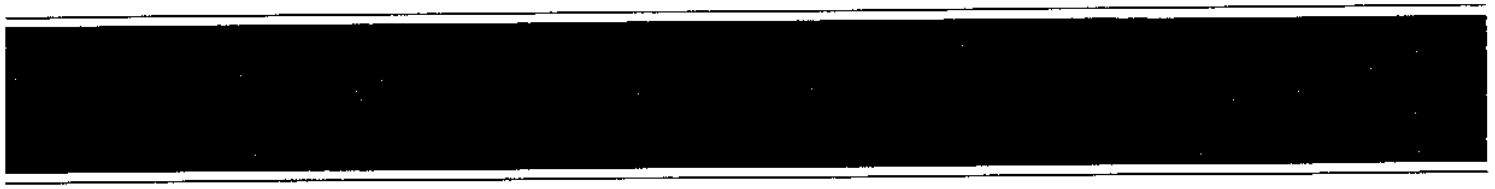
The Reliability of Clinical and Radiographic Examinations in the National Preventive Dentistry Demonstration Program

Stephen P. Klein, Robert M. Bell,
Harry M. Bohannan, Judith A. Disney,
Alisa Wilson



The research reported herein was performed under grants from The Robert Wood Johnson Foundation. The opinions and conclusions expressed herein are solely those of the authors, and should not be construed as representing the opinions or policy of The Robert Wood Johnson Foundation.

RAND PROJECT STAFF

Stephen P. Klein, Director
Robert M. Bell
Sally Carson
Craig B. Foch
John N. Grabowski
Stephanie A. Knapik
Frank H. Leone
Marian Oshiro

AFDH PROJECT STAFF

Harry M. Bohannan, Director
Paul E. Anderson
Judith A. Disney
Richard C. Graves
John W. Stamm

NATIONAL ADVISORY COMMITTEE

Alvin L. Morris, Chairman

James P. Carlos
Frank Dick
Donald J. Galagan
Charles W. Gish
Bernard G. Greenberg
Ronald Johnson
Robert I. Kaplan
Lynden M. Kennedy
Robert S. Morrison
Robert J. Murray
Irwin M. Rosenstock
Frederick J. Stare
Judith Weitz
Sumner Yaffee

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

R-3138-RWJ

The Reliability of Clinical and Radiographic Examinations in the National Preventive Dentistry Demonstration Program

Stephen P. Klein, Robert M. Bell,
Harry M. Bohannan, Judith A. Disney,
Alisa Wilson

June 1984

Prepared under a grant from
The Robert Wood Johnson Foundation

Rand
1700 MAIN STREET
P.O. BOX 2138
SANTA MONICA, CA 90406-2138

RESEARCH SERVICE
AFDH
EDUCATION

PREFACE

The National Preventive Dentistry Demonstration Program was conducted to assess the costs and effects of various types and combinations of school-based preventive dental care procedures. The Program was supported by two grants from The Robert Wood Johnson Foundation of Princeton, New Jersey. One grant went 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 went to The Rand Corporation, a nonprofit public institution that conducts policy research and analysis on problems of national security and domestic affairs.

The initial concept for the Program was developed and presented to The Robert Wood Johnson Foundation by AFDH. Within the Program, AFDH was responsible for providing and supervising the preventive care, collecting most of the data, and conducting annual dental examinations. Rand was 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 10 sites around the country at which it was conducted, preparing reports, and establishing the procedures for other areas of joint responsibility.

The first major activity at a site involved securing parental consent for participation in the Program. This was followed by clinical (visual plus tactile) and radiographic dental evaluations of all participants. Children were then assigned to treatment groups. The children who remained in the Program until its completion received four more annual clinical examinations and one more radiographic examination.

This report presents the results of analyses that assessed the degree to which the Program's 31 clinical examiners and 6 radiograph readers agreed with themselves and each other in their evaluations of the children's oral health. Almost 9,000 children received replicate, concurrent, clinical examinations, and almost 4,000 sets of radiographs were read twice. Longitudinal analyses of examiner agreement focused on a sample of 5,579 children. Thus, the Program's databases enabled us to conduct the most comprehensive study of dental examiner reliability that has ever been undertaken.

Other Program reports present a complete description of: the treatment and examination procedures (Klein and Bohannan, 1979), the procedures used to analyze the dental data (Klein and Bell, 1981), past and present procedures used to assess the costs of preventive dental care (Foch, 1981a and 1981b), surveys of teachers and principals who participated in the Program (Klein et al., 1984), the reasons some parents did not enroll their children in the Program (Leone et al., 1981), and the findings with the baseline examinations (Bell et al., 1982). Three additional reports describe the examination procedures (Bohannan et al., 1984) and the analysis of treatment and cost data (Bell et al., 1984; and Foch et al., 1984).

SUMMARY

The National Preventive Dentistry Demonstration Program was conducted to assess the costs and effects of various types and combinations of school-based preventive dental care procedures. The Program was operated at 10 sites across the country. During the first of its five years of operation, 20,491 first, second, and fifth graders received clinical (visual plus tactile) examinations. These baseline examinations were conducted by 31 examiners. Children who remained in the Program after baseline received four annual examinations. These examinations were conducted by a subset of the 31 examiners who participated at baseline.

About 11 percent of the 20,491 children at baseline received a second clinical examination within a few hours of their first one. About 50 percent of the replicate exams involved a child being seen twice by the same examiner (thereby providing data about concurrent intraexaminer consistency), whereas the remaining children in the replicate sample had one examiner for one exam and a different examiner for the other exam in order to provide information about interexaminer agreement.

Replicate interexaminer data were collected on each of the four annual examinations conducted after baseline. Intraexaminer data were collected for only the two annual exams following baseline because of high intraexaminer consistency and the greater utility of the interexaminer data. By the end of the Program, 5,736 sets of replicate interexaminer exams and 3,058 sets of replicate intraexaminer exams had been conducted.

Because of the reduction in the number of examiners who participated in the Program after baseline, and other logistical considerations, only about 10 percent of the children who were seen by a given examiner on one year's exam had that same examiner on a subsequent exam. This situation enabled us to assess what effect, if any, maintaining versus not maintaining the same examiner/child pairings over time had on the quality of the data collected.

In all five years, the children were examined for DMFS (number of decayed, missing due to decay, and filled tooth surfaces), gingivitis, and other measures of oral health. The Year 1 and 5 data also included 3,592 sets of bitewing radiographs that received replicate evaluations. All the children who had their films read twice also had a clinical exam. However, the six dentists who served as radiograph readers did not have access to the clinical results, just as the clinical examiners did not have access to the radiograph results.

The Program's database allowed us to conduct the most comprehensive study of examiner reliability that has ever been undertaken. This study differed from most past research on examiner reliability in the following ways:

- Both interexaminer and intraexaminer consistency were assessed with samples from the same population of subjects.
- Data were collected on examiner consistency within (concurrent) and across (longitudinal) years.
- All the data were collected as part of the Program's regular examination process, and in general the examiners did not know which children were in the samples used to assess reliability.
- The reliability of both clinical and radiograph examinations was investigated.
- Data were analyzed to assess the effects on reliability of combining clinical and radiograph results.

- Thousands of children received replicate examinations. Moreover, these children varied with respect to age, race, socioeconomic status, treatment regimen, fluoride level of their water supply, and geographic region.
- More than 30 dentists participated as clinical examiners and/or as radiograph readers.
- Reliability was studied for several oral health indexes in addition to DMFS.
- Several analytic techniques were used to assess reliability on each database.

The analyses of concurrent clinical DMF data indicated that examiners were slightly more consistent with themselves than they were with each other. Their overall level of reliability was comparable to that of examiners in other studies. About one-half of all the discrepancies among examiners involved disagreements as to whether surfaces were sound or decayed. However, disagreements did not occur often enough to seriously interfere with the detection of possible treatment effects.

The frequency of disagreements decreased over time. This occurred despite older children having more carious surfaces than younger ones and thereby having more opportunity for disagreements. This trend was probably due to improved examination conditions after Year 1, the selective retention of examiners, and the continued development of examiner expertise.

The analyses of concurrent clinical DMF data also showed that some examiners were slightly more consistent than others, but that there was usually very little difference in their mean scores and these differences were not especially stable from one year to the next.

The analyses of the longitudinal DMF data indicated that maintaining examiner/child pairings over time had essentially no effect on reversals (surfaces classified as carious one year and as sound two years later). Children in the Program had about as many reversals as children in other recent studies.

Examiners agreed far more with themselves than they did with each other in their evaluations of gingivitis. Moreover, differences in average scores among examiners tended to be consistent over time. Gingival scores will therefore have to be adjusted for systematic differences among examiners before they can be used to assess possible treatment effects.

The assessment of general oral health in the form of a referral code and the determination of whether sealants were present were generally highly reliable within and between examiners. The referral code also may be sensitive to treatment effects because of its close relationship with the D component of DMFS scores.

Radiograph readers generally agreed with themselves and each other as to whether a proximal surface was sound, decayed, filled, missing, unerupted, or unreadable. Most of the few inconsistencies that did occur involved disagreements as to whether a surface was sound or decayed. There were no major differences in average score among readers.

Very few surfaces classified as sound by the clinical examiners were classified as affected by decay by a radiograph reader. An analysis of the surfaces whose classification would have been affected by the addition of radiograph results to the clinical findings revealed that the readers were more likely to disagree than agree with each other as to whether a surface was sound or carious. Thus, the radiograph results probably should not be combined with the clinical findings for the purposes of measuring treatment effects.

ACKNOWLEDGMENTS

The preparation of this report benefited greatly from the advice and assistance of many persons. Rose Mitchell and Jon Whinnery in Rand's Library were extremely helpful in securing reference materials. Sally Carson, John Grabowski, Roger Madison, and Marian Oshiro assisted in cleaning and analyzing the data.

Dr. John W. Stamm of McGill University and Dr. John Ware of Rand reviewed earlier drafts of the report and provided many useful suggestions for its improvement. Becky Goodman provided expert editorial guidance.

Paul Anderson of AFDH and Frank Leone of Rand were especially helpful in coordinating and supervising the examination activities. These activities required the cooperation and assistance of the dentists who served as examiners and radiograph readers, radiograph technicians, the community volunteers who served as recorders and aides, principals and teachers, the Program's staff at each site, and the Westinghouse Learning Cooperation, which processed the examination forms. The universal good nature of the children who complied with our seemingly bizarre request that they undergo replicate examinations was essential to our research, and we are grateful to them.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS vii

TABLES xi

Chapter

1. INTRODUCTION 1
   - Outline of the Report 1
   - Reasons for Measuring Examiner Consistency 1
   - Purpose 2
   - Literature Review Procedures 2
   - Reliability of Clinical Evaluations of DMF 2
   - Reliability of Radiograph Evaluations 8
   - Gingivitis 9
   - Characteristics of Past Research 9
   - Research Questions 12

2. PROCEDURES 14
   - Program Description 14
   - Clinical Examinations 15
   - Examiner Training and Calibration 16
   - Clinical Examination Equipment, Staff, and Conditions 16
   - Replicate Clinical Examinations 17
   - Radiographic Examinations 21

3. RELIABILITY OF CLINICAL DMFS EXAMINATIONS 22
   - Introduction 22
   - Number and Types of Inconsistencies 23
   - Inconsistencies in Concurrent DMFS Calls 24
   - Inconsistencies in Longitudinal DMFS Calls 30
   - Effect of Examiner Inconsistencies on Detecting Treatment Effects 35

4. RELIABILITY OF OTHER INDEXES OF ORAL HEALTH 37
   - Gingivitis 37
   - Referral Code 40
   - Sealants 41
   - Opacity 43

5. RADIOGRAPH READER RELIABILITY 45
   - Reread Samples 45
   - Frequency of Reader Inconsistencies 46
   - Effect of Reader Inconsistencies on DFS Scores 46

6. DISCUSSION OF RESULTS 49
Clinical Evaluations of DMFS 49
Radiographic Evaluations of DFS 50
Other Indexes of Oral Health 51
Conclusions 51

Appendix
A. CLINICAL AND RADIOGRAPH EXAMINATION FORMS 53
B. DIAGNOSTIC CRITERIA FOR DMF SCORES 56
C. CRITERIA FOR RADIOGRAPH EXAMINATIONS 59
D. ESTIMATION OF INDIVIDUAL EXAMINER DIFFERENCES 61
E. PROCEDURE FOR ESTIMATING POWER TO DETECT TREATMENT EFFECTS 64

REFERENCES 67

TABLES

1.1 Studies Measuring Intraexaminer Consistency in Conducting Clinical Examinations 4
1.2 Studies Measuring Interexaminer Consistency in Conducting Clinical Examinations 5
1.3 Studies Measuring Examiner Consistency in Using Gingival Indexes 10
2.1 Sites Participating in the National Preventive Dentistry Demonstration Program 14
2.2 Number of Children and Examiners Participating in the Program 16
2.3 Number of Children Receiving Each Type of Clinical Reexamination Each Year, by Cohort 18
2.4 Mean Clinical DMFS Score in the Program and in Each Reexamination Sample, by Year 19
2.5 Standard Deviation of Clinical DMFS Scores in the Program and in Each Reexamination Sample, by Year 19
2.6 Number of Children Each Examiner Saw in Each Year's Interexaminer and Intraexaminer Samples 20
3.1 Data Sources for Examiner Reliability Analyses 22
3.2 Number and Percentage of Disagreements Relative to Number and Percentage of Children Examined 23
3.3 Percentage of Disagreements Within Each Type of Concurrent Examination, by Type of Disagreement and Cohort 24
3.4 Difference in Average DMFS Score Between the First and Second Concurrent Examination 25
3.5 Average Observed Differences, by Examiner and Year 26
3.6 Percentage of Variance in Observed DMFS Difference Scores Due to Systematic Differences Between Examiners, by Year and Cohort 27
3.7 Correlation Coefficients and Consistency Ratios Between the First and Second Concurrent Exam, by Year and Cohort 27
3.8 Percentage of Children with Each DMFS Absolute Difference Score 29
3.9 Percentage of Children with Each DMFT Absolute Difference Score 29
3.10 Average Absolute Difference Scores, by Cohort, Year, and Type of Reliability Sample 30
3.11 Average Absolute Difference Scores, by Examiner and Year 31
3.12 Number of Surfaces Available for Longitudinal Studies of Examiner Consistency 32
3.13 Percentage of Children Who Had the Same Examiner for a Given Pair of Exams 32
3.14 Percentage of Surfaces Classified as Affected by Decay on One Exam But Not Classified as Affected by Decay Two Years Later, by Cohort and Whether Examiner/Child Pairings Were or Were Not Maintained 33
3.15 Average Number of Reversals Per Child from Year 1 to 3 and from 3 to 5, by Cohort and Whether or Not Examiner/Child Pairings Were Maintained 33

3.16 Percentage of Children with a Non-Zero DMFS Score in One Year Who Had One or More Reversals Two Years Later, by Cohort and Whether or Not Examiner/Child Pairings Were Maintained 34
3.17 Reversal Rate for Surfaces Classified as Decayed in Base Year, by Surface Type . . 34
3.18 Percentage of Surfaces Reversed from Year 1 to 3 and 3 to 5, by Treatment Regimen, Cohort, and Fluoridation Level 35
3.19 Probabilities of Detecting Two- and Four-Year Treatment Effects when Variance Due to Examiner Inconsistencies Is and Is Not Present 36
4.1 Average DHC Score by Site in Year 1 38
4.2 Correlation Between DHC Scores Within Each Reliability Sample, by Cohort and Year 39
4.3 Percentage of Children on Which Examiners Agreed or Disagreed with Each Other in the Referral Codes that Should be Assigned 40
4.4 Percentage of Children on Which Examiners Agreed or Disagreed with Themselves in the Referral Codes that Should be Assigned 41
4.5 Percentage of Children in Each Reliability Sample on Which Examiners Agreed or Disagreed About Presence of Sealants 42
4.6 Percentage of Children in Each Reliability Sample on Which Examiners Agreed or Disagreed About Presence of Opacities 43
5.1 Number of Film Sets in Each Reread Sample 45
5.2 Percentage of Agreement Between Readings on Each Type of Call 46
5.3 Percentage of Children with Various Absolute DFS Difference Scores 46
5.4 Percentage of Year 1 Surfaces Classified as Affected on the Clinical Exam and/or on One or Both Radiograph Readings 47
5.5 Percentage of Year 5 Surfaces Classified as Affected on the Clinical Exam and/or on One or Both Radiograph Readings 48
E.1 Power to Detect Two-Year Treatment Effects in Cohort 5 with One- and Two-Tailed Tests when Variance Due to Examiner Inconsistencies Is and Is Not Present 66

Chapter 1

INTRODUCTION

OUTLINE OF THE REPORT

This chapter summarizes the reasons for measuring examiner reliability and reviews the characteristics and findings of recent studies that have assessed it. The chapter also discusses some of the important reliability issues that have not been addressed (or addressed adequately) by past research efforts, but which we were able to investigate with data collected in the National Preventive Dentistry Demonstration Program.

Chapter 2 describes the Program and the procedures that were used to collect the data discussed in later chapters. Chapter 3 presents data on interexaminer and intraexaminer consistency in assessing dental decay both within a year and across years. Chapter 4 contains reliability results with other indexes of oral health, such as the amount of gingivitis (inflammation of the gums) that was present. Chapter 5 describes the degree to which the Program's radiograph readers agreed with themselves and each other in interpreting the X-ray films. The final chapter discusses the implications of the findings for the Program's analysis of treatment effects.

REASONS FOR MEASURING EXAMINER CONSISTENCY

Clinical trials of the efficacy of preventive dental care procedures usually involve measuring whether these procedures reduced the number of decayed (D), missing due to decay (M), and filled (F) teeth (T) or tooth surfaces (S) that would have been present if the procedures had not been used. Epidemiologic studies also use DMFT and/or DMFS scores to assess the prevalence and incidence of dental decay. Both types of studies sometimes include other indexes of oral health, such as the amount of gingivitis that is present.

Information about the accuracy with which scores on oral health indexes are assigned is essential for determining the degree to which chance and systematic factors mask or bias treatment effects (Jackson, 1966). Chance (or accidental) errors, such as inconsistent use of examination criteria and random errors in marking recording forms, produce greater similarity in average scores among groups than the amount of decay actually present would justify. Systematic factors, such as one examiner tending to classify borderline conditions as carious and another calling them sound, could bias results toward greater or less similarity in average scores among groups. Chance and systematic inconsistencies also pose problems for interpreting epidemiologic data, such as the degree to which the presence or absence of differences in the prevalence of decay among sex, race, or regional groups is a function of examiner error.

There are three other ways in which data about examiner reliability could be used: (1) selecting examiners for further participation in a project, (2) assessing the effectiveness of alternative means of improving consistency, and (3) determining necessary sample sizes and other research procedures.

Although reliability data could be used for selecting examiners, we found only one study that used them for this purpose (Handjelman et al., 1981). One reason reliability data have not been used for examiner selection is that researchers usually attempt to maintain the same examiner/child pairings over time. Authoritative sources advise researchers that "Each subject should be examined on each follow-up survey by the same examiner who made the initial diagnoses at the time of the baseline survey so that additional interexaminer variances are not introduced into the data" (Horowitz, 1972, p. 97). This advice apparently rests on the unproved assumption that greater reliability will be achieved by maintaining the pairings than by retaining the examiners who are the most reliable.

A few studies have used examiner reliability data as a dependent variable to assess the effectiveness of various types of examination criteria (e.g., Davies, Horowitz, and Wada, 1973) or examiner training programs. For instance, Davies and Cadell (1963) have shown that differences in mean scores between examiners can be reduced through the use of standardized criteria and calibration sessions. Information about the relationship between examiner reliability and the characteristics of the subjects examined (such as their age and race) could be used in calculating necessary sample sizes. It could also be used in determining the number of independent exams that would be needed per child to obtain sufficiently reliable scores for measuring treatment effects (Kamen and Schmee, 1974).

PURPOSE

We measured examiner reliability in the National Preventive Dentistry Demonstration Program primarily because we wanted to determine: (1) whether we had to control for possible systematic differences among examiners in measuring treatment effects, and (2) whether examiner errors may have masked these effects. We were also interested in assessing whether the Program's examiners made their diagnoses more or less reliably than examiners in other studies, especially given the uniqueness of some of our procedures. This latter purpose led us to conduct a review of the dental literature to find out the typical level of examiner reliability in prior studies and the factors that might affect examiner consistency.

