RECEIVED
MAY 30 4 22 PM '00
POSTAL RATE COMMISSION
OFFICE OF THE SECRETARY

MPA-T-4

BEFORE THE
POSTAL RATE COMMISSION
WASHINGTON DC 20268-0001

POSTAL RATE AND FEE CHANGES, 2000	)	Docket No. R2000-1

DIRECT TESTIMONY
OF
KEITH HAY
ON BEHALF OF
MAGAZINE PUBLISHERS OF AMERICA, INC.
ADVO, INC.
ALLIANCE OF NONPROFIT MAILERS
AMERICAN BUSINESS MEDIA
ASSOCIATION FOR POSTAL COMMERCE
ASSOCIATION OF AMERICAN PUBLISHERS
COALITION OF RELIGIOUS PRESS ASSOCIATIONS
DIRECT MARKETING ASSOCIATION, INC.
DOW JONES & COMPANY, INC.
MAIL ORDER ASSOCIATION OF AMERICA
THE McGRAW-HILL COMPANIES, INC.
NATIONAL NEWSPAPER ASSOCIATION
PARCEL SHIPPERS ASSOCIATION
and
TIME WARNER INC.

I. Autobiographic Sketch

My name is Keith Hay, I am Professor of Economics at Carleton University in Ottawa, Canada. I am also the President of Econolynx International Ltd., a company specializing in economic research.

I was educated at the University of Southampton, in the United Kingdom; at the University of Toronto, in Canada; and Brown University, in the United States. I was a U.K. State Scholar, a Ford Foundation Fellow and a Killam Foundation Fellow. I am also a Fellow of the Foundation for Advanced Information Research in Japan. I have been "Visiting Professor" at the University of Southern California; York University, in Ontario Canada; and the University of Alberta (Japan Foundation), in Canada.

Over the last quarter century, I have undertaken some two hundred research assignments, often acting as an international consultant for such organizations as: the World Bank, the Asian Development Bank, the Inter-American Development Bank, the Bank of Canada, the Canadian International Development Agency, the Organization of American States, and numerous international corporations, trading companies and banks. I was executive assistant to Simon Reisman - - the "father" of the Canada-US Auto Pact and the Canada-US Free Trade Agreement - - during the period when Canada was formulating its modern free trade policies. Most recently, I have been working on the proposed Canada-Japan Free Trade Agreement, assessing the potential gains and losses.

I have worked for Canada Post on a number of assignments, most significantly, the development and maintenance of a large database of parcel competitor service standards, marketing incentives and customer rates. I serve as an adviser to several Canadian high-technology companies and I have been the CEO of a publicly quoted software company. I am a citizen of both Britain and Canada, and I live in Ottawa, Canada.

Perhaps the most significant experience I bring to these proceedings is the fact that A.T. Kearney employed me as the technical editor on the Data Quality Study. I was tasked with reading all the component studies compiled by the various experts to ensure that they read well individually, and that collectively they had some cohesion. As such I met often with the authors and discussed the various data quality issues at length. I believe this gives me an excellent insight into the subject of "Data Quality and Rate Making."

II. Purpose and Scope of Testimony

Data quality is fundamental to sound decision making based on sample statistics. Good decisions must therefore rely on good data. Postal rate making in the United States is viewed from the outside as setting "best practices" for the world, by adhering to the highest standards of scientific method and statistical application. Accountability, transparency, methodology and the ability of third parties to replicate statistical methods and sample results are the key-stones of the high standards desired by the Postal Rate Commission, the United States Postal Service and its end-user stakeholders.

The cost of mistaken decisions based upon inappropriate cost estimates could potentially be severe for the stakeholders and for the credibility of the US rate-makers. Moreover, there is no going back; once the standards for research integrity are lowered, the floodgates will open and science-based rationality will prove difficult to enforce in the future. While the desire for a quick answer or fix may be understandable,k the risk of making a mistaken decision is much greater to the shareholders and American consumers than any benefit of a quick answer derived from applying non-random and judgmental statistical procedures.

