DISCUSSION

I.P. Fellegi, Statistics Canada

This discussion is based on an interim report of the Subcommittee, issued by OMB on May 27, 1977. Wherever possible, I will discuss the papers as a group, they being (I presume) largely the product of the Subcommittee. I want to state at the outset that in my opinion the authors have carried out an extremely thorough job, producing a most comprehensive review of the "state of the art". Notwithstanding some of the critical comments made below, I largely agree with their findings and certainly admire their thoroughness.

1. My first, and most important point relates to the implication of the definition of disclosure accepted by the authors. This definition, originally proposed by Dalenius, is extremely broad. Working with such a broad definition is useful at least from one point of view: it enables them to provide an excellent and comprehensive discussion of every conceivable disclosure--of great educational value! It is also very limiting: in fact, the definition is so broad that in the case of quantitative variables clearly every tabulation cell is a "disclosure"--as defined.

Perhaps the most extreme illustration of the implication of the breadth of definition of disclosure relates to what Bell calls "probability-based disclosure". The example quoted refers to a county in which a table shows that over 80% of the persons are earning income in the range of $2,000+ -- the conclusion being that "it is very likely that a given person in the county has a monthly income in excess of $2,000" and that consequently probability-based disclosure would occur. It seems to me that this somewhat stretches the issue: if the income class $2,000+ were broken out in more detail, no clear majority would fall into any given class. So by showing more detail, the apparent probability-based disclosure can always be remedied -- a result which is not intuitively too appealing.

Starting with their very broad definition of disclosure, clearly the authors found it very difficult to formulate guidelines with respect to disclosure-avoiding approaches. In fact, the operational aspects of the guidelines can be summarized, somewhat simplistically, as follows: there are no federal guidelines, each agency should formulate its own policies, and internal procedures to implement them. These should be reasonable and should always prevent exact disclosure of financial and related information; in formulating their own guidelines agencies should be aware of the educational material developed by OMB. Summarized in this somewhat crude way, it might sound to some like an admission of failure. I would certainly not agree with that assessment. Instead, it is an honest admission of the fact that the legal framework does not provide an operationally useful definition of disclosure, that the logical framework of Dalenius is too broad to be of operational (as opposed to educational!) value, and that, therefore, the operationalization of the concept of disclosure must be based on pragmatic considerations.

One can certainly protest that relying on individual judgements is intellectually not satisfying, but I tend to agree with the authors that it is the only realistic course, given our current state of knowledge of the issues. By analogy, our legal system survived well without the concept of "guilt" ever having been precisely defined. Convictions are based on being found guilty beyond reasonable doubt. The formal legal framework identifies various actions as being "crime" or "tort", but the definitions involved are usually somewhat abstract, and few except the most obvious cases coming to courts represent "perfect fits"--thus the need for the personal judgement of judges and juries. Pursuing the analogy, an important aspect of the legal system is that it rests on precedents--similarly, the guidelines encourage agencies to document, on the one hand, the details of any alleged disclosure and, on the other hand, requests for tabulations or microdata which were refused on grounds of potential disclosure. This is clearly a sound recommendation. In the absence of anything analogous to the Supreme Court, the guidelines propose that the Statistical Policy Division of OMB assist and advise agencies in cases of allegation of either disclosure or unnecessarily restrictive disclosure avoidance policies.

The analogy with the judicial/legal system breaks down in two fundamental ways: judgements can usually be appealed and reversed. However, disclosure, once it occurs, cannot be reversed--published data cannot effectively be withdrawn, nor the resulting damage to the statistical system easily repaired. For this reason, I tend to disagree with the implied criterion for balancing the "right of privacy versus the need to know". Indeed, the paper of Michael et al argues that there has been no documented case of a person having been harmed as a result of statistical disclosure and that, by contrast, this does not appear to be the case with respect to companies. Based on this observation, the paper states that, with respect to population data, there appears to be an "imbalance where there have been no instances of harm to data subjects but several cases where requests for data have been denied"; and that in the business sector, "there is a better balance between the interests of data subjects and users". Thus, it would appear that the state of equilibrium recommended by the paper would occur where the dissemination program, through gradual liberalization, begins to result in documented harm being caused to persons. Of course, it may well be true that some agencies are too conservative with respect to their dissemination program--I would simply argue (quite strenuously) against the implied criterion of equilibrium.