LITERATURE REVIEW PROCEDURES

We began our review with a computer search for studies published after 1965 that involved clinical examiners and/or radiograph readers. We also consulted a bibliography prepared by Eklund (1981), checked references to previous research in recent articles, and reviewed all of the issues of the journal of *Community Dentistry and Oral Epidemiology* published since 1978 (because this journal contains many of the studies that reported on examiner consistency). Given the purposes of the review, we restricted it to studies that contained examiner reliability data on the variables investigated by the Program; i.e., DMFS, DMFT, and gingival scores.

RELIABILITY OF CLINICAL EVALUATIONS OF DMF

The studies we found that reported on the reliability of clinical (visual plus tactile) evaluations of tooth decay usually could be differentiated in terms of whether they investigated: (1) systematic biases among examiners by comparing the similarity of their mean scores

at one point in time, (2) the prevalence of chance errors by observing the frequency of logical inconsistencies between how a tooth (or surface) is classified at one point in time and how it is classified one year later, or (3) both systematic and chance effects by assessing the consistency between concurrent replicate examinations of the same group of subjects.

Several studies reported large and consistent differences in mean DMF scores among examiners even when these examiners had been trained to use the same criteria (e.g., Andlaw and Tucker, 1975; Beiswanger et al., 1981; Rytomaa et al., 1979; and Weisenstein and Zacherl, 1972). Other studies found only moderate to small differences (e.g., Lopes et al., 1973; and Radike et al., 1973). However, claims for high examiner consistency that are based solely on the comparability of means (e.g., Videroni et al., 1976) could be very misleading because there are other important ways in which examiners might disagree, such as in how they rank-order subjects.

Although many studies that compared examiner means also could have investigated other dimensions of agreement, most did not do so. Table 1.1 contains summary data on the studies published after 1965 that went beyond a comparison of means to provide a more thorough analysis of intraexaminer reliability in making clinical DMFS or DMFT evaluations. All the studies listed in this table involved replicate exams that were conducted under the same conditions; e.g., we excluded a study by Horowitz and Peterson (1966) in which only one of the two examiners used radiographs. Table 1.2 contains the corresponding data for the interexaminer studies that were found. A few articles are cited more than once because they contain separate studies.

Tables 1.1 and 1.2 show that in addition to a comparison of means, the most common indexes of reliability are reversal rate for longitudinal data, and consistency ratio, correlation, and distribution of difference scores for concurrent data. The findings with each type of index are discussed below.

Reversals

Reversals are usually defined as teeth (or surfaces) that are classified as carious (decayed, missing because of decay, or filled) on one exam and as noncarious on a subsequent exam (De Paola and Alman, 1972; Radike, 1972). The frequency with which reversals occur has been used to indicate examiner reliability because the decay process is assumed to be irreversible.

There are several problems with using reversals as the sole index of examiner agreement: (1) Reversals do not indicate all disagreements, because a surface classified as sound on the first exam and as carious on another exam may have been carious on both; (2) the probability of erroneously calling a surface sound is not necessarily the same as erroneously calling it carious (Poole et al., 1973); (3) reversal rates can be artificially reduced by shifting examination standards (De Paola and Alman, 1972) and/or by increasing the time between the first and follow-up examination (because there is a greater chance for surfaces incorrectly called carious on the first exam to become actually carious by the subsequent exam); (4) reversals focus on individual teeth or surfaces rather than the total score used in analyses of treatment effects (Slakter et al., 1976); and (5) some reversals may occur as a result of reversing the decay process (remineralization) rather than examiner error (Torell, 1967).

Another problem with reversal indexes is the way they are often reported in the literature. Ship et al. (1966) found there were almost as many "reversals" (decayed surfaces becoming sound) as there were "errors" (filled or missing surfaces becoming sound). Most articles do

Table 1.1

STUDIES MEASURING INTRAEXAMINER CONSISTENCY IN CONDUCTING CLINICAL EXAMINATIONS

Author(s)	DMF Index	Number of Examiners	Number of Subjects	Age at Baseline	Index of Reliability	Reported Reliability
Fischman et al. (1976)	DMFS*	2	1219	6th grade	Reversals per child	.37 average after 1 year
Horowitz & Peterson (1966)	DMFS*	3	5396	6 to 10	Reversals per child	.03 to .05
Ship et al. (1966)	DMFT+*	1	598	6 to 14	Reversals per child	.06 per year
Thomas & Jamison (1966)	DMFS+*	1	464	7 to 16	Reversals per child	1.3 after 1 or 2 years
Torell (1967)	DMFS*	2	About 1600	12	Cavities reversed	24% reversed (.19/child)
Fogels et al. (1979)	DMFS+	2	4019	6 to 11	Concurrent reversals	.19 to .24 per child
Andlaw & Tucker (1975)	DMFS	2	121	11 to 12	Consistency ratio	82 to 84
Andlaw et al. (1982)	DMFS	1 2	26 36	11 14	Consistency ratio	75 83 to 87
Jackson (1966)	DMFS & DMFT	1	70	12	Consistency ratio	62 for DMFS 66 for DMFT
Shaw & Murray (1975)	DMFS	2	50	11 to 12	Consistency ratio	Mean of 77
Fischman et al. (1976)	DMFS	2	283	6th grade	Number of inconsistent calls/child	.34 average per examiner per year
Rugg-Gunn et al. (1976)	DMFS+ DMFS+ DMFS	1 1 1	Study A=45 Study B=53 Study C=99	12	Correlation coefficient	A (.92 - .97) B (.94 - .98) C (.99)
Howat & Cannell (1979)	DMFS	2	Yr 1 = 200 Yr 3 = 178	11 to 12	Correlation coefficient	Pearson .97 to .99
Rugg-Gunn & Holloway (1974)	DMFS & DMFT	1	44-45	12	Correlation coefficients	Pearson=.95; intraclass=.97
Miller et al. (1981)	DMFS	10	About 5% of 37,956	6 to 18	Intraclass correlation	>.95 for all 10 examiners
Palin et al. (1982)	DMFS	1	35	9 to 10	Kappa	.89

NOTE: * = Longitudinal data, + = includes bitewing radiograph results.

Table 1.2

STUDIES MEASURING INTEREXAMINER CONSISTENCY IN CONDUCTING CLINICAL EXAMINATIONS

Study	DMF Index	Number of Examiners	Number of Subjects	Age when Examined	Index of Reliability	Reported Reliability
Shaw & Murray (1975)	DMFS	2	80	11 to 13	Consistency ratio	73
Andlaw et al. (1982)	DMFS	2	28	14	Consistency ratio	82
Miller et al. (1981)	DMFS	11	50	6 to 18	Intraclass correlation	>.81 for 7 of 10 examiners
Miller et al. (1981)	DMFS	11	8 class-rooms	Not provided	Intraclass correlation	>.65 for 8 of 10 examiners
Kelly & Scanlon (1971)	DMFT	5	393	6 to 11	Distribution of DMFT score differences	Agreed on 74% of subjects examined
Kelly & Harvey (1974)	DMFT	7	407	12 to 17	Distribution of DMFT score differences	Agreed on 48% of subjects examined
Burt et al. (1982)	DMFT	10	360	1 to 14	Distribution of DMFT score differences	Agreed on 44% of subjects examined

NOTE: In the studies using DMFS, all examiners saw all children. In the studies using DMFT, results are based on the pairing of a senior reference examiner with each regular or part-time examiner.

not state whether reversals include what Ship et al. defined as errors even though such errors affect DMF scores.

Reversal rates are usually calculated by dividing the total number of reversals by the number of children examined, including those who could not have a reversal because they had a DMF score of 0 at baseline. Thus, the amount of decay reported at baseline could have a tremendous effect on the reported rate; the more decay, the greater the opportunity for reversals. For instance, if there are 100 reversals in a group of 200 children and if 100 of these children have a DMFS score of 0 at baseline, then the reversal rate would be 0.50 if all children are counted, but 1.00 if the sample is limited to those who could exhibit a reversal.

Our concerns about reversal rates were not allayed when we observed the large variation in these rates (see Table 1.1). Horowitz et al. (1966) suggested that the unusually low rate in their study (.04) may have been attributable to their strict application of National Institute of Dental Research diagnostic standards for classifying a surface as carious. They also held daily calibration sessions for their examiners. The largest recent study of reversal rates involved one-year follow-up exams on children who were initially seen when they were sixth graders (Fischman et al., 1976). In that study, the average number of reversals per subject was .44 for one examiner and .30 for the other.

Although several authors have used reversal rates to assess intraexaminer consistency over time, our review did not uncover a single report in which this (or any other index of reliability) was used to measure interexaminer consistency from one year to the next. This void is

probably attributable to the fact that longitudinal studies generally use only one examiner, or maintain the same examiner/child pairings across years. Thus, prior to the research presented in this report, there may not have been an opportunity to assess whether using only one examiner or maintaining pairings has any effect on the reliability of longitudinal data.

Consistency and Reproducibility Ratios

Horowitz et al. (1973) recommend reporting concurrent examiner agreement in terms of a reproducibility ratio (RR). This ratio is computed by dividing the number of sites (teeth or surfaces) on which there is a sound versus carious disagreement by the number that are consistently classified as carious. De Paola and Alman (1972) suggest using a consistency ratio (CR) that is computed by dividing the number of sites consistently classified as carious by the number classified as carious on one or both exams, and then multiplying the resulting proportion by 100. In order to facilitate comparisons among studies that reported either RRs or CRs, we converted RR results to CRs as follows: $CR = [100]\{1 - [(RR)/(RR + 1)]\}$.

RRs and CRs only use data from subjects who had at least one surface classified as carious on at least one exam. They therefore permit more appropriate comparisons of reliability data across studies than reversals because they partially adjust for differences in the prevalence of disease among these studies. However, they can only be used with concurrent data.

The several studies reviewed by Shaw and Murray (1975), as well as their own research, yielded intraexaminer CRs ranging from 62 to 86, i.e., agreement on about three out of every four carious surfaces. Interexaminer CRs ranged from 62 to 82.

Correlation Coefficient

The correlation coefficient indicates the degree to which the subjects who receive high scores on one exam also tend to receive high scores on another exam. A 0.00 coefficient signifies that there is no relationship between a subject's score on one exam and that subject's score on another exam, whereas a 1.00 coefficient indicates there is complete agreement between the two exams in the relative standings of the subjects.

The intraclass and Pearson correlation coefficients are often used to indicate examiner agreement. The size of the intraclass coefficient depends on whether its computation takes into consideration differences in mean scores among examiners. The computational procedure also determines whether the coefficient reflects the reliability of a single examination or the reliability of the score derived from several examinations of the same child (Winer, 1962; Fleiss et al., 1979). The Pearson coefficient is a special case of the intraclass correlation. It shows the correlation between two examinations of the same children after controlling for possible differences in mean scores between these two sets of exams.

One recent study that computed intraclass correlations was the 1979–1980 National Dental Caries Prevalence Study (Miller et al., 1981). In this study, all 11 examiners (including 8 regular, 2 part-time, and 1 reference examiner) evaluated the same set of 50 children, i.e., there were 550 replicate exams. These exams were conducted immediately after the calibration session. The authors report that (p. 144):

> Intraclass correlation coefficients were used to measure the consistency of diagnoses of each examiner compared to the standard examiner. All correlations for permanent surfaces were above 0.81 except in the case of three of the ten examiners who showed some inconsistency in diagnosing caries (D). As a result, additional training was given to these examiners. By the end of the training period, all examiners were diagnosing consistently.

We are not told what type of intraclass correlation was used or what is meant by the last sentence, but we assume it implies that all examiners were at .81 and higher. The authors did note that four times during the project, "replicate examinations were performed on two classroom groups by the standard examiner and every other examiner." The resulting correlations were "above .65 for all but two of the ten examiners." The average correlation among examiners in this or in the original group of 50 is not provided even though it would have been more informative because the reference examiner saw only a few of the children in the total survey.

No explanation was given for the drop from .81 that was obtained with the original group, but it may have been due to an increased homogeneity of the examined sample and/or a drift in examination standards. This latter possibility is of particular interest because most of the few studies that report concurrent reliability data collect these data during or immediately after calibration rather than throughout the regular examination process.

Only two studies investigated the relationship between intraexaminer and interexaminer consistency with samples drawn from the same population. In the 1979–1980 caries prevalence study (Miller et al., 1981), intraexaminer correlations on a 5 percent sample of the total survey population remained above .95 throughout the project despite the previously noted drop in interexaminer consistency. Shaw and Murray (1975) found a similar (but not as strong) trend with a sample of 130 children (CRs on DMFS averaged 76 and 73 for intraexaminer and interexaminer comparisons, respectively).

Distribution of Difference Scores

As part of the 1971–1974 National Health Survey (Burt et al., 1982), 360 subjects who ranged in age from 6 to 74 received two examinations. The first exam was conducted by one of the survey's two senior dentists. This dentist then recorded the results for the second exam which was done by one of the survey's eight regular dentists. The senior examiners and the regular examiners disagreed on the score that should be assigned in over 50 percent of the cases. The percentage on which there was a disagreement of 1, 2, 3, or more than 3 DMFT were: 20.3, 16.4, 5.6, and 13.6, respectively. These disagreements occurred even though there was a close similarity in mean scores between the regular and reference examiners (the number of times a reference examiner assigned a higher score than a regular examiner was about as frequent as the opposite pattern).

Other health surveys (e.g., Kelly and Harvey, 1974; and Kelly and Scanlon, 1971) also have described examiner consistency in terms of the distribution of DMFT difference scores between reference and regular examiners. The relatively high agreement rate in the Kelly and Scanlon study (74 percent) was probably attributable to the large portion of the subjects in their survey not having any carious surfaces.

In all three surveys, disagreements occurred despite an extensive training program prior to data collection, and the examiners knew which children were in the replicate exam sample. Further, with the exception of this sample, the reference examiner(s) saw only a small percentage of all the children included in the survey. Thus, there is a serious question about whether the results with the replicate samples are indicative of examiner reliability in the full surveys.

RELIABILITY OF RADIOGRAPH EVALUATIONS

Radiographs of interproximal surfaces are usually obtained to supplement clinical findings. The general rule for combining data from clinical and radiograph exams is that a surface (or tooth) is classified as affected by decay if either or both types of exams have classified it as affected (e.g., Horowitz and Peterson, 1966). Using radiograph data in this way can significantly affect results because there is only a moderate relationship among clinical, radiograph, and clinical plus radiograph evaluations of a tooth (Marthaler and Germann, 1970; and Murray and Shaw, 1975). We were therefore surprised to discover that there has not been a single published report on whether adding radiograph results to clinical findings increases or decreases the reliability of the combined score.

Rugg-Gunn et al. (1976) reported slightly higher reliability coefficients with clinical exams than with the combination of clinical and radiograph exams. However, the coefficients were based on different (and small) samples. Clark et al. (1982) reviewed six studies that reported treatment effects based on clinical and the combination of clinical and radiograph data. They found that the inclusion of radiograph data had no systematic impact on statistical significance levels in tests of treatment effects.

Although many studies collected the data needed to examine the effect on reliability of adding radiographs to clinical findings, they chose not to investigate it. Horowitz et al. (1966) conducted clinical and radiograph examinations, but only provided data about the reliability of the clinical scores. Andlaw and Tucker (1975) investigated the reliability of clinical and radiograph scores, but they did not contrast the reliability of the combined score with the reliability of the separate scores. They did report an interreader CR of 78 with 150 film sets.

De Paola and Alman (1972) collected baseline and one-year follow-up clinical and radiograph data on 512 subjects, but their research design only permitted an assessment of intrareader reliability on the radiograph data for each exam. They found intrareader CRs of 64 to 70 and intraclass correlations of about .82. Intrareader agreement in classifying the degree of cavitation has been as high as 95 percent (Holst, 1976). Kjaerheim et al. (1980) reported intrareader reliability coefficients of .97 and .99, but each sample contained only about 32 children; and, the children in each sample varied markedly in age (7-year span) and thereby in decay level.

Shaw and Murray (1975) reviewed seven studies that contained radiograph reader reliability data. This review, their own research (Murray and Shaw, 1975), and other recent studies (e.g., Craig et al., 1981) indicate that intrareader CRs range from 72 to 89 whereas interreader CRs range from 63 to 94. Some but not all of this variation is due to whether carious versus unreadable disagreements are counted as inconsistencies. The Murray and Shaw findings are especially interesting because they found that interreader consistency on 1417 sets of radiographs was slightly higher than the readers' consistency with themselves.

Correlations between readers have been reported as high as .85 and .86 (Haugejorden, 1977; Handjelman et al., 1981); however, these correlations are biased upward by the comparison of one reader's results with the average score assigned by three or more other readers. In addition, the sample sizes in these studies were quite small.

Statistically significant variations in mean scores among readers have been found in some studies (Howat and Cannell, 1979; Leijon and Marken, 1968) but not in others (Birkeland et al., 1975). Poulsen et al. (1980) reported that most of the disagreements among their 18 readers were due to random error rather than systematic bias. They also noted that the reliability was not especially high (correlations ranged from .39 to .78) and that it increased with each increase in the degree of severity of the lesion that was required before a surface could be

classified as carious. Mileman et al. (1982) discovered large and stable differences among readers in their ability to make accurate diagnoses.

GINGIVITIS

Periodontal disease, rather than tooth decay, is the major reason for tooth loss among adults. Gingivitis (inflammation of the gums) may be one of the first signs of this disease. Although there is no one generally accepted index of gingivitis (Hazen, 1974), many clinical studies and health surveys try to assess it because it can be greatly reduced or eliminated by good oral hygiene.

Despite the potential importance of gingivitis and the frequency with which it is assessed by both practitioners and researchers, our review located only nine articles in the last 27 years that provided data on examiner agreement in using a gingival index (see Table 1.3). This lack of information is due partly to some studies collecting both gingival and DMF data, but only reporting examiner reliability results for the DMF portion of the examination (e.g., Andlaw and Tucker, 1975).

The studies that reported examiner consistency in using gingival indexes often found that reliability improved with experience and training in using an index (Alexander et al., 1971; Smith et al., 1970), that reliability was substantially lower than it was on DMF indexes, and that the examination itself may temporarily change the condition of the gingival tissue (Birkeland and Jorkjend, 1975).

CHARACTERISTICS OF PAST RESEARCH

Articles in the dental literature rarely provide information about examiner reliability. This situation exists despite the quasi-subjective nature of the examination process and empirical evidence that examiners are far from perfectly reliable. Between January 1978 and January 1983, 164 separate studies in the journal *Community Dentistry and Oral Epidemiology* relied on examiner evaluations. However, only 28 of these studies reported reliability data and only 12 of them had 30 or more replicate examinations.

A few studies, such as Poulsen et al. (1980), provide clear and complete accounts of what data were collected and how they were analyzed. More often, the reader encounters descriptions of procedures and results that raise more questions than they answer. A sample of these problems is presented below:

- Bille (1980) stated that "All examinations were performed by the author" (p. 28), and then went on to contrast interexaminer and intraexaminer results.
- Thomas and Jamison (1966) computed reversal rates by dividing the number of reversals that occurred after one year by the number of subjects retained after two years.
- In a study by Gibson et al. (1981), "The levels of repeatability of examiner observations were determined by calculation of the Pearson correlation coefficient. For . . . treatment need, r = .77" (p. 6). It is apparent from the authors' discussion of "treatment need" that scores on this variable do not refer to points along a continuum and therefore should not be used to compute a Pearson coefficient.
- Several studies use terms such as "reproducibility," "agreement," and "repeatability" without describing or providing a reference for how they were computed.