When there is no study design, a lack of pre-set confidence limits, weak adherence to consistent random sampling, no statistical cost study questionnaire, variable decision rules, no training manuals for enumerators or great concern about consistency of data collection, and only *ex post facto* attempts to get stake-holders to buy into results, then the interpretation of the arising results must be treated very warily. Recent work by Mr. Raymond and Mr. Baron reworking an Engineering Standards Study to produce inputs for cost-estimates appears to exhibit many of the afore-mentioned shortcomings.

In and of themselves, Engineering Standards studies have important roles to play in determining time and motion aspects of route performance. However, the data acquisition methods applied in ES research are quite different and often inappropriate for ratemaking purposes. This is a situation in which wrongly applied "any data" (arising from the ES study) may be worse than "no data" (from statistical cost studies) and the compounding of decision making errors cannot be justified. The budget costs – and delay -- of undertaking a scientifically sound, well designed, statistical study, as suggested by the Data Quality Study, in the immediate future are dwarfed by the likely value of the improvements in sample accuracy, data quality and avoidance of rate making errors.

In my testimony, I review some issues of statistical research in decision making; look at concerns about data collection methodology, and discuss the question "is any data better than no data?"

III. Value Of Research in Decision Making

a. Scientific Method in Statistical Studies

As is well known, there is a long history of the use of scientific method in survey research. Probability theory has been ably applied for almost a century to the issue of obtaining estimates of the parameters of a population based upon random sampling of that population. The structuring of the research project requires careful planning, which involves:

- consulting early with clients, end-users and decision-makers likely to be affected by study outcomes (stakeholders);
- reviewing previous studies/literature;
- determining a set of questions to be answered or objectives to be fulfilled;
- adopting the appropriate null hypotheses;
- establishing acceptable confidence limits for the desired results;

- selecting a random sample frame and method - - for instance stratified sampling, cluster sampling - - and/or multiple applications of these;
- developing a questionnaire with expert advice from the client, end-users and those who will base their decisions on the research outcomes;
- making sure that the answers fit the questions - - not that the questions fit the answers;
- running a pilot study to refine procedures;
- incorporating lessons learned from the pilot study;
- establishing decision rules to deal with sampling and data quality issues *before* they arise;
- devising manuals to guide enumerators and analysts;
- ensuring consistent methods of data collection across the sample strata or clusters by means of training, handbooks and logbooks;
- training the trainers and emphasizing continuity and consistency in quality control;
- recognizing the importance of moments of demarcation in activities subject to analysis and measuring them with a keen eye to precision;
- handling the data with care with a view to preserving the scientific integrity of the overall methodology; and
- presenting the results with suitable disclaimers as and when appropriate.

All of this, of course, to be achieved on a research budget which is always - - by definition - - too tight, and within a timeframe that is inevitably too short! These are not easy tasks. But in general, the stricter the adherence to the pre-designed research approach, the more likely are the results to be usable with known *confidence,* while the *quality* of the resulting data will more likely be acceptable to researchers, clients and end-users as a whole.

Research design, sample randomness, enumeration accuracy and overall transparency are fundamental to the ability to positively answer the question: "if another researcher independently undertook to answer these same questions with these same data, could the original results be *replicated*"?

b. Designing the Sample Frame

i. Randomness versus system

Statistical analysis is used to make accurate inferences about the parent population under examination. A sample is selected and observed for this purpose in order to know more about the population as a whole. Difficulties arise because of ever-present variation among elements of the population, such that successive samples are usually different. The task of the researcher is to come to appropriate and reasonable conclusions about the population while bearing in mind the issues associated with sampling variation.

The researcher must cope with *two* key requirements in carrying out the analytical task. The first is to design a sampling frame and undertake the sampling so that it is representative of the population, and the second is to use the sample results to draw correct inferences about the population. Clearly, it is most difficult to achieve the second objective if the first is not well done. Inferences are unlikely to be accurate unless the sample has been taken competently. Therefore, the sampling procedure must be acceptable before attributing to the population results arising from an analysis of the sample.

