Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 55–61

Comment

David Holt[1]

Platek and Särndal are concerned with very significant questions. Their primary focus is the quality of official statistics, both actual and perceived, and a recurring theme in the article is whether there could be a complete and coherent theory for national statistics production. Or is the best that can be aspired to a collection of practices, backed by some theory and empirical evaluation as a framework in which practitioners make judgements and sensible choices in the context of particular applications? They rightly focus on the needs of users: Can the data and statistics produced by the National Statistics Office be trusted, do they meet the users' needs, can the official statisticians deliver not just numbers but quality assurance, and what forms should that assurance take?

1. A Coherent Theory for All Aspects of the Survey Process

There seems to be an implication in the article that a coherent and integrated theory that covered all aspects of the survey process would in some way guarantee the quality of the resulting statistical outputs. An implication also that it would render official statisticians free from any challenge as to their conduct and as a result build greater confidence with users. Such a theory would have to cover the key elements of sample design and estimation, frame deficiencies, nonresponse, measurement error and the essential processing issues of edit and imputation. It could be argued that it should extend to cover coding and classification theory, and special topics such as index construction, seasonal adjustment and so on. And the theory should be presented in a coherent, integrated form.

From a theoretical standpoint many official statisticians would welcome a complete and coherent theory for national statistics. It would satisfy the statisticians' technical, and even aesthetic need but what impact would it have on users? Would it, as the authors contend, overcome the lack of a basis to ''challenge any particular way of operating, because no firm theory dictates the choice. Ad hoc solutions may be the result?'' I doubt that any theory, however coherent, would achieve this. However strong the theory, the application to any particular issue will require a large number of professional judgements that cannot be automatically determined. And this is not a characteristic of Official Statistics in particular but applies generally to statistics.

To illustrate the point, consider sample survey theory. This underpins survey design and estimation and is the most established aspect of a complete theory. Following Neyman (1934) the randomisation theory became established. Stratification, optimal allocation,

[1] University of Southampton, Department of Social Statistics, Southampton SO9 5NH, U.K.

cluster and multi-stage sampling, selection with unequal probability are universally used, and valuable extensions to cover two-phase sampling, rotating samples and multiplicity sampling have been developed. And yet, if we need to design a new survey, this well-established theoretical framework does not lead us, unerringly, to the optimal survey design for the new purpose. Many judgements have to be made: about the choice of stratification variables, about the use and definition of multi-stage units, about the availability and use of auxiliary information at the estimation stage. And these judgements will properly differ according to the social and economic environment in which the survey is to be conducted. Hence the ''best'' design for one country will not be the ''best'' for another.

The other factor that brings professional judgements into play is that surveys are usually designed with more than one objective in mind and this inevitably leads to compromise and a lack of optimal design for any single objective. Consider, for example, the design of a Labour Force Survey. The best designs for estimates of employment, unemployment or economic inactivity will not be the same. Furthermore the best designs for estimates of level, month on month or quarter on quarter change and of trend are all different. The proportion of rotation and the pattern of rotation will have different implications for each of the objectives and compromise choices will be necessary.

Thus even if the sampling frame is perfect and there is no nonresponse or measurement error the theory does not lead to a single ''best design.'' Of course the judgements made in arriving at a design are informed by the implications of the theory and by past experience and measures derived from previous surveys. But the judgements for the required design will depend upon characteristics of the population that are unavailable and are often estimated or ''guesstimated.'' Hence despite a very strong theoretical basis, new survey design is underpinned by the professional judgement of the official statisticians and the survey theory complements this rather than being a substitute. In short we depend on a mixture of theory and empirical evaluation and it is difficult to see how this could be different.

Hence, even though good survey design is supported by a well-established theory this does not provide any foolproof guarantee that the design chosen is free from criticism since extensive professional judgement is also required. However, in this case it could be argued that the professional judgements have an effect only on the efficiency of the final estimates. A poor design will result in larger standard errors and this effect will be apparent from the measures of quality that are generated once the survey is conducted and so users will not be mislead. The same is not true for nonresponse.