Table 1.3

STUDIES MEASURING EXAMINER CONSISTENCY IN USING GINGIVAL INDEXES

Study	Type of Study	Index[a]	Number of Examiners	Number of Subjects	Sample	Time Between Measurements	Index of Reliability	Reported Reliability
Davies et al. (1974)	Inter	PI	4	710	Age 6-44	Not stated	Comparison of means	Significant differences
Alexander et al. (1971)	Intra and Inter	GI PMA	3	50	Wide range of severity of inflammation	2 to 5 days	Comparison of means	Significant differences before training
Cutress et al. (1978)	Inter	PSI PI	5	2,109	Over 14 years old	No replicate exams	Comparison of means	Variation in means on PSI
Birkeland & Jorkjend (1975)	Intra	GI	1	31	Age 13	2 hours	Comparison of means	Mean higher on 2d exam
Axelsson & Lindhe (1974)	Intra	GI	1	18	Age 7-8	Within hours	Consistency ratio	94
Davies et al. (1967)	Intra	PI	13	13	Probably dentists	1 day	Consistency ratio	Average of 48
Smith et al. (1970)	Intra	DHC Ramfjord	1 per index	50	Male navy recruits	Within same half-day	Percent agreement	DHC=82% Ramfjord=84%
Davies et al. (1974)	Intra	PI	4	162	Age 6-44	At least 1/2 hour	Percent agreement	90%
Hamp et al. (1978)	Not clear	GI	2	13 to 27 per year	One classroom	Within 2 days	Percent agreement	Mean of 81%/yr over 4 years
Elliott & Clemmer (1972)	Intra and Inter	GI	2	22	Adults	1 week	Percent agreement	Intra = 78% Inter = 64%
Markkanen et al. (1979)	Intra	PTNS	1	23	Paper mill employees	2 weeks	Weighted kappa	.80 (84% agreement)
Shaw & Murray (1977)	Inter	GI	2	117	Age 11-13	4 to 6 hours	Correlation coefficient	.79, similar means
Shaw & Murray (1977)	Intra	GI	2	42 & 44	Age 11-13	4 to 6 hours	Correlation coefficient	.81-.84, similar means
Burt et al. (1978)	Intra	GI	1	89	Grades 3-4	2 to 7 days	Spearman rank order	.76, similar means

[a] A complete description of each index appears in Hazen (1974).

- Studies often contain only one- or two-sentence descriptions of the methods used to collect reliability data, such as "Examiner error was assessed by duplicate examinations of a randomly selected 10% of the sample" (Ritchie, 1979, p. 223), and then do not present the results with the replicate exam data.

Because of problems of the types illustrated above and the frequent lack of attention to reliability in the dental literature, there is almost no generally useful information about whether some examiners are significantly more reliable than others, whether systematic differences in mean scores among examiners are stable over time, whether examiners become more reliable as they become more experienced, whether maintaining examiner/child pairings across examination cycles increases the chances of detecting treatment effects, and whether reliability is related to the characteristics of the children examined.

When reliability data are collected, they are often gathered in ways that undermine their utility. For example, a slight modification in the design used by De Paola and Alman (1972) would have eliminated the confounding of reader and subject effects in their study as well as provided data about interreader consistency. Another design problem with many studies is that the replicate exams used to collect reliability data are conducted during calibration, immediately following calibration, or as part of a special study rather than throughout the regular examination process. And, with few exceptions, examiners know which subjects are seen for the purposes of measuring examiner consistency. Such studies may not reflect the level of reliability that would be obtained under typical conditions. This is an especially serious concern in some large health surveys.

Because several statistical techniques may have been used to describe the degree of examiner agreement it is often impossible to compare the results of one study with those of another. Even when the same general technique is used, comparisons are difficult because:

- Sample sizes are usually quite small. For example, Hetland et al. (1982) used only four subjects in their analysis of examiner consistency.
- Salient differences in sample characteristics affect the size of the index. For instance, a homogeneous sample (as indicated by a small standard deviation of DMFS scores) will tend to lower correlation coefficients more than a heterogeneous sample, and a sample with a low prevalence of decay will have fewer reversals than one with many surfaces that could be reversed.
- Statistical assumptions underlying a technique may be violated by the characteristics of the reliability sample. For instance, correlational techniques assume that the joint distribution of scores being correlated does not differ markedly from a bivariate normal distribution, but the distribution of DMFS scores in many studies, and especially those with children, do not satisfy this requirement because the score distributions are highly skewed (Bell et al., 1982).
- Important details about the particular method used to compute an index are often missing. For instance, articles that use reversal rates rarely report whether these rates are calculated by dividing all the reversals that occur by all the subjects examined (or by just those who had an initial score that could show a reversal) and whether reversals are calculated using the D; D and F; or D, M, and F components. These seemingly subtle distinctions have a profound effect on reported results.

The failure to report information that is essential to the interpretation of results is not limited to studies that rely on reversals. When intraclass correlations are used to describe

reliability, we are usually not advised whether the particular coefficient has been adjusted for mean differences between examiners or whether it represents the reliability of the typical examiner or the reliability of the average score assigned by all the examiners who saw the same set of children. Some studies only give information about the degree of relationship between the reference examiner and the regular examiners even when all subjects in the reliability sample are seen by all examiners. A more informative statistic would be the average (typical) correlation between any two regular examiners because the regular examiners conducted the bulk of the exams.

An even more serious problem is that the type of reliability data presented is not really relevant to how the data will be used. Although reversals and especially CRs have value for comparing studies, they certainly do not help us determine the effects of examiner inconsistencies on detecting treatment effects. And, although radiographs are often used to supplement clinical findings, none of the studies that used radiographs reported whether they tended to increase or decrease the reliability of the total scores obtained. Distributions of difference scores are cumbersome to present, but because they describe reliability in terms of how much the DMF scores differ they are more readily interpretable.

RESEARCH QUESTIONS

The preceding review of examiner reliability studies was undertaken to provide a basis for determining whether the examiners in the National Preventive Dentistry Demonstration Program were more or less reliable than examiners in other studies. The review disclosed that there are relatively few studies on which to base such a judgment and that most of those were limited to sixth graders (12-year-olds). Among the studies that do exist, there is no generally accepted index of examiner consistency (let alone standard to be achieved). Comparisons of the Program's examiners with those used in other studies are further hampered by significant differences in sample characteristics, examination criteria, and research procedures.

The review also revealed several important questions about examiner reliability that have not been answered or answered adequately by previous research in this area. The analyses presented in this report were designed to provide information regarding many of these issues as well as to determine the reliability of the Program's examiners.

This research addressed the following questions:

- To what degree do inconsistencies affect the likelihood of detecting actual differences among treatment groups?
- What are the most common types of examiner inconsistencies and how often do they occur?
- What proportion of the inconsistencies are attributable to systematic biases (such as one examiner being more likely than another to classify a surface as carious) versus chance or random effects (such as examiner fatigue or recorder error)? And, how much is a subject's score affected by systematic effects?
- Are examiners sufficiently more consistent with themselves than they are with each other to justify the additional costs of maintaining the same examiner/child pairings over time?
- Are some trained examiners significantly more reliable than others in how they apply diagnostic protocols?
- Does reliability improve as examiners become more experienced in conducting screening examinations?

- Is reliability related to a child's age, sex, race, decay level, or treatment group?
- How reliable are evaluations of gingivitis and other non-DMF indexes of oral health?
- How reliable are radiograph readers?
- Does adding radiograph data to a child's clinical exam results increase or decrease reliability?

Chapter 2 describes the databases that were used to answer these questions.

Chapter 2

PROCEDURES

PROGRAM DESCRIPTION

The examiner reliability studies presented in this report were conducted as part of the National Preventive Dentistry Demonstration Program which investigated the costs and effects of various types and combinations of school-based preventive dental care procedures. The Program operated at nine sites for four years and at one site (New York) for three years. Five sites were considered nonfluoridated (less than 0.2 parts of fluoride ion per million) and five had optimally fluoridated water (see Table 2.1). Enrollment in the Program was open to all the children who in the fall of 1977 were in grades 1, 2, and 5 at their site's participating schools.

Six regimens of preventive care were provided at each site. Four regimens included classroom procedures, such as lessons on oral hygiene and twice weekly plaque control exercises (brushing and flossing). Three of these regimens also involved weekly fluoride mouthrinse (and at nonfluoridated sites, daily fluoride tablets). Two of the last three regimens further involved applying an epoxy-like resin (called "sealants") to the chewing surfaces of the teeth that are most susceptible to decay and/or providing biannual fluoride paste prophylaxis and gel treatments. The fifth regimen involved just the sealant and prophy/gel procedures.

The children in Regimen 6 did not receive any of the foregoing procedures. However, these children, as well as those in the other five regimens, did receive a dental examination when the Program began at their site (called the baseline exam) and at one-year intervals thereafter for a total of five examinations (provided, of course, they did not leave the site or otherwise become unavailable).

Within a site, schools, rather than individual children, were assigned to regimens because certain types of care were more efficiently delivered when children were in classroom groups. A team of dental hygienists and assistants went from school to school to provide the two types

Table 2.1

SITES PARTICIPATING IN THE NATIONAL PREVENTIVE DENTISTRY DEMONSTRATION PROGRAM

Region	Nonfluoridated Sites	Fluoridated Sites
Northeast	Billerica, Massachusetts	New York, New York
Southeast	Tallahassee, Florida	Chattanooga, Tennessee
Central	Wichita, Kansas	Minneapolis, Minnesota
Southwest	Monroe, Louisiana	El Paso, Texas
Northwest	Pierce County, Washington	Hayward, California

of clinical care (sealants and prophy/gel); classroom teachers were usually responsible for the other procedures.

CLINICAL EXAMINATIONS

One part of each yearly examination consisted of a clinical (visual plus tactile) assessment of the number of decayed, missing due to decay, and filled surfaces on a child's permanent teeth (i.e., the child's DMFS score). The second part involved evaluating the degree of gingivitis (inflammation of the gums). The last part was devoted to determining whether there were any sealants present on the child's teeth, whether there was any discoloration of the teeth, and the urgency with which the child needed to see a dentist for possible restorative or corrective care. The examination process also provided an opportunity to observe and record each child's sex and racial/ethnic group.

The Program's site staff, parents, and other community volunteers served as recorders. They received about one hour of training and practice before being assigned to an examiner. In Year 1, a new set of recorders was trained each day. In subsequent years, a recorder generally stayed with the examination team for the entire week the team was at a site. Recorders marked onto machine scorable forms the code numbers and letters the examiners called out to indicate their evaluation of each tooth surface and the other characteristics noted above (see Appendix A).

The examinations began in the fall of 1977 and continued on a schedule of roughly one site every other week (excluding school vacation periods). In Years 1 and 5, an examiner stayed behind for a week to examine the children who were absent. About 9 to 10 members of the Program's cadre of 31 specially trained dentists conducted the baseline examinations at a site. Most of the examiners participated at three or four sites.

A large sample of children (N = 4,320) who were not enrolled in the Program were also examined at baseline. These cross-sectional children were in grades 3, 4, 6, 7, and 8 at the same schools as those attended by the children in the longitudinal cohorts. Another large sample (N = 4,746) of cross-sectional children (from grades 1–9) was examined in conjunction with the Year 5 examinations. None of the children in the two cross-sectional samples were used in the analyses of examiner reliability.

More examiners were used at baseline than in subsequent years because there were far fewer children to examine on the follow-up examinations (see Table 2.2). This decline in sample size was due to absences, attrition in the longitudinal cohorts, and not examining cross-sectional children each year.

There were 13,036 children with DMFS scores in Year 1 who also had Year 3 scores (not counting children whose teeth were banded or otherwise unexaminable), 8,696 who had both a Year 3 and 5 exam, and 9,566 who had a Year 1 and 5 exam. Cohort 5 represented about one-third of each group.

The substantially reduced sample sizes after baseline permitted completing each site's subsequent yearly exams within one week despite the reduced number of examiners at that site. However, the reduction in the number of examiners after baseline and other logistical factors frequently made it impossible to maintain the same examiner/child pairings across years. On the other hand, the reduced number of examiners per site facilitated the process of calibrating them with respect to the Program's diagnostic criteria (see Appendix B). The decision to retain an examiner was based on availability, degree of agreement with other examiners in DMFS calls, and examination speed.

Table 2.2

NUMBER OF CHILDREN AND EXAMINERS PARTICIPATING IN THE PROGRAM

Year	Number of Children Examined	Grade Levels	Total Number of Examiners	Average Number of Examiners Per Site	Average Number of Sites Per Examiner
1	24,811[a]	1 - 8	31	9.6	3.1
2	15,363	2-3,6	18	6.7	3.1
3	13,727	3-4,7	15	5.0	3.3
4	12,377	4-5,8	13	4.5	3.5
5	15,104[a]	1 - 9	16	6.6	3.7

[a]Years 1 and 5 include children from cross-sectional samples.

EXAMINER TRAINING AND CALIBRATION

An orientation session was conducted in May 1977 for 17 of the Program's cadre of 31 examiners. This session was provided by experienced dental staff from the National Institute of Dental Research (NIDR). It included a detailed slide lecture and clinical practice in applying the DMF criteria. A periodontist with extensive examining experience provided a similar program for the gingival component of the exam. The 14 examiners who did not attend the initial sessions were primarily west-coast based. Separate training and calibration sessions were held for them before they conducted examinations.

The orientation session was followed by a one-day calibration session in which three, four, or five examiners independently evaluated a child, compared their results, and then discussed and resolved differences under the guidance of the instructors. This process was repeated until all the examiners appeared to be applying the criteria in the same way.

The slide lecture was repeated at each site on the day before the regular examinations began at that site. In addition, a calibration session involving cross-examinations was conducted each day prior to beginning the regular examination process.

The reliability data presented in this report were obtained during the regular examination process. They were *not* obtained from training or calibration sessions, and they do not include data from children who participated in these special activities.

CLINICAL EXAMINATION EQUIPMENT, STAFF, AND CONDITIONS

The primary clinical examination equipment consisted of portable dental chairs, Rolux fiber optic lights, #5 mirrors, #23 explorers, and support equipment. The support equipment included a portable unit that was especially designed for the Program. This unit provided compressed air, water, and vacuum to as many as four examiners at one time.

A site's exam team included clinic staff to sterilize instruments, runners to bring children from their classrooms to the examining room, desk clerks to hand out examination forms and record sex and racial group data, monitors to supervise children while they brushed and waited

to be examined, and research staff to assign children to examiners so that examiners did not have to wait for children and so that the children designated to receive replicate exams got them from the proper examiners.

The examinations were usually conducted in a school's gymnasium, although libraries, cafeterias, classrooms, and other locations were also used. Children generally waited outside the examination room until it was close to their turn to be seen. When they came into the room, they were able to watch one or two other children being examined just before it was their turn to be seen. With few exceptions (and most of these were at baseline), examination rooms were relatively quiet and well lighted. An examiner usually saw 50 to 60 children per day.

A more complete description of the examination procedures appears in other reports (Klein and Bohannan, 1979; and Bohannan et al., 1984), and a film showing these and other procedures has been produced by Gittleman (1979).

REPLICATE CLINICAL EXAMINATIONS

Each year, approximately 10 percent of the children in each of the three longitudinal cohorts received a second clinical examination. The second exam was almost always conducted within a few hours of the first exam. About one-half of the reexamined children saw the same dentist twice in order to provide data about intraexaminer consistency. The remainder of each year's sample of reexamined children had one dentist for their first exam and a different dentist for their second exam. Examiners were not advised as to whether a child was in a replicate sample and the children in the replicate samples were intermixed with all the other children who were examined.

Intraexaminer checks were discontinued after Year 3 when it became evident examiners agreed highly with themselves. The intraexaminer checks were also especially disruptive to normal school activities and required special staff supervision because a child had to come to the examining room twice.

In Year 1, identification numbers were assigned randomly to children. All the children whose number ended in 5 were selected to receive either an interexaminer or an intraexaminer replicate exam. In subsequent years, replicate samples were selected randomly (by a computer program) from the population of children who had not received a replicate exam in the previous year. A small percentage of the children selected to receive a replicate exam did not receive one because they were absent on the day the examination team came to their school, the examiner or child left the school before the second exam could be completed, etc. In some instances, these children were replaced with children who were selected randomly from among those who were available.

Two to three teams of examiners were formed at a site. Usually one team conducted the exams at one school while another team visited a different school or set of schools. Team composition was changed daily. As a result of this procedure, a site's interexaminer replicate exam data consisted of each examiner being paired about the same number of times with every other examiner at that site. Additional procedures ensured that each examiner's replicate exams were balanced with respect to the sex and cohort of the children examined.

Table 2.3 shows the number of children receiving each type of reexamination each year. In this and all subsequent tables, children in Cohorts 1 and 2 have been combined into one group labeled Cohort 1+2 because of the similarity of their data. Tables 2.4 and 2.5 contain mean DMFS scores and standard deviations, respectively, in the Program and on the first of the two examinations each year in each reexamination sample. The columns labeled "All" in these tables refer to all the children in the Program. Analyses of the data in Table 2.4 revealed that a reliability sample's average DMFS score was comparable to the average score of other Program children at the same age level. A series of two sample t-tests indicated that only once did the mean DMFS score of the reexamined children differ significantly from the mean score of all the other children examined. This occurred in Cohort 1+2 in Year 1, but the difference was only 0.2 surfaces.

Table 2.3

NUMBER OF CHILDREN RECEIVING EACH TYPE OF CLINICAL REEXAMINATION EACH YEAR, BY COHORT

	Interexaminer			Intraexaminer		
Year	1+2	5	Total	1+2	5	Total
1	822	424	1246	695	356	1051
2	739	358	1097	718	333	1051
3	667	303	970	660	296	956
4	945	399	1344	(a)	(a)	(a)
5	787	292	1079	(a)	(a)	(a)

[a]Intraexaminer data were not collected in Years 4 and 5.

Table 2.6 presents the number of children each examiner saw in each year's interexaminer and intraexaminer samples. The numbers in this table are roughly proportional to the total number of children seen by each examiner in the total population. The sum of an interexaminer column contains twice the number of children seen because each interexaminer child had two examiners. The missing entries indicate examiners who did not participate in the Program after the first, second, or third year.

Table 2.4

MEAN CLINICAL DMFS SCORE IN THE PROGRAM AND IN EACH REEXAMINATION SAMPLE, BY YEAR

	Cohort 1+2			Cohort 5		
Year	All	Inter	Intra	All	Inter	Intra
1	1.01	1.14	1.20	4.11	4.31	4.24
2	(a)	1.57	1.46	(a)	4.67	4.47
3	2.05	1.98	2.13	5.81	5.38	5.42
4	(a)	2.70	(b)	(a)	7.17	(b)
5	3.09	3.05	(b)	7.77	8.00	(b)

[a]Years 2 and 4 data were not analyzed for the total sample.

[b]Intraexaminer data were not collected in Years 4 and 5.

Table 2.5

STANDARD DEVIATION OF CLINICAL DMFS SCORES IN THE PROGRAM AND IN EACH REEXAMINATION SAMPLE, BY YEAR

	Cohort 1+2			Cohort 5		
Year	All	Inter	Intra	All	Inter	Intra
1	1.91	1.94	2.23	4.33	4.56	4.13
2	(a)	2.34	2.32	(a)	4.89	4.47
3	2.86	2.78	2.81	6.22	5.73	5.17
4	(a)	3.28	(b)	(a)	7.67	(b)
5	3.92	3.92	(b)	7.94	7.29	(b)

[a]Years 2 and 4 data were not analyzed for the total sample.

[b]Intraexaminer data were not collected in Years 4 and 5.

Table 2.6

NUMBER OF CHILDREN EACH EXAMINER SAW IN EACH YEAR'S INTEREXAMINER AND INTRAEXAMINER[a] SAMPLES

Examiner Number	Year 1		Year 2		Year 3		Year 4	Year 5
	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Inter
1	209	97	61	34	284	158	501	118
2	119	40	139	70	153	74	272	288
3	85	40	26	12				34
4	12	4						
5	141	61	137	70	38	19		85
6	25	12						
7	164	66	82	48	116	53	279	141
8	12	4						
9	67	28						
10	50	21						
11	84	25	171	80	127	67	139	239
12	102	45	123	60	173	77	191	107
13	93	43						
14	97	40	180	95	176	90	209	217
15	48	22						
16	178	57	296	120	324	162	59	23
17	81	36	86	43	34	14	321	156
18	43	25	47	25	30	17	211	77
19	94	29	98	44				
20	103	57	356	162	155	73	132	194
21	77	34	78	38				
22	71	37	59	33	99	44	133	127
23	74	30	101	45	122	60		175
24	73	37						
25	76	30	57	29	77	33	180	90
26	70	40						
27	73	31						
28	13	9						
29	33	6						
30	73	29	97	43	32	15	61	87
31	52	16						

[a]Intraexaminer data were not collected in Years 4 and 5.

RADIOGRAPH EXAMINATIONS

Trained technicians took bitewing radiographs (X-rays) of the children's posterior teeth (i.e., the ones most susceptible to decay) at baseline and again at the end of the Program. A specially designed truck containing two X-ray machines was used to ensure adherence to strict safety requirements and to maintain standardization of procedures and equipment across sites. Exposed films were developed according to standard processing protocol.

Each child's films were evaluated by a dentist for the presence of interproximal decay and fillings on permanent, posterior teeth. The radiograph readers did not have access to the clinical examination results just as the clinical examiners did not have access to radiograph results.

The readers were trained and calibrated in the use of the Program's written protocol for film evaluations (see Appendix C). Using a standard view box and magnifying lens, they conducted the readings in a quiet room with reduced lighting. About 30 to 45 film sets were viewed per hour with at least a 10-minute break every 45 minutes.

