Title
Clinical judgment in orthodontics

Author
Miller, Ross J.

Publication Date
1998

Peer reviewed|Thesis/dissertation

Clinical Judgment in Orthodontics: Relationships Between Pretreatment Classification and Evaluations of Posttreatment Records

by

Ross J. Miller, D.D.S.

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Oral Biology

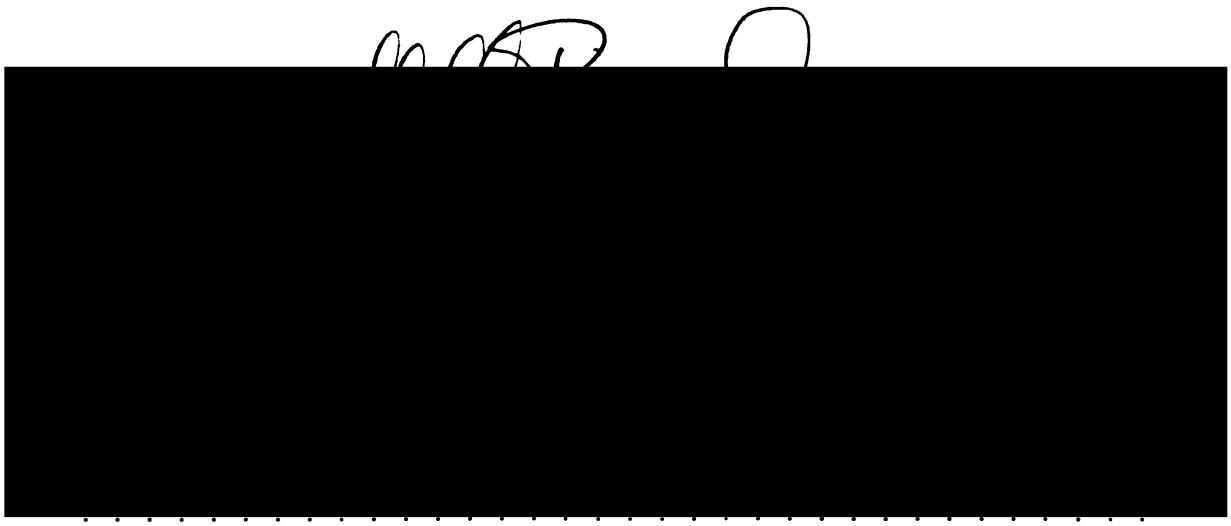
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco



Date University Librarian

Degree Conferred: . . . . . .

ACKNOWLEDGMENTS

I would like to thank my mentor and committee chairman Dr. Sheldon Baumrind for his help in this project and for giving me a deeper understanding of orthodontics. I would also like to thank Dr. Jane Weintraub and Dr. Grayson Marshall for serving on my thesis committee and for providing guidance and support.

A special thanks goes out to Dr. Stuart Gansky who helped with all the statistics on this project. Together we were able to educate me enough to give me a vague understanding of kappa.

Thanks also go out to Dr. Paul Sherick, Dr. Rick Herrmann, and Dr. Jack DuClos for their early work on this material.

I would also like to give thanks to Dr. John Gibbs who was gracious enough to offer the records of his treated patients for case analysis and to all the clinical faculty who participated by evaluating the displayed cases.

Thanks also go to my parents Troy and Rita Miller for their support over the years. And last but not least a very special thanks goes to my wife Cheryl who has shown her understanding and patience for my going back to school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

This study analyzes a number of the judgments clinicians make regarding the final outcome in orthodontic treatment. In order to understand how orthodontists make these judgments it is important to review the complexities of orthodontic treatment and of occlusion.

Orthodontists formulate their diagnosis, treatment planning, and improvement assessment from clinical evaluations and from patient's records. These diagnostic records include study models, photographs, x-rays and written medical and dental histories. The diagnostic process is complex and consisting many components and decisions.

During the diagnostic phase an orthodontist analyzes the study casts to evaluate a number of occlusal traits which contribute to a malocclusion. These include but are not limited to: Angle Classification, amount of crowding, tooth size discrepancies, overjet, overbite, and missing teeth. Not all clinicians evaluate these traits the same way. Uncertainty and ambiguity may hinder the precise diagnosis of a certain trait.

Photographs of the patient are also analyzed. Usually these include frontal and profile facial photographs, with the patient smiling and in repose, and intra-oral photographs. The facial or extra-oral photographs are primarily used to assess facial esthetics. The intra-oral photographs are used to assess the malocclusion, soft-tissues and individual teeth.

Cephalometric films and other radiographs are also a part of the diagnostic records. The cephalometric film can be used to evaluate pathology, but most orthodontists use it to measure various landmarks to evaluate skeletal and dental relationships that are used to reach a diagnosis.

Once the diagnostic data gathering process and evaluation is complete, treatment alternatives are devised. A wide range of treatment approaches may be used depending on the type of malocclusion and the age of the patient. The orthodontist must make a number of decisions regarding the treatment approaches he/she wishes to use. During this

treatment planning stage a decision regarding the need for extraction is made. This is possibly the most critical decision in the delivery of routine orthodontic care (Baumrind 1993). The extraction of teeth in orthodontics has been a long historically and interesting controversy that continues to this day (Baumrind, Korn et al. 1996).

Once treatment is complete a final set of records is taken. This usually includes study models, photographs, and radiographs.

In order to understand some of the complexities in orthodontics it is necessary to go into some depth into the methods used for evaluating malocclusions. Numerous methods have been developed over the past 40 years to assess treatment priority and malocclusion severity These usually attempt to systematize severity or treatment outcome. Many indices exist to evaluate malocclusion and they usually classify malocclusions into types (Angle 1907), those that record prevalence in epidemiological studies (Solow and Helm 1968), those that attempt to record treatment need or priority and those that assess the success of treatment (Richmond, Shaw et al. 1992). Incisor alignment can be measured using the irregularity index developed by Little (Little, Riedel et al. 1988). This scoring method involves measuring, in millimeters, the linear displacement of anatomic contact points, as distinguished from the clinical contact points, of each maxillary and mandibular incisor from the adjacent tooth anatomic point. A number of other indexes have also been developed to assess treatment priority: The Treatment Priority Index (TPI) (Grainger 1967), the Occlusal Index (OI) (Summers 1971; Summers 1972), the Peer Assessment Rating (PAR) index (Richmond, Shaw et al. 1992a; Richmond, Shaw et al. 1992b), Handicapping Labio-lingual Deviations (HLD) index (Draker 1960), the Handicapping Malocclusion Index, and the Index of Orthodontic Treatment Need (IOTN) (Brook and Shaw 1989). These indices are based on specifying the amount of deviation from normal occlusion and these assessments of malocclusion severity are based, more or less, upon a description of individual morphologic malocclusion traits.

Dentists and orthodontists may be unaware of the degree of reliability of the methods and data used in clinical practice (Bader and Shugars 1995). Reliability of clinical measures in dentistry (periodontal health assessment, caries detection) remains a difficult problem in diagnosis, treatment planning and assessing different outcomes. The Treatment Priority Index, Occlusal Index and Peer Assessment Rating index scores have been shown to have acceptable reliability. However, none of these is routinely used in describing malocclusions in the United States; therefore, terms describing individual malocclusion traits, such as Angle Classification, overjet, overbite, remain the language of clinical orthodontics.

To understand malocclusion it is important to understand normal occlusion. The literature offers several definitions for normal occlusion and malocclusion. Normal occlusion: What is today called a normal occlusion was described as early as the eighteenth century by the famous anatomist John Hunter. Carabelli in the nineteenth century was the first to describe abnormal relationships of the upper and lower dental arches in a systematic way. The term *orthodontics* was first used by Lefoulon of France (Weinberger 1926). Even though a number of treatises on orthodontics had already been written by the beginning of the twentieth century, including Kingsley (Kingsley 1880), there was no acceptable method of describing irregularities and abnormal relationships of the teeth and jaws.

Ideal occlusion (Kraus, Jordan et al. 1969) rarely exists in nature and Graber says it is better to call this concept the "imaginary ideal." Unfortunately, there is no clear-cut or acceptable definition of normal occlusion; therefore much of our diagnosis in orthodontics is based on this highly arbitrary concept of ideal (Graber 1994). A number of authors have had a profound impact on dentists' and orthodontists' concept of ideal occlusion.

The point of departure for modern orthodontics begins with Edward H. Angle with the start of the first specialty in dentistry. The Angle School of Orthodontia started in 1900 and was the first school of its kind. Angle created a systematic method of evaluating

malocclusion and defined what normal occlusion was. Angle's line of occlusion and his classification of malocclusion into three classes continues to be used today (Angle 1907). One of the most important characteristics in normal occlusion for Angle was that the mesiobuccal cusp of the maxillary first molar should occlude in the mesiobuccal groove of the mandibular first permanent molar.

A description of normal occlusion usually involves occlusal contacts alignment of teeth, overbite and overjet, arrangement and relationship of teeth within and between the arches and relationship of teeth to osseous structures (Graber 1986). Normal implies a situation commonly found in the absence of disease, and normal values in a biological system are given within an adaptive physiologic range. Normal occlusion , therefore, should imply more than a range of anatomically acceptable values; it should also indicate physiologic adaptability and the absence of recognizable pathologic manifestations.

How orthodontists evaluate occlusion is not without its controversies. Historically the orthodontic study model has been of the hand held variety. This allows the orthodontist to evaluate various aspects of how the teeth come together. Study models provide a three-dimensional record of the dentition and are essential for many reasons. Evaluating how the teeth come together with these hand held models is still acceptable, but there is a trend in orthodontics that finds them unacceptable for evaluating occlusion.

There are those that find an articulator absolutely necessary to evaluate the occlusion (Roth 1973). The idea of functional occlusion is that the joints, facial muscles, and teeth all work together in harmony. According to Roth the joints are the "determinants" of mandibular and condylar position during full closure, and the joints determine and guide the movement aided by the anterior teeth. Tooth positions are, thus, subservient to the dictates and guidance of the joints. Roth discusses "mutually protective" occlusion by this he means that the posterior teeth should contact equally and evenly upon closure into occlusion with no actual contact of the anterior teeth (.005 inch clearance) to avoid lateral stress to the anterior teeth and supporting structures (Beyron 1964). There should be

minimal overjet and overbite, but sufficient overbite so that upon any movement in any direction out of full occlusion, the anterior teeth act as a group to gently, but immediately, disengage or disclude the posterior teeth.

Without discussing this controversy further, in this study the models at our disposal are of the hand held variety. With this definition we are only looking at anatomical occlusion and not functional occlusion (Roth 1976).

In 1972 Andrews evaluated 120 casts of naturally optimal occlusion. Andrews described six keys to normal occlusion that these casts all had in common.: (1) molar relationship (Angle class I); (2) crown angulation (tip); (3) crown inclination (torque); (4) absence of rotations; (5) tight contacts; and (6) a flat occlusal plane or slight curve of Spee (Andrews 1972). From Andrews work came the straight wire appliance, it and appliances based on his work dominate the orthodontic appliance market. Absence of any of the six qualities result in an occlusion that is proportionally inferior to the naturally optimal sample. These six keys can serve as a base for evaluating occlusion and also may be used as treatment objectives for most patients (Andrews 1989).

Normal occlusion is made up of a number of physiological, anatomical and dental definitions. The study cast is just one way of looking at a small part of the overall occlusion. Occlusion is extremely complex and varies quite widely between individuals and age groups (Ash and Ramfjord 1995). Every dentist and orthodontist has their own personal beliefs they bring to the evaluation process.

Let's now turn our attention away from normal occlusion. Malocclusion: Historically, any deviation from ideal occlusion has represented what Guilford termed malocclusion (Guilford 1889). Occlusion in every age group offers the clinician a spectrum: from the imaginary ideal, to normal, to the most severe malocclusion. There are patients that vary from ideal occlusion only slightly in that they may have a rotation of an anterior tooth to those patients with severe craniofacial anomalies. Both of these patients

represent two points along this wide range of occlusal conditions that the clinician may encounter.

Severe malocclusion usually is accompanied by skeletal discrepancies of the face and jaws, often referred to as dentofacial deformities. However, malocclusion should not be thought of as a pathologic condition but merely as human morphologic variation. (Exceptions occur in syndromes and trauma).

The description of malocclusion is usually a list of problems that vary from normal or ideal. This includes: A description of the occlusion in the three planes of space; vertical, sagittal and transverse; a list of which individual teeth are out of alignment; a description of overbite and overjet, Angle classification; tooth size discrepancy (Bolton 1962), crowding or spacing, and many more.

Hellman's description was among the first and is still one of the few biometric studies of occlusion (Hellman 1921). He believed that in the maximum intercuspal position, specific landmarks in the opposing dental arches should contact. If the orthodontist's view of occlusion remains one of static descriptive morphology, a problem lies in being able to define the *individual norm*. This *individual norm* is difficult to define because function, physiological and psychological adaptation must be considered when evaluating whether an individual's occlusion is normal.

A current view is that a malocclusion exists when a misarrangement of the teeth creates a problem for the individual, whether functionally or psychosocially. This definition is not wholly adequate from a morphological standpoint in that it is, in part, a cultural one. With this definition the same arrangement of teeth could be a functional or psychosocial problem in one setting and not in another (Albino, Cunat et al. 1981). Because neither the reactions of the patient nor the impact on function can be predicted confidently from the morphology it is difficult to draw a line based on morphology alone that differentiates normal occlusion from malocclusion (Graber 1994). Although there is the individual's conception of there own occlusion, this is not a part of the present study, we

mention it because it can be a large part of what dictates treatment. What treatment is performed is decided by the clinician with the patients input. The end of treatment is usually decided by the clinician, but there may not be an absolute end point. It may be decided that stopping treatment before the goals of the clinician are met, is best for the patient. This is especially true when periodontal or concerns regarding carries are present.

