DATA, STATISTICS, INFORMATION - SOME ISSUES OF THE CANADIAN SOCIAL STATISTICS SCENE[1]

Ivan P. Fellegi[2]

This paper looks at the current state of development of social statistics in Canada. Some key concepts related to statistics and social information are defined and discussed. The availability and analysis of administrative data is highlighted, along with the need for social surveys. Suggestions are made about the types of data analysis needed for the development of social decision models to meet policy requirements. Finally, an outline of priorities for future work toward the effective use of social statistics is given.

1. INTRODUCTION

I should start by apologizing for my temerity to address an audience of demographers and sociologists on a topic about which I know so little. Perhaps, I should seek solace in the old Hungarian proverb which, roughly translated, says that "If God gives power to someone, He will be good enough to give some brains to go with it". At any rate, in my previous position at Statistics Canada, I was quite successful at being the best computer expert among survey designers, and vice versa. Perhaps, I can pull off some similar tricks in my new position - starting with this talk.

I have just spent nine months in the United States on a Presidential reorganization project to review the U.S. Federal Statistical System. Particularly, in light of my new responsibility, I could not help observing with some envy, the social statistics data base available there. At least in terms of subject coverage, social statistics in

---

[1] A verbatim reproduction of a paper presented at the Learned Societies Conference, Saskatoon, June 2, 1979.

[2] Ivan P. Fellegi, Assistant Chief Statistician, Social Statistics Field, Statistics Canada.

the U.S. are certainly much more developed than in Canada. It is tempting to speculate why.

You and I, after perhaps a few minutes of blaming one another, could probably quickly agree on blaming "the powers that be" for allocating insufficient resources to social statistics. Before we reach that inevitable conclusion, however, perhaps we can pause for a few minutes and ask ourselves: why should they invest more resources in social statistics? Suppose they invested considerable funds in social statistics; would they become better informed with respect to issues they can influence through social policies?

2. SOME KEY CONCEPTS

By and large, what public policy discussion needs is insight and information. We tend to offer statistics as a substitute. I think the difference between these notions is of key importance in trying to understand the present state of social statistics in Canada. Let me share with you my attempt to clarify in my own mind the difference between these and a few other related concepts. Those of you having a preference for visual presentations may trace an approximation of these concepts on Chart 1 at the end of the paper.

1. Datum: A datum is a quantity (e.g., the dollar value of sales) or a code (e.g., the numerical code identifying an industry, race, sex, occupation, etc.) which arises out of observation or measurement. A datum must have at least three components: a quantity or code value; a concept[1] which is quantified in the form of a code or quantitative measurement (e.g., the particular way "unemployment" is defined in the Labour Force Survey); and a

---

[1] Concepts are abstractions which can be made operational only by selecting some dimension of the "real world" as a measureable proxy to represent that concept.

reference entity, i.e. an entity to which the quantity or code refers. A reference entity may be a person, business, institution, etc. or it may be a group of such entities (e.g., all persons with a permanent residence in Nova Scotia as of January 15, 1978 and whose ages were between 14 and 21 years). A datum for a given reference entity can, of course, include the quantified measure or code of more than one concept - in which case we can talk about multi-dimensional data.

2. Statistic: A statistic is a summarization of data referring to a unique group of persons, businesses, events, phenomena, etc. The expression "unique" means that the members of the group are unambiguously identifiable: producers and users of statistics can apply the same test to a potential member of the group and come to the same conclusion as to whether it is or is not a member of the group. The group definition may take the form of a list of its members, or it may be in terms of their attributes (e.g., all residents of Nova Scotia on a given date). Thus, a person who does not know the identification of all members of the group (in the form of a list) can nevertheless determine, on the basis of its relevant attributes, whether a given entity (person, business, etc.) is or is not a member of the given group.

   A well-defined group can serve as a reference entity and a statistic can therefore be considered as a datum with that group as its reference entity. When we want to distinguish between a statistic and the data which it summarizes, we refer to the component data as micro data.

3. Interpretable message: A datum or statistic which has been communicated by a person or institution is defined as an interpretable message. For example, when Statistics Canada publishes the proportion of unemployed within a given age-sex group in Nova Scotia, this becomes an interpretable message. Before its publication,

the same rate would be a datum, in fact in this case a statistic. To quote a more special example, data we may beam into outer space are interpretable messages: we do not know whether they are received and interpreted by anyone.

4. Information: The term is derived from the verb "inform". To inform is a process. Information, for purposes of this paper, is defined as the process of conveying an interpretable message as a result of which the receiver of the message acquires knowledge, i.e., becomes better informed. Hence, information involves interpretation.

   An interpretable message has the potential of informing, but it needs a receiver. More precisely, an interpretable message becomes information if it is received by an intelligent receiver who interprets it; i.e., does not screen it out but rather stores it in his/her mind for the purpose of some expected future use. An intelligent receiver means a person with the knowledge needed to "decode" a statistic into the three components of data identified earlier and having the ability to relate the decoded statistic to other aspects of his/her knowledge or experience.

   Whether or not an interpretable message, when received, is screened out or stored can be influenced by the sender: through the medium and presentation used, through repetition, etc. More importantly, the sender can induce storage and interpretation of the message by calling the attention of the receiver to the fact that the message has some intrinsic interest or utility for the receiver.

5. Decision model: I am clearly leading up to the point that an interpretable message, duly sent and received, will generally become information (i.e., stored in the mind of the receiver) if the receiver judges the message to be of some interest, relevance or usefulness for him. Leaving aside the question of

interest, the message can be useful to a person if it relates to a phenomenon about which he/she wishes to make a judgement or decision and the result of that decision has some concrete utility to the decision maker.

Decisions typically have as their objectives the modification of the real world in some fashion, or they represent alternative strategies to respond most effectively to different conditions of the real world. In order to do this most effectively, the decision maker should ideally have some reasonably clear objectives and, further, some utility function whose maximization represents a reasonable trade-off between the costs and benefits of alternative decisions. He is interested in those factors which have a relationship with or impact on his utility, depending on decision or judgement to be made. The consideration of these relevant factors takes place within a formal or informal framework, or thought process that will be referred to as a decision model. Before going on, I want to emphasize that the decision model as understood here need not be a formalized mathematical or probabilistic model. It can range all the way from complex econometric or simulation models to an unstructured accumulation of experience. In fact, even where complex and formal models exist, few major decisions are made automatically on the basis of the model's predictions: typically these are tempered by judgements. My definition of a model, for purposes of the present discussion, therefore includes the entire mental process involved in decision making.

