REED: I'd like to answer Dr. Goldberg's questions. The first had to do with the fact that a lot of the numbers are very close. First of all, the Lu model and the Carlos-Senning models give almost identical incidence rates. If you look at the form for each of those estimators, you will see that the two estimators merged with each other as the proportion of reversals gets very small, as is the case with the set of data that I used for illustration here. The two approaches to analysis – randomization testing and the finite population approach – also gave very close numbers. I attribute that to the fact that the sample sizes are very large, and under those circumstances randomization testing has been known to approach normal theory distributions. The finite population approach that I used to which Dr. Goldberg referred made the assumption that the sampling fraction was zero. If you assume a different sampling fraction, you have a very good idea about the size of the population to which you are generalizing; you may then be encouraged to use the finite population approach. In general, however, I don't think that we know that much about the populations from which our samples are coming. Finally, I agree that there are many, many more aspects to the diagnostic error problem than the ones that are embodied in the assumptions and postulates on which the Carlos-Senning and the Poole-Clayton-Shah models are based. I didn't account to you what these assumptions are. The assumptions are constrained by the fact that they are only for a two-group comparison, and by the fact that there are three parameters with which to explain a whole myriad of factors that may result in examiner error. Personally, I don't think any one type of modeling should be done until a very, very good idea is attained of just how the examiner error comes about. At any rate, whatever model is eventually used, if any one is, the problem of cluster sampling will always have to be addressed.

BELL: I'd like to comment briefly on several points. First, I agree that it is a desirable goal to maintain examiner-subject pairings. My point was only to mention that it doesn't solve all the problems that are associated with systematic bias. In certain situations, other goals might take precedence over it. Concerning Dr. Smith's point about the trade-off between reproducibility and sensitivity, he suggested that he wanted to have examiners who would be sensitive to changes that did occur. I think that that would be good in the case of a single trial. However, among other things, when someone does report data, it is quite likely that other people will come along later and say, look, the level of caries in this community at this particular time *was* such and such, but now it is something different. They would like to be able to know that every attempt was made to use the typical criteria, and to hold the examiners as close to those criteria as possible. Also, I think that most of us here would feel more confident with a study where they understood the criteria used, rather than one in which the normal criteria were the same but where the examiners were allowed to interpret them any way they felt.

I may have used a little bit of overkill trying to make a point about the importance of calibrating and monitoring the examiners. However, Martyn is correct: I don't have any hard evidence that increasing the amount of monitoring and the amount of calibration does improve the reliability. I doubt that anyone has really studied this very much, but I would like to hear the views of the people here who have more experience with the examination process as to whether increased monitoring and calibration of the examiners make much of a difference in the outcome of the trial.

M. SMITH: I agree with Bob Bell wholeheartedly and feel it would be really nice if one of the things coming out of this Conference would be that there be sort of a minimum standard accepted in reporting reliability in clinical trials, and that this be encouraged in publishing the results in the journals.

POULSEN: I am only going to make one comment about the distinction between the systematic and the random error. When you do the trial, you have very little possibility of distinguishing whether there are differences in diagnostic criteria. And that's over a one- or three-year period. You have no possibility, actually, of determining such differences, so that error is going to be included in your variables. For that reason, I think it should be avoided as much as possible.

YANKELL: I would like to comment on the statement of Dr. Poulsen's, which goes along beautifully with the overall presentation made this morning by Dr. Carlos. He asked whether interim calculations are really necessary. I don't think that interim measurements need to be taken at one- or two-year intervals in long-term clinical trials of standard inorganic fluorides. I think where interim monitoring should be done is with the newer, perhaps more effective, agents, such as antibacterials, or with products that contain higher than 1000 parts per million fluoride in toothpastes. Especially important in studies of this nature might be studies of side-effects, primarily of staining. Here, short-term studies of 21 days are valuable, or even longer studies are very valuable to see what types of staining might be occurring, it is especially important to look at intensity of staining. If you do have staining as a side-effect, baseline values are important, because there are certain colors of teeth that prevent good staining results from being seen. Staining may also be a side-effect that people would not like, and therefore, you might have higher drop-out rates in the groups where you have staining. On the other hand, staining might be viewed as a positive item by some of the marketing people, since this might be an advantage because you know the product is working.

STURZENBERGER: From the manufacturer's point of view, it might be more important to have a more frequent examination or have examinations at a shorter period of time. The feasibility of a long-term study has to be assessed before a lot of effort can be spent. One wants to be reasonably sure the product has some benefit and if it can be assessed.