In general, for samples to contain worthwhile and reliable information about the population, *each* unit of the sample must be selected at *random,* requiring that each element of the population has a known probability of appearing in the sample. If selection is left to the judgement of the researcher, his/her associates or interested third parties and they exercise their own choices, then the probability surrounding these selections becomes unknown and the application of standard statistical procedures is confounded.[1]

A common procedure for ensuring randomness in a sample is to leave its drawing to a mechanical process, such as a random number generator, beyond the control of the research team and interested parties. This argument also applies when samples are stratified and/or clustered (as noted below). While we can admit that pure randomness is rarely attained in research practice, it is a fundamental aim of statistical research methodology, and invokes the mathematical model upon which the preponderance of statistical theory relies.

The closer the researcher can approximate randomness, the more nearly accurate will be the inferences drawn from the research study.

ii. Sample Size and Cost/Confidence Considerations

Given that procedures are in place to achieve a high degree of randomness in sample selection, a key issue is how large must be the sample size? If the sample size is too small – it may be too inaccurate to be reliable. Too large a sample may require the expenditure of too many resources while adding little extra information beyond what could be obtained from some smaller yet useful sample size. At issue is a determination of how large an error the researcher and his stakeholders can live with in the estimate. Moreover, the decision on an acceptable error also must take into account the uses to be made of the results and the potential cost and revenue consequences of different magnitudes of error - - for the client and other end-users who may ultimately be

[1] Non-probabilistic sampling procedures, such as *quota* sampling and *convenience* sampling, represent judgement samples, since they involve the selection of items in a sample on the basis of opinion, not randomness. When the population is small, or time/money will not allow collection of a random sample, or the study is strictly exploratory, then a judgement sample may be justified, but the statistical implications of abandoning random sample selection should be well understood, should be clearly flagged and should be expected to attract comment.

affected by the use of these results. This goes to the issue of *data quality* and the validity of inferences to be drawn from the data.

Put another way, the researcher should set up an allowable error, in terms of confidence limits, *before* designing the sample frame and deciding on the sample size - - overall, by strata and/or by cluster.

Once a decision of this type is made by the researcher and his stakeholders - - say that they are only willing to take a 5% chance that any error will exceed the allowable error in the sample mean - - then they have selected a 95% confidence limit for their study. With this decision in hand, there are then a number of ways to estimate what is an acceptable sample size for the research undertaking. These require bearing in mind prior information (from earlier studies or related populations), results of pilot studies, statistical methods for complex sampling, and budget constraints. Essentially, some advance estimates are needed of both the relative costs per unit of collection and expected variance in the strata and/or cluster under observation; rough estimates will often give sample size indications that are acceptably close to an optimum allocation.

Simple random sampling of a large population may be difficult to achieve, not least because it might prove very costly. More practical procedures may be employed recognizing that they will also be more restrictive and open to discussion and dispute. Among the methods that may be employed are:

- *Systematic sampling* – choosing a random starting point and selecting every *Kth* element to be an item in the sample;
- *Stratified sampling* – dividing a population into homogeneous groups or classes as *strata.* Each stratum is then randomly sampled;
- *Cluster sampling* – where the parent population is sub-divided into groups so as to design an efficient sample. These *clusters* ideally have the same characteristics as the parent population and are then randomly sampled.

iii. Stratified Sampling

The best method of selecting strata is to find groups *with a large variability between strata, but only a small variability within the strata.* Choice among and within these groups may then be based on a random selection method.

A *proportional* stratified sampling plan would use items from each stratum in proportion to the size of that stratum, to ensure that each stratum in the sample is weighted by the number of elements it contains, relative to the parent population. A *disproportionate* stratified sample may be an efficient device, if it is known that a particular stratum contains a high degree of variability that will yield a maximum amount of information for a given amount of research effort. The *weighting* of such results should reflect the proportionality or dis-proportionality of the sample strata.

iv. Cluster Sampling

The objective here is to obtain observations such that there is *little variability between clusters, but a high degree of variability - - representative of the parent population - - within each cluster.* If each cluster is assumed to be representative of the parent population, then the characteristics of the population can be estimated by *randomly* picking a cluster and *randomly* sampling elements within this cluster. Two-stage random sampling within a cluster is often effective and efficient.

v. Multiple and Sequential Sampling

When budget constraints impact sample design, it is often useful to frame a pilot study wherein only a small number of items are used to represent the parent population. If high variance is uncovered, then it may be valuable to undertake multiple-stage sampling, especially when the parent population is large.