Consider three examples. The first example involves a farmer's decision to apply X tons of fertilizer to a corn field so as to maximize net profits. The utility function can be quite well defined as a trade-off of costs and benefits to achieve the single objective. Data on the cost and performance characteristics of fertilizers are directly relevant. The second example

might involve a decision by a person to spend some amount of money either on a vacation or on improving the resale value of his house. The value to the person of the alternative decisions would be very difficult or artificial to incorporate into a formal model. Yet people do take such decisions, so a decision model is involved. Part of this model may be more precisely formulated: for example, the trade-offs involved in the length, location and level of luxury involved in alternative vacation plans. Clearly, data about travel costs are relevant. A third example, involving a more ambitious social decision, may relate to the objective of reducing the inequalities in the income distribution of Canadians. Here the utility function would be quite difficult to define, partly because the same decision makers are likely to be involved with other objectives and hence ideally the combined utility of all these objectives should guide the decisions, partly because even the single desirable outcome is not sufficiently well understood in terms of causalities. What are the factors affecting income inequality - government transfer payments, tax policies, general state of the economy, education, family backgrounds? In the case of factors thought to be relevant, what is their current state?

The real world is infinitely complex - even the limited segment of it which the decision maker wishes to modify or react to. Necessarily, whehter he is aware of it or not, he can only cope with this complexity within the framework of a simplifying model which, over time, he may be able to elaborate further. The model helps in sorting out the relevant factors that should be considered, i.e. those expected to have some predictable relationship with the outcome which he wants to modify. Further, the model would indicate how these relevant factors interrelate with one another and with the outcome, which of the factors can be modified, what is the likely effect on the outcome of modifying one or more of

the factors, what is the current state of the relevant factors. Note that only the last of these questions can possibly be answered by statistical data, and only the past interrelationship between some of the factors can be answered through statistical analysis.

So, facing the real world, the decision maker needs the simplification offered by his formally articulated or informal model. Another complexity where his model helps the decision maker is known as information overload - the reception of a great variety of data, only some of which is related to his decision problem. The model would indicate, in effect, which data he should be interested in and which can he afford to screen out. From the point of view of the sender of an interpretable message, this role of the decision-maker's model is particularly important: if the sender knows that his message is about a factor which is part of the decision-maker's model, he can be reasonably sure that his message will be interpreted. Further, if the sender is able to articulate a relationship between his message and some factor of known interest to the receiver, then in effect, the sender can cause the decision-maker to evaluate and possibly change his model by incorporating the new factor and thereby render the current message relevant for the receiver thus ensuring that it becomes information.

6. Validity: Data collection typically involves compromises between the concept a decision maker might wish to measure (the "ideal concept") and what is possible and practical to measure (the "operationalized concept"). One may have an ideal concept in mind as to how unemployment status should be defined in the context of a decision problem at hand. Different users faced with different decision problems may well lead to different ideal concepts. However, those involved in actually conducting a household survey may decide that

a concept, in order to be measurable with reasonable accuracy, must be related to some concrete activities of individuals which, if they are questioned about them, they are likely to remember. For example, this consideration was a significant reason for the development of the activity-based concept of unemployment used by the Labour Force Survey (LFS) over a long period of time. The respondent not only has to be able to remember his answers, he should also be disposed to respond, willing to accept the burden of response, etc. All of these considerations may lead a survey taker to accept compromises in the concept to be measured. The distance between a given users's "ideal" concept and the operationalized concept actually used measures the validity of the data for the given use. For example, the operationalized concept of unemployment used in the LFS is not ideal for the purpose of monitoring the number of persons suffering economic hardship as a result of unemployment, thus affecting the validity of the LFS for this purpose.

It is critical to understand that, if the resulting data are to have required validity for a decision maker, the underlying concept must be a close enough approximation within his decision model, of an aspect of the real world. Thus "ideal" concepts arise within a decision context. The unemployment concept mentioned above is decisively affected by the decision context in which this concept was defined - monitoring the labour market as opposed to the meaning of social or economic well-being. This has clearly major statistical policy implications: concepts have to be updated either as a result of changes in the real world, or changes in the decision problems addressed. Furthermore, often a single concept related to a particular phenomenon cannot fit exactly the needs of important but different decision problems; in such cases, the job of a statistical agency is either to find the best available compromise, or to collect (as, for example, in the case of unemployment) sufficient detail to permit the construction of estimates for alternative definitions of the concept. Given resource constraints, the latter alternative is only rarely feasible.

7. Relevance: As indicated above, data only have the potential of becoming information. If the utilization of data by a decision maker would reduce the uncertainty associated with his decision, we say that the data are of relevance to him. Clearly, relevance is a property of the data in relation to a class of users or uses, not a property of the data alone. It is a very broad concept. A decision maker with a well articulated decision problem may, for example, need data on the unemployed. In this case, data on the unemployed are quite likely to be of relevance to him, essentially depending upon the distance between the particular concept of unemployed he needs and the one that is available. Thus, in this particular case, relevance becomes synonymous with validity. However, relevance is the broader concept. A decision maker concerned with "general well-being" might consider data on health, income, housing, cultural activities, etc. all relevant, depending on the operationalized concepts used. A statistical agency wishing to render its data as widely relevant as possible must therefore acquire considerable knowledge of the decision issues and models of its users, as well as skills in operationalizing the concepts most useful to decision makers. Such knowledge is acquired - except in the case where data are collected by the end users themselves - through a variety of analytical activities shedding light on the end users' decision problems, at the very least by maintaining close dialogues with a wide cross-section of end users. Once again, the notion of relevance has major policy implications for a statistical agency.

A necessary (although not sufficient) condition for data to be relevant in a decision context is their explanatory power or relatedness with respect to the object of the decision. In the case of micro data, the inclusion of more than one (carefully chosen) variable can often exponentially increase the explanatory power (relevance) of the retrievable statistics. Data on the distribution of the unemployed by age are clearly vastly more relevant

for most purposes than separate data on the age distribution of the population plus the number of unemployed. Thus, the potential relevance of a micro data base is strongly affected not only by the choice of the concepts measured but also by the richness of the data base. Furthermore, given the fact that most models have to use data from several sources, the useability of a given datum in a model strongly depends on the ease with which it can be used jointly with other data. Thus, another prerequisite for increasing the relevance of data emerges: standardization of concepts.

8. Accuracy: The accuracy of data, broadly defined, is the extent to which the actual measurement of the operationalized concept any hypothetical, error-free counterpart are close to one another. It includes the well known components of measurement and, when applicable, sampling errors.