Each reader dictated his/her evaluations to a recorder who entered the results on a machine scorable form containing the child's name and identification number (see Appendix A). Both the recorder and the reader checked that the name and ID number on the envelope containing a child's films matched those on the child's form.

A sample of about 10 percent of the longitudinal cohort children with usable films were selected to have their films reread. About 40 percent of these children had their films reread by the same dentist who had evaluated their films previously. The remaining children had their films reassessed by a different dentist. The film sets that received replicate readings were cycled back into the regular reading process in a way that prevented a reader or recorder from knowing which sets were in the reread sample.

Chapter 3

RELIABILITY OF CLINICAL DMFS EXAMINATIONS

INTRODUCTION

This chapter presents data on the clinical examiners' reliability in assigning DMFS scores (i.e., classifying surfaces as decayed, missing due to decay, and filled). The analyses investigate interexaminer and intraexaminer consistency both within and across years. The within year (concurrent) analyses are based on the children in each year's 10 percent random sample who received a second examination shortly after their first one. The between year (longitudinal) analyses are based on the children who had both a Year 1 (baseline) and a Year 3 examination. Table 3.1 illustrates the four kinds of reliability discussed. The Year 3 examinations at a site were usually conducted about 25 months after the Year 1 examinations.

The degree to which systematic (rather than chance) differences among examiners are present in the data is indicated by the difference in interexaminer and intraexaminer consistency indexes. If examiners agreed with themselves but not with each other, then the observed differences could be due to examiners remembering previous calls on a subject and/or one examiner tending to call borderline conditions as carious and another examiner tending to call them sound. If the two indexes are quite similar, then it can be inferred that inconsistencies in DMFS calls were due mainly to chance rather than to systematic differences among examiners.

Even if concurrent data uncover substantial systematic differences among examiners, one should not assume that the differences will persist over a period of several years. An examiner who consistently calls high one week or one year may drift back into line after repeated calibration sessions. If that were the case, the number of intraexaminer inconsistencies would approach the number of interexaminer inconsistencies. Thus, longitudinal data provide a means for studying whether systematic differences among examiners persist, and thereby whether maintaining examiner pairings over time improves reliability.

Table 3.1

DATA SOURCES FOR EXAMINER RELIABILITY ANALYSES

Focus	Concurrent	Longitudinal
Interexaminer consistency	Child is seen by two examiners on the same day.	Child is seen by one examiner in one year and by another examiner in a subsequent year.
Intraexaminer consistency	Child is seen by the same examiner twice on the same day.	Child is seen by one examiner in one year and by the same examiner in a subsequent year.

We first discuss the most common types of interexaminer and intraexaminer inconsistencies that occurred and then contrast interexaminer with intraexaminer consistency rates in making concurrent DMFS calls. We describe the results of the longitudinal analyses of examiner consistency, especially with respect to the impact of maintaining or not maintaining the same examiner/child pairings across years. Finally, we address the issue of whether examiner inconsistencies may have interfered with the detection of treatment effects.

NUMBER AND TYPES OF INCONSISTENCIES

On any given examination, a tooth (or more precisely, a permanent tooth location) could be classified in one of eight ways, namely: sound, deciduous, unerupted, missing due to decay, carious, filled, both carious and filled, or excluded from analysis (e.g., a tooth that had been extracted because of trauma). There were therefore 28 ways in which two examinations of the same tooth might differ, such as its being classified as sound on one examination and as carious on another. Table 3.2, which combines data across years, shows that disagreements occurred on about 2 percent of all the teeth examined.

Table 3.2

NUMBER AND PERCENTAGE OF DISAGREEMENTS RELATIVE TO NUMBER AND PERCENTAGE OF CHILDREN EXAMINED

	Cohort 1+2		Cohort 5	
Item	Inter	Intra	Inter	Intra
Number of disagreements	2038	602	1507	454
Number of children examined	3960	2073	1776	985
Number of teeth examined	110880	58044	49728	27580
Percent of teeth with disagreements	1.8%	1.0%	3.0%	1.6%
Mean disagreements per child	.51	.29	.85	.46

Table 3.3 shows what percentage of the total number of disagreements on tooth classifications was due to each type of disagreement that could occur. The data for all five exams were combined in this table because there was essentially no difference in pattern across years.

The data in Table 3.3 show that (1) except for excluded teeth, the pattern of interexaminer disagreements was essentially identical to the pattern of intraexaminer disagreements, (2)

the most common disagreement between concurrent examinations was a tooth being classified as sound on one examination and as carious on the other examination, and (3) there were more apt to be disagreements involving deciduous and unerupted teeth with younger than with older children. The data in Table 3.3 also show that examiner inconsistencies could not have seriously biased the Program's analyses of tooth eruption patterns and rates (Bell et al., 1982).

Table 3.3
PERCENTAGE OF DISAGREEMENTS WITHIN EACH TYPE OF CONCURRENT EXAMINATION, BY TYPE OF DISAGREEMENT AND COHORT

	Cohort 1+2		Cohort 5	
Type of Disagreement	Inter	Intra	Inter	Intra
Sound - Carious	42	46	47	45
Sound - Deciduous	16	20	8	15
Sound - Unerupted	13	11	7	15
Filled - Carious & Filled	8	5	12	10
Deciduous - Unerupted	7	10	1	3
Sound - Excluded	5	1	10	2
Sound - Filled	3	2	3	5
All Other	7	5	11	6

About 45 percent of all the concurrent disagreements and over 90 percent of those that could affect DMFS scores were attributable to a tooth being classified as carious on one examination and as something other than carious, missing, or filled on another examination. Most of the few remaining discrepancies were due to the Filled component. Disagreements regarding whether a tooth was missing due to decay (versus unerupted or extracted for other reasons) were extremely rare.

The number of disagreements on occlusal, buccal/lingual, and proximal surfaces and on anterior and posterior teeth were proportionate to the number of carious surfaces on these types of surfaces and teeth.

INCONSISTENCIES IN CONCURRENT DMFS CALLS

Table 3.4 shows the difference between the average DMFS score for the first and the second concurrent examinations each year. There was a slight tendency for the second examination to have a higher score than the first examination—a trend just the opposite of what was anticipated. The diagnostic protocol specified that one factor that should be considered in classifying a surface as carious was whether the examiner could remove carious material from it (and thereby reduce the amount of material that could be removed on the second examination a few hours later).

Table 3.4

DIFFERENCE IN AVERAGE DMFS SCORE BETWEEN THE FIRST AND SECOND CONCURRENT EXAMINATION[a]

Year	Mean Inter Difference	Mean Intra Difference
	Cohort 1+2	
1	.02	.01
2	.02	.00
3	.06	.03
4	-.02	(b)
5	.02	(b)
	Cohort 5	
1	.00	.08
2	.10	.09
3	.17	.02
4	.02	(b)
5	-.07	(b)

[a]Difference = (2d Exam) - (1st Exam).

[b]Intraexaminer data were not collected in Years 4 and 5

The most important conclusion that can be drawn from Table 3.4 is that virtually none of the disagreements on concurrent DMFS calls was due to the first examination changing the characteristics of the surfaces in a way that would significantly affect the second examination. Thus any disagreement observed must have been due to chance or to systematic differences within or between examiners.

Table 3.5 shows each examiner's average observed difference (AOD) among examiners after controlling for the other examiners with whom he/she was paired and the age, sex, DMFS score, treatment regimen, and other relevant characteristics of the children seen by that examiner (see Appendix D for a description of the procedures used to institute these controls). In this and other tables by examiner, two-thirds of each score is based on Cohort 1+2 and one-third on Cohort 5. A one-way analysis of variance indicated there were statistically significant differences in AOD scores among examiners in Year 1 (alpha = .05). The Year 2 and 3 differences were not statistically significant.

There was a slight (but not statistically significant) relationship between the examiners' AOD scores across years. Spearman rank-order correlations between years ranged from −.14 to .34 (median was .04), but none were statistically significant.

Table 3.5

AVERAGE OBSERVED DIFFERENCES, BY EXAMINER AND YEAR

Examiner Number	Year 1	Year 2	Year 3	Year 4	Year 5
1	.08	-.02	.04	-.05	-.32
2	-.16	-.14	-.06	-.06	.05
3	-.24	.25			.05
4	-.25				
5	-.03	-.05	.13		-.09
6	-.14				
7	.11	-.14	-.21	.06	.11
8	.08				
9	-.19				
10	-.24				
11	.00	.12	.12	.10	.21
12	.07	-.05	-.01	.16	.23
13	-.05				
14	-.22	-.15	-.22	.19	.02
15	.20				
16	-.04	-.16	.05	-.04	.14
17	-.13	.01	.00	-.06	.00
18	-.01	.12	-.08	.08	-.16
19	-.04	.00			
20	-.09	.06	-.08	-.25	.05
21	.01	.08			
22	-.04	-.16	-.07	-.17	-.24
23	.15	.09	.01		-.13
24	.36				
25	-.12	.12	.17	-.01	.11
26	-.24				
27	-.14				
28	-.25				
29	.99				
30	.28	.02	.21	.05	-.03
31	.12				

Table 3.5 also illustrates the statistical finding that only a small percentage of the variation in observed DMFS difference scores was due to some examiners tending to call more surfaces carious than other examiners. For example, after Year 1, less than 5 percent of the variance in observed difference scores was due to systematic differences among examiners (see Table 3.6). Since usually less than 10 percent of the variance in DMFS scores was due to examiners, only a fraction of 1 percent ($.05 \times .10 = .005$) of the total variance was due to some examiners calling more surfaces carious than other examiners.

Table 3.6

PERCENTAGE OF VARIANCE IN OBSERVED DMFS DIFFERENCE SCORES DUE TO SYSTEMATIC DIFFERENCES BETWEEN EXAMINERS, BY YEAR AND COHORT

Year	Cohort 1+2	Cohort 5
1	5.7	8.0
2	3.1	4.8
3	0.7	4.8
4	1.6	1.6
5	5.5	5.4

Previous studies of concurrent examiner reliability have usually reported consistency in terms of correlation coefficients or categorical indexes, such as the number of surfaces classified as carious on both exams divided by the number classified as carious on one or both exams. Table 3.7 contains the correlation between the first and second DMFS score assigned to each child who received a replicate exam as well as the consistency ratios for the children with a non-zero DMFS score at baseline.

Table 3.7 shows very high correlations, but these correlations and those reported by others are very unstable due to the extreme positive skewness of the DMFS score distribution (see Bell et al., 1982). The intraexaminer consistency ratios obtained in the Program, especially those observed after Year 1, are comparable to those reported by others. For instance, the

Table 3.7

CORRELATION COEFFICIENTS AND CONSISTENCY RATIOS BETWEEN THE FIRST AND SECOND CONCURRENT EXAM, BY YEAR AND COHORT

	Correlation Coefficients				Consistency Ratios			
	Cohort 1+2		Cohort 5		Cohort 1+2		Cohort 5	
Year	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Intra
1	.87	.95	.93	.94	44	65	59	71
2	.96	.98	.97	.99	58	77	63	79
3	.95	.97	.95	.99	66	81	71	84
4	.97	(a)	.99	(a)	68	(a)	67	(a)
5	.98	(a)	.98	(a)	62	(a)	66	(a)

[a]Intraexaminer data were not collected in Years 4 and 5.

average index in the eight intraexaminer studies reviewed by Shaw and Murray (1975) was 75 (with about 60 replicate exams per study and with most of the examiners knowing which cases were in the replicate sample). The Program's results are based on 6,371 concurrent replicate exams embedded in the regular examination process.

Table 3.8 presents the distribution of DMFS absolute difference scores by year and cohort. For the purpose of this table, an absolute difference score is the difference in score between the first and second examination of a child regardless of the algebraic sign of this difference (if one examiner assigns a child a DMFS score of 3 and another examiner assigns that child a score of 2, the child would receive an absolute difference score of 1). Table 3.9 provides the corresponding data for DMFT.

Tables 3.8 and 3.9 show the same trends as those in Table 3.7: Examiners were slightly more consistent with themselves than with each other, they were more consistent with younger than with older children, and their consistency increased during the Program despite the children becoming older. Further, the degree of consistency is as good or better than that obtained in previous health surveys (see Table 1.2). Essentially all the differences were in the D score (the distribution of difference scores on this component almost mirrored that of the total score).

Perhaps the best indicator of examiner consistency is the average absolute difference (AAD) between the first and second score; i.e., the average difference between the scores on the two exams regardless of the algebraic sign of this difference (Klein and Bell, 1981). This statistic is similar to, but far less cumbersome than, the distribution of DMF difference scores presented above. It also makes it easier to conduct analyses of individual examiner characteristics, and it avoids many of the problems associated with reversal rates, consistency ratios, and correlation coefficients.

An analysis of AAD scores indicated that examiners tended to agree somewhat more with themselves than with each other. For example, in Cohort 5 in Year 1, there was an .83 average difference in DMFS score between interexaminer pairs but only a .46 average difference in DMFS score between intraexaminer pairs. Both types of disagreements were usually about 5 to 20 percent of the group's mean DMFS score (see Tables 2.3 and 3.10).

AAD scores showed relatively little variation across examiners within a given year; i.e., some examiners were not substantially more reliable than others in terms of either interexaminer or intraexaminer consistency indexes. Table 3.11 presents AAD scores by examiner and year after controlling for whom the examiner was paired with (in the case of interexaminer comparisons) and the characteristics of the children seen by that examiner.

The D component of a child's DMFS score was by far the major correlate of absolute difference scores, accounting for 20 to 30 percent of the variance in intraexaminer inconsistencies in DMFS and 40 to 50 percent of the variance in interexaminer inconsistencies. Age, sex, race, site, treatment regimen, and number of erupted teeth did not add anything to the prediction of absolute difference scores.

There was no consistent relationship among interexaminer or intraexaminer AAD or AOD scores across years (e.g., the Spearman rank-order correlation between an examiner's AOD score and interexaminer AAD score was .14 in Year 1, .06 in Year 2, and .26 in Year 3).

Table 3.8

PERCENTAGE OF CHILDREN WITH EACH DMFS ABSOLUTE DIFFERENCE SCORE

Size of Difference	Year 1		Year 2		Year 3		Year 4[a]	Year 5[a]
	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Inter
				Cohort 1+2				
0	72	79	78	89	83	90	81	80
1	18	16	17	9	13	8	14	13
2	6	4	4	2	3	2	3	4
3	2	1	1	(b)	1	(b)	1	1
>3	2	1	(b)	0	1	(b)	1	1
				Cohort 5				
0	58	72	67	81	70	86	68	62
1	23	19	22	15	20	11	21	22
2	9	5	8	4	5	1	6	8
3	5	2	1	1	2	0	4	4
>3	5	2	2	0	4	1	3	4

[a]Intraexaminer data were not collected in Years 4 and 5.

[b]Less than 0.5 percent.

Table 3.9

PERCENTAGE OF CHILDREN WITH EACH DMFT ABSOLUTE DIFFERENCE SCORE

Size of Difference	Year 1		Year 2		Year 3		Year 4[a]	Year 5[a]
	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Inter
				Cohort 1+2				
0	76	82	82	89	86	92	85	84
1	17	15	16	10	12	7	12	11
2	4	2	2	1	2	1	3	3
3	2	(b)	(b)	(b)	(b)	0	1	1
>3	1	0	0	0	(b)	(b)	0	0
				Cohort 5				
0	67	79	73	84	78	91	74	70
1	24	15	22	14	17	8	19	22
2	6	4	3	1	3	1	5	7
3	2	1	1	(b)	3	1	2	1
>3	1	1	0	0	1	(b)	(b)	1

[a]Intraexaminer data were not collected in Years 4 and 5.

[b]Less than 0.5 percent.

Table 3.10

AVERAGE ABSOLUTE DIFFERENCE SCORES, BY COHORT, YEAR, AND TYPE OF RELIABILITY SAMPLE

Year	Average Interexaminer Difference[a]	Average Intraexaminer Difference[a]
	Cohort 1+2	
1	.44	.29
2	.28	.14
3	.26	.15
4	.28	(b)
5	.30	(b)
	Cohort 5	
1	.83	.46
2	.52	.24
3	.63	.18
4	.58	(b)
5	.72	(b)

[a]Difference = |2d Exam DMFS - 1st Exam DMFS|.

[b]Intraexaminer data were not collected in Years 4 and 5.

Thus the examiners who tended to call more surfaces carious than other examiners were no more consistent with the other examiners than were those who had the opposite tendency.

INCONSISTENCIES IN LONGITUDINAL DMFS CALLS

The analyses of examiner consistency over time were based on a population of tooth surfaces that belonged to 5,579 children (3,821 from Cohort 1+2 and 1,938 from Cohort 5). These children were selected because (1) they had a non-zero DMFS score on at least one of the following exams: Year 1, 2, 3, and 5; and (2) they had a usable DMFS score on all four of those exams. A tooth surface was included in the population of surfaces if it was classified as D, M, or F on any one of these exams and if it was never classified as a "Y" (nonexaminable). Table 3.12 shows the number of surfaces that were used in the analyses.

The longitudinal analysis focused on whether maintaining or not maintaining the same examiner/child pairings over years affected the number of reversals that occurred. A "reversal" is defined as a surface that is classified as D, M, or F on one exam, but as free from any evidence of caries attack on a subsequent exam. The theory underlying this index is that the

Table 3.11

AVERAGE ABSOLUTE DIFFERENCE SCORES, BY EXAMINER AND YEAR

Examiner Number	Interexaminer					Intraexaminer		
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 1	Year 2	Year 3
1	.53	.35	.27	.48	.08	.42	.08	.06
2	.39	.28	.24	.44	.50	.30	.16	.28
3	.32	.61			.49	.29	.12	
4	.29					.31		
5	.72	.45	.50		.34	.36	.13	.14
6	1.00					1.02		
7	.65	.30	.41	.34	.56	.30	.08	.03
8	.24					.00		
9	.47					.21		
10	.72					.18		
11	.34	.30	.57	.33	.68	.67	.16	.28
12	.43	.32	.19	.52	.70	.58	.15	.26
13	.68					.26		
14	.40	.37	.37	.62	.46	.06	.28	.29
15	.68					.37		
16	.48	.31	.46	.18	.60	.17	.10	.15
17	.70	.17	.21	.40	.44	.18	.16	.05
18	.28	.37	.04	.30	.25	.20	.00	.00
19	.55	.75				.65	.26	
20	.72	.29	.28	.30	.50	.39	.12	.05
21	.45	.37				.24	.13	
22	.58	.36	.45	.43	.16	.68	.31	.02
23	.35	.38	.34		.30	.21	.39	.09
24	.56					.40		
25	.32	.27	.48	.28	.56	.30	.13	.18
26	.67					.11		
27	.87					.46		
28	.20					.07		
29	1.07					.39		
30	.54	.27	.16	.33	.41	.48	.53	.21
31	.63					.31		
Average	.54	.36	.33	.38	.44	.34	.18	.14

Table 3.12

NUMBER OF SURFACES AVAILABLE FOR LONGITUDINAL STUDIES OF EXAMINER CONSISTENCY

Year	Cohort 1+2	Cohort 5
1	5,598	8,200
2	8,302	10,508
3	11,148	12,553
5	16,544	16,998

decay process is irreversible; therefore, any reversals signify examiner error. Because there is now evidence that some reversals may occur as a result of remineralization, the number appropriately attributable to examiner error is slightly inflated.

Reversal rate is defined as the number of reversals that occur on one exam divided by the number of surfaces classified as D, M, or F on a prior exam. For instance, the denominator for computing the reversal rate in Cohort 5 between the Year 1 and 3 exams is 8,200.

Table 3.13 shows what percent of the 5,579 children had the same examiner for any given pair of exams. For example, 11 percent of the Cohort 1+2 children and 9 percent of the Cohort 5 children had the same examiner at baseline as they had at the end of the Program.

Table 3.14 contains reversal rates from Year 1 to 3 and from 3 to 5 by cohort and by whether or not the same examiner/child pairings were maintained over time. These data show that reversal rates were not affected by whether the same examiner/child pairings were maintained, and that the reversal rates between Year 3 and 5 were uniformly smaller than the rates between Year 1 and 3.

Table 3.13

PERCENTAGE OF CHILDREN WHO HAD THE SAME EXAMINER FOR A GIVEN PAIR OF EXAMS

Year	1	2	3	5
1		15	9	11
2	17		15	14
3	9	13		11
5	9	10	11	

NOTE: The data for Cohort 1+2 appear above the main diagonal, and the data for Cohort 5 appear below it.

