SOME REMARKS* ON THE REPORT OF THE TECHNICAL COMMITTEE** ON BROADCAST RATINGS ENTITLED "EVALUATION OF STATISTICAL METHODS USED IN OBTAINING BROADCAST RATINGS"

By William G. Madow, Stanford Research Institute

1. Introduction

On March 11, 1960, the Hon. Oren Harris, Chairman of the then existing Special Subcommittee on Legislative Oversight of the Committee on Interstate and Foreign Commerce and also Chairman of the latter Committee, in a letter to Dr. Morris H. Hansen, then President of the American Statistical Association, included the following paragraphs that summarize very well the motivation, objectives, and conditions of work of the Technical Committee on Broadcast Ratings organized by the American Statistical Association in response to Congressman Harris' letter:

"...the Subcommittee recently has had occasion to consider the existing statutes and regulations, or absence of them, applicable to the character of the programs which licensed radio and television stations are broadcasting over the public airways. It appears from the testimony that the choice of the kind of program broadcast over networks during prime viewing hours has often been predicated upon public acceptance of preference as indicated by certain 'ratings' ascribed to programs by certain 'rating services.'

As it is clear that the determination of any such ratings must be derived from statistical procedures involving sample surveys, our Committee has requested you to arrange for an examination and evaluation of the statistical methods used by the principal rating services. It is my understanding that you have taken this request up with your Council and that they, and you, recognizing the public interest and professional responsibility involved, have agreed to designate a group of scientists which would make an independent study for us. I understand that the group would act in its own capacity, being free to present its findings without your prior review or that of our Committee." (Report, p. 1)

The Technical Committee on Broadcast Ratings came into existence late in March, 1960 and submitted its report early in March 1961 to the American Statistical Association, which in turn transmitted the report to Mr. Harris. The Technical Committee, its task having been completed, ceased to exist in June 1961.

Seven rating services were considered: American Research Bureau, Inc.; C. H. Hooper, Inc.; A. C. Nielsen Co.; The Pulse, Inc; Sindlinger and Co., Inc; Trendex, Inc.; and Videodex, Inc. Without exception, they were helpful and cooperative.

Although the report is also concerned with radio ratings this paper will, for simplicity, consider primarily television ratings. This paper is intended to summarize my views on some aspects of the report.

2. The Setting of the Study

The quotation from Congressman Harris' letter makes it clear that while the desire to have an examination and evaluation of the statistical methods used by the rating services in estimating the ratings was motivated by problems of programming policies, the task of the Technical Committee included neither the examination nor the evaluation of programming policies, nor was the Technical Committee asked to estimate the ratings that would be received if different programs were available. Our responsibility was solely to examine and evaluate the statistical methods used by rating services in estimating the ratings. No limitation was placed on the extent of our examination of statistical methods.

As remarked above, the rating services cooperated with us fully throughout the study. I am sure they often wished that we might merely use material that had been given others and not press them further. However, they did cooperate. Whatever omissions there may be in our knowledge of their statistical procedures are, I am sure, due to our not inquiring rather than to their not being willing to provide the information.

There was one aspect of this study, however, that the Technical Committee felt to be important and for which we turned, not primarily to the rating services, but to those who are their actual or potential clients. For what purposes are ratings used and what ways are the ratings used?

To evaluate ratings by some absolute criterion such as, for example, a standard error of less than so and so taking into account all possible sources of error, might be desirable but it might also be that relations among ratings suggest conclusions even when standard errors are not available. Requirements on data are relative to their uses just as requirements on any communication system are relative to the uses to be made of that

---

*Prepared for a panel discussion of the report at a meeting of the American Statistical Association, December 27, 1961.

**The Technical Committee on Broadcast Ratings, a Committee of the American Statistical Association, consisted of William G. Madow, Chairman, Stanford Research Institute; Herbert H. Hyman, Columbia University; and Raymond J. Jessen, General Analysis Corporation (now CEIR, Inc.); assisted by Paul B. Sheatsley, National Opinion Research Center, and Charles R. Wright, University of California, Los Angeles.

The report of the Technical Committee on Broadcast Ratings was published as House Report No. 193 of the 87th Congress, First Session, entitled, "Evaluation of Statistical Methods Used in Obtaining Broadcast Ratings," A Report of the Committee on Interstate and Foreign Commerce, Oren Harris, Chairman. All references are to the report.

system. In practice, the uses of data depend on much more than the data themselves--and the requirements on the accuracy of data are naturally affected by both the importance of the decisions they influence, and the extent to which they influence the decisions. I am glad to say that, despite occasional grumbling, we received a great deal of cooperation from those sponsors, advertising agencies, networks, and broadcasting companies to whom we wrote about the uses, and I am sure that had we had more time and been able to accept some of the invitations we received to visit with those who replied, we could have obtained more information.