   Furthermore, accuracy is also affected by the extent to which the reference entity, in this case a group, is incorrectly identified, for example, by failing to include in the group, persons who according to the group definition, belong to it. Accuracy which is inadequate for a particular application may render data irrelevant. Put differently, accuracy commensurate with a given substantive objective is one of the many attributes of relevant data.

9. Misleading data: The notions of validity and accuracy lead us to other desirable properties of data. The concept which is measured is often described only very briefly (such as through the use of a term like "unemployment"). In that case, the receiver of the message may assume the concept to correspond to his or her notion of what "unemployment" is, which may or may not be the same as that which was actually and explicitly implemented. Similarly, unless an explicit statement about accuracy is provided, the receiver is free to assume any level for it, including "complete accuracy".

The result may clearly be potentially misleading. Thus potentially misleading data are data whose concepts and accuracy are inadequately or incompletely described. Misleading data are those whose concepts and accuracy are incorrectly described.

The implications of the concepts developed up to this point will be spelled out below in the context of social statistics in Canada.

3. SOCIAL SCIENCE AND SOCIAL STATISTICS

The notions outlined above, particularly the concepts of decision models, validity and relevance, help me to understand the scene of social statistics in Canada and how it developed. Let me share with you the highlights as I see them.

1. Predominance of administrative data in social statistics:
   A striking phenomenon of the social statistics scene in Canada is the relative predominance of data initially collected for administrative, as opposed to statistical purposes: vital statistics, statistics on health institutions, educational institutions, etc. Of course, this is not an accident. The fact is that each of these data sources (as indeed most data sources should!) came about in response to specific decision problems, typically related to the administration of particular social programs. This is not the time or place to discuss why and how the social programs themselves developed. Suffice it to say, without any value judgements being implied, that the programs themselves were largely initiated within decision models that were more political than quantitative, where the questions raised were more in the nature of "how fast can we afford to do it" rather than "what are the objectives, through what alternative means can we achieve them, and what impacts might each of the alternatives have in addition to the furtherance of the particular objective?" Thus, the launching or extension of programs,

such as universal health insurance or the large-scale expansion of higher education, did not represent decision problems[1] in support of which comprehensive statistical programs would have developed.

Having launched the programs, their efficient administration, of course, represents a continuing decision problem in need of data - the so-called administrative data. Typically, therefore, the definition of the concepts measured and the identification of the reference entities is determined within the framework of these particular decision models, e.g., the administration of hospital programs. The particular decision problems have a major impact on the data created to support them: for example, the problem of efficiently administering a set of hospitals is very different from the issue of how to improve the health of Canadians. Hospital statistics were largely developed in response to the administrative problem. While they are still relevant to a study of the general health of Canadians, their validity might clearly be impaired for some of the analyses which might be involved in such a study.

2. Problems of validity and relevance of administrative data for general purposes:

This point is implicit in the previous one. Neither the operationalized concepts used, nor the reference entities (coverage in more traditional terms) lend themselves easily to the purpose of developing or assessing general social policies, although a lot of work has been done by Statistics Canada to influence administrators to modify their concepts in such a way as to improve their validity for more general uses. A further problem of relevance of these data is the relative paucity of the data bases, i.e., the relative lack of

---

[1] In all fairness, it should be added, however, that in recent years there has been more interest and activity in the area of program evaluation, particularly before changes to major programs were introduced.

appropriate concomitant variables which would render the conceptual linkage of the data that are there with other data, and with a variety of decision models, more manageable. Record linkage, a technically difficult and often controversial undertaking, can often ameliorate this problem. On the positive side, administrative data are cheap, so long as one does not want to overcome their limitations, i.e., so long as one can use them more or less "as is". Attempts to overcome their limitations can be expensive. Given that they must be collected, any alternative collections (e.g. statistical surveys) impose not only extra costs but also response burden. And, finally, their accuracy is generally high, i.e., their accuracy as measures of their operationalized concepts, so long as one can accept the limitations of validity often resulting from those concpets.

Since the question of general relevance of data is of necessity of secondary importance to administrative agencies, they give only a passing recognition to the extent to which the resulting statistics fit decision models other than those they were initially designed to support. Nevertheless, given the relevance of administrative data sources for a broad class of users, overcoming their shortcomings for non-administrative uses is a major challenge for Statistics Canada, one to which we attach a high priority.

3. Social Surveys: If there are some decision models which are in need of statistical information not available as by-products of the administrative processes, particularly for analyses which require information cutting across the subject boundaries imposed by the institutional boundaries of administrative processes, these can only be filled by data collected specifically for statistical purposes, i.e., through statistical surveys.

As pointed out earlier, the relevance of data is partly determined by its accuracy. The requirements of accuracy, particularly if provincial or sub-provincial analyses are required, often impose a sample size which renders data collection prohibitively expensive for most non-Federal collectors. At the Federal level, the main survey vehicles used to collect social statistics are the quinquennial censuses of population and housing, the Labour Force Survey, surveys attached to it as supplementary questions whose subject matter has a largely irregular frequency, and the annual or less frequent surveys of Consumer Finance, Household Facilities and Family Expenditures. The latter three actually are either supplements to the Labour Force Survey or use its facilities some other way. During 1978-79, a new social survey, the Canada Health Survey, was launched but became a victim of budget cut-backs.

All of these surveys are widely used, at least by governments, in their decision models. A significant demonstration of the utility of the census could be seen in the reaction of data users when consideration was given to reduce the data coverage of the 1981 Census to legally mandated questions.

One can ask whether our current survey program is big, small or just about right. It is certainly small in comparison with the U.S. social statistics scene. We do not have surveys on health, comprehensive victimization, housing, time allocation, quality of life, a detailed survey to measure the impact of transfer payments, longitudinal surveys to measure the impact over time of "prevailing conditions" on cohorts which are in a state of significant transition - the pre-retirement group, or entrants into the labour force. This partial list could be extended considerably. However, I hope it is clear by now that I consider any such abstract questioning of the adequacy of our survey program to be rather futile. What are the decision models which cannot be formulated; what important analyses cannot be carried out with an

acceptable level of uncertainty without such data? These are the prior questions to answer. It is only by engaging in relevant analytical work that social scientists can help answer these questions.

4. Analysis of social data: In an outstanding paper delivered at the American Statistical Association, Robert Parke and Eleanor Sheldon identified several types of analysis of social data serving to render them relevant for major policy purposes. Although this section of my paper deals with highlights of the Canadian scene as I see it, nevertheless, I would like to recapitulate briefly the analytical categories they identified, using my own terminology, so as to illustrate the wide variety of important social decision models which can be assisted by suitably analyzed social data or, in fact, which can be modified by such data and analyses. The specific analytical examples are mostly American and drawn from their paper.