Table 3.14

PERCENTAGE OF SURFACES CLASSIFIED AS AFFECTED BY DECAY ON ONE EXAM BUT NOT CLASSIFIED AS AFFECTED BY DECAY TWO YEARS LATER, BY COHORT AND WHETHER EXAMINER/CHILD PAIRINGS WERE OR WERE NOT MAINTAINED

Examiner/child Pairing	Year 1 to 3		Year 3 to 5	
	1+2	5	1+2	5
Maintained	13	8	6	5
Not maintained	14	10	4	6
Mean	14	9	5	5

Table 3.15 shows the average number of reversals per child. These averages are misleading because they do not include children with a DMFS score of 0 in the base year. For instance, the Year 1 to 3 rates would be reduced by 65 percent in Cohort 1+2 and by 25 percent in Cohort 5 if the samples used to compute them included children with a baseline DMFS score of 0. Nevertheless, the rates presented in Table 3.15 are similar to those reported by others. For example, Fischman et al. (1976) obtained an average intraexaminer reversal rate of .37 per child after one year for children who were initially examined as sixth graders. It is not clear whether this rate used all the children examined as the denominator or just those who had a non-zero DMFS score on the first exam.

Table 3.16 shows that the percentage of children with reversals is almost the same in Cohort 5 as in Cohort 1+2.

Tables 3.14, 3.15, and 3.16 show that maintaining the same examiner/child pairings across years had little effect on reversals. The Not Maintained row for the Year 1 to 3 results in each table includes data from examiners who did not participate after baseline. These

Table 3.15

AVERAGE NUMBER OF REVERSALS PER CHILD FROM YEAR 1 TO 3 AND FROM 3 TO 5, BY COHORT AND WHETHER OR NOT EXAMINER/CHILD PAIRINGS WERE MAINTAINED[a]

Examiner/child Pairing	Year 1 to 3		Year 3 to 5	
	1+2	5	1+2	5
Maintained	.37	.45	.22	.29
Not maintained	.40	.51	.17	.44
Mean	.39	.48	.19	.37

[a]Averages based on just the children with a non-zero DMFS score in the base year.

Table 3.16

PERCENTAGE OF CHILDREN WITH A NON-ZERO DMFS SCORE IN ONE YEAR WHO HAD ONE OR MORE REVERSALS TWO YEARS LATER, BY COHORT AND WHETHER OR NOT EXAMINER/CHILD PAIRINGS WERE MAINTAINED

Examiner/child Pairing	Year 1 to 3		Year 3 to 5	
	1+2	5	1+2	5
Maintained	28.3	29.0	10.7	15.3
Not maintained	28.7	29.6	12.1	19.8
Mean	28.5	29.3	11.4	17.5

examiners were somewhat less reliable than those who did continue to participate. This lack of parallelism may account for some of the little difference between rows.

Tables 3.14 and 3.15 illustrate that the potentially important difference between inter-examiner and intraexaminer consistency rates observed in the concurrent studies disappeared in the longitudinal analysis. There is therefore little empirical support for the hypothesis that maintaining pairings over time reduces inconsistencies and thereby improves the chances of detecting treatment effects.

Analyses of just the surfaces that were classified as decayed (i.e., when missing due to decay and filled were not considered) indicated that the reversal rate on buccal/lingual surfaces was about twice as high as it was on occlusal surfaces (see Table 3.17). For instance, only 7 percent of the occlusal surfaces classified as decayed on the Year 3 exam were classified as sound on the Year 5 exam. This tendency for occlusal surfaces to have the lowest reversal rate runs counter to the theory that reversal rates were biased by the application of occlusal sealants after baseline.

The issue of whether the Program's reversal rates may have been affected by the various treatment regimens was investigated directly by computing the rate in each regimen. From the results of this analysis in Table 3.18 we can see that there was no systematic relationship between reversal rate and regimen. For instance, the sealant regimens (1 and 3 at

Table 3.17

REVERSAL RATE FOR SURFACES CLASSIFIED AS DECAYED IN BASE YEAR, BY SURFACE TYPE

Surface Type	Year 1 to 3	Year 3 to 5
Buccal/lingual	.30	.19
Interproximal	.19	.16
Occlusal	.15	.07

nonfluoridated sites and 1, 2, and 3 at fluoridated sites) did not have consistently higher or lower rates than the other regimens, although in Year 1 to 3 data, Cohort 1+2 children had a slightly higher reversal rate than Cohort 5 children.

EFFECT OF EXAMINER INCONSISTENCIES ON DETECTING TREATMENT EFFECTS

The primary reason for conducting the reliability analyses presented in this report was to assess the degree to which examiner inconsistencies might interfere with the detection of possible treatment effects. This issue was investigated by calculating the probability of detecting a given size difference between two treatment groups in their respective average DMFS increment scores between Years 1 and 3 when examiner inconsistencies were and were not present.

The calculation of probability estimates used data from both the concurrent and longitudinal analyses. The variance in increment scores between years indicated the variance in the increments when examiner inconsistencies were present. The variance due to examiner error in Year 1 was estimated by taking one-half of the average of the squared difference (ASD) scores between examiners in Year 1. The same technique was used to estimate the variance due to examiner error in Year 3. We then obtained an estimate of the variance in increment scores without examiner error by means of the formula below (where: Vwo = variance without examiner error, Vw = variance with examiner error, and ASD1 and ASD3 are the average of the squared absolute difference scores between examiners in Years 1 and 3).

$$Vwo = Vw - (0.5)(ASD1 + ASD3)$$

The two types of variances, Vw and Vwo, were used to compute the likelihood of detecting a difference in the average two-year DMFS increment scores between two groups (if a difference actually exists) when examiner inconsistencies are and are not present. These probabilities, presented in Table 3.19, indicate that the likelihood of detecting practically important treatment effects (at least one surface in two years) is not appreciably reduced by the level

Table 3.18

PERCENTAGE OF SURFACES REVERSED FROM YEAR 1 TO 3 AND 3 TO 5, BY TREATMENT REGIMEN, COHORT, AND FLUORIDATION LEVEL

	Year 1 to 3				Year 3 to 5			
Regimen	NF 1+2	NF 5	F 1+2	F 5	NF 1+2	NF 5	F 1+2	F 5
1	11	10	18	14	6	4	6	8
2	14	8	18	10	4	7	4	5
3	16	9	17	13	5	9	5	5
4	12	8	13	8	6	5	4	1
5	15	8	10	9	5	6	2	4
6	13	10	13	9	3	7	3	7
Average	14	9	15	11	5	6	4	5

of examiner inconsistency observed in the Program. For instance, in Cohort 5, the chances of detecting a one-surface difference in mean DMFS increment scores between any two of the Program's regimens are .95 when examiner inconsistencies are present and .99 when they are not present.

The probabilities in Table 3.19 are for a one-tailed test (alpha = .05) with 200 subjects per group and with none of the variance in increment scores explained by covariates (such as the child's site or baseline score). Probabilities would increase if there were more subjects per group and if some of the variance in increment scores could be predicted by covariates. The Program had far more than 200 subjects per group after two years, and we were able to predict slightly more than 10 percent of the variance in two-year increment scores. Thus, the actual probabilities of detecting true two-year effects were slightly higher than those presented in the left hand portion of Table 3.19.

The power of a statistical test (the probability of detecting a true difference between groups if one exists) also depends on the ratio of the size of the true difference we are trying to detect to the standard deviation of the increment scores. Thus, as may be seen from a comparison of the left and right sides of Table 3.19, we had more power in detecting a 1.0 surface difference in four years than we had in detecting a 0.5 surface difference in two years. Appendix E contains a more complete discussion of the procedures used to compute the probability estimates.

Table 3.19

PROBABILITIES OF DETECTING TWO- AND FOUR-YEAR TREATMENT EFFECTS WHEN VARIANCE DUE TO EXAMINER INCONSISTENCIES IS AND IS NOT PRESENT

Cohort	Examiner Error	Size of Difference Trying to Detect in Two Years			Size of Difference Trying to Detect in Four Years		
		0.5	1.0	1.5	0.5	1.0	1.5
1+2	Present	.80	1.00	1.00	.59	.98	1.00
	Absent	.88	1.00	1.00	.63	.99	1.00
5	Present	.49	.95	1.00	.32	.76	.97
	Absent	.64	.99	1.00	.35	.81	.98

NOTE: Probabilities are for a one-tailed test at alpha equal to .05 with 200 children per group.

Chapter 4

RELIABILITY OF OTHER INDEXES OF ORAL HEALTH

In addition to providing the data needed for calculating DMFS scores, the clinical examiners evaluated four oral conditions: (1) degree of gingivitis (inflammation of the gums), (2) the urgency with which the child needed to see a dentist for possible treatment (as indicated by a referral code), (3) whether sealants were present on a child's teeth, and (4) whether there were white spots on the child's teeth that could have been due to fluoride or other factors. This chapter provides information about concurrent examiner reliability on each of these indexes.

GINGIVITIS

After completing the DMFS portion of the clinical examination, the examiners assessed the degree of gingivitis that was present, using the San Francisco Public Health Service Dental Health Center (DHC) Index (Suomi et al., 1969). This index involved evaluating the degree of inflammation of the gingival tissue along both the lingual and buccal surfaces of six permanent teeth. Each of these twelve surfaces was 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—gingivae adjacent to the tooth surface were pale pink in color and firm in texture. Swelling was not evident and stippling could usually be noted.
- 1 = Inflammation not encompassing all tissue adjacent to the tooth surface (including papillae)—gingivae were a definite red or magenta color; color change did not extend from mesial to distal.
- 2 = Inflammation encompassed all tissue adjacent to the tooth surface (including papillae); color change extended from mesial to distal.
- NA = In this category were primary and missing teeth, and permanent teeth that were banded or less than 50 percent erupted.

A child's DHC score was the average score assigned to all the surfaces that were examined. Thus, the possible score ranged from 0.00 to 2.00. If all of a child's teeth were classified as NA, then that child did not receive a DHC score. And if one or both surfaces of a tooth were classified as NA, then that tooth was eliminated from the calculation of a child's DHC score. Thus the number of teeth used in computing a DHC score was not the same for all children.

The Program measured gingivitis because periodontal disease is the major source of tooth loss in adults, and gingivitis is one of the first signs of this disease. Moreover, one theory holds that gingivitis and tooth decay are related because dental plaque is assumed to be a factor in both of these oral diseases.

As noted in Table 1.3, our review of the dental literature located only one study that assessed examiner reliability with the DHC index (Smith et al., 1970). In this study, one

examiner conducted replicate DHC evaluations on 54 children (779 surface areas). This examiner agreed with himself about 82 percent of the time; i.e., the call made on the first exam (0, 1, or 2) deviated from the call made on the second exam on only 18 percent of the surfaces evaluated. No data were provided regarding the mean, standard deviation, or distribution of the scores (such as whether almost all of them were zeros). However, the authors noted that reliability improved over the three-day data collection period. Reliability coefficients with other indexes tended to run in the high .70's to low .80's; but the sample sizes were quite small.

Analyses of Year 1 data in the Program's total sample indicated that DHC scores were not related to a child's DMFS score, age, sex, or racial/ethnic group (Bell et al., 1982). However, the average DHC score in the first five sites visited (1.06) was significantly lower than the average DHC score in the remaining sites (1.38). In fact, none of the first five sites had a higher average DHC score than any of the remaining sites (see Table 4.1). The change in examination standards across sites is illustrated clearly by two groups of children, A and B, that were examined in Minneapolis. Although Group B's examinations were conducted only a few months after Group A's exams (and before any preventive care was provided), Group B's mean DHC score was 41 percent higher than Group A's mean.

Because most of the examiners at the first five sites did not visit the remaining ones, differences in average score between sites were more likely due to how different examiners interpreted the DHC criteria than to actual differences in the average level of gingivitis at these sites. There were no statistically significant differences in average DHC score between early and late sites on the Year 2 or Year 3 examinations.

Table 4.2 presents the correlation between DHC scores on the first and second examination each year within each reliability sample and Cohort. Examiners were far more consistent with themselves in making DHC calls than they were with each other; for instance, the mean interexaminer correlation across years in Cohort 5 was .42 as compared to .83 for the mean intraexaminer correlation. Each sample's mean score on the first exam differed by no more than .05 from its mean on the second exam. Thus the differences were not statistically significant.

Table 4.1

AVERAGE DHC SCORE BY SITE IN YEAR 1

Order Visited	Site	Average DHC Score
1st	Chattanooga	1.15
2d	Billerica	1.00
3d	New York	0.97
4th	Minneapolis A	1.02
5th	Wichita	1.18
6th	Pierce County	1.36
7th	Hayward	1.46
8th	El Paso	1.40
9th	Monroe	1.22
10th	Tallahassee	1.41
11th	Minneapolis B	1.42

Table 4.2

CORRELATION BETWEEN DHC SCORES WITHIN EACH RELIABILITY SAMPLE, BY COHORT AND YEAR

	Cohort 1+2		Cohort 5	
Year	Inter	Intra	Inter	Intra
1	.41	.80	.52	.85
2	.28	.79	.34	.84
3	.33	.87	.41	.82
4	.27	(a)	.29	(a)
5	.30	(a)	.18	(a)
Mean	.32	.82	.35	.83

[a]Intraexaminer data were not collected in Years 4 and 5.

Taken together, the data in Tables 4.1 and 4.2 point to large systematic differences among examiners in how they interpreted the DHC scoring criteria. A series of one-way analyses of variance also showed that within a year, some examiners had statistically significantly higher DHC means than others. Systematic differences explained 25 percent of the Year 1 variance of examiner differences in DHC scores, and from 48 to 57 percent in the Years 2 to 4, and 62 percent in Year 5. Examiners with higher than average mean scores in one year tended to have higher than average means in the other years. With one exception, Spearman rank-order correlations between years ranged from .59 to .84 (significant at $p = .05$). The one exception was between Years 1 and 5 where the correlation was .22.

The foregoing findings suggest that the low level of agreement between examiners can be attributed to a lack of adequate calibration rather than to any inherent problems in being able to apply the DHC criteria. The Program devoted far more time and effort to calibrating examiners on the DMFS index than it did on the DHC. Furthermore, examiners were retained after the first year on the basis of their consistency with other examiners on DMFS without regard to agreement on the DHC index. DMFS was emphasized because: (1) it had greater relevance to the policy issues addressed by the Program and (2) resources did not permit allocating the extensive amount of examiner time needed to obtain a high level of interexaminer agreement on both DMFS and DHC.

The DHC data imply that the DHC's reliability would improve if the same examiner/child pairings are maintained across years. The findings also indicate that DHC scores will have to be adjusted for systematic differences among examiners before analyses can be conducted to assess whether the Program's treatment regimens varied in their ability to reduce gingivitis.

REFERRAL CODE

Examiners assigned one of four codes to indicate the urgency with which a child needed to see a dentist. This code determined which one of four form letters was used to notify a child's parents of the general results of the examination. The four codes are:

1 = No problems, only routine dental care indicated.

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

3 = One or more serious problems requiring immediate attention. Parent should be contacted by telephone as well as notified in writing.

4 = A complete examination was not performed (e.g., more than 50 percent of the child's permanent teeth were banded). Children with this code were excluded from all analyses of examiner reliability.

Table 4.3 shows the percentage of children on which the examiners did and did not agree as to the code that should be assigned. For instance, in Year 3, different examiners agreed on the code that should be assigned to 88 percent of the children in Cohort 5; the remaining 12 percent were due to disagreements between either codes 1 and 2 or 2 and 3. Intraexaminer agreement rates are presented in Table 4.4.

Tables 4.3 and 4.4 show that agreement rates on referral codes improved over time, were higher within examiners than between them, and were the same with younger children as they were with older ones. A child classified as a "1" on one exam and as a "3" on another exam

Table 4.3

PERCENTAGE OF CHILDREN ON WHICH EXAMINERS AGREED OR DISAGREED WITH EACH OTHER IN THE REFERRAL CODES THAT SHOULD BE ASSIGNED

	Agreed on Code				Disagreed on Code			
Year	1	2	3	Total	1vs2	2vs3	1vs3	Total
				Cohort 1+2				
1	46	18	5	69	24	6	1	31
2	55	21	3	79	16	5	1	22
3	58	24	4	86	12	2	0	14
4	54	26	2	82	15	2	1	18
5	60	17	2	79	17	4	1	22
				Cohort 5				
1	38	23	6	67	24	7	2	33
2	47	26	3	76	20	4	0	24
3	54	27	7	88	8	4	0	12
4	45	33	3	81	15	5	0	20
5	45	31	3	79	14	5	1	20

Table 4.4

PERCENTAGE OF CHILDREN ON WHICH EXAMINERS AGREED OR DISAGREED WITH THEMSELVES IN THE REFERRAL CODES THAT SHOULD BE ASSIGNED

	Agreed on Code				Disagreed on Code			
Year[a]	1	2	3	Total	1vs2	2vs3	1vs3	Total
				Cohort 1+2				
1	54	24	8	86	10	3	0	13
2	61	27	3	91	7	2	0	9
3	63	28	3	94	5	1	0	6
				Cohort 5				
1	43	36	6	85	14	2	0	16
2	53	32	4	89	9	2	0	11
3	53	36	5	94	5	1	0	6

[a]Intraexaminer data were not collected in Years 4 and 5.

would constitute the most serious type of disagreement because parents might not be notified if the child had a serious condition. This type of disagreement occurred on less than 1 percent of the children in the interexaminer sample and on none of the children in the intraexaminer sample.

In Years 1 and 2, some examiners were statistically significantly more apt to make code 2 or 3 calls than were other examiners. The percentage of variance in referral code that was attributable to systematic differences among examiners was 7 percent and 5 percent in Years 1 and 2, respectively. In the remaining years, it was always less than 3 percent. The tendency for some examiners to call high codes more often than other examiners was correlated with these examiners' observed DMFS difference scores. The correlations ranged from .27 to .84 with a median of .51.

The high degree of consistency with which referral codes were assigned, especially after Year 1, and their close relationship with the D component of DMFS scores (Bell et al., 1982) suggest that the codes could provide a sensitive index of the effects of different treatment regimens on a child's general oral health status. In fact, although age was correlated with referral code, the percentage of children that received a code 1 in each Cohort steadily increased over time. This finding reflects a positive effect of the Program on the children's oral health.

SEALANTS

A sealant is a plastic-like coating which is applied to the occlusal (chewing) surfaces of the teeth most susceptible to decay. Each year, examiners were asked to indicate whether any sealants were present.

Year 1 results showed that less than 1 percent of the children had sealants. After the Year 1 exam, children in regimens 1, 2, and 3 at fluoridated sites and in regimens 1 and 3 at

nonfluoridated sites had sealants placed on their decay-free permanent posterior teeth. Because the type of sealant applied (Delton) was transparent it was harder to detect than an opaque-type sealant.

The treatment protocol specified that the clinic team could place sealants on surfaces not previously classified as carious by one of the Program's dental examiners. However, the clinic team or a child's private dentist may have placed a sealant on a surface diagnosed as carious on the Year 1 exam if the Year 2 examiner classified that surface as sound.

Table 4.5 shows the percentage of children on which the examiners agreed and disagreed as to whether sealants were present. Only the Year 2 and 3 results are presented because hardly any children had sealants before the Program provided them. The data show that in both Years 2 and 3, examiners agreed with themselves and each other on over 90 percent of the children. Because schools rather than children were assigned to regimens, examiners usually learned quickly whether a school was in a sealant regimen. By the Year 2 exam, several schools, and particularly junior highs, had children from more than one regimen. Thus, many of the disagreements probably stemmed from examiners developing a habit of saying sealants were present when they visited schools where most of the children had sealants, and developing the opposite tendency at other schools.

Agreement rates in Table 4.5 are higher than those reported by Burt et al. (1977) and Horowitz and Poulsen (1978) for classifying individual teeth. This comparison is not fair, however, because it is much easier to agree when an examiner only has to determine whether any tooth has a sealant.

Table 4.5

PERCENTAGE OF CHILDREN IN EACH RELIABILITY SAMPLE ON WHICH EXAMINERS AGREED OR DISAGREED ABOUT PRESENCE OF SEALANTS

	Interexaminer			Intraexaminer		
Year	Agreed present	Agreed absent	Disagreed	Agreed present	Agreed absent	Disagreed
			Cohort 1+2			
2	33	64	3	34	62	4
3	39	57	4	36	60	4
4	42	54	4	(a)	(a)	(a)
5	42	52	6	(a)	(a)	(a)
			Cohort 5			
2	32	60	8	38	56	6
3	33	61	6	41	55	4
4	41	55	4	(a)	(a)	(a)
5	43	51	6	(a)	(a)	(a)

[a]Intraexaminer data were not collected in Years 4 and 5.