GRAVES: If we are on the subject of interim examinations, even though I agree we shouldn't generally tamper or make changes with the calls, sometimes annual exams in studies over a long period of time give you some insight as to what may have gone wrong. For example, if you assess that a first molar is missing at baseline, and then at subsequent examinations you have the opportunity to see that calls were unerupted, sound, decayed, and filled, for example, you are pretty sure that your first call was wrong. In fact, missing calls carry such a heavy score that it defeats any possibility of demonstrating an increment. Particularly now that disease levels are as low as they are, we need to give some importance to longitudinal editing. I think that's one value of interim examinations.

CHILTON: There is also the possibility of changing criteria midstream. I'm sure more than one study has been undone because of the changed criteria used by the examiner.

ROSS: The sponsors have to know whether or not the study is progressing according to the protocol as it was originally written. It is necessary at these interim examinations to assure that the standards are the same as they were at baseline.

Is it within the scope of this Conference to answer the question whether we need to continue to run clinical trials for new formulations of fluoride compounds which had been clinically tested many times previously, with generally used fluoride concentrations and abrasive systems?

IMREY: I would like to make two suggestions about observer error, both of which go against the grain of what has been inculcated in all of us – the idea that we have to avoid bias at any cost in clinical trials. But in fact, avoiding bias of all forms is not necessarily the most important thing, as Dr. Greenberg has previously commented. Each of these suggestions involves the introduction of bias into a trial in a way that's not very harmful. The first possibility is that, in the assessment of the results of a trial, one look only at those surfaces which have been identified by the initial examiners as being at risk for caries. In other words, attention has only been paid to changes at risk of surfaces which were regarded as carious at the initial examination but which are not regarded as carious at the second examination. This procedure biases the estimation of change in a particular group, but it does not necessarily bias comparisons between groups treated in different ways, or at least the possibility that it does not bias such comparisons needs to be examined because it may produce increases in efficiency.

The second suggestion is something a little different. This deals with the mechanism of data collection, and it will be clearer if I explain it in the context of a study where it has been used for a different reason. A study was done at the University of Illinois involving tumorogenesis: Rats given anthracene develop breast tumors. The hypothesis of this study was that the development and maturation of these tumors, and possibly regression of these tumors, will be affected by dietary measurements which differ for the animals. A study was designed, and rats were palpated weekly over a period of many months to see if tumors developed and whether they progressed or regressed. In the re-examination of these rats, there was, as you can imagine, a great deal of observer variability possible. To combat this, what was done was something that's very similar to what's done in the clinician's office: A record was made of where the tumors were found, using a diagram of the various rat body areas. The examiner was then provided with this map to be used in subsequent examinations, so that tumors which were recorded previously could be noted – a procedure which would conceivably reduce the variables, particularly when the initial examinations are done by multiple observers. I wonder if this has been considered?

CHILTON: Dr. Imrey has touched on two points discussed for many years. The first one involves correcting reversals, and I am sure at least a dozen papers have been written on that particular subject. The problem is, will we accept remineralization of caries lesions, and do we consider white spots as caries lesions which are remineralizable? Thus, a change from a carious surface call to an intact call at subsequent examination can be a physiological reversal and not a statistical error or error of measurement. I think Dr. Imrey's point about this biasing in one direction but not affecting the outcome is an interesting one and should be discussed if not investigated.

The second point brings into question one of the sacred rules which so many of us have used in dental clinical trials – that is, not to have recourse to the previous examination records. This might bias the observer so that he doesn't make a fool of himself by showing that he is inconsistent. We also have such things as gross illogicals, when a tooth which is missing at the first session is now present in the second session – in other words, a false call of what tooth was present. Many of us have resolved that situation by later re-checking the mouth. If a tooth listed as missing the first time is present the second time, this is considered corrected for all future examinations.

IMREY: I would just like to make two further comments: One is, I'd like to protect myself by saying that I am not offering these as solutions to the problem, but I am offering them as things which are worthy of examination or re-examination. In relationship to your comment about not wanting to make fools of ourselves, a good way to state it is to say that it proceeds from the philosophy that it doesn't matter if we make fools of ourselves, providing we do it uniformly.

CHILTON: We don't want to make random fools of ourselves. Your point about a bias occurring which may not influence the final result is an important consideration. We have always felt that a bias coming in might influence the result favorably.