   (a) Cognitive information
   Studies which identify new problem areas of major social importance fall into this category. The outstanding examples are the studies linking smoking and a variety of health hazards, drinking and drug abuse, tension and auto accidents. They may not fit initially a direct programmatic decision model, but do fit the higher decision models of those responsible for health policies in general and may indirectly result in the articulation of more specific, lower level, problem-oriented decision models.

   (b) Identifying important external constraints
   The contribution here is to invite the attention of policy makers to developments not under their control that may alter the way they wish to conduct affairs that are under their control.

An example from outside the social sciences is the weather reports in which we are interested because they may alter our "adaptational strategy", even though we cannot manipulate the weather. A social science example quoted by the authors is a study of the impact of wide fluctuations in birth rates over the past 30 years on: the female labour force, low income and minority groups, geographic mobility of workers, unemployment, GNP, consumption patterns, schools, health services. Specific policies can affect each of these separate problem areas, but only within a margin determined by external constraints. The setting of realistic goals and the definition of success or failure of particular policies is an important indirect contribution.

(c) Projecting consequences

This category is somewhat similar to the previous one with the exception that manipulative, as opposed to adaptive, strategies are available. The illustration is a combination of the decline in the absolute number of births starting in 1961 with the rise in teachers' college enrollments. Good current statistics were available of both phenomena and could certainly have been suitably analyzed to forecase the inevitable over-supply of teachers. Had this been done with the appropriate penetration of the analytical results in the audiences concerned, educational authorities might well have taken appropriate actions to prevent a serious dislocation in the labour market.

(d) Analysis of specific decision options

The analysis here concentrates on specific decision options contemplated so as to assess them in terms of their contribution to the achievement of objectives. The example quoted by the authors relates to the policy objective of the early

'70's to reduce the population growth of metropolitan areas of over 1 million population by encouraging growth in smaller, less congested "growth centers". An analysis showed that between 1970 and 2000 even a 30 percent growth by such centers would only absorb about 10 million persons, leaving still 70 million persons to be absorbed by the larger metropolitan areas, mostly as a result of local births and immigration from abroad. Clearly, growth centers would not contribute significantly to the achievement of the stated objective. The decision model would need to be amended.

(e) Communicating the meaning of data

This is the well-known but insufficiently often practiced data analysis which draws out the story of the data that will not be told without it. That "story" might transform the data into information; without it the data is in danger of remaining an interpretable message. The analysis here intends to show the relevance of data for a variety of not necessarily explicitly specified decision models. Examples are: the transformation of mortality data into life expectancy tables; the presentation of income data showing the fraction of total personal income received by the one-tenth of families receiving the smallest incomes; the transformation of current marriage and divorce data into cohorts having different divorce experiences; more elaborate multivariate analyses involving standardization of populations over time with respect to some characteristics so as to study the impact of others; the use of loglinear or other techniques to study the relationship between geographic mobility and socio-demographic characteristics; etc. In case I might be misunderstood, I want to emphasize that this "story that the data tells" is <u>not</u> equivalent to the verbalization of tables, or even some analytic material that seems to be contrived and "l'art pour l'art" - the latter often strikes me as a declaration "here is the answer" without first asking "What is the question".

The reason I dwelt on the Parke-Sheldon paper at length is partly to illustrate the enormous analytical potential in social statistics, but also to make more plausible to you my belief that there is a great relative paucity of comparable analytical work in Canada. In fact, I think this is a striking aspect of the Canadian scene as compared to the U.S. One may argue that the difference is a function of the relative paucity of social statistics to analyze. With due respect, I must submit that there appears to be a shortage of analyses of even the statistics that are available. Comparing again the Canadian scene with the U.S., the federal statistical system here seems to be doing at least as good a job in disseminating its data in useable form as its U.S. counterpart. Most of our household surveys are, for example, available on tape in micro-data form, ready for analyses. The Labour Force Survey or the income surveys are examples. In the U.S., the so-called March supplements of the CPS are used by scores of users. Workshops held on them are typically over-subscribed. By contrast, as one of my colleagues (who shall remain unnamed) put it: "I wish that a few Canadian social researchers would discover the potential of the Labour Force Survey. I would estimate that an untenured assistant professor in Canada who has a logical mind and quantitative tools at his command could easily acquire tenure, promotion and a scholarly reputation by just mining the LFS".

Similar comments could be made of a number of our data bases, particularly with respect to the potential of the joint exploitation of several of our data sources.

5. A climate of scarce resources and relatively low priorities: Without doubt, a major aspect of the Canadian scene of Federal social statistics is a climate of extremely scarce, in fact diminishing resources. It would be foolish of all of us not to recognize it and to pretend that the degrees of freedom open to us are significant.

This climate of scarce resources is general within the Federal government and certainly within Statistics Canada. The general scarcity of resources is exacerbated by the relatively lower priority accorded to social statistics compared to economic statistics. This is not a result of some mysterious internal struggle within Statistics Canada, or lack of even-handedness in the distribution of cuts. It is a direct consequence of government priorities.

Governmental priority is, of course, a complex process. It is partly established by the government's own policy agenda which usually is a judicious mix of what it believes is good for the country and what it believes the country or some important groups in it want. In the case of social statistics, government priorities can be impacted upon either by demonstrating effectively that a certain level and type of service is needed for it to be a more effective government, or that this service is vitally important for other levels of government or groups of the population which in turn are considered to be important to the government. I want to emphasize that the above should not be construed as lobbying for your vocal support. Such lobbying for more social statistics would, I believe, be ineffective at any rate. The challenge I want to put to you, and indeed to my social scientist colleagues in Statistics Canada, is for social science to make a greater and more readily visible contribution to Canadian society.

6. Social science and social statistics: The five previous highlights of the social statistics scene in Canada indicate a relatively low level of activity and even a low level of interest. The immediate question is: if there is a low level of activity and a low level of interest, perhaps supply and demand are well balanced and everything is quite alright in this best of all good worlds. Or, to quote a little poem of the Danish poet Piet Hein: "The universe may be as great as they say, but it wouldn't be missed if it didn't exist".

However, when I consider the highly relevant work quoted by Parke and Sheldon, I feel really sad about the opportunity cost to Canadian society.