OPACITY

Year 1 examinations at the first few sites visited revealed that a few children had opaque, chalky white spots on their teeth. Such opacities could be caused by the children receiving more than an optimal amount of fluoride. Other sources include systemic disturbances (such as febrile conditions), tetracycline staining, dehydration of the enamel due to respiratory habits, and trauma. Because several of the Program's treatment regimens involved applying fluoride and because of concern that preexisting opacities might be attributed to the Program, we decided to monitor whether opacities were present on a child's teeth. This monitoring began at the Wichita site in Year 1 and was conducted at all sites in each of the following years.

Table 4.6 shows the percentage of children in each reliability sample on which the examiners agreed or disagreed as to whether opacities were present. The 93 to 98 percent intraexaminer agreement rates in this table are higher than those reported by Davies et al. (1967), Olsson (1979), and Thylstrup and Fejerskov (1978).

Even though older children had more opportunity to develop opacities than younger ones, there was no statistically significant difference between age groups on this index. There was an increase over time in the percentage of children with opacities and a parallel drop in interexaminer agreement rate. Analyses of total sample data suggest that changes over time

Table 4.6

PERCENTAGE OF CHILDREN IN EACH RELIABILITY SAMPLE ON WHICH EXAMINERS AGREED OR DISAGREED ABOUT PRESENCE OF OPACITIES

	Interexaminer			Intraexaminer		
Year	Agreed present	Agreed absent	Disagreed	Agreed present	Agreed absent	Disagreed
			Cohort 1+2			
1[a]	1	95	4	2	96	2
2	6	77	17	9	84	7
3	15	65	20	19	74	7
4	17	52	30	(b)	(b)	(b)
5	18	48	34	(b)	(b)	(b)
			Cohort 5			
1[a]	2	94	4	3	93	4
2	9	77	14	9	85	6
3	12	70	18	15	81	4
4	14	52	34	(b)	(b)	(b)
5	13	59	28	(b)	(b)	(b)

[a]Year 1 results are restricted to just the sites at which opacities were assessed.

[b]Intraexaminer data were not collected in Years 4 and 5.

were probably attributable to subtle modifications in the examination criteria rather than to any real change in the prevalence of the opacities. There was no systematic relationship between a child's treatment regimen and whether that child was classified as having opacities. A child receiving one or several types of fluoride through the Program was therefore no more or no less likely to develop opacities than a child who did not receive these fluoride treatments.

Chapter 5

RADIOGRAPH READER RELIABILITY

Two bitewing radiographs of posterior teeth, one of the left side and one of the right, were taken per child in Year 1 and again in Year 5. The six dentists who read the films evaluated the two interproximal surfaces (distal and mesial) on 16 teeth, classifying a surface as sound, unerupted (including deciduous), missing (because of decay or trauma), nonreadable, carious, or filled. Only the last two calls could affect a child's DMFS score because readers had no way of determining the reason for a missing permanent tooth. A film set was considered readable if at least one surface on each film was judged readable.

This chapter discusses which children were selected to have their films reread, the frequency of interreader and intrareader disagreements, whether some readers were more consistent than others in their evaluations, and how reader inconsistencies affected the accuracy of a child's DMFS score when radiograph data were considered in calculating that score.

REREAD SAMPLES

About 12 percent of the children in Cohorts 1, 2, and 5 were selected to have their film sets reread. Almost all these children had their sets classified as readable on the first reading. The other variables used in selecting a child's set for reread were site, school, Cohort, and the reader who initially evaluated the set. Within each Cohort's group of selected sets, each reader evaluated about the same number of sets per site and each site had about the same number of sets.

A selected set was assigned to an intrareader or an interreader sample. Each intrareader set was reread by the same dentist who had initially evaluated it. In the interreader sample, the first and second readings were done by different dentists. Table 5.1 shows the number of children (and thereby film sets) in each reread sample each year.

Table 5.1

NUMBER OF FILM SETS IN EACH REREAD SAMPLE

	Year 1			Year 5		
	Cohort 1+2	Cohort 5	Total	Cohort 1+2	Cohort 5	Total
Interreader	945	476	1421	715	241	956
Intrareader	710	327	1037	110	68	178
Total	1655	803	2458	825	309	1134

FREQUENCY OF READER INCONSISTENCIES

Table 5.2 shows how often one reading of a surface agreed with another reading of that surface. These data indicate intrareader agreement rates were only slightly higher than interreader rates; and, with the important exception of decayed calls, both rates were 77 percent or higher. Almost all the carious call disagreements involved a surface being classified as decayed on one reading and as sound on another reading. No percentages are reported for filled surfaces in Cohort 1+2 in Year 1 because there were fewer than 10 filled interproximal surfaces in their reread samples.

EFFECT OF READER INCONSISTENCIES ON DFS SCORES

The number of decayed and filled surfaces (DFS) on each reading was calculated for each child in the reread samples. Table 5.3 shows the distribution of absolute difference scores between readings. These data indicate that the degree of agreement between readings declined with age level. For instance, there was no difference in DFS score between readings on slightly more than 80 percent of the fifth graders in both Year 1 (Cohort 5) and Year 5 (Cohort 1+2).

Table 5.2

PERCENTAGE OF AGREEMENT BETWEEN READINGS ON EACH TYPE OF CALL

	Year 1				Year 5			
	Cohort 1+2		Cohort 5		Cohort 1+2		Cohort 5	
Type of Call	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Intra
Sound	83	89	88	94	90	95	88	96
Unerupted	80	88	94	96	95	97	72	87
Decayed	53	67	57	74	68	85	70	81
Filled	--	--	92	93	94	96	96	97
Nonreadable	83	90	77	86	78	90	77	92

Table 5.3

PERCENTAGE OF CHILDREN WITH VARIOUS ABSOLUTE DFS DIFFERENCE SCORES

	Year 1				Year 5			
Size of Absolute Difference Score	Cohort 1+2		Cohort 5		Cohort 1+2		Cohort 5	
Between Readings	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Intra
0	95	95	80	85	81	89	62	66
1	4	4	15	12	13	11	20	19
2	1	(a)	3	3	4	0	10	10
>2	0	(a)	2	(a)	2	0	9	5

[a]Less than 0.5 percent of the children.

At first glance, the data in Table 5.3 suggest a very high degree of consistency both within and between readers in making radiograph calls that could affect a child's DFS score. However, the especially high degree of consistency with Cohort 1+2 children in Year 1 was largely because almost all of them did not have any interproximal decay (i.e., they received a DFS score of 0 on both readings). In addition, radiograph DFS calls are not used by themselves but as a supplement to clinical DMFS calls. The rule for combining clinical and radiograph calls to obtain total DMFS scores is: "A surface is classified as decayed if it is called decayed by a clinical examiner and/or a radiograph reader." A parallel rule applies to filled surfaces. Thus, the real issue is whether radiograph calls provide reliable data when they are used to *supplement* clinical results; i.e., when they affect a child's total DMFS score.

In Cohort 1+2's interreader sample in Year 1, 112 surfaces were classified as affected (decayed or filled) by either or both of the radiograph readers and/or by the clinical examiner. There were 95 such surfaces in their intrareader sample. The corresponding values in Cohort 5's samples in Year 1 were 365 and 244. The values for Year 5 were: 565, 71, 697, and 200.

Table 5.4 shows the percentage of Year 1 surfaces classified as affected on each radiograph reading relative to how these surfaces were classified on the clinical exam. For instance, on 83 percent of the 112 surfaces in Cohort 1+2's interreader sample, one or both of the radiograph readers said the surface was affected but the clinical examiner called it not affected. However, this 83 percent is made up of 29 percent where the radiograph readers agreed and 54 percent where they disagreed. An almost 2 to 1 ratio of disagreements to agreements also was obtained in the interreader sample in Cohort 5. In short, on those surfaces where the radiograph results would modify the clinical findings, the readers were about twice as likely to disagree as they were to agree.

Table 5.5 shows that in Year 5, the readers were just as likely to disagree with each other as they were to agree on the surfaces that were classified as "not affected" by the clinical examiners. This is a slight improvement over the Year 1 agreement rate. In both years, how-

Table 5.4

PERCENTAGE OF YEAR 1 SURFACES CLASSIFIED AS AFFECTED ON THE CLINICAL EXAM AND/OR ON ONE OR BOTH RADIOGRAPH READINGS

Clinical Call	First and Second Radiograph Calls	Cohort 1+2		Cohort 5	
		Inter	Intra	Inter	Intra
Not affected	Agreed affected	29	38	23	38
	Disagreed	54	43	41	28
		83	81	64	66
Affected	Agreed affected	4	7	25	26
	Disagreed	2	1	4	2
	Agreed not affected	12	11	7	6
		18	19	36	34

Table 5.5

PERCENTAGE OF YEAR 5 SURFACES CLASSIFIED AS AFFECTED ON THE CLINICAL EXAM AND/OR ON ONE OR BOTH RADIOGRAPH READINGS

Clinical Call	First and Second Radiograph Calls	Cohort 1+2		Cohort 5	
		Inter	Intra	Inter	Intra
Not affected	Agreed affected	36	61	36	40
	Disagreed	37	21	32	24
		73	82	68	63
Affected	Agreed affected	23	17	29	35
	Disagreed	2	1	1	0
	Agreed not affected	2	0	2	3
		27	18	32	37

ever, radiograph calls were very unreliable in those important instances where they might be used to change a child's DMFS score. These findings called into question the utility of radiograph data and led to a decision not to rely on radiograph results in the analyses of treatment effects.

Chapter 6

DISCUSSION OF RESULTS

The analyses presented in this report were conducted to determine whether clinical examiners and radiograph readers in the National Preventive Dentistry Demonstration Program were more or less reliable than those in other clinical and epidemiological studies, whether inconsistencies in diagnosing Program children may have interfered with the detection of possible treatment effects, and whether reliability was related to various characteristics of the children examined. Three types of data were analyzed: (1) clinical (visual plus tactile) evaluations of the number of decayed, missing due to decay, and filled tooth surfaces (DMFS scores); (2) readings of bitewing radiographs; and (3) assessments of other indicators of oral health, including the degree of gingivitis and the presence of sealants and opacities. The remainder of this section discusses the major findings with each type of data.

CLINICAL EVALUATIONS OF DMFS

Our review of the dental literature revealed that it would be extremely difficult to compare examiner consistency across studies because of important differences in sample characteristics, methods of calculating reliability, and examination procedures and criteria. For example, studies varied in terms of whether examiners knew or did not know which subjects were scheduled to receive concurrent replicate examinations. Most past studies focused on 11- to 12-year-olds (sixth graders) and therefore did not provide information about examiner reliability with samples that may have differed in terms of race, age, decay level, and average DMFS score (see Table 1.1). However, we were able to ascertain that at least with respect to scoring DMFS, the Program's examiners were not substantially more or less reliable than examiners in other studies. Program examiners usually had comparable means, correlations between exams in excess of .95, consistency ratios over 70, and reversal rates of about .11 (see Tables 1.1, 1.2, 3.5, 3.7, and 3.13).

We studied how examiner inconsistencies affected the likelihood of detecting treatment effects by computing the probability of obtaining a difference in average increment scores between two groups when examiner inconsistencies were and were not present. These analyses showed that chance errors or systematic biases in the examination process would not seriously affect the detection of important differences in dental decay rates among the Program's treatment groups.

The analysis of concurrent replicate examination data indicated that examiners were more consistent with themselves than they were with each other (see Tables 3.10 and 3.11). However, maintaining the same examiner/child pairings from Exam 1 to Exam 3 had essentially no effect on reversal rate (see Tables 3.13, 3.14, and 3.15). This finding and other considerations suggest that maintaining pairings would not increase the likelihood of detecting actual treatment effects. Also, by not requiring the maintenance of pairings we reduced the costs and time needed to conduct the examinations.

A child's decay level and age were related to examiner consistency. But, whether higher decay levels were related to higher or lower agreement rates depended on what index of

agreement was used and how it was calculated. Correlation coefficients and consistency ratios were higher with older children than with younger ones, thereby suggesting that reliability increased with the age of the child examined. However, just the opposite conclusion would be drawn from the findings that average absolute difference (AAD) scores and the number of reversals per child were higher with older than with younger children. These discrepancies stem from the presence of a large number of young children who did not have any decay. Table 3.15 shows almost no difference among age groups in the percentage of children having reversals when the samples are limited to children with non-zero DMFS scores.

Interexaminer and intraexaminer consistency rates improved over time regardless of which index of reliability was used. This trend occurred despite older children tending to have higher DMFS scores than younger ones and thereby greater opportunity for disagreements. For instance, interexaminer AAD scores in Cohort 1+2 dropped from .44 in Year 1 to .26 in Year 3, while in Cohort 5 the drop was from .83 to .63.

One explanation of the increased examiner consistency over time is that the examiners became more adept in using the Program's diagnostic protocol. There is certainly ample evidence in the dental literature for such improvement as examiners devote more and more time to calibration. The increased reliability also may be due to the Program's policy of retaining after Year 1 just those examiners who proved to be the most consistent, i.e., had interexaminer AAD scores less than .75 (see Table 3.11).

A third explanation of the increase in reliability from one year to the next is that an examiner's task may have become easier over time as a result of the increased prevalence of sealants and fillings in the study population. An examiner who saw that a surface was sealed called it sound automatically. Because of the increase in the number of children with sealants from one year to the next, this policy may have indirectly increased concurrent agreement rates. It may also have caused children in the sealant regimens to have a slightly higher reversal rate than those not in the regimens (see Table 3.18). Moreover, filled surfaces represented a much larger proportion of the total DMFS score in Year 3 than they did in Year 1. Some groups even showed an average decline in decay levels. This trend may have led to increased consistency because most of the disagreements within and between examiners had to do with whether a surface should be classified as sound or decayed (rather than whether it was filled).

RADIOGRAPHIC EVALUATIONS OF DFS

A superficial analysis of the radiograph data suggested that the readers were highly consistent with themselves and each other. There was 95 percent agreement in the score that should be assigned in Cohort 1+2 and 80 to 85 percent agreement in Cohort 5 (see Table 5.3).

As we noted in our brief review of the radiograph literature, overall agreement rates among or within readers are not especially relevant because radiograph results are usually used to supplement clinical findings. We therefore focused our analysis on those surfaces where the radiograph findings would have affected a combined clinical plus radiograph score. We found that two readers were almost twice as likely to disagree with one another as they were to agree; we obtained this ratio in both Cohort 1+2 and Cohort 5 (see Table 5.4). Previous analyses also indicated that the relationships between DMFS scores and other variables—such as age, race, sex, and socioeconomic status (Bell et al., 1982)—were not changed by adding radiograph findings to clinical scores. Thus, we see no reason to use radiographs in studies of the effects of preventive care with children.

OTHER INDEXES OF ORAL HEALTH

The clinical examiners evaluated four oral conditions in addition to providing the data needed for calculating DMFS scores. These other conditions were: (1) degree of gingivitis, (2) the urgency with which the child needed to see a dentist for possible treatment, (3) whether sealants were present on a child's teeth, and (4) whether there were white spots on the child's teeth that could have been due to fluoride or other factors.

We found that the examiners were generally consistent with themselves but not with each other in measuring gingivitis, apparently because it was difficult to achieve standardization on the Program's gingival index (probably because it did not involve any test for bleeding) and also because examiners were insufficiently calibrated in the use of this index. The fact that some examiners tended to have much higher average scores than others and that these differences were relatively stable over time indicates that gingival scores will have to be adjusted for examiner effects before we can assess whether they were affected by the Program's treatment regimens. And, even with these adjustments, the chances of finding treatment effects with the index are at best marginal because of the reliability problem.

Each child was assigned a referral code on each examination. This code determined which one of four form letters would be sent to his/her parents to give them the general results of their child's clinical examination. The codes ranged from "no problems, only routine care required" to "serious problem, immediate attention needed." We did not intend for this index to be an important dependent variable and accordingly did not devote much attention to calibrating examiners on it. However, the examiners had relatively high agreement rates in using it and there were almost no instances in which a serious disagreement arose. Thus, it may prove to be a gross, but sensitive indicator of treatment effects.

The remaining two indexes, sealants and opacities, also were not intended as major dependent variables and therefore we did not devote as much time to calibrating examiners in their use as we did with the DMFS criteria.

The examiners had no trouble identifying whether a child did or did not have sealants. However, we were unable to determine whether they were any more or less reliable than examiners in other studies because most other studies report examiner consistency in recording sealant retention on a tooth-by-tooth (rather than on a child-by-child) basis.

Opacities was the only index we used that showed a decline in consistency from one exam to the next. In Year 1, for example, interexaminer agreement in Cohorts 1+2 and 5 was 96 percent, but by Year 3 it had dropped to 80 and 82 percent, respectively. We also noticed that scores rose with time but there was no difference in means between Cohorts (see Table 4.6). Although some antifluoridationists might argue that the Program's use of fluoride led to an increased number of opacities, there was no evidence that the children who received fluoride treatments had any more opacities than the children who did not. Thus the most likely explanation is that there was a gradual shift in examination criteria over time.

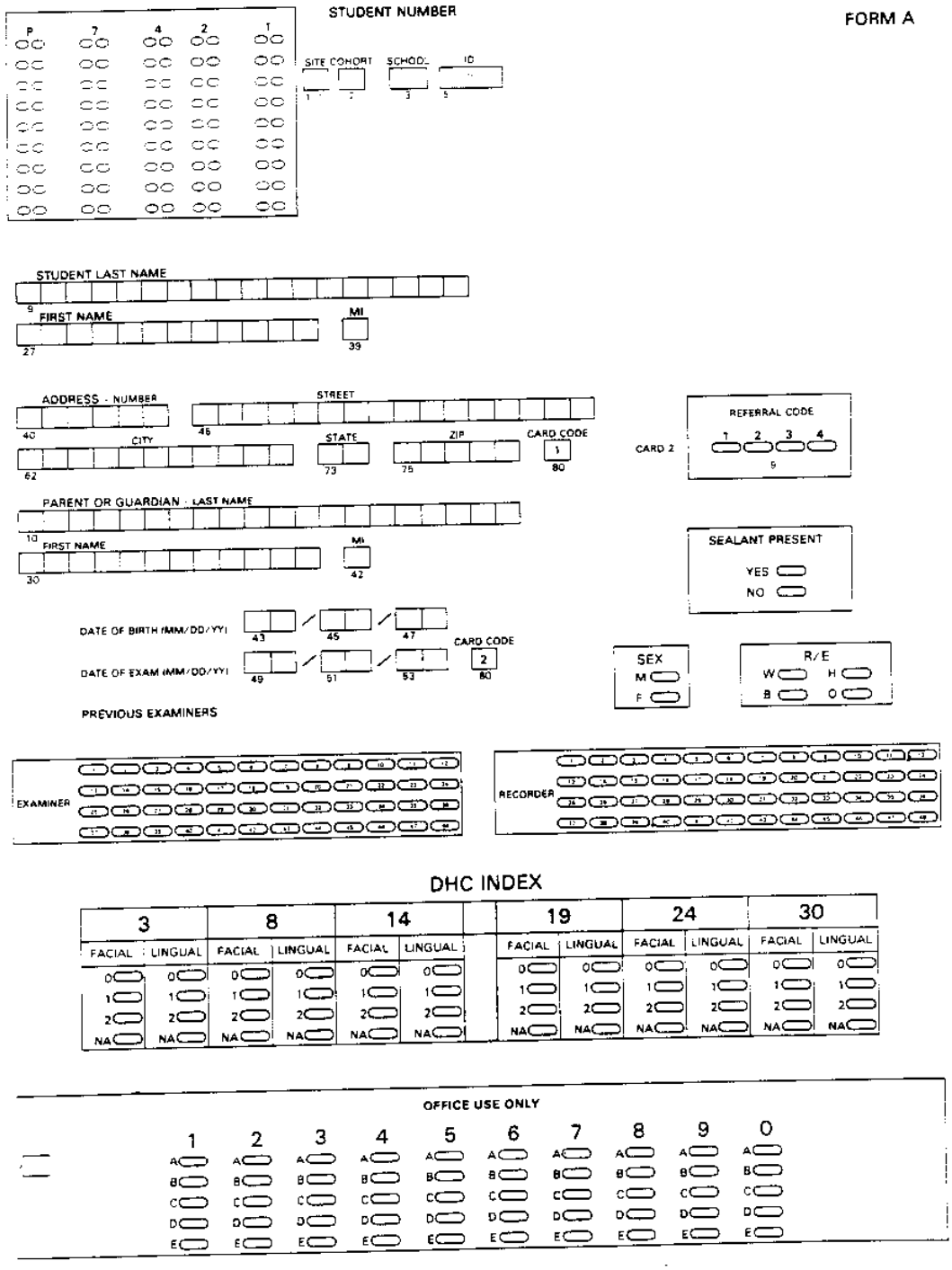
CONCLUSIONS

The results presented in this report indicate that inconsistencies due to clinical examiner errors and systematic biases will not seriously interfere with the detection of practically significant treatment effects on DMFS scores. Adding radiograph results to the clinical findings would certainly not improve the reliability of the total combined score and may even lower it.