The advantage of sequential sampling is reflected in the savings that result when fewer items than usual must be observed, say from a cluster within a cluster.

vi. Choice of Sample Methods

Selection from among several types of random sampling plans depends on the researchers prior knowledge of the parent population (and the results of previous research); namely the likely validity of *stratified* and *cluster* sampling to achieve efficient and confident parameter estimates of the population. Issues to address include:

1) What is the most cost effective method to collect samples that best ensures that the samples are representative of the parent population?
2) How reliable are the inferences and conclusions about the parent population likely to be drawn from sample information?
3) What are the best ways of describing sample information usefully while not-overstating the predictive power of the results?

It is the decisions resulting from incorrect inferences that can be costly, not the incorrect inferences themselves. Thus, there is a requirement on behalf of the client and stakeholders that the sampling methods employed minimize the cost of making an incorrect decision, or error.

At the end of the day, a primary objective of sample design is to balance the potential costs of making an error against the costs of undertaking sampling.

vii. Trial Sample Testing and Lessons Learned

Pilot studies, which use the overall research design, sampling methodology and questionnaire set-up, yield valuable insights. Discussion of pilot results with clients and stakeholders often refines the issues, tightens the project focus, and sharpens the statistical tools. It also ensures a higher degree of stakeholder acceptance of the research end results. Moreover, information on sample statistics gleaned from the pilot can be very helpful in deciding on the optimal overall sample size to achieve best value for money within the confidence limits acceptable to the clients and stakeholders. It also allows the researcher to test the decision – rules adopted concerning data quality, data inclusion/exclusion, and analytical methods. In summary, results from a pilot or trial sample usually reveal potential pitfalls in avoiding bias in the final results. Studies which neither explain the choice or and rationale for one of these methods of sampling, nor provide the target confidence limits should be viewed with concern. Only if the study is exploratory, or its conclusions regarding the parent population unnecessary, should these rigorous standards be relaxed,

IV. Data Collection Issues

i. Questionnaire design – "Answers to Questions" v. "Questions made to fit the answers."

It almost goes without saying that questionnaire design is very important to achieving useful results. Clients and stakeholders should be consulted. Badly designed questions elicit difficult to interpret answers. Any Canadian will give you plenty of examples concerning questions about "Quebec Separation" – how distorted do you want the answers to be?! Pilot surveys usually reveal unexpected questionnaire responses due to a poorly framed interrogatory. Re-wording will usually remove potential response biases. Perhaps the most alarming approach is to design the questions *after* the survey has been conducted (for another purpose) and attempt to make "the questions fit the answers" in some fashion. Since the enumerators *did not know* these post-survey questions, how could they exercise any quality control over what was being measured *ex.post*, or recognize any data deficiencies – random or endemic? This "cart before the horse" procedure leaves in tatters all the issues of errors in data collection, data exclusion/inclusion and decision rules, since the relevant questionnaire and its objectives were unknown to the research designers and the enumerators *until after* the data had been collected.

For example, none of the questions that Mr. Raymond answers in his cost study were posed to enumerators.[2] All answers recorded were based on a different "unspecified" set of criteria. This is an instance of a researcher fitting the observation tallies, i.e. "the answers" into a new set of questions – the six cost categories. How well he has done this is a matter of conjecture and divination. It appears as if the researcher is doing the complete exercise *backwards*. For reasons earlier discussed, it is not possible to offer any level of confidence in the sample or the parameter estimates arising therefrom.