I believe social science itself must become more relevant to decision makers in order for social statistics to be given higher priority. What I have in mind is the identification of social problems of recognized importance; the identification through analysis of factors related to such problems, including those which can be influenced through decisions; the determination of the extent to which the decision models cannot be articulated adequately without additional social statistics; and finally, effective communication of the fruits of social science in the language of the "man on the street".

In effect, my hypothesis is that the main issue is the image that decision makers have of social science, rather than that of social statistics. Some evidence to support this hypothesis is provided by a recent study of the Institute for Social Research, University of Michigan. The study involved "204 interviews on social science research utilization and policy formation with persons holding important positions in various departments, major agencies, and commissions of the executive branch of government". One of the results of the study is shown by Chart 2 at the end of the paper. It indicates that sociology is not rated very highly in terms of its "validity and reliability" - not only far below the "hard" sciences of physics and biology, but considerably below economics, and even below psychology. This is <u>not</u> due to the tools most frequently used in social statistics, as shown by Chart 3. Indeed, survey research is considered to have higher validity and reliability than controlled laboratory experiments. It does not even seem to be due to perceived limitations of social statistics. As Chart 4 shows, "population statistics" and the "unemployment rate" are rated

as very valid and reliable, more so even than economic trend data, even while economics itself is rated in Chart 2 as considerably more reliable than sociology.

The authors of the study go on to point out that "the heavier users of social science information consistently rates the social science items higher than the less frequent users and non-users. These differences in ratings across utilization score levels were sizeable and statistically significant". While one can never know, of course, the direction of causality between higher ratings and more utilization, I have sufficient confidence in social science to believe that more and more meaningful exposure by policy makers to the best that social sciences can offer will have significant benefits for both communities.

It would appear, therefore, that a prerequisite of a richer social statistics data base is greater involvement by social scientists with social problems perceived to be important by policy makers, particularly at the federal and provincial level. Put bluntly, it is not very effective for social scientists to point out to Statistics Canada that more social statistics are needed. Such demands become really effective when they come to Statistics Canada in the form of needs by public policy makers or at least for public policy articulation. An examination of the history of our major "general purpose" social statistics programs - the Census, Labour Survey, Survey of Consumer Finance, Family Expenditure Survey - will readily confirm the validity of this statement.

Let me add, in case I might be misunderstood, that there is a sharp distinction between political interference with statistics and the role of the policy process in setting statistical priorities. Political interference in statistics, however subtle, which may affect statistical data, their timeliness, or unrestricted availability is, of course, totally unacceptable. Setting statistical

priorities in response to recognized public policy needs is not only proper but the only feasible course. The provision of statistics to shed light on the priority policy issues to be tackled or monitored by our elected policy makers assists not only the government which makes the decisions, but also all others who want to participate in debates concerning those issues, or those who wish to monitor the performance of the decision makers. This is true so long as statistics, once collected, are made generally available - a condition clearly adhered to by Statistics Canada and with strong support from a long succession of governments.

4. SOME CONCLUDING REMARKS - WHERE SHOULD WE GO FROM HERE?

Much of what I said can be construed as a Statistics Canada official telling social scientists what they should do to improve the state of social statistics - if indeed it is not judged to be adequately developed now. This is certainly not the impression I would like to leave with you. The question is how can we, together, be of greatest assistance to those who have to grapple with the social issues facing Canadians. The following is an indication of what I see as some of the ways we can, together, be more effective.

1. First of all it bears emphasis that what I said about social scientists in general certainly applies to those of them who are employees of Statistics Canada. However, as the main source of nationally comparable social statistics, we have some additional responsibilities. So I will start this section by outlining my priorities for the next period of time as Assistant Chief Statistician of the Census and Household Surveys Field.

   (a) Statistics Canada will put a high priority on ensuring that our statistics should not be "potentially misleading". I am using this term as defined earlier, i.e. in the sense that

the concepts used and the accuracy of statistics are described and disseminated to the best of our ability. Actually, in the case of our censuses and household surveys, our record is not at all bad in this regard: in the last decade such information has consistently been produced and made available. This practice will continue, but we will put more effort on calling attention to the availability of such information as opposed to simply making it available. I have the feeling that a good deal of material we produce on the accuracy of our data or their validity for different purposes would qualify as "interpretable messages", duly put into the public domain but becoming "information" to only a few users.

(b) I will attempt to strike a new balance between resources spent on generating data and those devoted to transforming them into issue-oriented social information. This applies not only to the so-called general purpose statistics, but also to statistics derived from administrative records. As indicated earlier, many of the latter files have high potential relevance in relation to a number of issues, even though their concepts and coverage diminish their validity for some purposes. Our analytical work should not consist of bland verbalizations of tables or purposeless researching of relationships which are not consciously designed to shed light on important hypotheses, issues or data problems. At the time time our analytical work must be carefully balanced and monitored to ensure that it remain objective, i.e. that policy relevance is maintained without policy advocacy. The distinction is subtle but, for a government statistical agency, extremely important: for example, we could analyze the extent to which unemployment, as measured, represents a hardship for different subgroups of the population - not the impact of alternative unemployment insurance schemes. To preserve

their objectivity, similar rules will apply to data analysis as to statistical data: they will be publicly available without preference to any group of users, and their assumptions, methods, and data sources will be clearly stated.

(c) In order to facilitate analytical work to be carried out by the academic community and others outside of Statistics Canada, I will bend every effort to make our data available and accessible. Subject to the very real resource constraint described earlier, I will stress the building of bridges with our user communities, particularly those who will help to render our data more meaningful and relevant to the public. Much of our data is available on micro-data tapes; all of it (subject to confidentiality constraints) is available for special retrievals. Perhaps, we can do more to demonstrate the potential of such data bases or otherwise assist analysts outside of Statistics Canada. I have a number of concrete ideas in mind and am also more than willing to listen to suggestions.

2. It is presumptuous of me to tell you what you should do. Nevertheless, being concerned with the status of social science and social statistics in Canada, I know what I would do if I were a social scientist. I would identify for myself a set of socially important issues, formulate some hypotheses which might shed some new light on them, perhaps point the way to some alternative remedies or show the unworkability of others, in other words engage in one of the types of policy-relevant analytical work outlined in the Parke-Sheldon paper.