The Technical Committee was not asked to investigate, evaluate, or compare the rating services with one another. In the study, we tried to avoid making comparisons of the rating services. I should like to caution against using parts of the report to compare the services with one another. We did not write or review the report in such a way that the statements we made about different rating services are comparable, except perhaps for the populations covered in rating surveys. Each rating service was considered within its own framework. Thus, a service that attempted little might have less criticism in the report than a service that undertook to and did accomplish more, but in so doing provided more occasions for criticism. This should not be taken to imply that the rating services are, in our judgment, equally good sources of estimates of ratings.

I should also like to stress, as we did in the report, that there may be errors in some of the details we give concerning the practices of the services. While the errors are not important for our evaluation, they obviously may lead to erroneous conclusions if used to judge or compare rating services.

When we began this study, the Committee had hoped to submit its report in draft form to the rating services for their comments before the report became final. For various reasons, including the fact that neither the American Statistical Association nor the Congressional Committee was to review the report prior to publication, it was deemed best at the end of the study to issue the report without comments in advance from the rating services.

The report would certainly have benefited from being submitted to others, including the rating services, for comment. We did ask the rating services for comments after the report was published. Those that we received did not disagree with our conclusions although, as intimated above, there were some errors in details and there is some feeling that the report may appear to have been more critical of one service than another at various points. On the whole, the comments have not been unfavorable.

It is worth noting that the rating services are well aware of the defects in their surveys and that they seem to inform their clients. My impression is that they face severe pressures in the timing and costs of their surveys, and it is these pressures rather than lack of interest or knowledge of improvements that lead to the existence of the defects to which we call attention. Also, my impression is that many of their clients are well informed of the possible defects of rating surveys, and use them with these in mind. Again, I want to caution against any implication that all rating surveys are equally good or bad.

In the following remarks I have selected some of the major aspects of the report for comment. Further information is given in Dr. Jessen's paper. Details will be found in the report.

3. A Summary View of the Uses of the Ratings in Relation to the Statistical Methods Used by the Rating Services

To obtain material on uses, we wrote to some sponsors, some advertising agencies, some broadcasting stations, including chains, and to the three major networks. No sample was selected since we were primarily concerned with whether a picture of uses emerged that could be synthesized.

Some parts of the industry claim that ratings are not the major factor in their decisions but only a portion of the evidence they use, some parts claim that ratings or cost per thousand are a major factor if not the sole factor, in decisions affecting them. Both statements are correct; they apply to different parts of the industry. But they can be effectively combined for our purposes in the statement: Other factors being equal, the program having highest rating, or lowest cost per thousand, is preferred. The other factors considered vary from none to some that are possibly more important in certain cases than the ratings themselves.

Omitting non-numerical factors in preference, the other numerical variables most often cited besides ratings, cost per thousand, circulation, share of the audience, coverage, and so on were the composition of the audience by sex, age, income, geographic location, and other demographic variables. Many users stressed the importance of trends as well as level of ratings*.

Another way of looking at the ratings is that the organization that makes many small decisions, for example, an organization that buys many spot announcements, will be likely to base the decision primarily on the major readily available data such as ratings and cost per thousand. If the organization deals with smaller markets, it cannot obtain trends very easily; the surveys aren't made often enough.

The organization that makes large decisions spends more on each and may bring in many factors such as, for example, balance of programs, strategy, image-product compatibility, and the results of special surveys. It does not rely so heavily on ratings, and cost per thousand, and trends in them as the organization that makes

*We shall use the word "ratings" to refer to the many different measures cited above.

many small decisions. But no organization ignores ratings, cost per thousand, and the other data cited above.

To summarize, ratings, cost per thousand, and related data including audience composition and trends in these data, are important to all parts of the broadcast industry. The more costly or rewarding the decision, the more that is spent on auxiliary information that reduces dependence on the ratings. Organizations with greater resources, knowing the possible faults in ratings and either depending on large smaples or on many smaller samples for the economic consequences of their decisions, treat the ratings as one part of the information on which they base their decisions. Organizations such as stations or advertisers in small communities have smaller resources and are more dependent on the results of a single rating survey--and therefore on the statistical methods used in a single local survey rather than on those used in national samples or on the average effects of the statistical methods used in many local surveys.