2. My next comment concerns the treatment in the Bell paper of the issue of sensitivity of data and the assurances given to respondents. I would be wary of classifying variables into "sensitive" and "non-sensitive" classes, presumably with the intention of being more liberal with

respect to the disclosure of non-sensitive variables. There are few variables, at least relating to people, which can safely be assumed to be non-sensitive. Even such basic demographic variables as age and relationship to head can be extremely sensitive: they can have a significant impact on, for example, social welfare eligibility. Moreover, when we promise confidentiality to respondents, we do not restrict our promise to some unspecified "sensitive" variables. We can hardly have a dissemination policy which is in conflict with our promise to the public at the time of collection.

3. My next point relates to the treatment in the papers of disclosure within the complex of federal government departments. One of the guidelines in the paper by Michael et al deals with the release of micro data files which do not meet the criteria of public-use microdata files. The same proposals surface also in the paper by Zeisset. The guideline, in effect, states that such files can be released if the receiving agency has the authority and obligation to protect the microdata files, with appropriate sanctions for violation of confidentiality provisions. Not being totally familiar with the legal framework under which U.S. federal statistical agencies work, I can only express a visitor's opinion that without an umbrella Statistics Act, which would establish "statistical enclaves" (to use Mr. Duncan's terminology) within the different departments, all subject to the same confidentiality protection statutes, this guideline might not be particularly useable. In the absence of such a Statistics Act, it is important to regard potential disclosure within the federal establishment as being just as serious as disclosure to non-governmental bodies or persons. At least at one place in the paper of Zeisset I could detect a distinction being made in favour of federal departments. The paper argues that in order to recognize unidentified persons on a microdata file, an extensive population register is required. It goes on to state that "in this country the best lists would be in the hands of the Internal Revenue Service and the Social Security Administration, but these are not available to the public". I find this argument quite unconvincing: the administrative (as opposed to the statistical research) arms of SSA and IRS might be precisely the agencies which the public might most strenuously wish to ensure do not get access to identifiable statistical records of other agencies.

4. One of the few areas where the educational material of the papers is, I believe, relatively incomplete relates to complementary disclosure. Very little is said about it in any of the papers except that by Dr. Cox. The proposed guidelines suggest only that agency policies should deal with situations where sets of tables can be algebraically manipulated in such a fashion that the result is an unacceptable disclosure. The truth of the matter is that, as demonstrated in my 1972 paper, the detection of such disclosure is mathematically equivalent to the comparison of the ranks of two typically hugh matrices--in other words not feasible in general. In spite of the very great difficulties involved, most statistical offices carry out a valiant effort to check their publication programmes for residual disclosure. This effort, although undoubtedly not complete, has nevertheless been largely successful so far--at least if the absence of complaints can be accepted as a yardstick. Thus, although agencies could not guarantee that all residual disclosure is detected, they managed to keep at least one step ahead of the risk. I think the educational value of the papers could be significantly enhanced by the inclusion of a substantive discussion of the problems related to residual disclosure, together with a documentation of the best agency practices in the field.

The paper by Dr. Cox deals with a particular procedure designed to prevent residual disclosure in business surveys. It is a description of a proposed algorithm--thus it is not, nor is it designed to be, a substitute for the educational type discussion mentioned above. In fact, the detection and avoidance of complementary disclosure can be considered as a process involving three steps. The first is the detection of complementary disclosure. The paper avoids this problem since it assumes that the classificatory variables which define statistical tables are sufficiently small in number so that all possible logical tables can explicitly be displayed and considered. For example, in business surveys if all tabulation cells are defined strictly in terms of, say, geography and SIC, then the maximum disaggregation of the data is defined by the finest level of geography cross-classified by the finest level of SIC. If there is no disclosure at this level of disaggregation, then of course there can be no disclosure at higher levels of aggregation. The next step involves checking the disclosure status of any proposed or derivable tabulation cell. This is a relatively easy step. The last is the remedial step. In other words, should a potential tabulation cell be a disclosure, it would have to be suppressed, together with enough other cells sufficient to prevent the calculation of the suppressed cell as a linear combination of the published ones. It is this last, and very difficult step, which Cox addresses explicitly. The author describes an algorithm designed to create a suppression pattern within a predetermined set of publications so as to protect against all would-be disclosures, while taking great pains to avoid over-protection (i.e. over-suppression). The great advantage of the algorithm is that it seems to work. However, its theoretical properties are as yet largely unexplored: is all residual disclosure indeed avoided, and is it avoided at minimal cost in terms of unnecessary suppressions? A more practical question relates to the dimensionality of tables involved in the publication program: the algorithm can deal with tables of relatively low dimensions, such as those defined by geography and SIC. What if other classificatory variables are involved in the definition of tables: such as employment size groups, assets in terms of ranges, use of different forms of energy, etc. Conceptually, every one of the questions on the Economic Census forms is a candidate for defining an additional dimension of the tables. At what point would the algorithm break down or become prohibitively expensive to apply? This question is of considerable interest: in the Population Census publications almost every question on the