Agreement: There is much information on what normal occlusion is and what malocclusion is and there are a number of methods for quantifying treatment need and treatment results. What about agreement between those doing the evaluations? Keeling (Keeling, S et al. 1996) found that there was poor reliability in determining maxillary and mandibular anteroposterior positions in screenings of children, but found excellent reliability for judging posterior crossbites. Fields evaluated orthodontists' agreement of soft tissue profiles in children and found them to be acceptable (Fields, Vann et al. 1982). Han looked at agreement of various combinations of diagnostic record groups compared to the complete sets of records and found that in the majority of cases, 55%, study models alone provided adequate information for treatment planning (Han, Vig et al. 1991). The problem with agreement and using it as measure of reliability or correctness is that even if everyone agrees with some diagnosis, everyone can be incorrect.

Specific areas of the diagnosis have associated with it varying degrees of reliability that may effect the amount of agreement. For example, Angle Classification has been examined for its reliability, and in a recent study disagreement on Angle Classification was found to be 25% (Baumrind, Korn et al. 1996). Gravely found a wide range of agreement on Angle Classification depending on what classification the patient was (Gravely and Johnson 1969). In a questionnaire Katz found that many in the orthodontic community realize the diagnostic problems with Angle Classification (Katz 1992).

The question of agreement is an important one in many areas of health care for a variety of reasons. It is important to the patient that their diagnosis is correct, it is important to the provider in order to decrease errors and it is important to the party paying

the fees. The more agreement decreases the chance that the diagnosis or treatment is incorrect. Many insurance companies require a second opinion when surgery is being performed for a number of reasons, but one important one is it increases the chances that the diagnosis is correct (McSherry, Chen et al. 1997).

As part of this study we will be looking into this question of agreement. Specifically how the initial Angle Classification and extraction decision effect the agreement on the quality of a group of records: posttreatment study casts, posttreatment photographs, and complete pre- and posttreatment records. Knowing how the initial state of the case effects the amount of agreement about the case is important for a number of reasons. It can help in deciding what treatment approach will have the greatest benefit for the patient. It can point to areas of difficulty. It can help the clinician know which cases are likely to have a favorable outcome and which ones are more likely to have poor outcome. Knowing what cases have a better chance of success with a specific type of treatment can be very helpful to the clinician.

Another reason the overall question of agreement to orthodontics is important is that at this time we really do not know how much information we get out of each part of the record. We do not know how much information we get from the study casts, photographs, and radiographs. By evaluating the amount of agreement we can begin to understand how important each part of the record is to us.

This study hopes to add more information to the literature regarding where we as clinicians can expect to find high levels of agreement or low levels of agreement with our peers on various parts of diagnostic records. Is the amount of agreement intrinsic to the Angle Classification or extraction decision? Can we expect to find more agreement on the study casts compared to the complete set of records? These are some of the areas of interest to this study. One implication of disagreement may be that the nomenclature used is ambiguous. The definitions of clinical terms may be too imprecise. Different types of education or orthodontic practice may emphasize characteristics of the malocclusion

differently. There may be many different approaches to the same problem. We will now discuss what our aims are in this study.

GENERAL AIM

The general aim in this study was to quantify agreement among clinicians in assigning judgments of the degree of success of orthodontic treatment and the relationship of these judgments to Angle Classification and extraction decision.

SPECIFIC AIMS

Aim 1) To evaluate what association Angle Classification and extraction decision has on the judgment of treatment outcome based on viewing posttreatment study casts. The following order is expected to be present, from best to poorest: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction. I expect that treatment outcomes associated Class I will be considered better than Class II's and nonextraction treatment outcomes to be considered better than extraction.

Aim 2) To evaluate what association Angle Classification and extraction decision has on the judgment of treatment outcomes based on viewing posttreatment extra oral photographs. The following order is expected to be present, from best to poorest: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction. I expect treatment outcomes associated with Class I to be better than Class II and nonextraction to be better than extraction.

Aim 3) To evaluate what association Angle Classification and extraction decision has on treatment improvement when judges are asked to examine complete pre- and posttreatment records. The following order is expected to be present, from best to poorest: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction. I expect treatment outcomes associated with Class I to be better than Class II and nonextraction to be better than extraction.
The same order is expected through the first three aims because Class I is considered normal occlusion and nonextraction treatment leaves the patient with a full compliment of teeth, for what may be a better occlusion.

Aim 4) To assess the correlation between the evaluation of posttreatment study casts and the evaluation of posttreatment photographs, the correlation between the evaluation of posttreatment study casts and the evaluations of treatment improvement, and the evaluation of posttreatment photographs and the evaluations of treatment improvement. I expect there to be a high degree of correlation between these three groups.

Aim 5) Categorize the reasons given for the "best" and "poorest" responses for this sample, and determine if there is a difference in the reasons given between the Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction, Class I vs Class II and extraction vs. nonextraction treatment groups. The hypothesis is there will be differences among these treatment groups.

Aim 6) Determine the agreement among the judges over the three sets of records and within each of the treatment groups, Class I and Class II, extraction and nonextraction, Class I nonextraction, Class I extraction, Class II nonextraction and

Class II extraction. The hypothesis is that there will be differences among these groups.

MATERIALS AND METHODS

The sample of orthodontic patient records and associated judgments were collected from the participating clinicians by Drs. Rick Herrmann and Jack DuClos, under the direction of Dr. Sheldon Baumrind. Part of this data set is the input to the present study. First I will discuss how the treating clinician was chosen, how the actual patient records were collected, and then how the "judges" were selected. Samples of the forms that "judges" completed are also included.

SAMPLE SELECTION

The cases in this sample received orthodontic treatment by a single university trained orthodontic expert. The requirement for classification as an expert was that the clinician satisfy at least four of the following criteria:

1) More than 10 years of clinical experience.
2) Certification by the American Board of Orthodontics.
3) Membership in the Edward H. Angle Orthodontic Society.
4) More than five years as an orthodontic instructor at an American Dental Association accredited orthodontic program.
5) Publication of one or more original papers in a peer-reviewed orthodontic journal.

Dr. John Gibbs of San Mateo, who provided the cases for this study, is a clinical professor of Orthodontics at the University of the Pacific and met the other requirements above.

PRESELECTION OF ORTHODONTIC CASES

The sample was gathered in the following manner: A random list of all patients for whom treatment had begun after Dr. Gibbs' fifth year of practice and for whom treatment had been completed by January 1, 1990 was generated. Cases were selected based upon the completeness of their treatment records and on the following criteria: 1) subjects between the ages of ten and fifteen at the commencement of orthodontic treatment that high certainty of having growth remaining. 2) The presence of a mild to moderately severe Angle Class I or Angle Class II malocclusion. 3) Presence of a normal complement of teeth anterior to the second molar in each quadrant. 4) Patient treated without orthognathic surgery. 5) Patient records were to have the following: medical and dental history forms, examination forms, treatment cards, pretreatment study casts, lateral cephalograms, radiographs, cephalometric tracings and photographs. 6) Absence of chronic disease. 7) No history of prior orthodontic treatment, and 8) Availability of similar posttreatment records.

All subjects who met these criteria were sequentially numbered and placed into the following four categories:

1) Class I nonextraction (25 subjects),
2) Class I extraction (13)
3) Class II nonextraction (25)
4) Class II extraction (24)

These four categories do not represent the true proportions in the general population. Although this is the case, the sample is reasonably representative of the problem being assessed. In the United States the approximate percentages of Angle Classifications are: 30% at most have a normal occlusion, 50% have a Class I malocclusion, 15% have a Class II malocclusion, and 5% have a Class III malocclusion (El-Mongoury and Mostafa 1990). Extraction treatment is performed at a lower rate today

than it was 20 years ago and in one study there were slightly more extraction cases for the Class I malocclusion children (26.7%) than either Class II (23.1%) or Class III (24.1%) (Kuthy, Antkowiak et al. 1994).

The rondomly selected cases were assigned numbers. Class I nonextraction cases were numbered from 100-125, Class I extraction cases from 200-213, Class II nonextraction cases from 300-325, and Class II extraction cases from 400-424. From among the randomly selected cases, 12 were chosen from each of the four categories (48 total) via use of a computer generated random number list. Twelve cases were chosen because it was thought 3 "best", 6 "middle", and 3 "poorest" would best simulate a normal distribution in the population and make the analysis easier, but it was found that the "middle" groups was overrepresented. The 48 cases were then divided into four groups of 12 coded A through D and each assigned a number from 1 through 12. Figure 1 shows a diagram of the random groupings of cases. The sample was stratified to contain 12 Class I nonextraction subjects, 12 Class I extraction subjects, 12 Class I nonextraction subjects and 12 Class II extraction subjects. No effort was made to balance them for gender. All of the 48 patients in this study were adolescent patients between the ages of 7.59 to 15.39 years (mean = 10.94±1.76) at the time of pretreatment records, and between the ages of 12.54 to 18.04 years (mean = 14.92±1.50) at the time of posttreatment records. Females outnumber males by 29 to 19. The characteristics of the data can be found Table 1.

Figure 1. Stratified Random Sample of 48 cases, by Angle Classification and Extraction status.

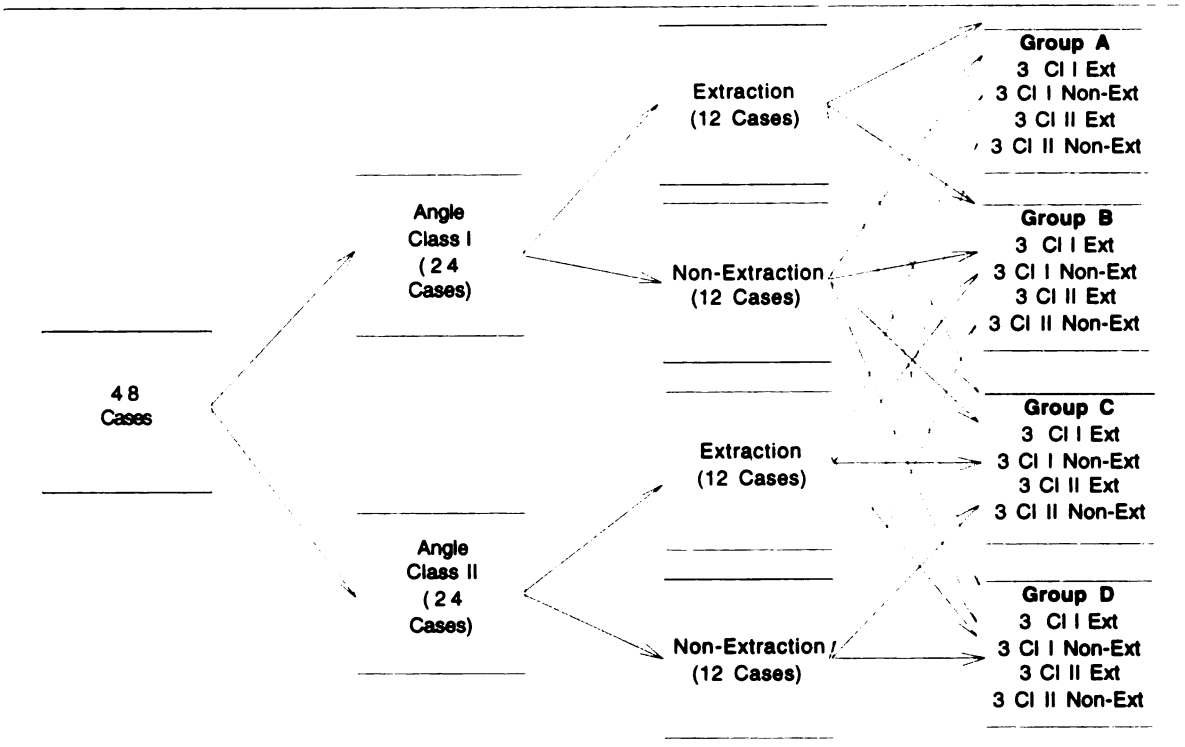


Table 1. Sample Characteristics**

	Nonextraction			*Extraction*			*Nonextraction*		*Extracti*	
	N	*Mean**	*SD**	*N*	*Mean**	*SD**	*M*	*F*	*M*	*F*
1. Class I and Class II Cases	24			24			8	16	11	1
a. Age at T1		10.88	1.61		11.01	1.93				
b. Elapsed Time-T1 to T2		3.75	1.60		4.20	1.87				
2. Class I Cases	12			12			2	10	5	7
a. Age at TI		11.02	1.80		11.57	2.07				
b. Elapsed Time-T1 to T2		3.43	1.75		3.35	1.46				
3. Class II Cases	12			12			6	6	6	6
a. Age at TI		10.74	1.46		10.45	1.68				
b. Elapsed Time-T1 to T2		4.07	1.43		5.04	1.89				

*In years.

**Distributed by angle class, extraction/nonextraction category, and gender.

SOLICITATION OF CLINICAL INSTRUCTOR PARTICIPATION (JUDGES)

A set of clinicians was invited to participate in this study. In order to satisfy the design requirement of having five independent evaluations for each subject, it was necessary to involve a total of fourteen members of the UCSF Division of Orthodontics clinical teaching staff. This was necessary since the teaching staff was usually only at UCSF one day a week, and it was not possible to have the same five clinicians evaluate all the records, which would have been desirable. The principal investigators did participate in the judging.

The credentials of these clinicians are summarized in Appendix A. The time in clinical practice ranged from 4 to 44 years and the length of time as a clinical instructor ranged from 2 to 44 years. Like the great majority of clinicians who trained in their age cohorts, all of them were male. In terms of ethnic background, eleven were European, two were Asian, and one was African American. Although all fourteen clinicians received their specialty education at UCSF, it is believed that these clinicians approximately reflect the distribution of education and experience of contemporary university-trained clinical orthodontists. Although all were graduates of this institution most of these clinicians have different practice philosophies and treatment methods. They are, therefore, not so similar as might be anticipated. A round robin scheme was used in order to facilitate more equal involvement of as many clinicians as possible and to help avoid the potential for clinician burnout during the study.

COLLECTION OF JUDGMENTS

Four tasks were performed by the judges. The decision to extract (Task 1), judgments of posttreatment study casts (Task 2.1), posttreatment photographs (Task 2.2) and complete pretreatment and posttreatment records were obtained (Task 3). The study was planned so that the clinical instructors ("judges") could participate in the study during the normal time period in which they saw patients with the residents in the orthodontic

clinic. It was therefore important to make the completion of the task forms as simple and as expeditious as possible.