3. It might be argued that the notion of more, and more relevant, analytical work as a prerequisite for a richer social data base may appear to be somewhat "upside down": if important data sources are missing, how can we do relevant analyses? I believe we have a long way to go before we can say that the existing data base is reasonably well exploited in terms of its analytical potential. Furthermore, it is only in the context of developing relevant decision models that the importance of data gaps can be convincingly demonstrated: by showing how a competent, relevant analysis could have been even better if specific statistics had been available. In the current tight economic climate, I do not want to give you the impression that even if a good case is made for certain types of statistics, the funds will be made available for them. However, without such a case the funds will certainly not be made available. Furthermore, I foresee a good deal of gradualism. When the need for new statistics is convincingly demonstrated, including the social issues they are supposed to help address, we will have to confirm (or otherwise) such findings through intensively exploited small scale surveys. Let such small scale surveys begin to be useful and relevant and let the limitation of sample size and/or frequency of measurement be seen to be a deterrent to more effective decision making before such surveys are expanded.

4. I was struck in the United States by the close ties, both formal and informal, which exist between the Federal statistical system and the academic community. It is a healthy relationship, not of the "you scratch my back, I'll scratch yours" variety. As the case may be, the academic community is either a vociferous critic or a most influential supporter of the statistical system. One thing it is not: a neutral bystander. My early impression is that, with a few notable exceptions, the academic community is far too passive in Canada in relation to the Federal statistical system. We need to have strong, fair, informed and influential critics. Without them the pressure to improve the quality of our

product is largely internally generated. Other users put pressures on us to improve the timeliness and quantity of statistics produced, and to improve our methods of dissemination. It is my impression that not enough external pressure is put on us to improve the accuracy of our data. Even the most professional organization is strongly tempted to respond more readily to external rather than internal pressures. I should add that the most useful critics are those who understand thoroughly our statistics, "warts and all", as well as the constraints under which we operate.

We also need to have your strong support when the system as a whole is in danger - for whatever reason. It must be kept in mind that the value of information accrues only through its use; but the cost of it is entirely front loaded. Moreover, information cannot be divided up so that each unit of it could be marketed at a relatively small cost: the entire cost must be prepaid. Under these circumstances, the "buyer" of information, basically, must rely on the reputation of the producer. Other facts are also relevant: large-scale statistical data collection, because of its cost, tends to be a government monopoly; we are asking the public to give us of their time and to share with us information that many consider to be confidential; and we are asking those participating in democratic debates to accept statistical information as a factual base for such debates. All of these factors together put a high premium on preserving the reputation of, and confidence in, the statistical system. I submit to you that when that is put into question, it is not in your own interest to be idle bystanders. To borrow a Latin phrase, "nostra res agitur" or, loosely translated, we are all involved together. If there is a real problem, to into battle to isolate and remove it; if there isn't, let it be known loud and clear. Should the public confidence in the statistical system ever be shaken, then to talk about future developments of social statistics amounts to fiddling while Rome burns.

5. Finally, I want to sound a note of caution: in some sense we can be too successful in rendering statistics policy relevant. I have in mind the disconcerting trend, particularly in the U.S. but to some extent here as well, of using statistics for formula-based decisions. Throughout this paper, at several points, I emphasized that statistical data become information through interpretation. I made several references to decision models. Without interpretation, statistics become just data; without judgement, decision models become formulae. Yet in a society in which decision making is increasingly subject to public scrutiny - by itself a largely healthy development - but which also is increasingly reluctant to accept judgements as the bases of public decisions, there is something very attractive in formula-based decisions. The tendency to try to render decisions unchallengeable is not restricted to the use of statistics. To give just two other examples, one finds instances of it in the unthinking use of high school marks as the basis of university admissions, or in the tendency of some medical doctors to obtain "irrefutable" evidence for their diagnoses by prescribing more laboratory tests than might be necessary without the increasing risk of malpractice suits. What concerns me here, however, is the use of statistics in formula decisions for purposes they were not designed to support. In the U.S., such uses often resulted in some extra funding support for new or expanded statistical data collections. However, I believe that the resulting politicization of statistics is too high a price to pay for the benefits involved.

In closing, I hope you will forgive my frankness. However, I believe that so shortly after having accepted my new position, I have a unique opportunity: to the extent you agree with some of my points, I am off to a good start; to the extent I might get into hot water with you, I can always claim ignorance. This opportunity may never return.

RESUME

Le présent document examine la situation actuelle de la statistique sociale au Canada. Il définit et analyse certains concepts fondamentaux relatifs à la statistique et à l'information sociale. L'accent est mis sur la disponibilité et l'analyse des données administratives, ainsi que sur la nécessité des enquêtes sociales. L'auteur propose différentes analyses de données susceptibles de permettre l'élaboration de modèles de décision sociale permettant l'application des politiques. Enfin, le document décrit dans les grandes lignes le travail à accomplir en priorité en vue d'une meilleure utilisation de la statistique sociale.