I should like to stress that I am not speaking just about a large enough sample versus a small sample, but about the greater dependence of the small user on a single chance event. Rating surveys are not made weekly or monthly in smaller market areas.

Finally, although some users doubtless use technical statistical, or operations research, or mathematical-model approaches to decision making to some extent, no such methods were mentioned by the users.

4. Over-all Evaluation and Recommendations

The Technical Committee found many details and a few fundamental matters to criticize concerning the statistical methods used by the rating services and the descriptions of methods and quality of the ratings published by the rating services.

Let me first consider two major policy issues that are often raised concerning the ratings and that are directly related to the statistical methods used by the rating services. These issues require an over-all evaluation of the effects of the various criticisms that we made of the statistical methods used by the rating services--and it is to over-all evaluations and recommendations that this section is devoted.

I should like to present my own interpretation of some of our conclusions.

For the big policy question--whether the ratings are sufficiently accurate to reflect the preferences of the audiences nationally for cultural as compared with non-cultural programs from among the programs available to the audiences --the ratings are sufficiently accurate, when, despite the different methods used, they are in agreement, as they seem to be on this issue. If anything, the incomplete coverage of the population is likely to increase the proportion of households claiming to prefer cultural programs, except that at least some services do not report the audience of educational television stations.

A second important statistical issue is whether the ratings are sufficiently accurate for comparing programs of the same type with one another, or stations in the same market area with one another. Ratings of so many programs are published, and the ratings of so many stations are compared, that we can be certain that even though many comparisons of programs and stations are correct, many are incorrect. Many apparent differences of ratings, many rankings of ratings, and many apparent trends in ratings are, in fact, just results that could occur by chance. Where samples are small, as in small market areas, or where ratings are close to one another there is greater likelihood of error.

"Error" tables, including both sampling and non-sampling errors, are needed for the comparison of stations and programs. The Technical Committee felt that the rating services were not publishing or otherwise making available sufficient information to permit satisfactory use of their "error" tables, nor are their "error" tables sufficiently extensive. In the report, we listed various recommendations to remedy what we considered to be inadequacies in this area. The inadequacies were primarily in the omission of estimates of non-sampling errors, and the omission of any treatment of many comparisons rather than individual estimates. To rectify these inadequacies will probably require some research. In addition, the rating services should publish more educational material and guides for the use of ratings, particularly as measures of trends.

In connection with both of the above conclusions, we should like to point out that the statistical defects of the rating surveys are likely to have a much more serious effect on the so-called qualitative information, namely, age, sex, size of family, income level, and other demographic characteristics, than on the ratings themselves. In view of the frequent statements of increasing reliance on these qualitative characteristics we feel it important to call attention to this danger.

These are two "big" conclusions from the Technical Committee's report. Now, what did the Technical Committee propose be done that would improve the situation?

The Technical Committee felt that even though the ratings as currently made were useful for many purposes, there was much that could be done to improve the ratings and the understanding of how to use ratings of individual programs and stations. Also, it seemed clear that in view of the frequent changes in the kinds of ratings issued and the new problems and uses of ratings that would occur, the interest in the quality of ratings should be permanent.

One source of improvement in the ratings is competition among the rating services themselves. But we felt that improvements might be more

rapidly made if certain recommendations were adopted. Three major recommendations were made. These are given in the report in greater detail, but are restated here:

(a) The rating services should increase the information they furnish on how they estimate the ratings and related data and on measures of the quality of the ratings and related data, and they should regularly publish such information. Obviously, as the rating services increase the information on quality, their clients will have greater means of understanding the quality, improving the uses made of ratings, and achieving the increases in quality required by those uses. "Error" tables should be constructed for the demographic characteristics of ratings as well as for the ratings themselves.

(b) The rating services should not only do research, but should publish research on estimating and using ratings. The services now do methodological research, and they make some research available to their clients, but it would improve the quality of ratings and their uses if methodological research were published.

(c) The broadcast industry should develop an industry-wide Office of Research Methodology--either to conduct or support a program of research in making and using audience measurements of quantitative and qualitative types.

My own summary is that we thought the ratings on the whole are useful but could and should be improved, and we suggested some approaches that should result in improved ratings. How the individual rating services respond to our recommendations will depend on the pressures on them and on the economics of the situation. I think we suggested how the broadcast industry can proceed so that it can expect steady improvement in the estimates, uses, and understanding of the ratings.