In Task 1 (The decision to extract), information was gathered concerning the necessity of extraction from the complete pretreatment records. This data is not used in the present study, but is mentioned for completeness.

EVALUATION I: EVALUATION OF POSTTREATMENT STUDY CASTS

In this task (see Task 2.1, Figure 2.), clinicians were asked to select the three posttreatment cases with the "best" occlusion and the three cases with the "poorest" occlusion from among a set of twelve cases based solely on the appearance of posttreatment study casts. Such a system was used rather than have the clinicians rank order all 12 cases. This was necessary because of time constraints, the potential for burnout of the clinicians, and in order to force the distribution of the cases.

It should be kept in mind, that there will be an unevaluated middle group consisting of six subjects for each of the twelve groups evaluated. This is a group that by definition is in the area where the judgments were not sought. We will refer to this as the "middle group" for simplicity. These are the subjects that were neither good enough to be in the "best group" or "most improved group" nor "poor" enough to be placed in the "poorest" or "least improved group."

Figure 2. Example of Data Collection Sheet used for Posttreatment Study Casts.

Investigator Name: _____ Task 2.1
Sample Group:_____ Date:_____

UCSF Division of Orthodontics - Clinical Decision Making.- Study

EVALUATION OF POSTTREATMENT STUDY CASTS

Posttreatment study casts for twelve subjects who have received full banded orthodontic therapy are arrayed on this table.

Please Examine each set of casts. Then indicate below the I.D. numbers of the three sets of casts representing the best occlusion and the three representing the poorest occlusion. For each of the six 'extreme' cases, please indicate briefly no more than three reasons leading- to your decision.

A. "BEST" OCCLUSIONS

I.D. #_____	I.D. # _____	I.D.#_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

B. "POOREST" OCCLUSIONS

I.D. #_____	I.D. # _____	I.D.#_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

This system was also chosen over a system in which only one case would be chosen as either "best" or "poorest" as this would have constrained the clinician's choices and made it difficult to rank-order the cases of the five clinicians' judgments. In order not to constrain the clinicians by giving predetermined choices, three lines were given below each case for the clinicians to indicate briefly in open-ended form, no more than three reasons leading to their decision. This strategy produces a data set which is free from our biases, but is relatively difficult to interpret.

EVALUATION 2: EVALUATION OF POSTTREATMENT PHOTOGRAPHS

This task (see Task 2.2, Figure 3) is analogous in design to Task 2.1, except that clinicians were asked to select from a set of twelve cases, the three posttreatment cases with the "best" facial appearance and the three cases with the "poorest" facial appearance based solely on the appearance of posttreatment frontal and lateral facial photographs. Again, following the same strategy, clinicians were to briefly indicate no more than three reasons leading to their decision on the lines below each case choice.

EVALUATION 3: EVALUATION OF TREATMENT IMPROVEMENT

This task(see Task 3.0 form: Figure 4) is also analogous in design to Task 2.1. In this task, each clinician was asked to evaluate pre-and posttreatment records for a set of twelve cases and choose the three cases that had the "most" improvement from treatment and to choose the three cases which had the "least" improvement following treatment. Again, following the same strategy, clinicians were asked to indicate no more than three reasons leading to their decision on the lines below each selected case.

Figure 3. Example of Data Collection Sheet used for Posttreatment Photographs.

Investigator Name: _____ Task 2.2
Sample Group:_____ Date:_____

UCSF Division of Orthodontics - Clinical Decision Making.- Study

EVALUATION OF POSTTREATMENT PHOTOGRAPHS

Posttreatment photographs for twelve subjects who have received full banded orthodontic therapy are arrayed on this table.

Please examine each set of frontal and lateral photographs. Then indicate below the I.D. numbers of the three sets photographs representing the best appearance and the three representing the poorest appearance. For each of the 'extreme' sets, please indicate briefly no more than three reasons leading to your decision.

A. "BEST"" APPEARANCE

I.D. #_____	I.D. # _____	I.D.#_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

B. "POOREST" APPEARANCE

I.D. #_____	I.D. # _____	I.D.#_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

Figure 4. Example of Data Collection Sheet used for Improvement Assessment.

Investigator Name:_____ Task 3
Sample Group: _____ Date:_____

UCSF Division of Orthodontics - Clinical Decision Making Study

EVALUATION OF TREATMENT OUTCOMES

Pre- and posttreatment records for twelve subjects who have received full banded orthodontic therapy are arrayed on this table.

We suggest that you first spend 2-3 minutes glancing at the records of each of the twelve subjects. Then, after you have a general idea about all 12 cases, review them as needed and identify the three cases representing the greatest improvement and the three cases representing least improvement. Finally, for each of these six " extreme" cases, please indicate briefly no more than three reasons for your decision.

A. **GREATEST IMPROVEMENT**

I.D. #_____	I.D. # _____	I.D.#_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

B. **LEAST IMPROVEMENT**

I.D. #_____	I.D. # _____	I.D.#_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

STUDY DESIGN STRATEGY

The strategy for presenting these tasks to the evaluating clinicians was designed to minimize clinician bias. That is, the display of the records had to be done in such a way that individual subjects would not be remembered. The clinicians did not know how many times they would be asked to look at different sets of records or what part of the records would be evaluated. Therefore there would be no reason to try to retain information about any of the cases, which hopefully decreased some bias. Whether or not this attempt at decreasing bias was successful is not known.

All photographs were lab mounted with intra- and extra-oral photographs. Because the aim was to determine how clinicians make decisions from posttreatment facial photographs independent of other records, the mounted photographs were placed into individual x-ray envelopes, cut so that only the non-smiling posttreatment frontal and lateral facial photographs were displayed. To prevent the display of teeth the smiling frontal facial photograph was covered.

Task 2.1 was always completed for a given set before proceeding to Task 2.2 for the same set. Clinicians were asked not to draw conclusions between the cases for each task . Tasks 2.1 and 2.2 were completed for the entire sample before proceeding with Task 3.

There was potential for some bias when evaluation of the pretreatment records was done, as some of the pretreatment faces might have appeared similar to the posttreatment faces. This possibility was not tracked during this study. It was also important to complete this task before proceeding with Task 3.0, when the complete pre- and posttreatment records were displayed so that bias concerning how the subjects were actually treated and actually grew did not affect the clinicians decisions regarding each case.

For Task 3.0, the evaluation of treatment outcomes, clinicians were asked to evaluate complete pre- and posttreatment records: extraoral and intraoral photographs, radiographs, cephalometric tracings and study casts. This section of the study was ordered

last to prevent the clinicians from seeing complete pre- and posttreatment records first. This should minimize bias when examining a portion of the records for the other tasks. It was thought that if something was remembered from the pretreatment records, it might influence the evaluation of the posttreatment records.

GENERAL DISPLAY OF CASES

Names of both patient and the expert clinician who treated the case were blocked out with black tape on the photographs, radiographs, and cephalometric tracings. Cases were displayed in the part-time faculty office in the Division of Orthodontics on a large counter by set (A-D), and in ascending order, left to right. Typically, only one set was shown at a time due to available counter space and the expected time needed for evaluating the cases. However, for Task 2.1 and 2.2 different sets of cases were shown at one time to facilitate doing more than one task during a session whenever sequencing permitted.

The case code was marked on Avery adhesive labels and placed on the mounted photographs and on top of the maxillary study casts for Task 2.1 and 2.2, the evaluation of posttreatment study casts and photographs.

For Task 3.0, the evaluation of treatment improvement, pretreatment extraoral and intraoral photographs, radiographs, and cephalometric tracings for a given case were placed into one x-ray envelope and marked with the case code, sex, and age of the patient at the time of the records. The corresponding pretreatment study casts were then placed on top of the envelope containing the rest of the pretreatment records and displayed as previously discussed. This was repeated for the posttreatment records and then they were placed in front of their corresponding pretreatment records.

Each set of cases was displayed for a period of time, typically from one to three weeks, until five clinical instructors in the Division of Orthodontics had independently examined all 12 cases and filled out the proper form (s) for each task. These would usually be the first five clinical instructors available, i.e. a convenience sample.

AIM 1 to 3: EVALUATION OF POSTTREATMENT STUDY CASTS, POSTTREATMENT PHOTOS AND IMPROVEMENT.

The general evaluation approach was to determine the number of "best" or "most improved" and "poorest" or "least improved" responses. Then, using SAS and Microsoft Excel, calculate Chi-Square statistics for nominal versus ordinal data. Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction were compared against the ordinal responses of "best", "middle", or "poorest" in the posttreatment study casts group and posttreatment photographs group.

In the evaluation of treatment improvement the ordinal responses were: "greatest improvement", "middle", or "least improvement."

To achieve an order for the four treatment groups a score of 1 was assigned the poorest group, a 2 for the middle group and a 3 for the best group. Then these scores were summed and used to order the treatment groups.

AIM 4: CORRELATION OF JUDGMENTS

Judgments from five clinicians were performed on each of the study casts, photos, and improvement assessments. To evaluate the correlation among the different groups, a score of 1 was assigned the poorest group, a 2 for the middle group and a 3 for the best group. For each case the scores of the five clinicians was summed for a final rating of the case. A score of 5 would indicate all judgments were in the poorest category and a 15 would indicate all judgments were in the best category. A Spearman partial correlation coefficient was obtained for all the 48 cases and also in the groups of Class I extraction, Class I nonextraction, Class II extraction, and Class II nonextraction using SAS. The partial variables were Class and Extraction.

AIM 5: REASONS

240 evaluations were performed for each of the posttreatment study casts, posttreatment photos, and complete records. There were 48 subjects, each evaluated by 5

judges. One quarter or 60 of these would be placed in the "best" category, one half or 120 into the middle group and one quarter or 60 into the "poorest" group. Three reasons for the judges decisions were asked for in the "best" and "poorest" categories. This yields a maximum of 180 "best" reasons and 180 "poorest" reasons or a total of 360 possible reasons for each of the three records groups (study casts, photos and complete records). To evaluate the reasons, they were placed into categories (Table 2,Table 3 and Table 4).

Table 2. Study Casts Legend

Category	Number
Anterior Occlusion (3-3)	1
Posterior Occlusion (4-8)	2
Occlusion (Unspecified where)	3
Overjet	4
Overbite/vertical	5
Archform/Transverse	6
Angle Classification	7
Other	8
Blank	9

Table 3. Photos Legend

Category	Number
Chin/mandible	1
Profile full/flat	2
Vertical	3
Lips	4
Angle Class./Submental fold/Mentalis	5
Nose/Nasolabial angle	6
Symmetry/Proportions/Balance	7
Other (maxilla, cheeks)	8
Blank	9

Table 4. Improvement legend

Category	Number
Occlusion	1
Facial Harmony/Esthetics/Profile/Lips/Smile	2
Vertical	3
OJ/OB	4
Difficulty/No Change/Or Worse/Amount of Improvement	5
Angle Classification	6
Transverse/Midlines/Incisor Torque	7
Other (space, crowding, Archlength, decalcifications)	8
Blank	9

Categorization was based on the investigator's interpretation of similarity of reasons. For statistical purposes these categories were created in such a way as to obtain a minimum of 20 in each category. No analysis was performed before the categories were determined.

AIM 6: AGREEMENT

Part of the difficulty in assessing agreement with this data set is that the same judges did not view all the same cases. There were 14 clinicians judging the records in various groups of five. The Kappa statistic (Cohen 1960; Cohen 1968) is generally used to assess the amount of agreement between a pair of judges. This data set is not well suited for this analysis, because there were only a few cases in which the same judge looked at all three record groups for the same subject. However, a generalization of unweighted kappa can be used to measure agreement among any constant number of raters where there is no connection between the raters judging the various subjects. This generalization can be used for the case of more than two raters and for the case where the raters judging one subject are not necessarily the same as those judging another (Fleiss, Nee et al. 1979; Fleiss 1981; Fleiss, Mann et al. 1991).

Kappa generally has an associated standard error (Landis and Koch 1977) and the statistical significance of kappa may be tested by referring the computed quantity "z" score

to that of the standard normal distribution. This is usually done to test for the null hypothesis that the agreement seen is the same as would be expected by chance(i.e. k=0). What we are interested in here instead is the level of agreement over the three sets of records: posttreatment study casts, posttreatment photographs, and the improvement assessment.

The Kappa statistic was be generated across the four treatment groups and three record sets. It will also be computed for the original sets of records(A-D) that were displayed to the judges to evaluate this agreement.