Examiner inconsistencies also may interfere with observing treatment effects when gingivitis and opacities are the dependent variables, but referral code may prove to be a more sensitive index than anticipated.

Appendix A

CLINICAL AND RADIOGRAPH EXAMINATION FORMS



DMFS - VISUAL

UPPER LEFT	UPPER RIGHT		LOWER LEFT	LOWER RIGHT
S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	1 CENTRAL	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis
S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	2 LATERAL	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis
S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	3 CUSPID	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis
S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	4 1st BICUSPID	S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis
S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	5 2nd BICUSPID	S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ D O 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis
S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	6 1st MOLAR	S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis
S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	7 2nd MOLAR	S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis	S O X O 5 O occ 0 O 6 O lin U O 1 O 7 O buc E O 2 O 8 O mes Y O 3 O 9 O dis

FORM B

T	2	4	7	P

STUDENT NUMBER

LAST NAME

FIRST NAME MI

DATE OF BIRTH (MM/DD/YY) / /

TECHNICIAN

READER

DMF SURFACES INDEX
Radiograph Examination

UPPER LEFT	UPPER RIGHT		LOWER LEFT	LOWER RIGHT
S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis	1st BICUSPID	S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis
S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis	2nd BICUSPID	S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis
S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis	1st MOLAR	S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis
S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis	2nd MOLAR	S U/M 2 B mes 3 9 dis	S U/M 2 B mes 3 9 dis

MRC FORM NO. 5713

Appendix B

DIAGNOSTIC CRITERIA FOR DMF SCORES[1]

The diagnostic criteria for dental caries used in the Program are those adopted by the Caries Measurement Task Group at their 1968 Conference on Clinical Testing of Cariostatic Agents (Radike, 1972). These criteria are used by the National Institute of Dental Research in its National Caries Program, as well as in numerous other studies.

GENERAL CRITERIA

Carious lesions may be subdivided into frank and borderline lesions. Frank lesions present few problems in diagnosis; they are detected on the basis of gross cavitation (where cavitation is defined as a usually irregular discontinuity of the enamel's surface). The cavitation produced by dental caries must be distinguished from fractures of the teeth and the smooth defects produced by erosion and abrasion.

Borderline lesions lack easily discernible, frank cavitation. The criteria for determining a surface is carious depend upon surface type. For pit and fissures on occlusal, vertibular, and lingual surfaces, the area is carious when the explorer "catches" after insertion with moderate to firm pressure and when the "catch" is accompanied by one or more of the following signs of decay: (1) softness at the base of the area, (2) opacity adjacent to the area provides evidence of undermining or demineralization, and (3) softened enamel adjacent to the area which may be scraped away with the explorer.

Smooth areas on facial or lingual surfaces are carious if they are etched or if there is a white spot as evidence of subsurface demineralization and if the area is found to be soft by penetration or scraping away the enamel with the explorer. These areas are diagnosed as sound when there is apparent evidence of demineralization, but no evidence of softness.

The smooth surface criteria are also used for proximal surfaces exposed to direct visual and tactile examination. For proximal areas not available to direct visual-tactile examination, any of the following criteria may serve as an indication of decay.

- Visual—If a marginal ridge shows an opacity as evidence of undermined enamel, the proximal surface beneath that marginal ridge is carious.
- Tactile—A discontinuity of the enamel in which an explorer will "catch" is carious, if there is softness.

No lesion is to be diagnosed as carious unless it is demonstrably carious beyond the possibility of a doubt.

---

[1]From A. D. Radike, *Proceedings of the Conference on the Clinical Testing of Cariogenic Agents*, as reported by H. Horowitz, *J. Pub. Health Dent.*, Vol. 32, No. 4, Fall 1972.

SPECIFIC CRITERIA

In recording dental caries, a deep pit or fissure will not be recorded as carious unless there is more evidence than a catch of the explorer pressure, visual evidence of backward decay of the enamel, or visual evidence of deterioration of the pit or fissure wall. "Catches" will be tested by rocking the explorer after it enters the pit or fissure. Deep pits or fissures per se are not to be considered carious or questionably carious. There must be an obvious break of surface continuity demonstrable; for example, visual evidence alone of undermining under a marginal ridge is acceptable evidence of a proximal lesion only when the examiner is certain that caries is, in fact, present. In such cases, an effort should be made to confirm the lesion by entering with an explorer.

Do not hesitate; record first impressions. If a lesion is doubtful, record the surface as caries free. A proximal cavity is a one-surface lesion if the occlusal is intact (even though its restoration will eventually require an occlusal extension).

The following general rules also apply:

1. Stain or pigmentation alone should not be regarded as evidence of decay inasmuch as either can occur on sound teeth.
2. A decision between soundness or decay of a surface must be made for borderline conditions; therefore, an intermediate category of "questionable" should not be used because its overuse for difficult decisions can become a convenience, which reduces sensitivity or the ability to detect a real effect.
3. Each student should be examined in the same manner. An examiner, for example, should avoid the temptation to examine one child more thoroughly who appears highly susceptible to dental caries and another child less thoroughly who is relatively free of apparent decay.

The following rules apply to specific conditions in all children:

- Only roots remaining—Call all surfaces carious.
- Full crown restoration—Call all surfaces filled. When a tooth has been crowned for obvious reasons of trauma, call it (Y).
- Erupted teeth—A tooth is considered to be in eruption when any part of its crown projects through the gum.
- Supernumerary teeth—Only one tooth is to be called. It is left to the examiner's judgment to decide which tooth is the "legitimate" occupant of the tooth space.
- "Split cells"—Where both a deciduous and a permanent tooth occupy the same tooth space, the condition of the permanent tooth is to be called.

The following rules apply when orthodontic bands are present:

- First- and second-grade children with two or more permanent teeth banded will not receive full examination.
- Fifth-grade children with six or more permanent teeth banded will not receive the full examination.
- Any child in a cross-sectional sample with orthodontic bands will not be examined.

Children who do not receive a complete examination because of orthodontic bands will be assigned a referral code of 4.

EXAMINER CALLS AND PROCEDURES

For each tooth, the examiner will either make one call to describe the overall status of the tooth as follows:

S = Sound permanent
D = Deciduous (regardless of caries status)
U = Unerupted permanent
E = Extracted due to caries
Y = Extracted or crowned for noncarious reasons

or make one or more calls to indicate carious or filled status on each surface as follows:

Surface	Carious Call	Filled Call
Occlusal	x	5
Lingual	0	6
Buccal	1	7
Mesial	2	8
Distal	3	9

If a given tooth surface is both carious and filled, then both calls should be made (e.g., "X and 5").

Although the child will have brushed before the examination, the mouth may be rinsed to remove loose material. In any event, dry the teeth with compressed air before examining for caries.

The teeth are to be examined in the order in which they appear on the examination form.

Appendix C

CRITERIA FOR RADIOGRAPH EXAMINATIONS

GENERAL DESCRIPTION OF FILMS

The first judgment the reader will render will be a general call stating the possibility of evaluating the proximal surfaces of erupted permanent posterior teeth. One of the following three calls must be made for each film set:

All unerupted—There are no permanent posterior teeth erupted. Any permanent posterior teeth visible on the film will be in a stage of development within the jawbone and/or gingivae.

All non-readable—Reader will note why the films are not readable, such as: no films exposed, films lost, films bent or contain other artifacts making them both completely unreadable, films under exposed or overexposed, there was motion when the films were exposed, a diagnostic image is not present due to a cone-cut or some other beam/film malalignment—i.e., overlapping contact areas (any degree of proximal overlap must result in a recording of NR unless a carious lesion is evident), or a combination of any of the above.

Readable—This box will be checked when the radiographic films are of acceptable diagnostic quality and one or more permanent teeth may be evaluated. The contact areas between the proximal surfaces must be open or without overlap, and the image and landmarks well defined, reflecting adequate exposure and development.

READABLE FILMS

The sequence for reading will be similar to that used for the clinical examination; start with the upper left first bicuspid, and proceed sequentially through the upper left second molar. Then call the lower left, and finish with the lower right bicuspids and molars. The reader will evaluate only the mesial and distal proximal surfaces. When appropriate, one of the following general codes will be used.

SOUND—Both the mesial and distal proximal surfaces are without restorations or dental caries. The proximal surfaces are visible below the contact area of greatest convexity.

UNERUPTED—The permanent tooth has not erupted above the gingival line. There is evidence of the developing tooth on the radiograph.

MISSING—The permanent tooth is missing; that is, there is sufficient information on the film to deduce that the tooth is missing and not simply unerupted. For example: the first permanent molar has tipped mesially into the second bicuspid. Or, there is a radiolucent image of a socket due to a recent extraction. The call "missing" includes a tooth missing due to caries, orthodontic extraction, or partial anodontia.

If a call of "SOUND," "UNERUPTED," OR "MISSING" is made, then no other calls will be made for that tooth. Proceed to the next tooth. If the tooth is not sound, unerupted, or missing, then one or more of the following calls will be made to describe it and the information recorded in the middle and bottom rows of ovals:

NR—NOT READABLE—Either the mesial or the distal surface or both are not readable for any or all of the reasons described previously.

2 & 3—CARIES—The figure 2 is the code for mesial caries, and the figure 3 is the code for distal caries.

Any definite interproximal radiolucency indicating a break in the continuity of the enamel surface is to be regarded as caries and recorded as either 2 or 3. To be recorded as caries, the radiolucency does not have to extend midway or to the DEJ. In the event of some proximal overlap, proximal caries may be recorded only when the radiolucency extends beyond the overlap or if the reader is certain beyond any doubt that caries is present.

The figure 8 is the code for a mesial proximal restoration and 9 is the code for a distal proximal restoration. When a metal alloy is used as the restorative material, interpretation is less difficult than when a composite resin is used. A radiopaque base may aid accurate determination and recording.

Appendix D

ESTIMATION OF INDIVIDUAL EXAMINER DIFFERENCES

Table 3.5 displays estimated average observed differences (AOD) and Table 3.11 shows average absolute differences (AAD) for the individual examiners in the Program. This appendix describes the analysis method used to derive the results in those tables. The final part of the appendix indicates the modifications made for analysis of individual differences in assessing referral code and the DHC gingivitis index.

There were two main purposes for this segment of the reliability study. First, we wished to discover whether certain examiners were significantly better or worse in agreeing with themselves and each other, as measured by AAD. This served as the main criterion for selection of examiners in subsequent examination cycles. Note that AAD measures both chance and systematic differences. Second, we searched separately for the presence of systematic bias, using the AOD measure. This allowed assessment of the degree to which systematic differences among examiners in their mean levels caused differences measured by the AAD and other measures. This assessment indicated whether it would help to adjust for examiners in the study of treatment effects.

Problems Addressed by the DMFS Analysis

Accurately assessing individual examiner differences faced several obstacles requiring special treatment. All these problems involved the potential for attributing to the examiners differences caused by other factors. The analysis described in this appendix accounted for the following problems:

- The number of potential disagreements depended on characteristics of the particular child, specifically the child's dental condition. For children with few erupted permanent teeth or no unfilled decay, few disagreements generally occurred. In contrast, certain other children with much decay, and usually many borderline lesions, represented difficult cases. Because some examiners worked at only one or two sites, that might have both been above (or below) average in caries prevalence, it is important to account for these factors.
- For interexaminer pairings, the frequency and type of disagreements that an examiner had could have been affected by which other examiner saw the same child.
- Because examiners agree perfectly on the large majority of children, the number of disagreements has a very long-tailed distribution relative to its mean. Thus, for an examiner who saw a limited number of children, one large discrepancy in the absolute difference could make the examiner look unreliable.

Major Disagreements in DMFS Scores

The analysis began by identifying and modifying major disagreements between examiners in the DMFS score. For these purposes, we defined a major disagreement as one that exceeded

five surfaces. Because major disagreements seemed to occur randomly with respect to the examiners involved, we limited the influence of those cases in the individual examiner ratings.

Of over 8500 pairs of replicate concurrent examinations in the Program, 35 pairs had disagreements that exceeded five surfaces in the DMFS score. Two anticipated patterns emerged: 69 percent of these cases occurred in Cohort 5 (despite the fact that only one-third of the pairs came from that Cohort) and 85 percent occurred on interexaminer pairs (counting only Exams 1 to 3, those with both type of examinations). In contrast, there was no evidence of a relationship with the examiner, i.e., no examiners were involved in an unexpectedly large or small number of major disagreements.

Based on the above evidence and the fact that many examiners saw only a few of the children in the reliability sample each year, we chose to study individual examiner reliability using a method that was robust to major disagreements. Any differences that exceeded five surfaces in absolute value were reduced to five surfaces. These modified differences were used in the analyses of both observed differences (ODs) and absolute differences (ADs). Because small sample sizes posed a major problem only in analyses of individual examiner differences, the modified differences were not used for any other tables in this report.

Adjustment for Child Characteristics and Examiner Pairings

As a first step, we developed a linear prediction equation for the modified AD in DMFS as a function of child characteristics and the average D, M, and F components from the two concurrent examinations. Weighted least squares was used to fit four separate equations, one for each combination of Cohort and pairing type (inter vs. intra). These regression models explained from 40 to 50 percent of the variance in interexaminer ADs and 20 to 30 percent of the variance in intraexaminer ADs. After controlling for the average D, M, and F components, no other variables (site, age, race, etc.) were statistically significant. In each case, the average D component accounted for at least 85 percent of the variance that was explained. Put roughly, the AAD increased linearly with the square root of the D component. Any other variance explained by the model was usually attributable to the M component.

To adjust for the difficulty of intraexaminer cases seen by an examiner, we computed the ratio of the actual AD to the predicted AD for each child.[1] Besides adjusting for case difficulty, this ratio provides a stable dependent variable with homogeneous variance. That is, the ratio had about the same distribution for both easy and difficult to examine children. For an examiner of average reliability, the mean value of this ratio should have been near 1, independent of the type of children the examiner saw. If the mean ratio was significantly above or below 1, it would indicate that the examiner was less or more consistent with himself than other examiners were. One-way analyses of variance were used to indicate whether the mean ratios differed significantly from each other. Table 3.11 reports scores that have been retransformed to the raw scale. For example, if an examiner's mean ratio was 20 percent higher than average for all the examiners in a given year, then his/her reported AAD would be 1.20 times the AAD for the overall AAD that year (combined across Cohorts).

For interexaminer AD scores, we also adjusted a given examiner's scores for the other examiners with whom that examiner had been paired. A ratio score was computed exactly as for intraexaminer ADs; however, a least squares regression estimated each examiner's contribu-

[1]The actual divisor was the maximum of 0.15 and (0.10 + the predicted AD for the child). This divisor was used to limit the impact of minor disagreements on children with very small predicted ADs.

tion to the AD ratio. In this way, no examiner's AAD score was affected by being paired more or less often with examiners who exhibited below average reliability.

For observed differences in DMFS (ODs), we computed a ratio of the OD to the predicted AD for the child. As with AD scores, the ratio improves stability and homogeneity of variance. Here a positive mean ratio indicates that the examiner tended to assign higher scores than the average examiner did. Again, regression analysis adjusted for the examiner pairings.

Analysis of individual examiner differences in the assignment of referral codes and DHC scores paralleled the analysis for DMFS scores. The exception is that all analyses for these variables used raw differences rather than modified differences or ratios.

Appendix E

PROCEDURE FOR ESTIMATING POWER TO DETECT TREATMENT EFFECTS

Table 3.19 shows the degree to which the elimination of examiner inconsistencies would improve the ability of a study like this Program to detect treatment effects of various sizes. This appendix describes the calculations used to create that table.

Formula for Power of a Test

The power of a statistical significance test is the probability of rejecting the hypothesis of no effect (or equal response in each group) given that a certain sized effect actually does exist. Obviously, when undertaking a large study, one would like a high probability of finding some effect when the true effect has practical significance.

The power of a test depends on the nature of both the test and the experimental design. In the Program, the test for treatment effects is essentially a two-sample t-test comparing a specific treatment regimen to the control regimen, using the DMFS increment as the dependent variable.[1] Because any effects are expected to be beneficial, we compute the power for a one-tailed test. The power of this test depends on these factors:

- α = the significance level of the test; i.e., the probability of finding an effect when none exists,
- c = the size of the true effect (0.5, 1.0, 1.5),
- σ^2 = the variance of the dependent variable, and
- n_1, n_2 = the samples sizes in the two groups (200 in each).

The formula for power is

$$\text{Power} = 1 - \Phi(x), \tag{1}$$

where

$$x = z_\alpha - \frac{c}{[\sigma(1/n_1 + 1/n_2)^{1/2}]}, \tag{2}$$

and for a given value of x, $\Phi(x)$ is the probability that a standard normal random variable (zero mean and unit variance) is less than x. The constant z_α is the critical value for the t-test; i.e., there is an α probability that a standard normal exceeds z_α.

[1]Although two-year DMFS increments, the dependent variable for Table 3.18, have a very skewed distribution, the mean of a large number of increments has a distribution not too dissimilar from that of a normal random variable. Thus the test should have the correct significance level, and the power formula should provide a good approximation to the truth.

Variance of DMFS Increments With and Without Examiner Error

For comparison of power with and without examiner inconsistency, any difference is due to a difference in the variance of the dependent variable because the other factors remain constant.[2] The variance of DMFS increments in the presence of examiner inconsistency was computed directly from the observed increment scores. We used the following relationships to compute the variance without examiner inconsistency: Let T_t be a child's true DMFS score at Time t (without examiner inconsistency), for $t = 1$ and 2; and let X_{ti} be the child's observed score on his i^{th} examination at Time t, for $i = 1$ and 2. We can write

$$X_{ti} = T_t + e_{ti},$$

where e_{t1} and e_{t2} are independent and identically distributed error terms. From the equation above, we have that the expected value of the squared difference in concurrent scores is

$$E(X_{t1} - X_{t2}) = Var(e_{t1} - e_{t2}) = 2Var(e_{t1}).$$

This relationship implies that one-half of the average squared difference (ASD) in concurrent scores is an unbiased estimate of the examiner variance in that year. Also note that

$$Var(X_{21} - X_{11}) = Var(T_{21} - T_{11} + e_{21} - e_{11})$$

$$= Var(T_{21} - T_{11}) + Var(e_{21}) + Var(e_{11})$$

$$= Var(T_{21} - T_{11}) + (1/2)[E(X_{t1} - X_{t2})^2 + E(X_{t1} - X_{t2})^2].$$

Thus we estimated the variance without examiner inconsistency by the variance in observed increments minus one-half the sum of the ASDs at Times 1 and 2. For two-year increments from Year 1 to Year 3 in Cohort 5, this yielded

$$\sigma^2 = 9.45 - (1/2)[2.95 + 3.33] = 6.31$$

as an unbiased estimate of the variance without examiner error.

Numerical Example

The following computations indicate the steps leading to the first entry of Table 3.19 for Cohort 5. For that case, we used $c = 0.5$, $n_1 = n_2 = 200$, and $\alpha = 0.05$; thus $z_\alpha = 1.645$. For two-year DMFS increments measured with examiner error, σ was $(9.45)^{1/2} = 3.07$. Substituting these values in equation (2) yields $x = [1.645] - [0.5/(3.07)(0.1)] = 0.02$. A table of the standard normal distribution gives that Φ (0.02) is 0.51, so that the power is $1.00 - 0.51 = 0.49$. In the absence of examiner error, $\sigma^2 = 6.31$ implies that $x = -0.35$ and $\Phi(x) = 0.36$. Thus, the power is $1 - 0.36 = 0.64$.

---

[2]In theory, one could argue that examiner inconsistency reduces the expected difference between groups because errors tend to raise true low scores and lower true high scores (see Poole et al., 1973). However, there is no empirical evidence to substantiate this hypothesis or to estimate the size of the reduction.