ROSS: We all recognize it. It has been traditional that we re-monitor and re-calibrate examiners at pre-determined intervals during the study. The question has come up today of the value of this, whether it is a positive or negative. It is probably both. We have had it suggested that an examiner, knowing that he is being re-calibrated, sometimes tries to recall which subjects he has previously examined. Knowing that he is undergoing examination himself, this places stress on him by not wanting to make a fool of himself, and not wanting to appear inconsistent. Does anyone know of any study to indicate that the re-calibration of examiners has decreased the sensitivity of that examiner and perhaps decreased the range in which he is willing to make his calls?

LU: My name has been called in vain several times today for a paper I presented in 1968. I blush for that paper, because that was my thinking many years ago. I have long since changed my mind about many aspects of that situation. Nobody says in the assumption of that paper that the error is independent of the condition of the decay extant on the surface. All I was saying is that there are two extremes: One extreme is completely healthy, and even I can tell you that, and on the other side, I couldn't mistake it either. In the middle, that's where you have a gray area. A good examiner has a narrow gray area. A bad examiner has a wider one. The problem is, we talk about reliability, reproducibility, dependability, and so on, but these are not synonymous for the word "accuracy". If you are accurate, by definition, you are reliable and reproducible, not *vice versa*. For example, suppose we have one patient and two examiners. They each examine this patient's mouth, and there are ten DMF. They are in agreement, but they each come with a different ten teeth. That's terrible. Then they look at reliability – that is a very dangerous practice. You can be guaranteed to be reliably wrong all the time. Reliability would be useful if, and only if, you have very clearly defined diagnostic criteria and a very faithful adherence to them. Regarding the concern with reversals, I want to put this issue to rest. There is no shame in having reversals. In fact, you expect more reversals in the treatment group if the agent is worthwhile. If it is effective, it would produce more uncertainty.

HOROWITZ: I want to go back to Dr. Imrey's using the

analogy of detecting breast tumors, that maybe we should re-think having previous examination results available when we do follow-up examinations. This has always been a no-no, and I think for a very good reason. We have error in terms of reversal in diagnosis. We also have error going in the other direction of something that is called sound on the first examination, which becomes diseased on the second, erroneously called. If we have previous examinations and have our level of detection of dental caries or breast tumors increased by having this previous information, if the examinations are done blindly, we may not bias *per se* the examinations of any particular individual, but if we are dealing with treatments of those breast tumors or dental caries, as Kuo Hwa pointed out, we will be getting more breast tumors or more caries developed in a placebo-treated group, or a less effectively treated group than the most effective group. Therefore, there is no way of having previous examinations that will not bias the data when your level of acuity is increased by having previous examination results.

MITROPOULOS: There are a couple of studies which I have done which might help to elucidate a couple of points you were talking about.

The first one dealt with having previous data available. We looked at radiographs as a series as opposed to looking at them separately at each annual examination, so I was looking at four sets of radiographs together, rather than separately. My findings were that by looking at the radiographs all at the same time, I increased the ability to detect the two groups, and to evaluate the reversals. In the second study, I studied the effect of different light sources on the examination for dental caries. I did the study where I examined the subjects, and we looked at the effect of memory on examinations. We found that we were less able to detect caries in natural daylight as compared with fiberoptics. Regardless of which light source we looked at, though, when we looked at the child for the second time, there was more diagnosed caries.

O'NEILL: Something that has concerned me for a while is particularly important in view of the caries prevalence and incidence rates being so low now. If you look at it as the number of surfaces at risk, it is perhaps on the order of 5%. The misclassification rates seem to be at least at that level. I think Dr. Goldberg referred to it earlier when looking at which model might be most appropriate to evaluate the data. You have to wonder at what point the misclassification rate starts to demonstrate the actual effect you are trying to see. It has a particular impact with accurate control studies being bandied about. There is essentially no price one pays for bad performance or misclassification, except that it might be reflected in a variance component which then either gets reflected in a poor power of the test – which essentially goes in favor of not finding the difference – or it goes in terms of "blowing up" the width of the confidence interval. If you take an approach which restricts the width of the confidence interval as an acceptable clinical and meaningful difference between treatments, then I think you really have to start to address the issue of the active control studies and the role of misclassification. It is not something to treat lightly. I would like to hear some discussion of this, if there is any thought on how high the misclassification rate has to be to fall within the finite rates that you are now seeing.