RESULTS

AIM 1: EVALUATION OF POSTTREATMENT STUDY CASTS

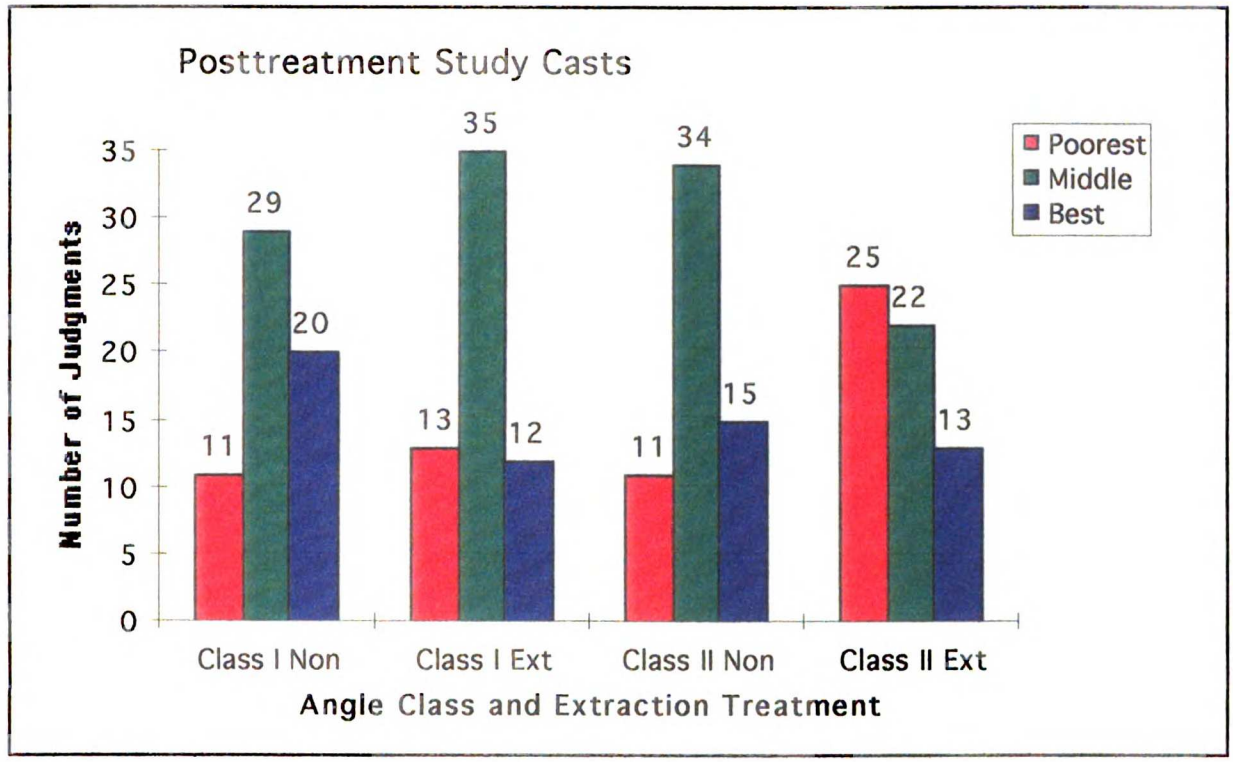
For this part of the study only the posttreatment study casts were evaluated by the judges. The judges were not given any pretreatment information regarding Angle Classification although the judges could tell which cases were treated with extraction treatment. We expected the order from best to poorest to be: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction. We expected that treatment outcomes associated with Class I would be considered better than Class II's and nonextraction treatment outcomes to be considered better than extraction. The data are presented given in Table 5. Figure 5 shows these data in graphical form.

Table 5. Evaluation of Posttreatment Study Casts

	Poorest	Middle	Best
Class I Non	11	29	20
Class I Ext	13	35	12
Class II Non	11	34	15
Class II Ext	25	22	13

It appears that the Class I nonextraction was significantly different, but this was not found.

Figure 5. Angle and Extraction Judgments for Posttreatment Study Casts.



Using the Chi-Square test, we found that there was a significant difference in the proportions of patients' treatment outcomes by Angle/extraction category (p=.019). Also comparison of pairs of occlusion/treatment categories showed that the Class II extraction patients had significantly different outcomes than each of the other categories or the other three categories combined (p=.020). Class II extraction was found to be evaluated more often poorer than the other three classes. No significant differences were detected between Class I extraction, Class I nonextraction and Class II nonextraction in this sample. No significant differences between extraction and nonextraction (p=.292), and Class I and Class II (p=.081) were found (Figure 6 and Figure 7).

Figure 6. Posttreatment Study Casts: Extraction and Nonextraction.

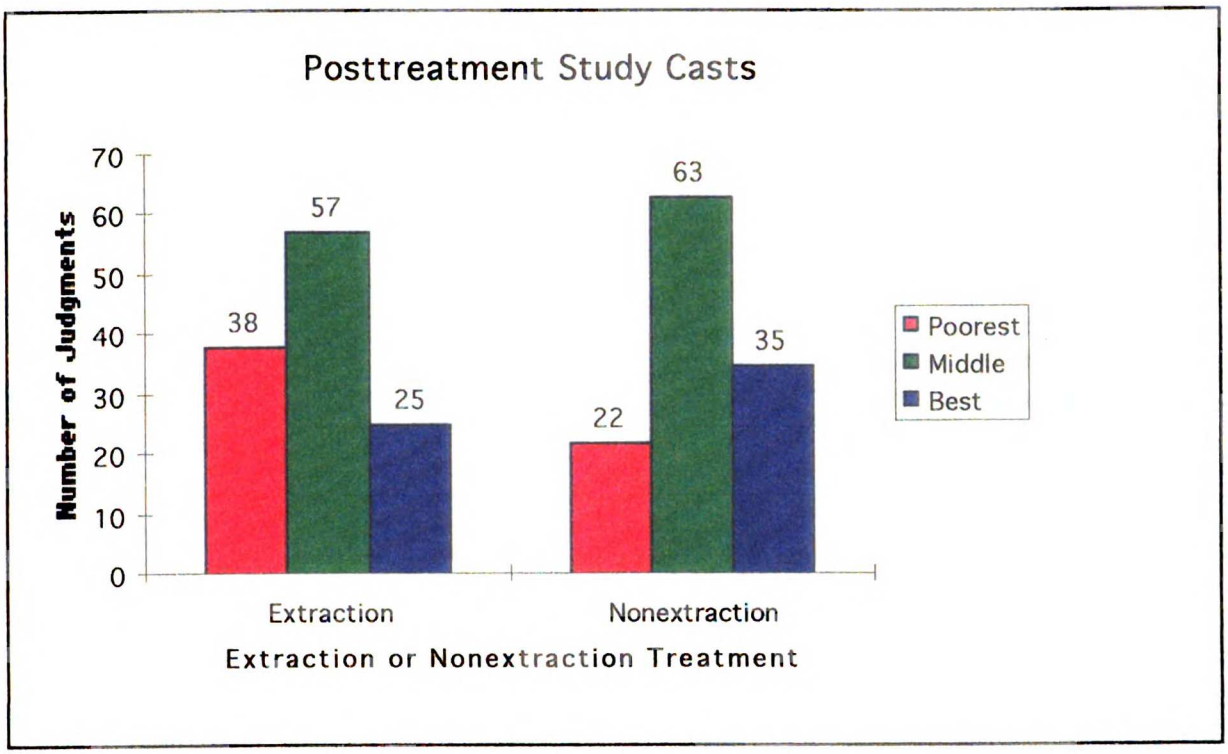


Figure 7. Posttreatment Study Casts: Class I vs Class II.

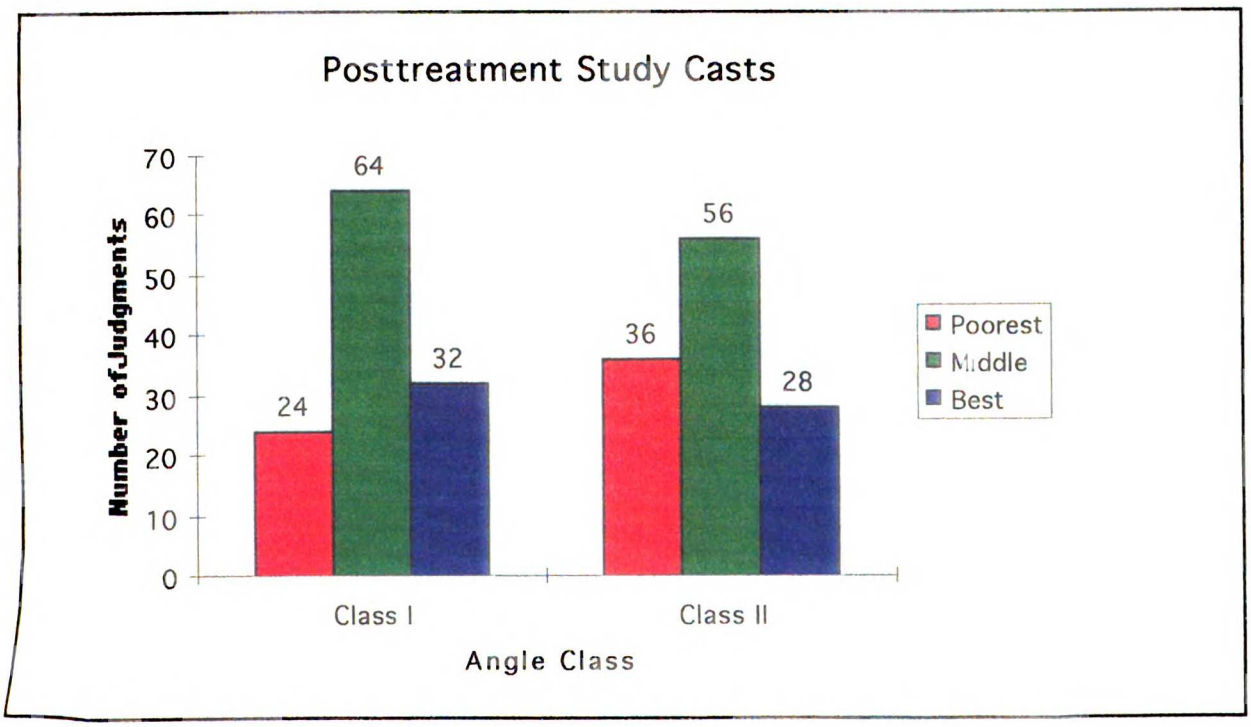


Figure 8. Posttreatment Photograph Evaluation by Angle Classification and Extraction/Nonextraction.

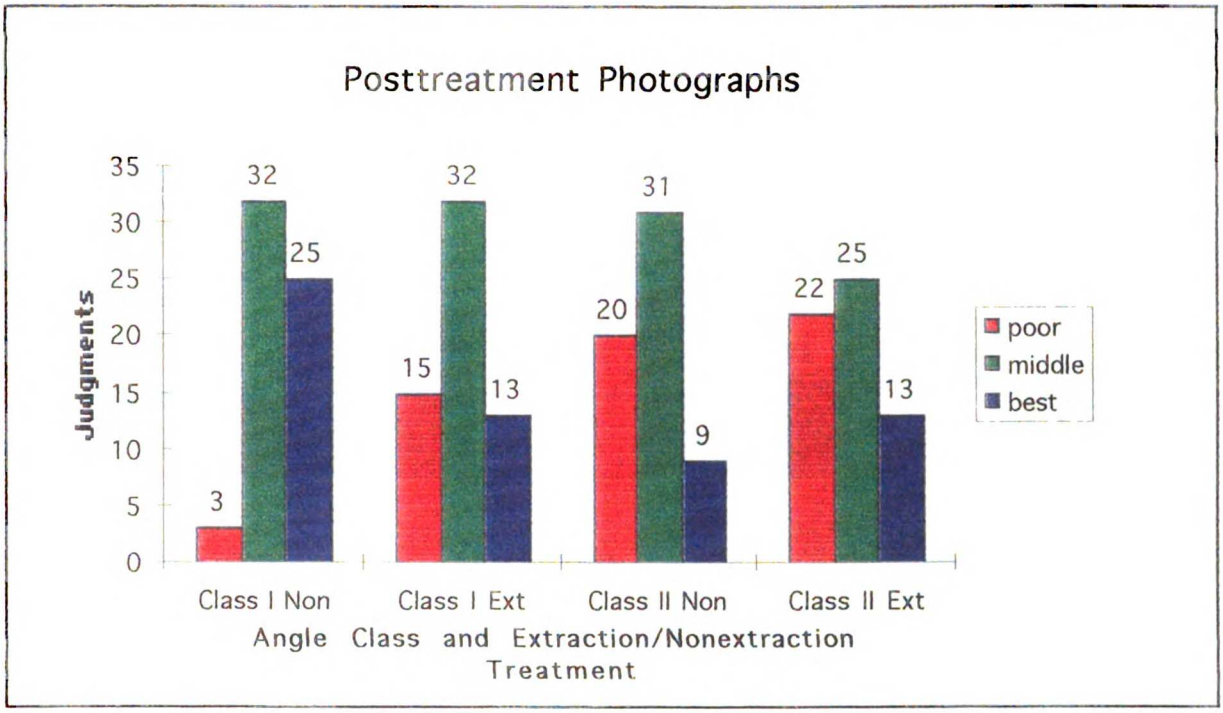


Figure 9. Posttreatment Photograph Evaluation by Extraction and Nonextraction Treatment.