CHART 1

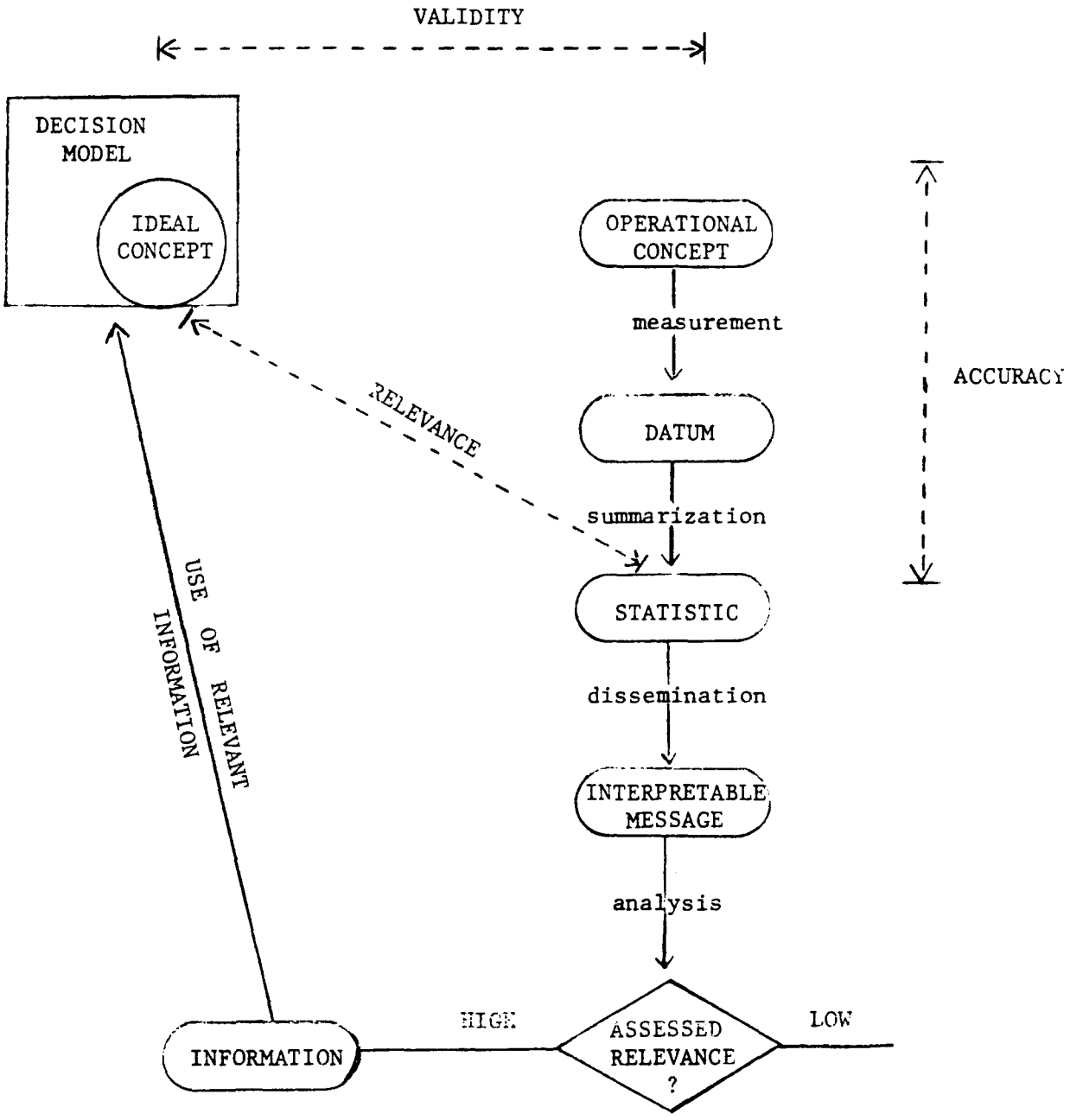
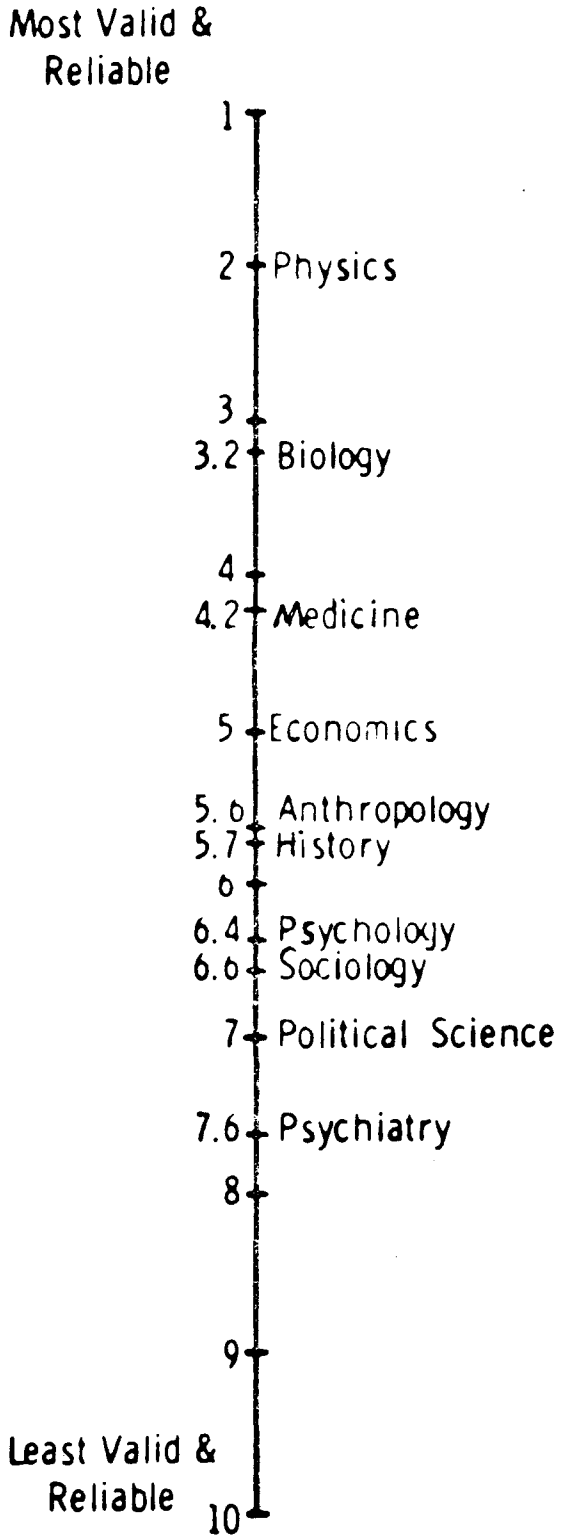
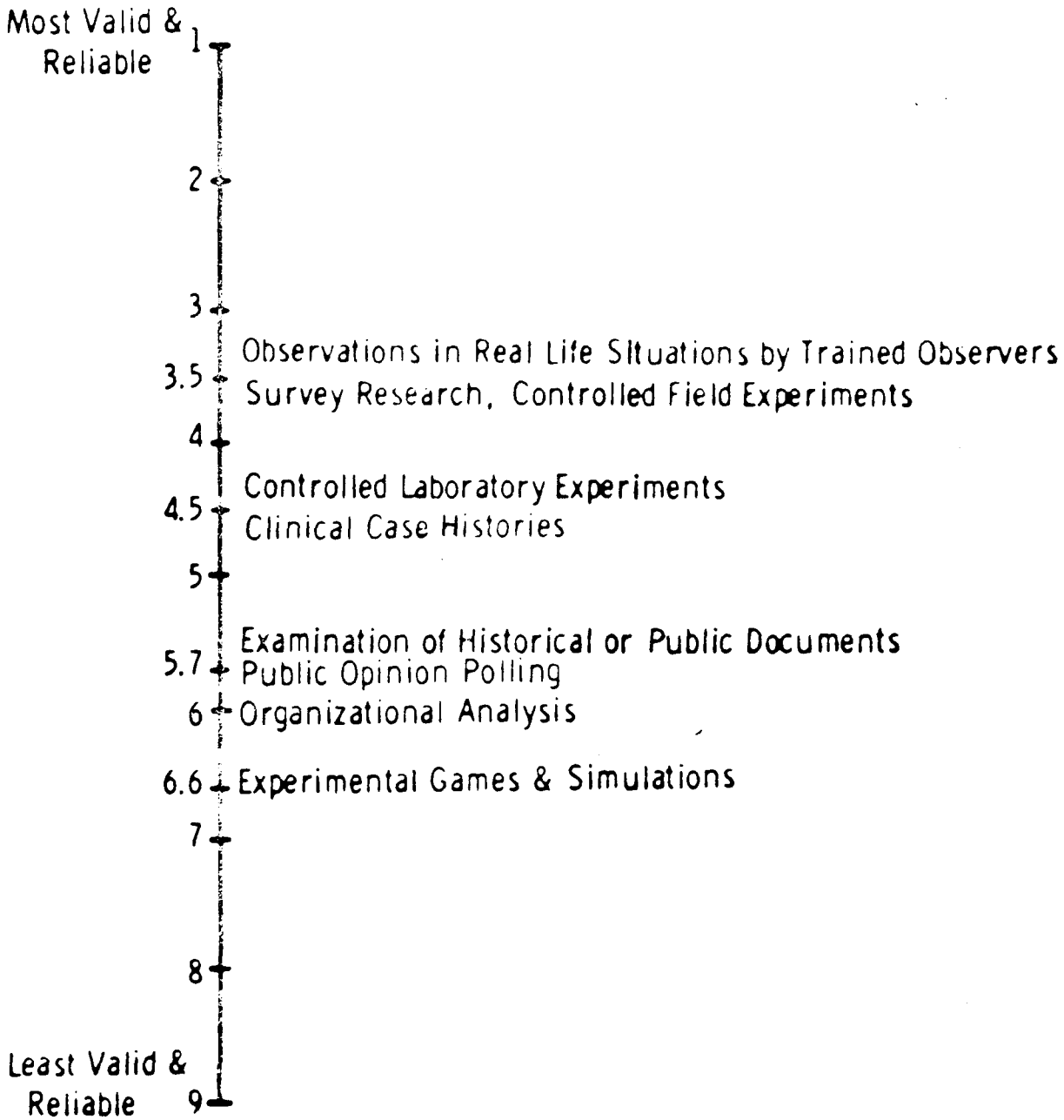