CHILTON: I seem to recall Albert Russell stating about 25 years ago that if the rate of reversals went above about 15% he began to look askance at the acuity of the examiner. Clinically, the impression has been that if there is a relatively large number of reversals, the examiner should be recalibrated.

M. SMITH: I'd like to comment on Norton Ross's original question about new areas and our research. We have done most of our studies to maximize the effect over a very broad category of individuals. I wonder if it may not be time to try to maximize the fluoride levels in terms of benefit to smaller subgroups than we have looked at previously. It doesn't seem to me that the same dosage of whatever regimen we are using should be a maximum for a 14-year-old female as for a six-year-old male. I think we need to do some research into what levels and what compounds are applicable to individual subgroups and search for some of the covariables that may tell us that. Maybe it is time now, really, to fine-tune the fluoride machine.

CHILTON: That might be discussed tomorrow in "Improvement in Selection of Study Participants", by Drs. Downer and Mitropoulos. Perhaps we can hold that until tomorrow.

DISNEY: I'd like to make a comment on the monitoring of examiner reliability. Doing a clinical trial or a demonstration program is very expensive and time-consuming. There is an enormous amount of resources that are used. I simply can't imagine anyone expecting to do something and not have his work evaluated. During our studies, we have attrition, and we simply didn't need as many examiners each year. Nobody wanted to quit because he was being monitored. I think, in our experience, examiners' egos aren't quite that delicate. Secondly, I don't think that the monitoring necessarily biased the results terribly in one direction. If there was a disagreement, there was an open exchange between what one examiner saw and his interpretation criteria. So I don't see it quite as negatively as I am hearing it discussed. In discussions of analyzing data from caries clinical trials, reversals have all been examiner error, and I hope there is going to be some discussion of reversals due to remineralization.

CHILTON: Since I am the only one here who participated in the first (and second) Conference on Caries Clinical Trials, I can state that this same point on how to analyze data based on remineralization of white spots has been discussed at each of the two previous Conferences. My view is to analyze the white spot data separately and also report the results separately.

WORTHINGTON: I have a point on the interim examination papers by Professors Poulsen and Greenberg. One of the points incorporates the stopping of clinical trials so that the data can be tested and the trials stopped if a result is found. One of the advantages in doing that would be to make sure the members of the control group weren't deprived of a beneficial agent.

O'NEILL: There has been a considerable amount of work subsequent to McPherson's cited by Professor Greenberg. I would put the Roe and Pocock repeated-measures work in the same bucket. There has been work done by the NIH on the problem of stopping procedures in clinical trials, and they have taken some work from O'Brien/Fleming, who published a paper in *Biometrics* about 1½ or two years ago. Essentially, the difference between their approach and the approach of McPherson is that you (as suggested by Helen Worthington today) use a constant alpha level every time you look at the data. So, were you to say that you were going to look at the data four times, you would decide that beforehand, and then use the Z value you get every time you look at the data, against a constant critical level. It may not be 1.96 but 2.4 to control for that, The NIH approach, based on O'Brien and Fleming's, is that you have a sliding scale of critical values which were much higher early on. You really must have an extremely large

difference to justify terminating the trial early on, so that when you get to the end of the study, you essentially have the same critical level as you would have had if you started off with one fixed look at it. One of the problems is that with the approach you are suggesting, it is hard to get that across to clinicians. If you were going to finish the trials and then you are on your fourth look, you have a Z value which essentially, under the fixed one look, would give you 1.96. Then you tell a clinician that you can't reject a known hypothesis because you looked at it four times; that's a tough thing to get across. Why should you pay the price if you didn't really look at it the first three times? Here I am at the end of the trial, and I let the trial go the entire time, resulting in a statistically significant difference. Another way of looking at it is to have a sliding scale. Intuitively that's a good feeling. There has been subsequent work done by NIH and published in *Biometrics* by Mitchell and Larry Rubenstein and others in the past year.

CASH: I have a reply for Dr. O'Neill concerning his comments on the role of misclassification. I hope it is not quite as bad as he may have thought it was. I don't think that the 5% of caries surfaces over the total surfaces at risk can be compared to the rate of misclassification, which may be 5%. I think there are two different percentages. The misclassification is based upon what's actually present. Its denominator is the 5%.