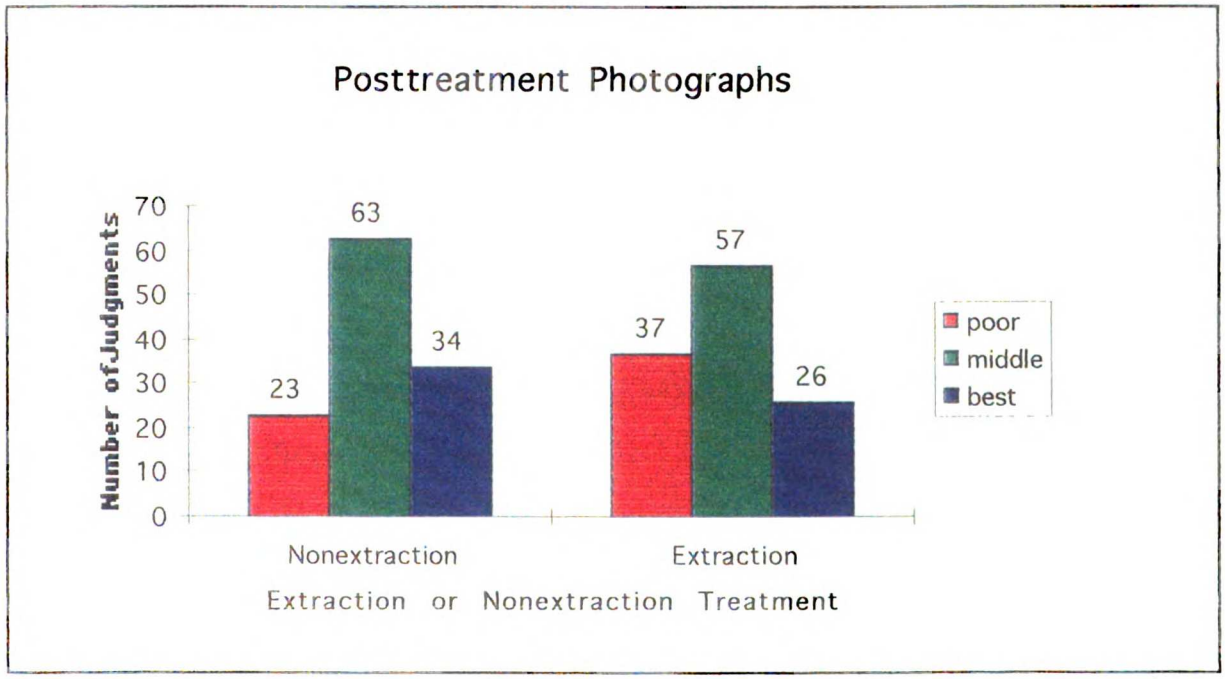
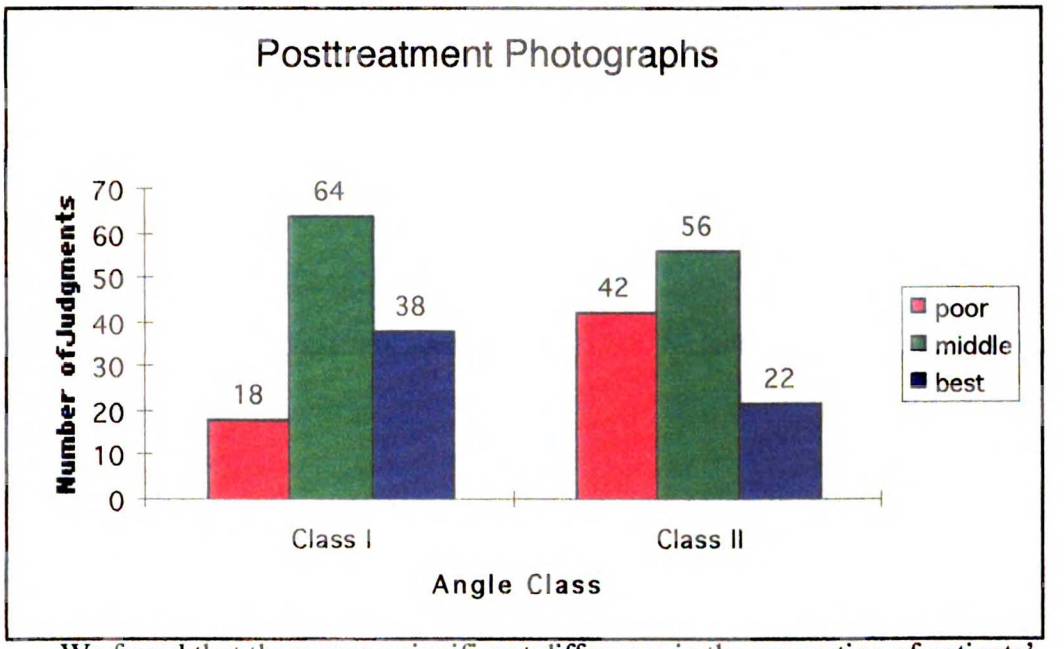


Figure 10. Posttreatment Photographs by Angle Class.



We found that there was a significant difference in the proportion of patients' treatment outcomes by Angle/extraction category (p=.001). There was a clear difference between Class I and Class II photographs. The Class II photographs were judged more often poor and the Class I more often better (p=.001). There was not a great difference when nonextraction was compared to extraction (p=.093).

In the Class I group nonextraction patients were more often judged better than extraction patients (p=.009). In the Class II's there was no statistical difference between nonextraction and extraction noted (p=1).

In the nonextraction group the Class I's were considered better than Class II's (p=.0001). But in the extraction group there was no significant difference detected between. Class I and Class II (p=0.444)

When comparing nonextraction Class I's against all the other groups combined there was quite a significant difference in favor of the Class I's being different (p=.0001). Unlike the study casts the Class II extraction photographs compared to all others were not judged to be significantly different.

The following order was found for the posttreatment photos from best to worst: Class I nonextraction (142), Class I extraction (118) Class II extraction (111) Class II nonextraction (109). Class I (260) was judged better than Class II (220) and nonextraction (251) better than extraction (229). The original hypothesis that the following order would be found from best to poorest: Class I extraction, Class I nonextraction, Class II nonextraction and Class II extraction was not found.

The hypothesis that Class I would be found to be better than Class II was found and it was statistically significant. The hypothesis that nonextraction would be better than extraction was not found.

AIM 3: IMPROVEMENT ASSESSMENT

In this part of the study the pre- and posttreatment records were evaluated by the judges in terms of improvement resulting from the treatment. As was stated earlier we expected an order of quality for these, from best to poorest: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction. The data are presented in Table 7. Figure 11 shows this data in graphical form. Figure 12 and Figure 13 show improvement assessment by extraction and Angle Class, respectively.

Table 7. Evaluation of Improvement Assessment.

	Poorest	Middle	Best
Class I Non	22	30	8
Class I Ext	7	33	20
Class II Non	20	27	13
Class II Ext	11	30	19

Figure 11. Improvement by Angle and Extraction/Nonextraction Treatment.

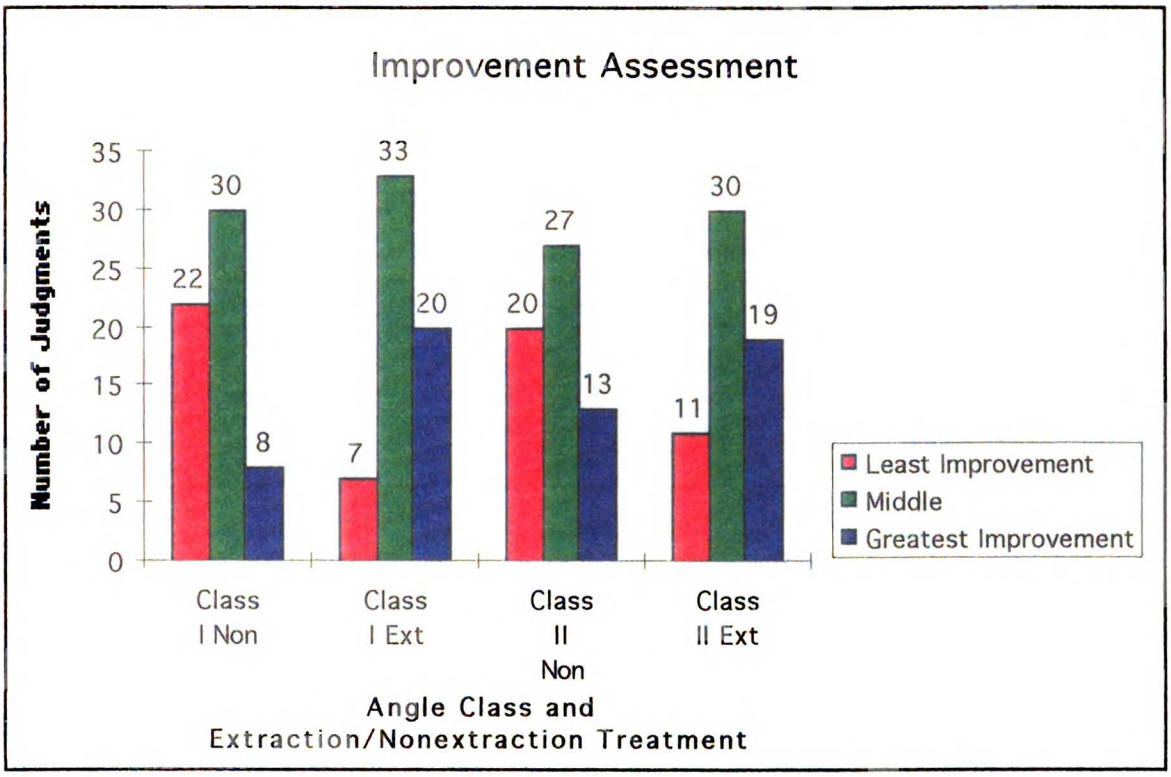


Figure 12. Improvement Assessment by Nonextraction/Extraction Treatment.

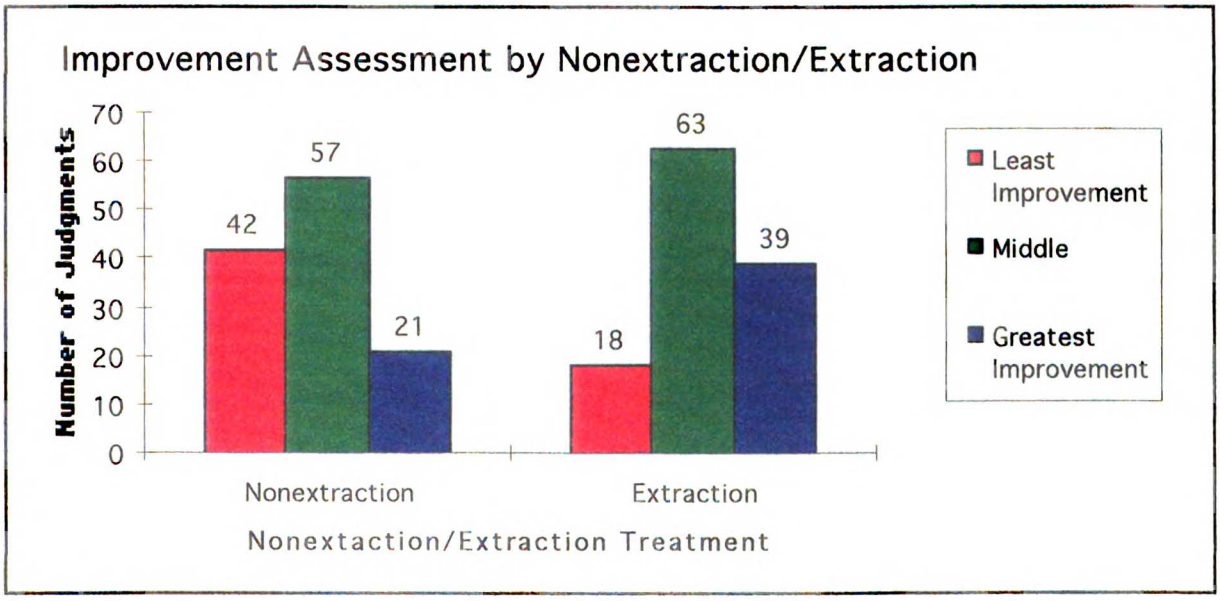
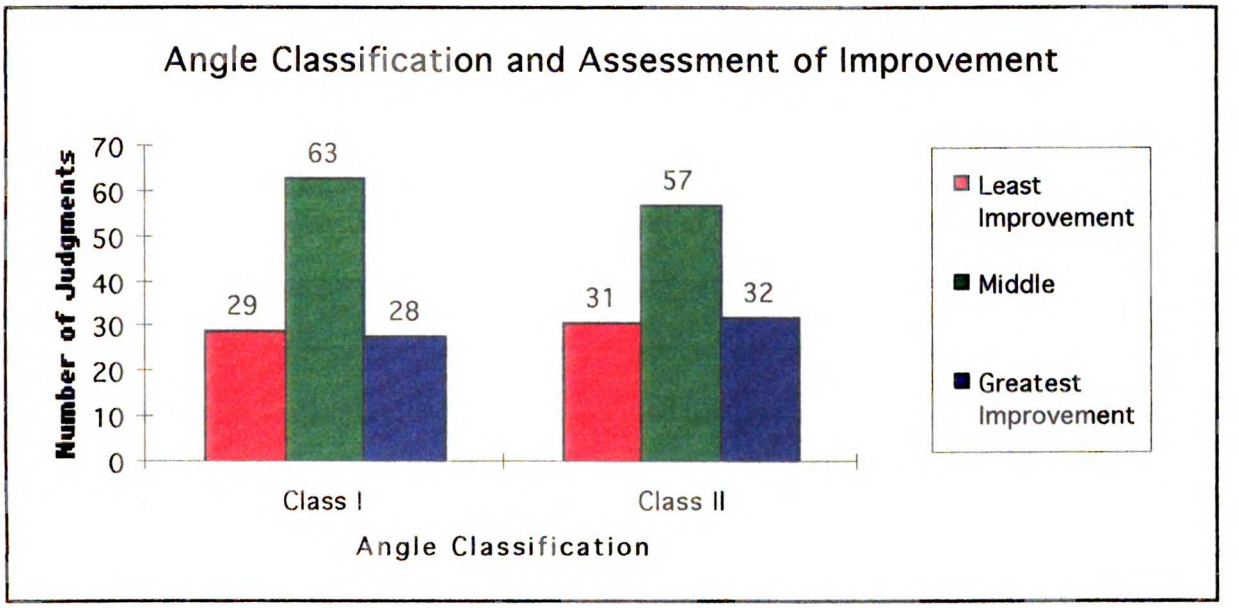


Figure 13. Angle Classification and Assessment of Improvement.



There were some differences in the evaluations of improvement. There were more extraction cases that were judged to have greater improvement than nonextraction cases (p=.006). There was no difference found between Class I and Class II when it came to improvement (p=1.000). Within the Class I group the extraction group was found to be most improved(p=.003). The same could not be said for the Class II's. If one looks at Figure 11 one can see that there might be a trend toward the Class II extraction group having more improvement than the Class II nonextraction group, but it was not statistically significant (p=.200).

The following order was found for treatment improvement from "greatest improvement" to "least improvement": Class I extraction (133), Class II extraction (128), Class II nonextraction (113), and Class I nonextraction (106). Class II (241) was judged more improved than Class I (239) and extraction (261) more improved than nonextraction (219). The original hypothesis that the following order would be found from most improved to least improved: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction was not found. The hypothesis that Class I would be

judged "greater improvement" than Class II was not found and the hypothesis that nonextraction would be judged better than extraction was also not found. In fact the reverse was found, extraction was considered more improved.