A typical cost study questionnaire design would clearly specify the activity to be observed and the *points at which it begins and ends*. No such questionnaire exists for these data nor are there any relevant observational standards.

ii. Engineering estimates versus cost estimates

There is a remarkable difference between quantifying the number of sufficient time and motion segments for an *engineering study* of time use, versus quantifying the appropriate number of routes, by route type to develop a

[2] See L. Raymond, Direct Testimony on behalf of the United States Postal Service before the Postal Rate Commission, Washington, D.C. 20268-0001, Docket No. R2000-1, USPS-T-13 and his Library Reference to USPS-LR-1-163, Engineered Standards Database.

statistically valid sample for purposes of cost estimation and rate-making decisions.

Industrial Engineers (I.E.) use sampling techniques to measure distinct pieces of work, which are not necessarily the same as those used in cost estimates. The I.E. advantage, from a statistical perspective, is that the individual errors are not cumulative, so as estimates are added together, provided no inherent bias exists, the total error reduces. This enables them to measure individual work elements to a lower degree of accuracy than is called for in statistical cost studies.

In addition, I.E. estimates often exclude any time measure for inefficiencies or low productivity. As cost estimates capture these two elements it is essential that the sampling for cost studies be constructed so as to avoid any bias from these factors. The various aspects and distinct elements of load time cannot be merged together - - as in I.E. - - without recognizing that there will be significant losses in accuracy and variability for cost estimation purposes.

iii. Enumeration methods

The method by which Mr. Raymond conducted his enumeration of data for the Engineering Standards study was generally acceptable for that species of study. A systematic time interval occurring frequently enough to minimize the affects of regular break times, cyclical activities, was measured. However, Mr. Raymond had his enumerators also doing a variety of other activities, such as taking video pictures, recording paces walked, at the same time as tallying the observations. Tallies were given a lower priority than these other activities, with the enumerator entering the information from memory some minutes later. This procedure is unacceptable in a typical cost estimate study because potentially it magnifies the probability of error.

iv. Training manuals and log-books

Mr. Raymond has consistently said that no training manuals for his data collectors exist and that the only logs kept were the notes made by the enumerators on the daily records that are buried in volumes of other raw data sheets.

In a typical cost study all data collectors would pass the same training course to ensure consistency between enumerators, and each would commence work with a training manual to use as a reference document during the study. A logbook is normally kept in which work times, numbers of observations and anomalies, are recorded - - together with any changes that are made to the observations after-the-fact. These manuals and logs are key elements of any well-designed statistical survey.

v. Training the trainers

Where it is necessary for more than one trainer to be involved in training the enumerators, it is essential to identify the key points that must be focused on to ensure subsequent consistent observations by the various trainees, e.g. the load time begins at the moment that the letter carrier's feet stop moving at the end of a walk and ends at the moment that the foot is lifted to start away from a stop.

It should also be remembered that the majority of the training for Mr. Raymond's study focused on factors of importance to the Engineering Study, i.e. video training, how to enter the information with the bar code reader, how to identify the various activities and types of mail receptacle rather than maintaining the consistency and accuracy of cost-related data collection.

vi. Training the enumerators

Enumerators all need to be trained to the same observational standards if data are to be consistent across strata or clusters. In Mr. Raymond's Engineering Study a variety of different training methods were used, which were certainly acceptable for the work being undertaken – observations of the work activities for industrial engineering time estimates, frequencies, and percentage occurrence of various different activities. This training however, was inconsistent and woefully inadequate for data collectors working on a statistical study to allocate costs.

vii. Decision Rules on Data Acceptance

Elimination of any sampled data should only occur in extremely vexed cases, e.g. violent weather, power failure and the like, and in accordance with pre-determined decision rules. Excluded data are usually presented for review by clients and to other researchers attempting to replicate the study results. These procedures are not necessarily adhered to in Engineering Studies. They appear not to have been subscribed to fully by Mr. Raymond when using engineering data to make cost estimates.

viii. Data Quality Maintenance

Throughout this discussion, the emphasis has been on efficiently obtaining usable research results, without sacrificing data quality. Researchers, clients and stakeholders all have interests in getting the best (accurate) and most up-to-date sample statistics concerning the key cost parameters in the parent population, in this case the route operations of the United States Postal Service.

