on an index are in the form of phi coefficients, then a z-score transformation will be used prior to computing the matrix. The matrix also will include other data on examiners, such as the average number of children they saw per day, the total number of children they examined, and the number of sites at which they worked (as indicators of experience), etc.

Appendix A

HANDLING THE LONG TAIL OF DMFS SCORES

As our best measure of dental health, the DMFS score (or increment) will be the focus of most analyses. This appendix discusses the analytic problems resulting from the long upper tail of the distribution of DMFS scores. Classical data analysis methods--such as the sample mean, analysis of covariance, and the significance tests resulting from them--are derived for use with normally distributed dependent variables. When one or both tails of the dependent variable are long, extreme observations have too much influence on the conclusions of the analyses. Thus, more robust techniques (i.e., ones that are less sensitive to extreme observations) will often be preferable. Choice of the proper analytic method requires balancing the degree of the problems caused by the long-tailed distribution against the advantages of the classical procedures. The factors which influence this decision are discussed below.

Many classical data analysis methods use a least squares fit. The fitted model is the one among all models which minimizes the sum of squared deviations between the observations and the corresponding predictions. The simplest least squares estimate is the sample mean; other least squares procedures include the standard ways of doing analysis of variance, analysis of covariance, and linear regression. These techniques, which are designed for use with a normal dependent variable, may give misleading results on data from a long-tailed distribution such as that of DMFS scores. One or two very large values will have more effect on the results than a large number of moderate values. Thus, in small samples particularly, least squares estimates tend to be unstable.

Another problem is that standard inference techniques -- t-tests, F-tests, and normal-based confidence intervals -- are derived by using the assumption that the dependent variable is normally distributed. If this assumption is grossly violated, the presumed significance levels may be incorrect. Fortunately, the normality assumption for the dependent variable is not required. Instead, theoretical results suggest that as the sample size increases relative to the complexity of the model, the departure from normality becomes less important.* Computer simulations by Glass et.al. (1972) support this result for highly skewed data from a caries increment distribution. They found that most statistical tests were slightly conservative in rejecting statistical hypotheses.

The magnitude of the problems discussed above is a matter of degree, depending on (1) the extent of the departure from normality in the tails of the dependent variable, (2) the number of estimated parameters in the model (relative to the sample size), and (3) the distributions of the predictor variables. The important aspect of the dependent variable is the frequency and magnitude of extreme observations. If neither tail is especially long, other departures from normality are unlikely to cause much trouble. However, as the tails grow, the resulting problems can increase quickly. We will carefully check the distribution of DMFS scores and increments before finally selecting analytic techniques.

---

* G.E.P. Box and G.S. Watson (1962).

The degree of the problem posed by long-tailed distributions also depends on the complexity of the model being fit. In an ANOCOVA with only one or two factors, and thus only a few cells, there may be enough observations in each cell to overcome the influence of a few outliers. Also, if the sample sizes in the cells are large, the adjusted means will have approximately normal distributions even for a long-tailed dependent variable. This fact is enough to guarantee approximately correct levels for statistical tests of significance. However, as models become more complex in attempts to discover more relationships, the sample is spread more thinly, and the threat from extreme values increases. Of course, what constitutes a large enough sample or a simple enough model to be safe, is all a matter of degree.

A closely related aspect is the distribution of the predictor variables. When the predictor variables are discrete (such as sex or site) or tightly bounded (such as age within an age range), the averaging process has a chance to reduce the unreliability of estimates caused by outliers. However, when the values of both the dependent and a predictor variable vary widely, the problem of unreliable estimates is greatly exacerbated. This is the reason for not using correlational techniques to study the reliability of DMFS scores; any child with very large scores on both examinations (whether they agree closely or not) could have much too large an effect on the correlation coefficient.

Only one predictor variable, baseline DMFS, has an especially long tail. When this variable is used, special care will be taken to ensure that no observations unduly influence the analysis.

While robust alternatives exist, there are tradeoffs in using these methods. If one wants to know the mean DMFS for a group, the only unbiased estimator of that quantity is the sample mean. Any other estimator, such as the sample median, would converge to the wrong number as the sample size grew. In the case of DMFS, a more robust estimator would tend to discount cavities of children with large DMFS scores or increments. Transformations of the dependent variable would have the same effect. While both of these methods have their advantages, we will hesitate to use them if the mean is the outcome of real interest.

Appendix B

EFFECT OF EXAMINER ERROR ON PRECISION OF THE ESTIMATES

To illustrate how examiner error affects the analysis results, the report on examiner reliability will contain tables contrasting our actual ability to make inferences about treatment effects with those which would be possible if there were no examiner error. Using examiner variances estimated from the reliability samples and approximate sample sizes for the continuous residence sample, the tables will give standard errors for estimated treatment effects, both with and without examiner errors. These standard errors determine the ability of tests of hypothesis to find statistically significant differences among the treatments.

Suppose that T_1 and T_2 are the true beginning and end values of some dental health index. In the absence of errors, we might use the increment in scores, (T_2-T_1), as the dependent variable, in which case we have

$$\mathrm{Var}(T_2-T_1) = \mathrm{Var}(T_2) + \mathrm{Var}(T_1) - 2\mathrm{Cov}(T_1,T_2) \ .$$

If, instead, we observe $Y_1 = T_1+u_1$ and $Y_2 = T_2+u_2$, with uncorrelated errors u_1 and u_2, we would have to use the dependent variable (Y_2-Y_1), which has variance

$$\mathrm{Var}(Y_2-Y_1) = \mathrm{Var}(T_2-T_1) + \mathrm{Var}(u_1) + \mathrm{Var}(u_2) \ . \ *$$

The ratio of the variance of true scores over that of the observed scores,

$$r = \frac{\mathrm{Var}(T_2-T_1)}{\mathrm{Var}(T_2-T_1) + \mathrm{Var}(u_1) + \mathrm{Var}(u_2)} \ ,$$

is called the reliability of the increment. The actual experiment provides r times the information of the same-sized experiment without examiner error.