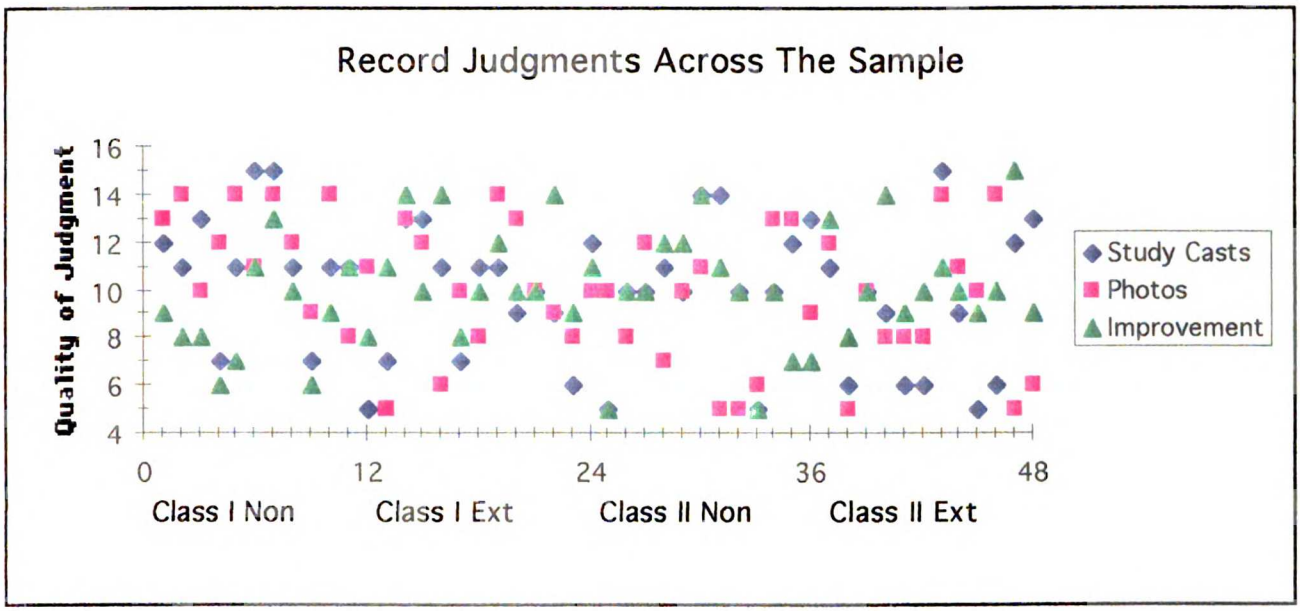
AIM 4: CORRELATION OF JUDGMENTS:

The three questions we asked in this part of the study is: Is there a correlation between the judgments of the posttreatment study casts and the posttreatment photos? Is there a correlation between the posttreatment study casts and the assessment of improvement? Is there a correlation between the posttreatment photographs and the assessment of improvement? The raw data are shown in Appendix B. The plotted data are shown in

Figure 14. If one looks at this graph one can see the variability among judgments. (Unanimous agreement of "poorest" would yield a five, unanimous agreement of "best" would yield a 15). Only in rare instances do we see the judgments coming close to one another. In the 48 case sample, it was found that there was a Spearman Correlation of r_S=.50 (p=.0004) between the posttreatment study casts and improvement assessment. This is a moderate association, which shows the importance of the posttreatment study cast in the evaluation of orthodontic treatment. The posttreatment photos did not correlate well with either the posttreatment study casts or the improvement assessment. These had an r_S=.19 (p=.208) and .02 (p=.91), respectively.

Correlations within the groups of Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction were also analyzed. In the Class I nonextraction cases there was a Spearman Correlation of R=.67613 (p=.0158) between posttreatment study casts and the improvement assessment. This is also a moderate association.

Figure 14. Judgments for all Patients.



The Class I nonextraction group was the only group to have a moderate association between the records sets. The Class I extraction, Class II nonextraction and the Class II extraction group failed to show an association between the records available in this sample.

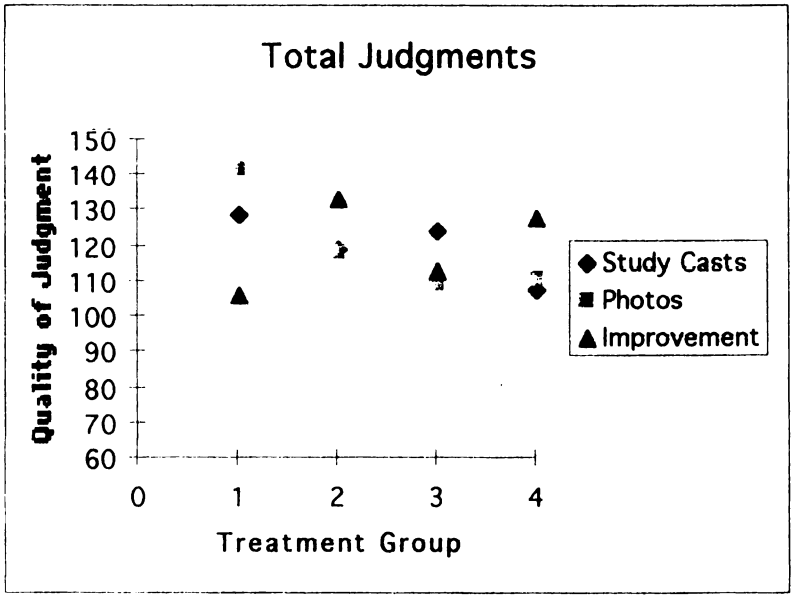
The sum for the judgments across the four treatment groups is shown below in Table 8. Figure 15 shows this data in a plot. The greater the number, the better or greater improvement, of the 12 subject records judged.

Table 8. Total Number of Judgments.

		Study Casts	Photos	Improvement
1	Class I Non	129	142	106
2	Class I Ext	119	118	133
3	Class II Non	124	109	113
4	Class II Ext	108	111	128

(For this set of data unanimous agreement on poorest would yield a 60, and a unanimous agreement on best would yield a 180.)

Figure 15. Judgment (Summary for Four Treatment Groups*)



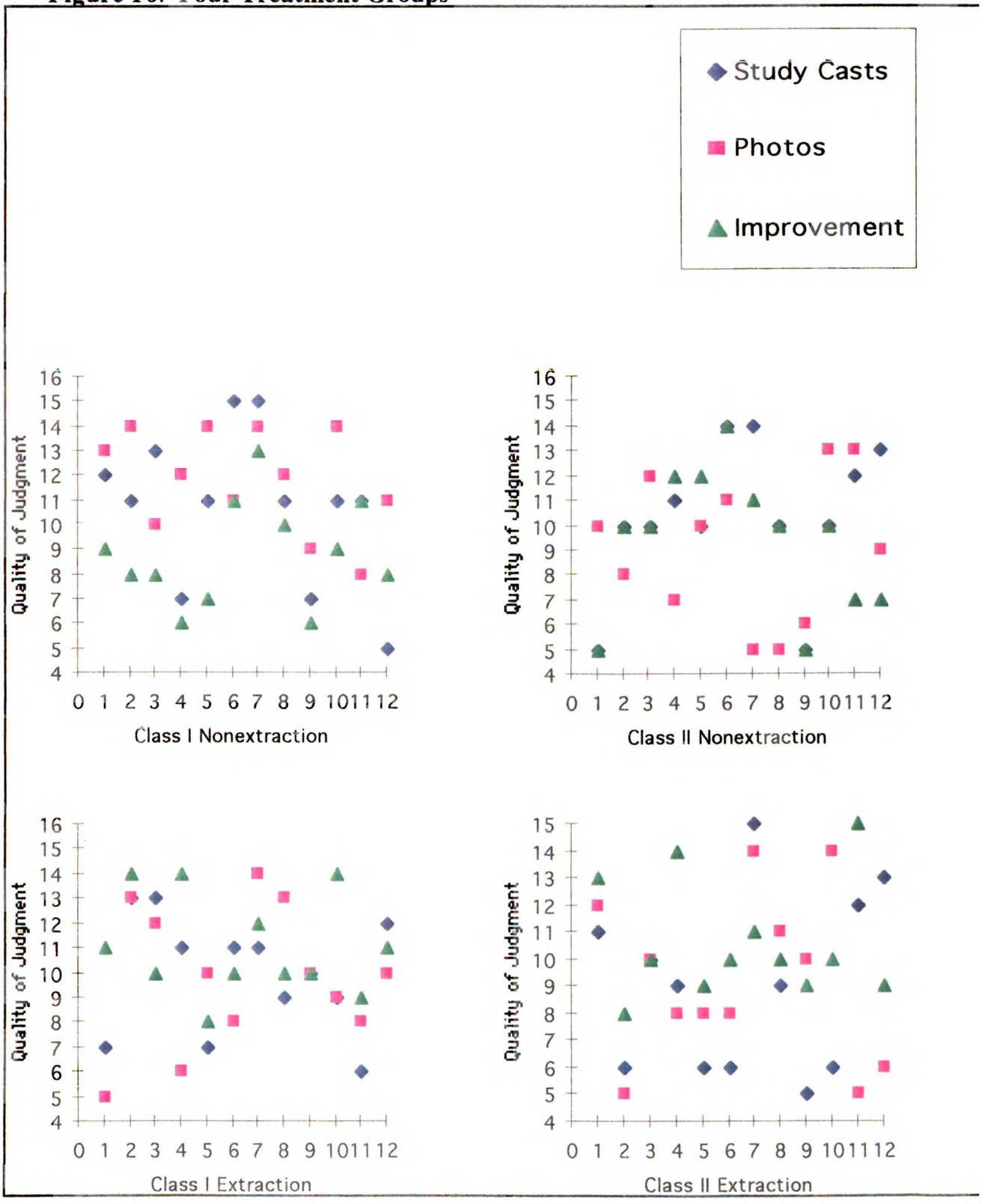
*(1=Class I non, 2=Class I ext, 3=Class II non, 4=Class II ext.)

It is interesting to note from Figure 15 that the greatest amount of improvement was seen in the extraction groups. This isn't too surprising, one would expect the extraction cases to be more difficult or perhaps show a more dramatic change.

Figure 16 takes a closer look at the individual treatment groups. The only group that had a significant correlation was Class I extraction and it was between posttreatment study casts and treatment improvement. The Class II non extraction group looks like there would be a correlation but none was found.

Appendix D shows these graphs with a running average line that helps display the correlation between these plots. The lines associated with posttreatment study casts and treatment groups appear to have similar shapes in the Class I extraction group.

Figure 16. Four Treatment Groups



AIM 5: REASONS

In this part of the study the reasons given for each assessment for the four extraction/ nonextraction/Angle groups are analyzed. If we refer to the original questionnaires (Figures 2-4), one can see that the three reasons asked for were open ended. Each judge was given freedom to answer any way the judge pleased. The judge was to give three reasons for the best three cases and the poorest three cases. Table 9 presents the number of different reasons given in each of the records groups.. Reasons were placed into categories and these categories can be seen in appendix C.

Table 9. Number of Different Reasons

	Posttreatment Study Casts	Posttreatment Photos	Improvement Assessment
Poorest/Least Improved	125	104	117
Best/Most Improved	92	72	119

We are interested only in the reasons that are given, not in the blanks or the middle group. A Chi-Square test was calculated to compare reasons among posttreatment study models, posttreatment photos and assessment of improvement.

For the posttreatment study models we found the following: For the 8 reason categories of Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction treatment groups there was no significant difference found in the reasons given (p=.649). For Class I vs Class II there was no significant difference (p=.198). For the extraction vs. nonextraction groups there was also no difference (p=.701). No differences were detected for the posttreatment study models. The hypothesis that there would be differences was not found.

Photos were the next group and when all 8 categories were looked across the four treatment groups no significant difference was found (p=.056), however this is borderline. There was also no difference between extraction and nonextraction (p=.268). There was a difference between Class I and Class II (p=.047). We can accept the hypothesis that there

were differences in the reasons given. There is a difference in what the judges were evaluating between Class I and Class II on the posttreatment Photographs.

The category with the most difference between Class I and Class II appears to be the chin/mandible category. Figure 17 shows the differences in numbers for each category. Table 10 shows the legend for the reason categories for photographs.

Figure 17. Posttreatment Photographs

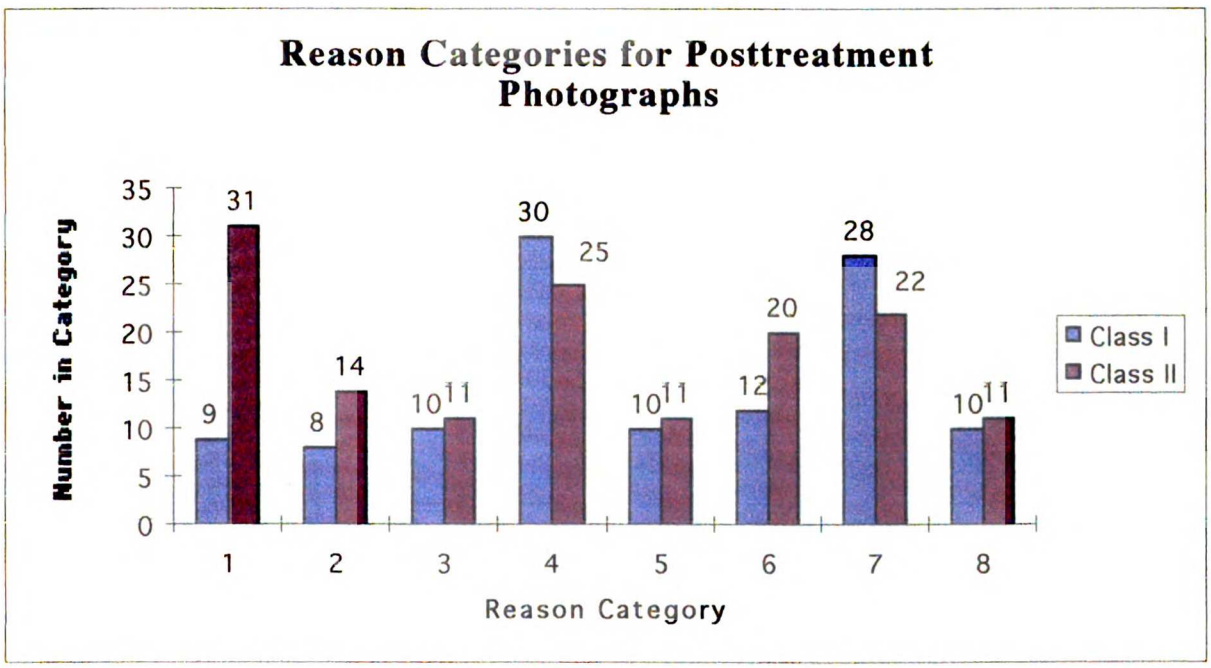


Table 10. Photos Legend

Category	Number
chin/mandible	1
profile full/flat	2
vertical	3
lips	4
Angle Class./Submental Fold/Mentalis	5
Nose/Nasolabial angle	6
Symmetry/Proportions/Balance	7
Other (maxilla, cheeks)	8

It also appears that the profile and Nose/Nasolabial angle category contributed to this difference in Class I compared to Class II

For the assessment of improvement, when the judges had all the pre- and posttreatment records, there were no significant differences found between the four treatment groups (p=.306) or Class I and Class II (p=.884). There was a difference between nonextraction and extraction in reasons given (p=.017). Figure 18 displays the number of different reasons in each category. Table 11 presents the legend for the categories.