O'NEILL: That's true, but I guess what I am talking about is maybe the increment. If you start off with a prevalence rate of something on the order of, let's say, ten surfaces per 100 that are carious at baseline, at the end of the study you have 15 surfaces that are carious. Essentially, the difference between those is five surfaces, essentially a difference of 5%. But if every one of those 100 surfaces was looked at and it had a chance of being classified carious or not carious, and you are talking about the rate of misclassification on a surface-by-surface basis, I guess my question is, at what stage do those two, the misclassification rates, the false positive, and false negative start to get in the way of the true difference?

CASH: I can't answer that because I don't know what has been done to classify the rate of misclassification or to estimate the rate of misclassification. I know there must be studies around.

GOLDBERG: Actually I think that's a very serious problem in the range of prevalences that you are looking at. I have looked at the effects of false negative and false positive rates on estimation, and on the difference between two proportions as well as on some other things in the range of low prevalences or low incidences. Basically, if the misclassifications (false negative and false positive rates) are equal in the two groups you studied, what you do is always look at the true difference. On the other hand, however, if the rates differ in the treatment and control groups, the effect on the estimate depends primarily on the false positive rates. The effect on the estimate of the difference between two rates depends primarily on the difference between the two false positive rates. Since that gets somewhat large, even though the absolute rates are very small, you wind up turning your study around, in effect. I don't remember the exact levels, but I had a paper in the *Journal of the American Statistical Association* in 1975 that details the effects on incidence. It is an incredible problem. Once it comes to that, at 5% positive rates, this can turn the study around.

GREENBERG: I had a double problem with that. That is the concept in the way you are doing the trials now. You have no standard, basically. It is a question of what you are calling false positive and false negative. That was raised before by several people who are looking at it. It is reproducibility, but if you are reproducing wrong, you are not accomplishing anything.

CHILTON: How would you classify it if a surface is classified as carious, which would be positive, and then at three subsequent examinations it is classified as negative?

GREENBERG: I would say that's negative. I have several other comments. For purposes of evaluating errors, you can do them on a sampling basis. You can sample some of the subjects for re-study to get estimates of the error rates for certain observers at certain times. You can have panel agreements and review the data while the patient is still there. Along the line of the thing that Dr. Imrey was talking about earlier, on that second examination the method would be to do it blind but then have an immediate feedback with the initial record and that subsequent one. Then you can make some consensus, maybe call in someone else to review it right at the time, so you can walk back into the room and see what the story is before it is too late and get a handle on the problem that way.

CHILTON: You take a small sample of the group, of course.

GREENBERG: The two depend on having a lot of available resources. There are some new kinds of approaches that you can take toward estimating those quantities.

SCHEININ: Mr. Chairman, with regard to the white spots, a few comments. First of all, I would like to argue about whether to include or not to include. They can be handled in any way once they have been recorded. Obviously, there is a prognostic value when recording the white spots from the clinical sense. There have been studies where a high diagnostic prediction value occurred with regard to development into clinically overt lesions. A caries index score (CIS) was developed, using white spot lesions found on surfaces by use of the stereo microscope. At 160 times' magnification, they diagnosed these lesions in terms of numbers ranging from 0 to 3. Three was the white spot with cavitation. I have personal experience with using this index. Unfortunately, it suffers from not being very sensitive when just one number is assigned to assess a total surface. A further development would be to carry out parametric analysis based on clinical photographs, which is possible and which has been done, but which is extremely time-consuming. A further refinement would be to use the Koulourides enamel slab technique and see how the treatment regimen will affect these slabs over a period of time.

I am still waiting for a comment on my question this morning: Are blind studies really possible?

GLASS: Some time ago, we carried out a clinical trial that involved a stannous compound. We were concerned about possible staining destroying the "blindness" of the study, so we gave every subject a prophylaxis immediately before the examination. I think that this particular approach, although very expensive and perhaps impractical, provides a possible answer to this question about providing a more or less uniform hygienic scheme prior to the examination, which would be a step in that direction, although perhaps it does not answer Professor Scheinin's question 100%.

CHILTON: A number of the suggestions Professor Scheinin made might be very worthwhile to add to a clinical trial by taking a small subsample and performing these detailed procedures. Instead of examining only using our standard manner, special examinations might be included

in concomitant studies using subsamples of the various "treatment" groups.

RIPA: When Dr. Disney discussed reversals, this was interpreted as primarily concerning white spot lesions. When I think about reversals these days, based upon the type of caries that we are seeing, I think about reversals in pit and fissure areas. While we may have been discussing reversals since 1955, I think it is still a very perplexing subject. We didn't have the prevalence levels that we have now. It is certainly conceivable that whatever background factor is affecting the prevalence levels may also be affecting reversal rates. People are saying it is due to a fluoride effect, and it is very conceivable that reversals we may be encountering are true reversals and are related to whatever the background factors may be. With respect to the intervention agent, I think the prospect of having true reversals must be seriously considered before we get into a discussion on the mathematical models that are used to resolve it. I think it has to be resolved first at the clinical level.