Certainly, there are many examples where the budget or available time-frame has driven the sample size and the confidence in the results obtained has suffered accordingly. It is however, critical to recognize that decisions that have far-reaching cost and revenue implications may not be best served if they are based upon results obtained from subsidiary studies in which corner-cutting considerations have perforce led to a series of deviations from "best-practice" statistical methodologies for cost studies.

Indeed, the Data Quality Study (1998) emphasized the importance of improving *methodological standards* rather than abandoning them. As world leaders in postal ratemaking practice, the Postal Rate Commission continues to require the highest standards of research performance – given the available resources – to enhance its deliberations and inform its decision-making.

IV. Are any data better than no data?

i. The need for new USPS cost data.

There is general agreement about the long-standing need to up-date and improve the USPS cost data. This need was highlighted in several parts of the Data Quality Study which unearthed "rules of thumb" dating back to the 1920's which are still being applied in the twenty-first century. Moreover, the client (USPS) and the stakeholders (the mailers) recognize that the familiar cost parameters dating from the past two decades have been overtaken by technical change, productivity shifts, traffic patterns, work methods and many other extraneous forces. Nevertheless, any shifts away from these long established "traditional" cost parameters should be gradual, well founded and widely "bought-into" by both the USPS and the stakeholders.

ii. Quality Data for Quality Decisions

In the balance, there is far too much revenue/expenditure at stake for ratemaking decisions to be based on inadequate new data or flawed research procedures. A robust and scientifically defensible innovative cost study needs to be done and the USPS needs to find the budget to commission it, as a matter of priority. Band-Aid solutions and half measures are simply not acceptable - - what would "Big One" lottery ticket holders have thought if their numbers were not included in the recent $360 million lottery drawing?! All the data from the parent population must be available for a random sampling process and professional vetting must be done when the research is designed, implemented and reported upon.

iii. What the Data Quality Study said about Letter Carrier Costs

The Data Quality Study, commissioned jointly by the PRC, USPS and the General Accounting Office, was quite specific in its recommendations with regard to Delivery Costs. Pages 53 to 56 of the Technical Report #4 are provided in an Appendix. These recommendations include:

- Redesign and update the relatively old and highly imprecise Delivery special studies.
- Review the data being developed by the Delivery Re-design project to assess if this information is a possible *long-term* (my emphasis) replacement for IOCS and some special study data.

These imply an extensive discussion of what the Re-design project was doing and what the Postal Service should do with it. The recommendation was qualified with the following important statement: "Reviewing this data now can also allow the rate making forces within the Postal Service to impact the quality of data to be collected in this new system."

It most certainly did not say: "Dig into what has already been done and see if you can fit some previous observations into something to replace the special studies." In fact no-one could be better placed than A.T. Kearney to understand whether the work by Mr. Raymond - - already completed when reviewed by the Data Quality Study - - could be used for rate-making, since A.T. Kearney was responsible for both the Data Quality Study and the Engineering Study managed by Mr. Raymond. The forward-looking nature of the suggested solution speaks volumes.

iv. Is the Engineering Study data better than no data?

Great caution should be exercised in considering whether to use the Engineering Study data results as a basis for developing new cost results guiding ratemaking. There is no criticism here of the Engineering Standards study *per se.* However, there is extreme reticence to use the reworked data from this study for purposes for which it was never designed or collected.

No confidence levels can be ascribed to these data because no sample design was made. The best we can say is that we have information on a number of pre-selected postal stations. How these relate to the total universe we are unable to say. The resulting cost data, calculated by Mr. Baron, may be indicative and even enjoy a degree of accuracy, but no one can say with any *confidence* what value to put on these sample estimates because of the unacceptable fashion in which they were obtained. The one thing that these results do achieve, is to underscore how important it is to undertake a transparent, replicable and scientifically defensible study of relevant cost parameters in the USPS route system at the earliest opportunity.

CERTIFICATE OF SERVICE

I hereby certify that I have this date served the foregoing document upon all participants of record in this proceeding in accordance with the Commission's Rules of Practice.

Anne R. Noble

Washington DC
May 30, 2000