Figure 18. Improvement Assessment

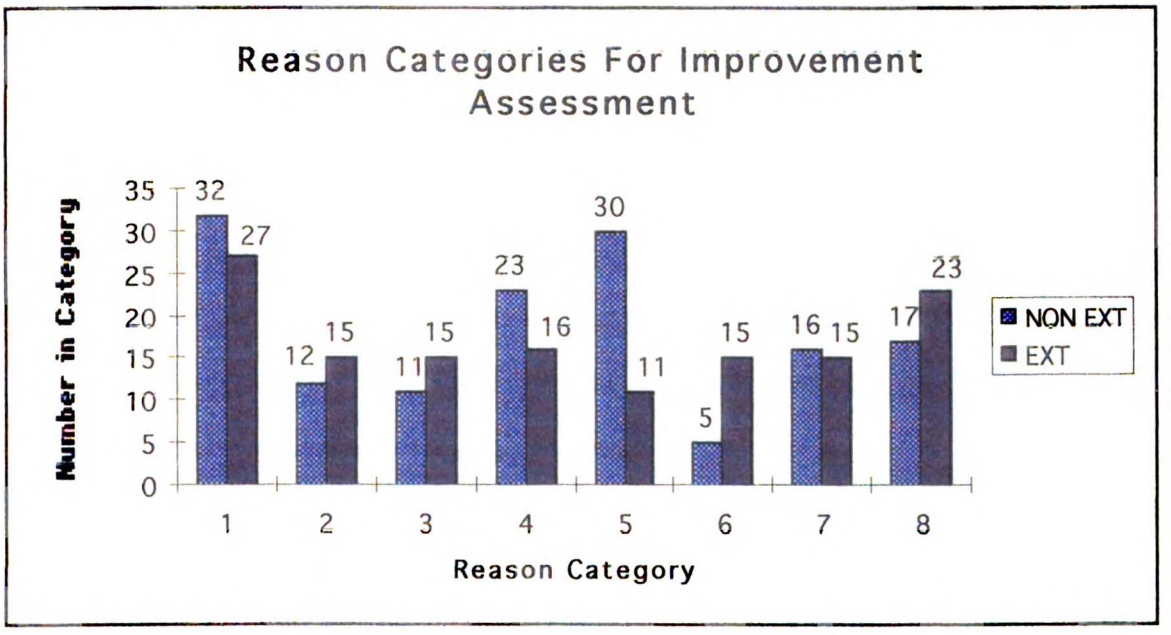


Table 11. Improvement Legend

Category	Number
Occlusion	1
Facial Harmony/Esthetics/Profile/Lips/Smile	2
Vertical	3
OJ/OB	4
Difficulty/No Change/Or Worse/Amount of Improvement	5
Angle Classification	6
Transverse/Midlines/Incisor Torque	7
Other (space, crowding, Archlength, decalcifications)	8

The hypothesis that there would be a difference between the four treatment groups was not supported. But the hypothesis that there would be a difference between extraction and nonextraction can be supported. The categories with the greatest differences appear to be Difficulty/Amount of improvement and also Angle Classification.

AIM 6: AGREEMENT

Strength of agreement determined by kappa statistics is usually presented in manner displayed in Table 12.

Table 12. Kappa Convention

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

When describing the relative strength of agreement associated with kappa statistics, the strength of agreement labels are usually assigned to the corresponding ranges of kappa. Although these divisions are clearly arbitrary, they do provide a useful "benchmark" for the discussion of agreement (Landis and Koch 1977).

Agreement for the 48 cases as displayed to the judges is presented in Table 13. From this table can be seen that the overall agreement for the posttreatment study casts and posttreatment photos was greater than improvement assessment.

Table 13. Kappa For Displayed Cases

	Posttreatment Study Casts	Posttreatment Photos	Improvement Assessment
A1-A12	0.51	0.65	0.56
B1-B12	0.63	0.49	0.38
C1-C12	0.74	0.69	0.51
D1-D12	0.49	0.59	0.51
Average	0.59	0.61	0.49

Within each of the original set of cases displayed before the judges the agreement appears to run from fair to substantial. The unweighted kappas for posttreatment study

casts, posttreatment photos and the improvement assessment for the treatment groups are shown in Table 14:

Table 14. Treatment Groups Agreement

Study Casts		
Category	Kappa	Strength of Agreement
Over 48 Cases	0.39	Fair
24 Class I	0.30	Fair
24 Class II	0.46	Moderate
24 Non-Extraction	0.48	Moderate
24 Extraction	0.28	Fair
12 Class I Non	0.41	Moderate
12 Class I Ext	0.16	Slight
12 Class II Non	0.56	Moderate
12 Class II Ext	0.34	Fair

Photos		
Category	Kappa	Strength of Agreement
Over 48 Cases	0.38	Fair
24 Class I	0.29	Fair
24 Class II	0.44	Moderate
24 Non-Extraction	0.36	Fair
24 Extraction	0.40	Fair
12 Class I Non	0.20	Slight
12 Class I Ext	0.33	Fair
12 Class II Non	0.42	Moderate
12 Class II Ext	0.46	Moderate

Improvement Assessment		
Category	Kappa	Strength of Agreement
Over 48 Cases	0.20	Slight
24 Class I	0.18	Slight
24 Class II	0.22	Fair
24 Non-Extraction	0.20	Slight
24 Extraction	0.16	Slight
12 Class I Non	0.12	Slight
12 Class I Ext	0.16	Slight
12 Class II Non	0.26	Fair
12 Class II Ext	0.16	Slight

Table 14 shows that there is more agreement for the posttreatment study casts (Fair) and posttreatment Photographs (Fair) compared to the improvement assessment (Slight). This order was also found in the judgments of the cases during the initial display (Table 13).

For the posttreatment study casts there was more agreement in nonextraction group than the extraction group. There was also more agreement in the Class II group and nonextraction group.

In the posttreatment photos there was more agreement in the Class II group.

For the improvement assessment group there was no agreement greater than fair and this was found only in the Class II nonextraction group. The least amount of agreement was found for the improvement assessment. Statistical significance was not determined for kappa in this study.

DISCUSSION

In this section we will briefly summarize the findings with respect to the six original hypotheses and touch on the clinical importance of each. We will also consider the strengths and limitations of the present study.

AIM 1: EVALUATION OF POSTTREATMENT STUDY CASTS

The order for the posttreatment study casts found in this study from best to worst was: Class I nonextraction, Class II nonextraction, Class I extraction and Class II extraction. Class I was judged better than Class II and nonextraction better than extraction. The original hypothesis that the following order would be found from best to poorest: Class I extraction, Class I nonextraction, Class II nonextraction and Class II extraction was not found.

Class I was found to be better than Class II and nonextraction was found better than extraction like our original hypothesis, but the differences were not statistically significant.

The most important finding from this section of the study was that Class II extraction group was considered to be poorest overall and this was statistically significant. This finding has a number of clinical implications. When looking only at the posttreatment study models, the Class II extraction cases were clearly judged to be poorer in this sample. Does this mean that these cases are more difficult and therefore more likely to have a worse outcome? Or does this mean that there might have been a better treatment approach for this group such as surgery? If the latter is the case, this finding may imply that everything possible should be done to avoid extractions in a Class II patient and by extension adds more evidence to those who favor arch development at an earlier age.

AIM 2: EVALUATION OF POSTTREATMENT PHOTOS

The following order was found for the posttreatment photos from best to worst: Class I nonextraction, Class I extraction, Class II extraction, Class II nonextraction. Class I was judged better than Class II and nonextraction better than extraction. The original hypothesis that the order would be found from best to poorest (Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction) was not found.

The hypothesis that Class I would be found to be better than Class II was found and it was statistically significant. The hypothesis that nonextraction would be better than extraction was not found.

The most important finding from this part of the study is that the Class II patients were more often judged "poorer" than Class I's. Class II is a dental relationship only, but is often used to refer to a person with a protrusive maxilla or a retrognathic mandible. This implies that a Class II dental relationship is associated with a facial type that will more often be judged "poorer" than Class I's. The Class I's were clearly better, as shown in Figure 10. Extraction did not seem to have an effect on the posttreatment photos. This does not support the idea that extraction treatment can adversely effect the profile.

AIM 3: IMPROVEMENT ASSESSMENT

The following order was found for treatment improvement from "greatest improvement" to "least improvement": Class I extraction, Class II extraction, Class II nonextraction , and Class I nonextraction. Class II was judged to have greater improvement than Class I and extraction to have "greater improvement" than nonextraction. The original hypothesis that the following order would be found from most improved to least improved: Class I nonextraction, Class I extraction, Class II nonextraction and Class II extraction was not found. The hypothesis that Class I would be judged more improved than Class II was not found and the hypothesis that nonextraction would be judged "more improved" than extraction was also not found. The Class I extraction group was found to be more improved than the Class I nonextraction group. This implies that those cases that have extractions performed as part of the treatment will improve the most; that is they may have a more dramatic change.

The most important finding in this part of the study was that Class I extraction cases are more likely to be judged to have "greater improvement" than nonextraction. These were the only two groups that had a significant difference between them. The Class II had a similar trend, but it was not significantly different.

AIM 4: CORRELATION OF JUDGMENTS

There was a correlation between the posttreatment study casts and the improvement assessment. Although it was a moderate association, it does show how important the posttreatment study casts are in the overall evaluation of a case. Other studies have also found the importance of the study models to orthodontic treatment decisions (Han, Vig et al. 1991).

There was no correlation between the posttreatment photos and the improvement assessment. This was somewhat surprising. Orthodontists tend to believe that they affect the lips (Bravo 1994) and this finding implies that the change in the lips on the photograph

isn't great enough to influence the judgment of improvement assessment. It could be that the change in the malocclusion may overwhelm the influence of the photos.

Another interesting finding is that there was no significant correlation between posttreatment study casts and posttreatment photos. This implies that orthodontic treatment does not affect the judgment on photographs.

AIM 5: REASONS

When it came to the reasons for a judge's response, they were placed into categories that were somewhat arbitrary. The categories here were the result of evaluating the reasons given to open-ended questions and combining them to facilitate analysis and interpretation. What categories were made influences the outcome of this part of the study. It must be kept in mind that the construction of categories is difficult and imperfect at best. Some interesting findings are found in spite of these limitations.

For posttreatment study casts there was no significant difference for the reasons given for the four treatment groups. The significance of this is the judges were looking at similar occlusal characteristics when judging these cases.

For posttreatment photographs there was a difference in the reason categories. This difference was found between Class I and Class II patients. The greatest differences were in the chin/mandible, profile/full/flat and the nose/nasolabial angle category. The Class II patients all had more reasons in these categories. This finding is not too surprising, since Class II subjects can have an underlying skeletal discrepancy. It is interesting that vertical reasons played very little role in the differentiation of Class I's from Class II's.

The improvement assessment differences in reasons were found between the extraction and nonextraction groups. Out of 41 reasons given in the category for difficulty/no change/or worse/amount of improvement 30 were in the nonextraction category and only 11 in the extraction category. Judges may have assessed a case negatively based on the relative lack of change or lack of difficulty which may be inherent

in a nonextraction case versus an extraction case. The improvement on the study casts seems to overwhelm the input from the posttreatment photos into the judgment of improvement.

The original hypothesis that there would be a clear order to the four treatment groups was not found. That does not mean it does not exist. It could mean that our sample size was not large enough to detect it. Which leads us to another limitation of this study. It is important to know what questions you are going to ask at the beginning of the study so you know how large a sample you will need to do your analysis. As the data is broken into smaller and smaller groups your ability to detect differences decreases.

AIM 6: AGREEMENT

There was less agreement in the improvement assessment than there was in the posttreatment study casts. This finding implies that individual pieces of the record are easier to find agreement on. This could be due to the fact that looking at just a part of the record decreases the complexity of the problem and makes it easier for clinicians to agree. The posttreatment study casts appeared to have more agreement in the nonextraction group.

The posttreatment photos had more agreement among the Class II cases. This was also the group that tended to be considered "poorer". These two separate findings indicate that clinicians may agree more on those factors that are considered negative. Other studies have also found low levels of agreement on photographs (Phillips, Bailey et al. 1994).

The improvement assessments tended all to have relatively low agreement among the four treatment groups. Only the Class II cases had fair agreement.

Clinicians appear to have relatively low rates of agreement over the three records groups. What are the implication for the patient? The patient may find that orthodontists have a wide range of opinions when seeking consults. A lack of agreement also implies that the language of orthodontics is not as precise as it could be.

STRENGTHS AND LIMITATIONS

Strengths: This is one of the few studies that attempts to analyze these three record sets. These records were evaluated by judges who did not know which cases started out Class I or Class II. This data set was collected in a rigorous manner. The kappa statistic was used to evaluate agreement among the judges and it tends to be conservative. Most of the data collected was either ordinal or nominal and non-parametric statistical methods were used where appropriate.

Limitations: There are a number of limitations regarding this study. A major one is that this study looked at an existing database that was not collected for the questions being asked in this study. Every effort was made to ask questions that the data could answer without the data influencing possible conclusions.