CHILTON: In a number of clinical trials, we have seen errors made by the failure of the examiner to recognize that there has been something placed in that fissure, making it look normal. Knowing Dr. Ripa's great acuity, plus his experience working on adhesives, I'm sure he is not one who found the reversal where an adhesive had been placed.

BURCHELL: Returning to the question of interim examinations, most new caries occurs on teeth which are erupting during the trial, particularly when we are dealing with 11- and 12-year-old children. If we have the interim examinations, we can find out when they erupt and take other parameters aimed at modeling the data. In modeling these data, we find out not only about product differences but also about how and where they are occurring. I put in strongly for maintaining the intermediate examinations.

CHILTON: It is a very interesting point that you presented, particularly since you come from an organization that pays for these studies. If a sponsor is willing to pay for the examination, that's a very important factor.

HOLLOWAY: I would like to support one of the previous discussants who made a plea for a definitive document on the need for the measurement of examiner reliability and the methods for doing this. For example, it seems to me that you should take into account that we are taking up the subjects' time in bringing them back and examining them. I think this is an ethical point that we need to examine. Again, we are increasing the resources of the study by requiring re-examination of the subjects. It seems to me that today we have talked about two entirely different situations. On the one side, we had the multiple examiner studies, where I think measurement reliability is very important indeed, and on the other side we have this sort of one-examiner study, which can be done by a very experienced examiner – for example, somebody like Lou Ripa. I wonder how long that sort of examiner has to go on repeating and repeating reliability studies in study after study after study. I wonder if there was ever a case in which an experienced single examiner quoted previous reliability data in support of the fact that he is an experienced and reliable person.

CHILTON: How many years does it take to make a Lou Ripa?

RIPA: I will ask my mom.

CHILTON: You should ask your dad, also.

GLASS: Bob O'Neill raised an interesting point when he spoke about incidence rates of 5% during a trial. John Peterson and I have just reported a clinical trial carried out in a fluoride area with incidence rates on the order of 2-2.5%. This is half the rate that Bob mentioned. But when we report clinical trials, we do include a measure of the so-called error rate or reversals. In this particular case, it was around 0.05-0.07%. As caries incidence decreases, the error rate may approach the actual incidence rate, and that is what Bob O'Neill was commenting about. As the late Arthur Radike used to say, an examiner is either a good examiner or he is not a good examiner, and no ounce of calibration is going to turn a poor examiner into a good examiner. Since our 1968 Conference, I have participated in a number of studies in which there were, thank goodness, two examiners, and had there been just one, on some occasions the study would have gone down the drain, like so many others. But I think the concept of negotiating or having a committee make every individual diagnosis is somewhat on the absurd side. We have to rely here on the law of large numbers. Again, in relation to errors (a term I prefer to reversals), it is impossible to differentiate between those errors which might be actual physiological reversals and those that are true errors. On the other hand, in the case of the errors that we do detect, we see only those errors which are false positives, for want of a better term. We do not see those errors which should have been diagnosed as carious but were in fact false.

I attempted to make the plea this morning to attempt to extend the sensitivity of our diagnostic criteria, at least on an experimental basis, in the meantime, by using the white spot technique within its proper context. There is something to be learned about this technique as well as possibly developing a severity score on which one develops an algorithm for weighting the different types of new caries observed, according to the relative susceptibility of the different surfaces involved.

CHILTON: I noticed your comment, Bob, about the errors caused by not calling "caries" caries. That would be a negative error. As Dr. Goldberg mentioned, she has found that positive errors throw off the study much more than do negative errors, so I feel better about that.

GOLDBERG: I wanted to add something about a single examiner. I was working on a study of liver biopsies with four authorities who wrote the textbook. Each one agreed with himself and not with the others. We are now trying to figure out why. It was shocking to all of us. The advantage of using the one perfectly consistent observer is that whatever error is happening is happening consistently in both groups, so that whatever there is going on is occurring in exactly the same way if it is in no way confounded with the treatment. But we don't know that. So, therefore, there is a little virtue when only one observer is involved.

CHILTON: There being no further points of discussion, therefore everything is settled.