Let me now turn to some of our specific criticisms.

5. The Statistical Methods Used in Estimating Ratings

Different definitions are used by the rating services in estimating ratings. None is necessarily better than the others and the variability resulting from differences in definitions is probably not a major cause of differences among ratings.

However defined, a rating of a program is usually the percentage of households with television sets in a specified area who were part of the audience of that program.

In putting such a general definition into practice many variations occur that affect the measurement.

The area might be the United States or a metropolitan area, or a market area, or an area served by television stations on a coaxial cable permitting the simultaneous broadcast of the programs.

Being in the audience may mean that the set was tuned to the program, or that at least one person in the household was viewing or listening for at least so many minutes. The audience may be the average audience say, per minute, or the total audience in the time period.

Ratings may be based on households and persons satisfying socioeconomic criteria, such as being in an income class or being teen-agers, as well as on all households or persons.

For purposes of the present comments, it will be desirable to omit consideration of the many different definitions of ratings and of related measures such as share of the audience and coverage. It is important, however, to recognize that differences in concept, area, and population surveys and in measuring instruments all contribute to the variability of estimates in addition to sampling.

Our primary criticism of rating surveys is in the difference between population to which ratings are applied and the populations from samples of which the services actually obtain information.

We do not have data with which to estimate the biases in ratings and related data resulting from this cause. But, it is probably important and should be estimated, at least from time to time. These biases should be discussed or approximated in connection with "error" tables.

We found that all the methods in use at the time we made our study about a year ago, resulted in effectively taking samples from between 50 and 70 percent of the population.

If diaries or meters are used, then there are difficulties in obtaining cooperation, and not all of those who agree to these methods actually produce usable data.

If telephones are used, then only about three-fourths of the households have access to a telephone and the proportions with access to a telephone differ considerably in various geographic and socioeconomic categories. Furthermore, some households with access to a telephone are not listed in telephone books. Telephone calls to toll-call areas may be omitted, or call-back telephone calls to such areas may not be made.

When personal interviews are made, there may be no call-backs possible without being involved in a recall period of more than 24 hours.

Now the audience characteristics of those who provide data in these surveys may well differ from those who do not. We know that for some of these methods their demographic characteristics do differ (Report, Ch. 4), and we believe that such differences exist to some extent for all methods.

While the requirements of rapid reporting may lead the rating services to continue to sample from these pseudo-populations, they should at least from time to time make surveys to determine the effects of their use.

All the rating services are aware of probability sampling methods and all use them--or at least use methods that should be equivalent to probability sampling--but I felt that they tended to cut corners more than desirable. In particular, there is some tendency to omit "error tables" or to use "error tables" that do not correspond to the sampling methods and estimation equations. Not all rating services are equally to be criticized. At least one has made great efforts not only to use probability sampling but also to have its "error tables" correspond to the procedures it uses. But there is much for the rating services as a whole to clean up here--even though, as mentioned above, we feel that the ratings are adequate for many of the uses to which they are put.

Another major criticism of the rating services is in the lack of objective evidence of quality control of the process of obtaining information. Respondents, enumerators, and data processors all make errors. They make them often, and the errors may be serious. While the cost of estimating the biases and variability due to such errors may be too great for a single survey, the rating services are continuously collecting data, and are using methods subject to error. However carefully they are organized, there is no substitute for the planned study and reporting of biases and variability arising from human or machine errors.

6. Summary

It is certainly unfair to some rating services to generalize about all of them. And obviously only a small part of the report has been discussed here.

My own summary is something like the following:

If you don't like the major programming policies, don't blame the ratings. If anything, I would expect the preference picture to be worse for cultural programs than the ratings show.

If you think that the rating services should use larger national samples, well, any sample is better when larger if no loss in quality occurs, but it will not have much effect either on preferences of types of programs or on major programs. Current national sample sizes, although not too large, are not too small.

If you think that the sometimes erratic behavior of local market area rating surveys is evidence of unethical behavior on the part of the rating services, well, we didn't investigate them. But even without any unethical behavior, there could be large errors in the data from some of those surveys--errors that might balance out over many stations or areas, or over time.

If you think that the above statements imply that we approve the statistical methods used by the rating surveys, please note that we think we have made some severe criticisms and some recommendations that we feel the industry can and should take into account.

Like most of us, the rating services have been doing a job good enough for many needs, but they could do better. We have tried to suggest steps that would ensure their doing so.