questionnaire is actually used as a classificatory variable in at least some of the tables.

Raising these questions should not be conceived as a criticism of Dr. Cox's achievement: he has taken a giant step toward the absolutely necessary development of mass production residual disclosure analysis, corresponding to the mass production of statistical tables. I am looking forward with great anticipation to further contributions from him.

5. This brings me to my next point. With a few exceptions, the material of the papers, taken together, deals with two kinds of dissemination programs: the usual printed publications, and public use tapes. A third kind of dissemination will, I believe, enjoy increasing importance in the future: ad hoc, custom-made retrievals. As indicated elsewhere, I strongly believe that the nature of surveys and censuses will change in an important way: instead of being vehicles for the production of some predetermined tabulations, they will be viewed as sources of statistical tabulations to be used and reused. Thus the relative importance of user-requested ad hoc retrievals will increase. If I am correct in this assumption, then some important consequences follow. First of all, as the amount of information in the public domain increases, the problem of detecting residual disclosure will increase exponentially. Second, each released data point represents a potential restriction placed on future retrievals, therefore posing for statistical offices a whole new class of problems: how to balance the extent of planned publications in relation to future, and therefore unspecified, ad hoc retrieval requests.

At least in the case of our 1971 and 1976 Census dissemination program, we came to the conclusion that the only way we could deal with this problem is to literally eliminate it. In effect, by random rounding every data aggregate disseminated from the census, the problem of residual disclosure largely disappears--whether in the context of pre-planned publications or with respect to subsequent ad hoc retrievals. Of course, this introduces another trade-off over and above that of "the right to privacy vs. the need to know": namely that of the amount of data that can be disseminated before residual disclosure de facto chokes off the data supply, versus a marginal increase in the mean squared error for each disseminated data point. In light of the basic importance of this trade-off, I fully support the recommendation of Michael et al relating to a program of research and development on "the impact of deliberately introduced random noise on statistical analysis as well as on disclosure risk". I also welcome the proposed research on "software systems for providing controlled on-line access to microdata files". The provision of such on-line access would truly unlock federal statistical micro-data for extensive utilization going far beyond the pre-planned publication program, provided that software can be developed which would prevent the retrieval of data involving statistical disclosure. Having said this, I disagree with Bell with respect to the somewhat simplistic treatment of the impact of random noise on the reliability of the published data: it deals with this additional source of error in isolation rather than in the context of the overall MSE. It may well be that random rounding has a rather small effect on the MSE of reasonably large aggregates (because for large numbers the relative rounding error is small), and has a moderate effect on even relatively small aggregates because for these the sampling and non-sampling errors are generally large to begin with.

6. My last point relates to the issue of statistical matching, discussed by Radner and Muller. I largely agree with their discussion. I would want to be a little more cautious then they are with respect to this procedure. In a situation where social scientists are so hungrily looking for increasingly rich data bases, statistical matching is a dangerously attractive procedure for creating files containing the logical union of the variables involved in either of the component files. Of course, the issue is not the marginal distribution of any single variable: the two files separately can produce these. If statistical matching is carried out, it is to create a file from which the joint distribution of the variables in the component files can be studied. But it is precisely here where statistical matching, at the present time, is largely based on typically unsubstantiated assumptions. I would like to see a good deal of empirical evaluation of the validity of such joint distributions before I would suggest removing the label from this procedure: "DANGEROUS - USE WITH CAUTION".

In conclusion, I must emphasize once again my admiration of the authors and of the Statistical Policy Division of OMB for having undertaken this study. The subcommittee is dealing with one of the truly most difficult conceptual issues facing statistical offices. It is dealing with the problem with great insight and sensitivity and is clearly in the process of producing educational material of the highest quality.