It would have been bast to have the same five judges evaluate all the records. As was stated earlier this was not possible given the nature of the clinic at the time of this study. There were 14 separate judges evaluating the records. This made it impossible to evaluate the judges agreement across the three record groups. Analysis of individual judges was not possible in this study. This also added a great amount of variability to this study - a problem that we never fully overcame.

For task 3.0 (Figure 4) those cases that had the "greatest improvement" or "least improvement" were to be evaluated. It would have been interesting to evaluate the cases that had the best "outcomes". Improvement and outcome are two separate, and possibly overlapping features of treatment that should be elucidated in the future.

Placing the reasons into categories is a difficult problem. It is a very subjective and complicated task. It would be interesting to give the reasons in this study to a number of orthodontists and make a study out of the different ways the reasons are categorized.

The statistical significance of kappa was not determined in this study. We cannot comment on the statistical significance of the agreement seen in this study. We can only point out possible trends. One comment can sum up the difficulty of kappa: "Many human

endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third (Fleiss, Nee et al. 1979)." Other authors are critical of kappa overemphasizing disagreement between categories (Hutchinson 1993).

In future studies the value of each individual part of the record should be scrutinized in order to bring to light what its value is to the diagnostic process and outcome of treatment. Future studies are also needed in the area of agreement.

SUMMARY

There are a number of clinically important findings that are brought to light by this study.

1) Judged without other records, the Class II extraction posttreatment study casts were most often judged poorest in a group of study casts.

2) Judged without other records, the Class II posttreatment photograph were more often judged poorer than its Class I counterparts.

3) When all the records are evaluated for treatment improvement the cases that on average were considered most improved were the extraction cases.

4) The posttreatment study cast tends to have the greatest amount of correlation with the degree of treatment improvement.

5) The reasons given to why Class II posttreatment photographs are poorer than Class I patients were often related to chin, mandible, nose, and overall facial profile.

6) On posttreatment study casts there was more agreement on the Class II cases and nonextraction cases.

7) There was more agreement on the Class II patient's posttreatment photographs.

8) There was less agreement on the improvement assessment than on individual parts of the record.

REFERENCES

Albino, J. E., Cunat, J. J., Fox, R. N.,Lewis, E. A.,Slakter, M. J.,Tedesco, L. A.. (1981). "Variables discriminating individuals who seek orthodontic treatment." Journal of Dental Research **60**(9): 1661-7.

Andrews, L. (1972). "The six keys to normal occlusion." American Journal of Orthodontics **62**: 296-309.

Andrews, L. (1989). Straightwire: The concept and appliance. San Diego, LA Wells.

Angle, E. (1907). The malocclusion of the teeth. Philadelphia, SS White Dental Manufacturing.

Ash, M. M. and S. Ramfjord (1995). Occlusion. Philadelphia, Saunders.

Bader, J. and D. Shugars (1995). "Variation, treatment outcomes, and practice guidelines in dental practice." Journal od Dental Education **59**(1): 61-95.

Baumrind, S. (1993). "The role of clinical research in orthodontics." Angle Orthodontist 63(3): 235-40.

Baumrind, S., Korn, E. L., Boyd, R. L., Maxwell, R. (1996). "The decision to extract: Part 1--Interclinician agreement." American Journal of Orthodontics and Dentofacial Orthopedics **109**(3): 297-309.

Baumrind, S.,Korn, E. L.,Boyd, R. L.,Maxwell, R. (1996). "The decision to extract: part II. Analysis of clinicians' stated reasons for extraction." American Journal of Orthodontics and Dentofacial Orthopedics **109**(4): 393-402.

Beyron, H. (1964). "Occlusal relations and mastication in Australian Aborigines." Acta. Odont. Scand. **22**: 597.

Bolton, W. (1962). "The clinical application of a tooth-size analysis." American Journal of Orthodontics **48**(7): 504-529.

Bravo, L. A. (1994). "Soft tissue facial profile changes after orthodontic treatment with four premolars extracted." Angle Orthodontist **64**(1): 31-42.

Brook, P. H. and W. C. Shaw (1989). "The development of an index of orthodontic treatment priority." European Journal of Orthodontics **11**(3): 309-20.

Cohen, J. (1960). "A coefficient of agreement for nominal scales." Educ Psych Meas **20**: 37-46.

Cohen, J. (1968). "Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit." Psych Bull **70**: 213-220.

Draker, D. (1960). "Handicapping labio-lingual deviations: A proposed index for public health puposes." Am J Orthod **46**: 295-305.

El-Mongoury, N. H. and Y. A. Mostafa (1990). "Epidemiologic panorama of malocclusion." Angle Orthod **60**: 207-214.

Fields, H. W., Vann, Jr., W. F.,Vig, K.W (1982). "Reliability of soft tissue profile analysis in children." Angle Orthodontist **52**(2): 159-65.

Fleiss, J. (1981). Statistical methods for rates and proportions. New York, John Wiley and sons.

Fleiss, J.,Nee J, Landis, JR. (1979). "Large sample variance of kappa in the case of different sets of raters." Psych Bull **86**(5): 974-977.

Fleiss, J. L.,Mann, J., Paik, M., Goultchin, J., Chilton, N. W. (1991). "A study of inter- and intra-examiner reliability of pocket depth and attachment level." Journal of Periodontal Research **26**(2): 122-8.

Graber, T. (1986). Orthodontics. State of the Art. Essence of the Science. St. Louis, C. V. Mosby.

Graber, T. a. V., RL (1994). Orthodontics: Current principles and techniques. St. Lois, Mosby.

Grainger, R. M. (1967). "Orthodontic treatment priority index." Vital and Health Statistics. Series 1: Programs and Collection Procedures **2**(25): 1-49.

Gravely, J. and D. Johnson (1969). "Angle's classification of maloccclusion: an Assessment of Reliability." British Journal of Orthodontics **1**(3): 79-86.

Guilford, S. (1889). Orthodontia or malposition of the human teeth, prevention and remedy. Philadelphia, Spangler.

Han, U. K., Vig, K. W., Weintraub, J. A., Vig, P. S., Kowalski, C. J.. (1991). "Consistency of orthodontic treatment decisions relative to diagnostic records [see comments]." American Journal of Orthodontics and Dentofacial Orthopedics **100**(3): 212-9.

Hellman, M. (1921). "Variations in occlusion." Dental Cosmos **63**: 608.

Hutchinson, T. P. (1993). "Focus on Psychometrics-Kappa Muddles Together Two Sources of Disagreement: Tetrachoric Correlation is Preferable." Research in Nursing and Health **16**: 313-315.

Kat, M.I.(1992). "Angle classification revisited 1: is current use reliable?" Am J Orthod Dentofac Orthop **102**: 173-9.

Keeling, S.D., McGorray S., Wheeler, T.T., King, G.(1996). "Imprecision in orthodontic diagnosis: Reliability of clinical measures of malocclusion." Angle Orthod **66**(5): 381-392.

Kingsley, N. (1880). Treatise on oral deformities as a branch of mechanical surgery. New York, Appleton.

Kraus, B.,Jordan, R.E, Abrams, L. (1969). Dental anotomy and occlusion: a study of the masticatory system. Baltimore, Williams and Wilkins.

Kuthy, R. A., Antkowiak, M.F., Clive, J.M. (1994). "Extractions prior to comprehensive orthodontic treatment in the mixed dentition." Pediatric Dentistry: 16.

Landis, R. and G. Koch (1977). "The measurement of observer agreement for categorical data." Biometrics **33**: 159-174.

Little, R., Riedel R.A., Artun, J. (1988). "An evaluation of changes in mandibular anterior alignment from 10 to 20 years postretention." Am J. Orthod **93**: 423-428.

McSherry, C. K.,. Chen P. J, (1997). "Second surgical opinion programs: dead or alive?" Journal of the American College of Surgeons **185**(5): 451-6.

Phillips C. L., Bailey, J. (1994). "Level of agreement in clinicians' perceptions of Class II malocclusions." Journal of Oral and Maxillofacial Surgery **52**(6): 565-71; discussion 572-3.

Richmond, S.,. Shaw, W. C (1992). "The development of the PAR Index Peer Assessment Rating : reliability and validity." European Journal of Orthodontics **14**(2): 125-39.

Richmond, S. Shaw W. C. (1992). "The PAR Index Peer Assessment Rating : methods to determine outcome of orthodontic treatment in terms of improvement and standards." European Journal of Orthodontics **14**(3): 180-7.

Roth, R. H. (1973). "Temporomandibular pain-dysfunction and occlusal relationships." Angle Orthodontist **43**(2): 136-53.

Roth, R. H. (1976). "The maintenance system and occlusal dynamics." Dental Clinics of North America **20**(4): 761-88.

Solow, B. and. Helm, S (1968). "A method for tabulation and statistical evaluation of epidemiological malocclusion data." Acta Odont Scand **26**: 63-88.

Summers, C. J. (1971). "The occlusal index: a system for identifying and scoring occlusal disorders." American Journal of Orthodontics **59**(6): 552-67.

Summers, C. J. (1972). "Tests of validity of indices of occlusion." American Journal of Orthodontics **62**(4): 428-9.

Weinberger, B. (1926). Orthodontics: an historical review of its origin and evolution. St. Louis, CV Mosby.

APPENDIX A: Characteristics of the Judges

JUDGE	SPECIALTY TRAINING AT	YEARS IN PRACTICE	YEARS AS ORTHO CLINICAL INSTR
1	USC	29	25
2	UCSF	7	6
3	UCSF	28	27
4	UCSF	6	4
5	UCSF	8	5
6	U OKLAHOMA	5	5
7	UCSF	8	7
8	ROYAL DENTAL COL/COPENHAGEN	23	23
9	U C S F	33	13
10	U C S F	19	19
11	U C S F	33	30
12	U C S F	44	44
13	U C S F	12	10
14	U C S F	4	2

APPENDIX B: Judgments of the Three Record Groups.

		Posttreatment	Posttreatment	
No.	Class/Extraction	Study Casts	Photos	Improvement
1	Angel Class I/Non	12	13	9
2	Angel Class I/Non	11	14	8
3	Angel Class I/Non	13	10	8
4	Angel Class I/Non	7	12	6
5	Angel Class I/Non	11	14	7
6	Angel Class I/Non	15	11	11
7	Angel Class I/Non	15	14	13
8	Angel Class I/Non	11	12	10
9	Angel Class I/Non	7	9	6
10	Angel Class I/Non	11	14	9
11	Angel Class I/Non	11	8	11
12	Angel Class I/Non	5	11	8
13	Angel Class I/Ext	7	5	11
14	Angel Class I/Ext	13	13	14
15	Angel Class I/Ext	13	12	10
16	Angel Class I/Ext	11	6	14
17	Angel Class I/Ext	7	10	8
18	Angel Class I/Ext	11	8	10
19	Angel Class I/Ext	11	14	12
20	Angel Class I/Ext	9	13	10
21	Angel Class I/Ext	10	10	10
22	Angel Class I/Ext	9	9	14
23	Angel Class I/Ext	6	8	9
24	Angel Class I/Ext	12	10	11
25	Angel Class II/Non	5	10	5
26	Angel Class II/Non	10	8	10
27	Angel Class II/Non	10	12	10
28	Angel Class II/Non	11	7	12
29	Angel Class II/Non	10	10	12
30	Angel Class II/Non	14	11	14
31	Angel Class II/Non	14	5	11
32	Angel Class II/Non	10	5	10
33	Angel Class II/Non	5	6	5
34	Angel Class II/Non	10	13	10
35	Angel Class II/Non	12	13	7
36	Angel Class II/Non	13	9	7
37	Angel Class II/Ext	11	12	13
38	Angel Class II/Ext	6	5	8
39	Angel Class II/Ext	10	10	10
40	Angel Class II/Ext	9	8	14
41	Angel Class II/Ext	6	8	9
42	Angel Class II/Ext	6	8	10
43	Angel Class II/Ext	15	14	11
44	Angel Class II/Ext	9	11	10
45	Angel Class II/Ext	5	10	9
46	Angel Class II/Ext	6	14	10
47	Angel Class II/Ext	12	5	15
48	Angel Class II/Ext	13	6	9

APPENDIX C: Reason Categories

Study Casts						
Category	Number	Class I NON	Cl I Ext	Cl II Non	Cl II Ext	Total
Anterior Occlusion (3-3)	1	9	8	5	15	37
Posterior Occlusion (4-8)	2	15	14	13	21	63
Occlusion (Unspecified where)	3	11	7	9	11	38
Overjet	4	11	5	9	12	37
Overbite/vertical	5	13	7	8	14	42
Archform/Transverse	6	6	5	3	6	20
Angle Classification	7	4	1	8	6	19
Other	8	8	8	2	4	22
Blank	9	16	23	18	25	82
	middle	87	102	105	66	360
	Grand Total	180	180	180	180	720

Photos						
Category	Number	Class I NON	Cl I Ext	Cl II Non	Cl II Ext	Total
Chin/mandible	1	4	5	14	17	40
Profile full/flat	2	3	5	7	7	22
Vertical	3	1	9	5	6	21
Lips	4	16	14	15	10	55
Angle class./Submental fold/Mentalis	5	6	4	3	8	21
Nose/Nasolabial angle	6	6	6	9	11	32
Symmetry/proportions/balance	7	20	8	11	11	50
Other (maxilla, cheeks)	8	4	6	6	5	21
Blank	9	24	27	17	30	98
	middle	96	96	93	75	360
	Grand Total	180	180	180	180	720

Improvement						
Category	Number	Class I NON	Cl I ext	Cl II NON	Cl II ext	Total
Occlusion	1	14	13	18	14	59
Facial Harmony/Esthetics/Profile/Lips/Smile	2	7	6	5	9	27
Vertical	3	3	8	8	7	26
OJ/OB	4	10	8	13	8	39
Difficulty/No Change/Or Worse/Amount of Improvement	5	17	5	13	6	41
Angle Classification	6	2	7	3	8	20
Transverse/Midlines/Incisor Torque	7	6	7	10	8	31
Other (space, crowding, Archlength, decalcifications)	8	10	13	7	10	40
Blank	9	21	14	22	20	77
	middle	90	99	81	90	360
	Grand Total	180	180	180	180	720

APPENDIX D: Correlation Running Average