Effects of Factors Other than Variance on Power

Several assumptions affected the magnitudes, but not the relative differences, of the numbers in Table 3.19. For example, all the powers would drop if two-tailed tests or lower significance levels were used. On the other hand, the covariates predict about 10 to 20 percent of the variance in DMFS increments, thereby reducing σ^2 in (2) by the same percentage. This would increase powers. Also, actual Program sample sizes uniformly exceeded 200 per group. Other likely complicating factors, such as unequal variances in the different groups, have been ignored. However, none of these factors would substantially change the relative effect of examiner inconsistency on the ability to detect treatment differences.

To illustrate the point that the relative importance of examiner inconsistency is insensitive to the above assumptions, Table E.1 compares powers for one- and two-tailed tests. For $c = 0.5$, the drop in power due to examiner inconsistency is 0.15 for a one-tailed test and 0.14 for a two-tailed test.

Table E.1

POWER TO DETECT TWO-YEAR TREATMENT EFFECTS IN COHORT 5 WITH ONE- AND TWO-TAILED TESTS WHEN VARIANCE DUE TO EXAMINER INCONSISTENCIES IS AND IS NOT PRESENT

Test	Examiner Error	Size of Difference To Be Detected 0.5	1.0	1.5
One-tailed	Present	.49	.95	1.00
	Absent	.64	.99	1.00
Two-tailed	Present	.37	.90	1.00
	Absent	.51	.98	1.00

NOTE: Probabilities are for a two-tailed test at alpha = .05 with 200 children per group.

REFERENCES

Alexander, A. G., A. R. Leon, J. W. Ribbons, and S. I. Morganstein, "An Assessment of the Inter- and Intraexaminer Agreement in Scoring Gingivitis Clinically," *J. Periodont. Res.*, Vol. 6, 1971, pp. 146–151.

Andlaw, R. J., and G. J. Tucker, "A Dentifrice Containing 0.8 per cent Sodium Monofluorophosphate in an Aluminium Oxide Trihydrate Base," *Brit. Dent. J.*, Vol. 138, 1975, pp. 426–432.

Andlaw, R. J., C. K. Burchell, and G. J. Tucker, "Comparison of Dental Health of 11-Year-Old Children in 1970 and 1979, and of 14-Year-Old Children in 1973 and 1979: Studies in Bristol, England," *Caries Res.*, Vol. 16, No. 3, 1982, pp. 257–264.

Axelsson, P., and J. Lindhe, "The Effect of a Preventive Programme On Dental Plaque, Gingivitis and Caries in School Children. Results After One and Two Years," *Journal of Clinical Periodontology*, 1, 1974, pp. 126–128.

Beiswanger, B. B., C. W. Gish, and M. E. Mallatt, "A Three-Year Study of the Effect of a Sodium Fluoride-Silica Abrasive Dentifrice on Dental Caries," *Pharmacology and Therapeutics in Dentistry*, Vol. 6, 1981, pp. 9–16.

Bell, R. M., S. P. Klein, H. M. Bohannan, et al., *Treatment Effects in the National Preventive Dentistry Demonstration Program*, The Rand Corporation, R-3072-RWJ, February 1984.

Bell, R. M., S. P. Klein, H. M. Bohannan, R. C. Graves, and J. A. Disney, *The Results of Baseline Dental Examinations in the National Preventive Dentistry Demonstration Program*, The Rand Corporation, R-2862-RWJ, April 1982.

Bille, J., "Dental Caries in a Group of 20-Year-Olds After Previous Participation in Public Child Dental Health Services in Copenhagen, Denmark," *Community Dent. and Oral Epidemiol.*, Vol. 8, No. 1, 1980, pp. 27–32.

Birkeland, J. M., and L. Jorkjend, "The Influence of Examination on the Assessment of the Intraexaminer Error by Using the Plaque and Gingival Index Systems," *Community Dent. and Oral Epidemiol.*, Vol. 3, No. 5, September 1975, pp. 214–216.

Birkeland, J. M., L. Jorkjend, and L. Broch, "An Approach to Objective Assessment of Caries Prophylactic Measures in a Dental Health Program," *Community Dent. and Oral Epidemiol.*, Vol. 3, No. 3, May 1975, pp. 126–131.

Bohannan, H. M., J. A. Disney, S. P. Klein, et al., *Examination Procedures in the National Preventive Dentistry Demonstration Program*, The Rand Corporation, N-2129-RWJ, 1984.

Burt, B. A., D. M. Roder, J. C. Cecil, and S. A. Eklund, "Saliva-based Colorimetric Test as an Index of Gingival Inflammation in Epidemiologic Studies," *Community Dent. and Oral Epidemiol.*, Vol. 6, No. 6, November 1978, pp. 290–295.

Burt, B. A., D. S. Berman, and L. M. Silverstone, "Sealant Retention and Effects on Occlusal Caries After Two Years in a Public Program," *Community Dent. and Oral Epidemiol.*, Vol. 5, No. 1, January 1977, pp. 15–21.

Burt, B. A., S. A. Eklund, J. R. Landis, F. A. Larkin, K. E. Guire, and F. E. Thompson, "Diet and Dental Health, A Study of Relationships: United States, 1971–1974," *National Health Survey, 1971–1974, USDHEW*, Series 11, No. 225, U.S. Government Printing Office, Washington, D.C., January 1982.

Clark, D. C., M. Mechanic, and D. W. Banting, "A Further Look at the Role of Radiographs in Trials of Caries-Prophylactic Agents," *J. Pub. Health Dent.*, Vol. 42, No. 3, 1982, pp. 237–250.

Craig, G. G., K. R. Powell, and M. H. Cooper, "Caries Progression in Primary Molars: 24-Month Results from a Minimal Treatment Programme," *Community Dent. and Oral Epidemiol.*, Vol. 9, No. 6, 1981, pp. 260–265.

Crowley, R. E., "A Method for Evaluating Consistency in Diagnosis of Gingivitis," *Oral Surgery*, Vol. 8, 1955, pp. 1128–1133.

Cutress, T. W., P.B.V. Hunter, D. J. Beck, and P. de Souza, "A Comparison of WHO Periodontal Status Index with the Periodontal and Oral Hygiene Indices," *Community Dent. and Oral Epidemiol.*, Vol. 6, No. 5, September 1978, pp. 245–252.

Davies, G. N., and P. B. Cadell, "Four Investigations to Determine the Reliability of Caries-Recording Methods," *Arch. Oral Biol.*, Vol. 8, 1963, pp. 331–348.

Davies, G. N., B. J. Kruger, and B. T. Homan, "An Epidemiological Training Course," *Australian Dental Journal*, Vol. 12, February 1967, pp. 17–28.

Davies, G. N., H. S. Horowitz, and W. Wada, "The Assessment of Dental Caries for Public Health Purposes," *Community Dent. and Oral Epidemiol.*, Vol. 1, 1973, pp. 68–73.

Davies, G. N., H. S. Horowitz, and W. Wada, "The Assessment of Periodontal Disease," *J. Periodont. Res.*, Vol. 9, 1974, pp. 62–70.

De Paola, P. F., and J. Alman, "Assessment of the Reliability of Radiographic Diagnosis in a Clinical Caries Trial," *J. Dent. Res.*, Vol. 51, No. 5, 1972, pp. 1431–1437.

Eklund, S., personal communication, 1981.

Elliott, J. R., and B. A. Clemmer, "On the Reproducibility of Periodontal Indices," *U.S. Navy Med.*, Vol. 59, 1972, pp. 41–43.

Fischman, S., A. Picozzi, D. Juliano, M. Slakter, and J. English, "Examiner Standardization for Caries Studies," *J. Dent. Res.*, Vol. 55, No. 6, 1976, pp. 926–929.

Fleiss, J. L., M. J. Slakter, S. L. Fischman, M. H. Park, and N. W. Chilton, "Interexaminer Reliability in Caries Trials," *J. Dent. Res.*, Vol. 58, 1979, pp. 604–609.

Foch, C. B., *A Cost Analysis Plan for the National Preventive Dentistry Demonstration Program*, The Rand Corporation, N-1732-RWJ, 1981a.

Foch, C. B., *The Costs and Benefits of Preventive Dental Care: A Literature Review*, The Rand Corporation, N-1670-RWJ, 1981b.

Foch, C. B., S. P. Klein, H. M. Bohannan, et al., *Cost of Treatment Procedures in the National Preventive Dentistry Demonstration Program*, The Rand Corporation, R-3034-RWJ, February 1984.

Fogels, H. R., J. E. Alman, J. J. Meade, and J. P. O'Donnell, "The Relative Caries-Inhibiting Effects of a Stannous Fluoride Dentifrice in a Silica Gel Base," *Journal of the American Dental Association*, Vol. 99, September 1979, pp. 456–459.

Gibson, A., S. Gelbier, and S. Bhatia, "Dental Health and Treatment Needs of 5-Year-Old Children in the Health Area of Lambeth, Southwark and Lewisham, England," *Community Dent. and Oral Epidemiol.*, Vol. 9, No. 1, 1981, pp. 5–9.

Gittleman Film Associates, Inc., "To Protect a Child," William E. Walch, Executive Producer, The Robert Wood Johnson Foundation, Princeton, New Jersey, March 1979.

Hamp, S. E., J. Lindhe, J. Fornell, L. A. Johansson, and R. Karlsson, "Effect of a Field Program Based on Systematic Plaque Control on Caries and Gingivitis in School Children After Three Years," *Community Dent. and Oral Epidemiol.*, 6, 1978, pp. 17–23.

Handjelman, S. L., D. H. Leverett, E. S. Solomon, and C. M. Brenner, "Use of Adhesive Sealants over Occlusal Carious Lesions," *Community Dent. and Oral Epidemiol.*, Vol. 9, No. 6, 1981, pp. 256–259.

Haugejorden, O., "The Construction and Use of Diagnostic Standards for Primary Radiographic Caries," *Acta Odont. Scand.*, Vol. 34, 1976, pp. 69–77.

Haugejorden, O., "The Construction and Use of Diagnostic Standards for Radiographic Caries Incidence Scores," *Acta Odont. Scand.*, Vol. 35, 1977, pp. 95–102.

Hazen, S. P., "Indices for the Measurement of Gingival Inflammation in Clinical Studies of Oral Hygiene and Periodontal Disease," *J. Periodont. Res.*, Vol. 9, Suppl. 14, 1974, pp. 61–69.

Hetland, L., N. Midtun, and T. Kristoffersen, "Effect of Oral Hygiene Instructions Given by Paraprofessional Personnel," *Community Dent. and Oral Epidemiol.*, 10, 1982, pp. 8–14.

Holst, D., "A Comparative Radiographic Investigation of Caries Prevalence in Danish Schoolchildren," *Community Dent. and Oral Epidemiol.*, Vol. 4, No. 6, November 1976, pp. 254–258.

Horowitz, H. S., "Inter- and Intraexaminer Variability," in *Proceedings of the Conference on the Clinical Testing of Cariostatic Agents,* held by the American Dental Association, October 14–16, 1968, published by American Dental Association, Chicago, Illinois, 1972.

Horowitz, H. S., J. L. Baume, et al., "Principal Requirements For Controlled Clinical Trials of Caries Preventive Agents and Procedures," *Internat. Dent. J.*, 23, 1973, pp. 506–516.

Horowitz, H. S., F. E. Law, M. B. Thompson, and S. R. Chamberlin, "Evaluation of a Stannous Fluoride Dentifrice for Use in Dental Public Health Programs: I. Basic Findings," *Journal of the American Dental Association,* Vol. 72, 1966, pp. 408–422.

Horowitz, H. S., and J. K. Peterson, "Evaluation of Examiner Variability and the Use of Radiographs in Determining the Efficacy of Community Fluoridation," *Arch. Oral Biol.*, Vol. 11, 1966, pp. 867–875.

Horowitz, H. S., and S. Poulsen, "Comparison of Blind and Non-blind Assessments of Occlusal Sealant Retention," *Community Dent. and Oral Epidemiol.*, Vol. 6, No. 1, January 1978, pp. 24–26.

Howat, A. P., and S. J. Cannell, "Value of Operating Auxiliary Personnel for Caries Diagnosis in Dental Epidemiology," *Community Dent. and Oral Epidemiol.*, Vol. 7, No. 3, June 1979, pp. 165–169.

Jackson, D., "Errors in Clinical Trials," *Advances in Fluorine Research and Dental Caries Prevention,* Vol. 4, 1966, pp. 23–32.

Kamen, A., and J. Schmee, "Diagnostic Errors and Multiple Examiners in Anticariogenic Studies," *J. Dent. Res.*, Vol. 53, No. 5, 1974, p. 1500.

Kelly, J. E., and C. R. Harvey, "Decayed, Missing, and Filled Teeth Among Youths 12–17 Years," *National Health Survey, 1966-1970, USDHEW,* Series 11, No. 11, U.S. Government Printing Office, Washington, D.C., October 1974, pp. 28–30.

Kelly, J. E., and J. V. Scanlon, "Decayed, Missing, and Filled Teeth Among Children," *National Health Survey, 1963-1965, USDHEW,* Series 11, No. 106, U.S. Government Printing Office, Washington, D.C., August 1971, pp. 35–37.

Kjaerheim, V., F. R. von der Fehr, and S. Poulsen, "Two-Year Study on the Effect of Professional Tooth Cleaning on the School Children in Oppegard, Norway, *Community Dent. and Oral Epidemiol.*, 8, 1980, pp. 401–406.

Klein, S. P., F. H. Leone, J. N. Grabowski, et al., *Teacher and Principal Survey Results in the National Preventive Dentistry Demonstration Program,* The Rand Corporation, N-1921-RWJ, 1984.

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

Klein, S. P., and R. M. Bell, *Plan for the Analysis of Dental Examination Data in the National Preventive Dentistry Demonstration Program,* The Rand Corporation, N-1658-RWJ, 1981.

Leijon, G., and K. E. Marken, "Roentgenological Diagnosis of Proximal Caries. Deviations Between Observers and Comparison Between the Recordings from Periapical and Bite-Wing Roentgenograms," *Acta Odont. Scand.,* Vol. 26, May 1968, pp. 35–61.

Leone, F. H., S. P. Klein, and H. M. Bohannan, *A Profile of Non-Participants in a School Based Preventive Dentistry Program,* The Rand Corporation, P-6677, 1981.

Lopes, E. S., H. N. Filho, J. Chiodi, and O. Tavano, "Interexaminer Variability when Using Viegas' Simplified Caries Index," *J. Pub. Health Dent.,* Vol. 33, No. 1, Winter 1973, pp. 7–12.

Markkanen, H., M. Rajala, and I. Paunio, "Evaluation of Periodontal Status in a Finnish Industrial Population," *Community Dent. and Oral Epidemiol.,* Vol. 7, No. 2, 1979, pp. 96–100.

Marthaler, T. M., and M. Germann, "Radiographic and Visual Appearance of Small Smooth Surface Caries Lesions Studied on Extracted Teeth," *Caries Res.,* Vol. 4, 1970, pp. 224–242.

Mileman, P., D. Purdell-Lewis, and L. van der Weele, "Variation in Radiographic Caries Diagnosis and Treatment Decisions Among University Teachers," *Community Dent. and Oral Epidemiol.,* Vol. 10, No. 6, December 1982, pp. 329–334.

Miller, A. J., J. A. Brunelle, J. P. Carlos, and D. B. Scott, "The Prevalence of Dental Caries in United States Children," *USDHHS,* NIH Publ. No. 82-2245, U.S. Government Printing Office, Washington, D.C., December 1981.

Murray, J. J., and L. Shaw, "Errors in Diagnosis of Approximal Caries on Bitewing Radiographs," *Community Dent. and Oral Epidemiol.,* Vol. 3, No. 6, November 1975, pp. 276–282.

Olsson, B., "Dental Findings in High-Fluoride Areas in Ethiopia," *Community Dent. and Oral Epidemiol.,* 7, 1979, pp. 51–56.

Palin, T., H. Hausen, L. Alvesalo, and O. Heininen, "Dental Health of Mentally Retarded Children in Eastern Finland," *Community Dent. and Oral Epidemiol.,* Vol. 10, No. 2, April 1982, pp. 86–90.

Poole, W. K., C. A. Clayton, and B. V. Shah, "The Estimation of Examiner Error and the True Transition Probabilities for Teeth or Surfaces in Dental Clinical Trials," *Arch. Oral Biol.,* Vol. 18, No. 10, 1973, pp. 1291–1302.

Poulsen, S., J. Bille, and A. J. Rugg-Gunn, "Evaluation of a Calibration Trial to Increase Interexaminer Reliability of Radiographic Diagnosis of Approximal Carious Lesions," *Community Dent. and Oral Epidemiol.,* Vol. 8, No. 3, June 1980, pp. 135–138.

Radike, A. W., "Examiner Error and Reversals in Diagnosis," in *Proceedings of the Conference on the Clinical Testing of Cariostatic Agents,* held by the American Dental Association, October 14–16, 1968. Published by the American Dental Association, Chicago, Illinois, 1972.

Radike, A. W., C. W. Gish, J. K. Peterson, J. D. King, and V. A. Segreto, "Clinical Evaluation of Stannous Fluoride as an Anticaries Mouthrinse," *Journal of the American Dental Association,* Vol. 86, February 1973, pp. 404–408.

Ritchie, J., "Oral Health of Rhodesia African First Year Student Teachers," *Community Dent. and Oral Epidemiol.,* 7, 1979, pp. 222–226.

Rugg-Gunn, A. J., and P. J. Holloway, "Methods of Measuring the Reliability of Caries Prevalence and Incremental Data," *Community Dent. and Oral Epidemiol.,* Vol. 2, 1974, pp. 287–294.

Rugg-Gunn, A. J., M. C. Downer, and F. P. Ashley, "Reliability of Caries Data in Three Clinical Trials," *Community Dent. and Oral Epidemiol.,* Vol. 4, 1976, pp. 15–18.

Rytomaa, I., V. Jarvinen, and J. Jarvinen, "Variation in Caries Recording and Restorative Treatment Plan Among University Teachers," *Community Dent. and Oral Epidemiol.,* Vol. 7, No. 6, 1979, pp. 335–339.

Shaw, L., and J. J. Murray, "Diagnostic Reproducibility of Periodontal Indices," *J. Periodont. Res.,* Vol. 12, May 1977, pp. 141–147.

Shaw, L., and J. J. Murray, "Inter-examiner and Intra-examiner Reproducibility in Clinical and Radiographic Diagnosis," *Internat. Dent. J.,* Vol. 25, No. 4, 1975, pp. 280–288.

Ship, I. I., E. M. Jones, and L. L. Laster, "Errors, Reversals, and Questionable Dental Caries Diagnoses Based on a Longitudinal Study in Children," *J. Dent. Res.,* Vol. 45, No. 2, 1966, pp. 323–331.

Slakter, M., D. B. Juliano, and S. L. Fischman, "Estimating Examiner Consistency with DMFS Measures," *J. Dent. Res.,* Vol. 55, No. 6, 1976, pp. 930–934.

Smith, L. W., J. D. Suomi, J. C. Greene, and J. P. Barbano, "A Study of Intra-Examiner Variation in Scoring Oral Hygiene Status, Gingival Inflammation and Epithelial Attachment Level," *J. Periodontics,* Vol. 41, December 1970, pp. 671–674.

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

Thomas, A. E., and H. C. Jamison, "Effect of SnF2 Dentifrices on Caries in Children: Two-year Clinical Study of Supervised Brushing in Children's Homes," *Journal of the American Dental Association,* Vol. 73, 1966, pp. 844–852.

Thylstrup, A., and O. Fejerskov, "Clinical Appearance of Dental Fluorosis in Permanent Teeth in Relation to Histologic Changes," *Community Dent. and Oral Epidemiol.,* 6, 1978, pp. 315–328.

Torell, P., "Reversals in Caries Diagnosis," *Acta Odont. Scand.,* Vol. 25, 1967, pp. 191–203.

Videroni, W., G. S. Sternberg, and G. N. Davies, "Effect on Caries Experience of Lifetime Residents After 10 Years of Fluoridation in Townsville, Australia," *Community Dent. and Oral Epidemiol.,* Vol. 4, No. 6, 1976, pp. 248–253.

Weisenstein, P. R., and W. A. Zacherl, "A Multiple-Examiner Clinical Evaluation of a Sodium Fluoride Dentifrice," *Journal of the American Dental Association,* Vol. 84, No. 3, March 1972, pp. 621–623.

Winer, B. J., *Statistical Principles in Experimental Design,* McGraw-Hill, New York, 1962.

[redacted]