---

* The variances and covariances in these first two equations and the one to follow are really conditional on any predictor variables used to estimate increment scores. For T_2-T_1, the conditional variance will be less than the unconditional variance of the increment. In contrast, since u_1 and u_2 are random errors, their variances will not change by conditioning on predictor variables.

In the treatment analysis, key estimates will be the adjusted mean increments for each treatment regimen. The standard error of the estimate for regimen i depends on the variance of (Y_2-Y_1) in regimen i (conditional on any covariates used to adjust the dependent variable) and the sample size for regimen i. If the sample being used is the fifth grade longitudinal cohort for one of the water supplies, then the sample size, n_i, should be in the range 200 to 300. For the combined first and second grade cohort, the sample size would be about 400-600. The standard error is given by

$$\text{s.e.(mean increment in regimen i)} = [\text{Var}(Y_2-Y_1)/n_i]^{1/2} .$$

Contrasts between two regimens, especially when one is the control group, will also be of interest. Since the estimate of the effect of regimen i is the difference of the (adjusted) mean increments in regimen 6 and regimen i, the standard error of this effect is

$$\text{s.e.(regimen i effect)} = [\text{Var}(Y_2-Y_1)\cdot(1/n_i+1/n_6)]^{1/2} .$$

For each regimen, an important step of the analysis will be to test the hypothesis that the regimen has no effect on the increment of the score. Assuming a true mean increment equal to certain hypothetical constants, denoted by c, we will calculate the probability that a statistically significant difference would be found. Using a one-sided t-test with 5 percent significance level, this probability would be

$$P(\text{test finds an effect}) = F[c/\text{s.e.(regimen i effect)} - 1.645] ,$$

where F(x) is the probability that a standard normal random variable is less than z. * When c = 0 (no treatment effect), the probability of finding one is 0.05. As c increases or the standard error of the regimen i effect decreases, the probability increases that the hypothesis of no effect will be rejected.

The examiner-related quantities which affect the above standard errors and the powers to detect differences are the examiner error variances at the beginning and end of the program, u_1 and u_2. Fortunately, the average squared difference (ASD) provides unbiased estimates of these quantities. Let $Y_{1a} = T_1+u_{1a}$ and $Y_{1b} = T_1+u_{1b}$ be two concurrent examinations of the same child at baseline. The difference between the two scores is $Y_{1a}-Y_{1b} = u_{1a}-u_{1b}$. If the errors, u_{1a} and u_{1b}, are independent and identically distributed, then the expected value of the squared

* Since most interesting contrasts will be between nested treatments, we will only care about differencs in one direction; thus one-sided tests will be appropriate.

difference (Y_{1a}-Y_{1b}) equals $Var(u_{1a}) + Var(u_{1b}) = 2Var(u_1)$. Thus one-half the ASD at baseline is an unbiased estimate of the variance due to examiner error at baseline.

REFERENCES

1. Box, G. E. P., and G. S. Watson, "Robustness to Non-normality of Regression Tests," Biometrika, Vol. 49, 1962, pp. 93-106.

2. Foch, C. B., A Cost Analysis Plan for the National Preventive Dentistry Demonstration Program, The Rand Corporation, N-1670-RWJ, 1981.

3. DePaola, P. F., and J. E. Alman, "Assessment of the Reliability of Radiographic Diagnosis in a Clinical Caries Trial," J. Dent. Res., Vol. 51, 1972, pp. 1431-1437.

4. Glass, R. L., J. E. Alman, S. Fleisch, and R. B. D'Agostino, "The Appropriateness of Analysis of Variance to the Statistical Analysis of Dental Caries Data," Arch. Oral Biol., Vol. 17, 1972, pp. 633-643.

5. Granger, R. M., and D. B. W. Reid, "Distribution of Dental Caries in Children," J. Dent. Res., Vol. 33, 1954, pp. 613-623.

6. Greene, J. C., and J. R. Vermillion, "The Simplified Oral Hygiene Index," J. Amer. Dent. Assoc., Vol. 63, 1964, pp. 7-13.

7. Haugejorden, O., and G. L. Slack, "A Study of Intra-examiner Error Associated with Recording of Radiographic Caries at Different Diagnostic Levels," Acta. Odont. Scand., Vol. 33, 1975, pp. 169-181.

8. Klein, S. P., and H. M. Bohannon, The First Year of Field Activities in the National Preventive Dentistry Demonstration Program, The Rand Corporation, R-2536/1&2-RWJ, 1979.

9. Radike, A. W., "Examiner Errors and Reversals in Diagnosis," Proceedings of the Conference on Clinical Testing of Cariostatic Agents, American Dental Association, 1972, pp. 92-95.

10. Schrader, R. M., and T. P. Hettmansperger, "Robust Analysis of Variance Based on a Likelihood Ratio Criterion," Biometrika, Vol. 67, 1980, pp. 93-101.

11. Slakter, M. J., D. B. Juliano, and S. L. Fischman, "Estimating Examiner Consistency with DMFS Measures," J. Dent. Res., Vol. 55, 1976, pp. 930-934.

12. Suomi, J. D., J. C. Greene, J. R. Vermillion, J. J. Chang, and E. C. Leathwood, "The Effect of Controlled Oral Hygiene Procedures on the Progression of Periodontal Disease in Adults: Results after Two Years," J. Periodontics, Vol. 40, 1969, pp. 416-420.