VALIDITY
DECISION MODEL
IDEAL CONCEPT
OPERATIONAL CONCEPT
measurement
DATUM
summarization
STATISTIC
dissemination
INTERPRETABLE MESSAGE
analysis
ASSESSED RELEVANCE ?
HIGH
LOW
INFORMATION
RELEVANCE
USE OF RELEVANT INFORMATION
ACCURACY

Chart 2



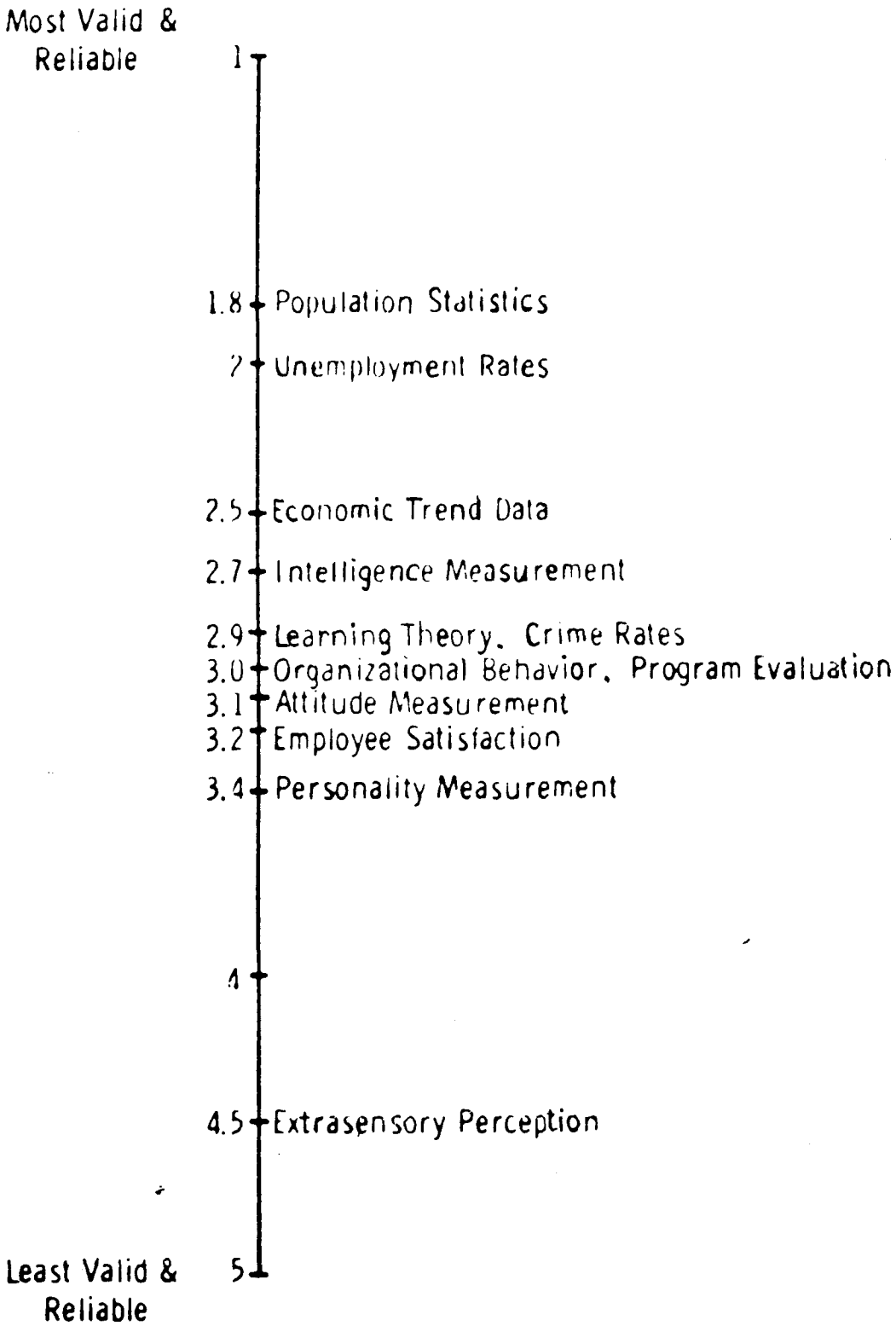
Mean Ranking of Validity and Reliability for Scientific Disciplines

Chart 3



Mean Rankings of Various Data Gathering Techniques

Chart 4



Mean Validity & Reliability Ratings for Various Types of Data