2. A Theoretical Framework for Nonresponse

It is possible to develop a theoretical framework that treats response as a random process for each member of the selected sample and which can lead to an integrated theoretical framework for both sample selection and nonresponse. One could, for example, assume that each member of the population has a probability of responding conditional on having been selected into the sample $p(r_i|s_i, x_i) > 0$ for all $i = 1, N$ where x_i is a set of auxiliary variables that affect the probability of response given selection for the i'th subject. If one also assumes that response is uninformative given s and x [i.e., $\mathrm{f}(y|r, s, x) = \mathrm{f}(y|s, x)$] then one may view both selection and response given selection as a two-phase selection process

for the achieved sample. Inference may be based on the joint randomisation distribution taken over repeated realisations of s and r with y treated as a fixed value. This theoretical framework would integrate sampling theory and nonresponse within the randomisation theory framework but not within the design-based paradigm since the probabilities of response are not determined by the sample design but depend on characteristics of the population members. This framework can be extended to informative response mechanisms although the theoretical framework becomes more complex and more dependent on underlying assumptions. Nonetheless, it is possible to develop a coherent, integrated theory for both sample design and nonresponse that will lead to unbiased estimates of population parameters (where bias is with reference to the joint randomisation distribution of s and r). Indeed commonly used methods of nonresponse adjustment such as poststratification and hot-deck imputation conditional on matching key variables are consistent with this theoretical framework.

The question is whether this integrated framework helps the official statistician or the user. We have an integrated theory, but is it applicable in the particular situation? To apply the theory the key issues are whether the auxiliary variables x that determine the probability of response have been properly identified, whether the resulting probability of response $p(r|s,x)$ is estimable and also whether response is noninformative. These assumptions are not fully testable and if they are false then the resulting estimates will be biased according to the true response mechanism even though they may be unbiased with reference to the assumed theoretical framework. And this is the nub of the issue since it is bias rather than variance that is usually the underlying concern of users in the context of nonresponse. Our theoretical framework has converted a potential bias into a component of variance but this theoretical comfort blanket will not add to the confidence of users since it depends on untestable assumptions. In the final analysis the user is dependent on the experience and professional judgement of the official statistician, underpinned by empirical studies of response mechanisms to identify as well as possible the assumptions that will lead to a plausible adjustment for nonresponse. However, official statisticians could do more to investigate the plausible size of any residual bias that may remain. For example sensitivity analyses could be used to explore the impact of different choices of imputation group or the impact of plausible levels of informative nonresponse. Such methods will not provide an estimate of the residual bias (if they could then we would use it to remove the bias) but they may provide some idea of the sensitivity of the results to the assumptions made.

3. Total Survey Error Model

And perhaps this description provides some explanation for the failure of the total survey error model to reach the full potential that many of us hoped for. Its greatest success was the contribution that the theory made to recognise the additional source of variance due to the correlated component of response variance. The use of interpenetrating designs allowed such components to be measured for both components of response variance and coder variance. However, very plausible components of variance models can be proposed for these sorts of measurement error. The difficulty comes when errors that are sources of bias are converted into components of mean squared error and are either not

estimable or depend on untestable assumptions. At this stage the theory, however elegant, may not be of practical benefit to the specific application.

This is not to say that official statisticians have done enough to measure the relative size of the various sources of error and then devote effort to those that are more significant. The problem is that such studies are not easy and often not cheap. They do not emerge as a convenient by-product of the survey process itself but require different approaches. For example, the total survey error model requires significant changes to the survey design such as interpenetrating samples and these can complicate, and even detract from, the fundamental purposes of the survey. The more components of variance one may try to disentangle, the more complex will be the required design. The results may bring benefit to future surveys or future redesigns of the current survey but the direct impact on the results of the current survey is often marginal. And the problem of empirical investigation of the size of sources of error becomes even more difficult where sources of bias are concerned. Frame deficiency studies, nonresponse studies or measurement error studies that compare the observed (or imputed) variable value with the true value depend on a source of truth that is often unavailable or that calls for very different research design to the survey itself. The research literature is not over-stocked with such examples. The growth of administrative sources and the technological advances of recent years present new opportunities to explore this vital area.

4. Quality as a Characteristic of the Data Source or the Statistical Use

Very often the concept of quality, the various sources of error and measures of survey performance are discussed in the context of a data source. We refer to sampling variance, nonresponse rates and measurement error as applied to a particular survey. We treat these as characteristics of the quality of the data source as if they were independent of the use to which the data is to be put. The implicit assumption is that if we can get the survey design and survey process ''right'' then the quality of the data is assured for the subsequent uses. But we know that the impact of almost every form of error will vary with the way that the data is used to make estimates.

Consider a typical repeated household survey design involving a stratified multi-stage design with sample rotation of a proportion of the respondents between each repetition of the survey. This is the sort of design that is used for Labour Force Surveys and other household surveys in many countries.

If we consider some of the sources of error we know that their impact will vary with the uses to which the data are put.

- Positive intra-cluster correlation will give rise to the cluster effects exhibited in the population and reflected in the multi-stage design. These will tend to increase the sampling variance for estimates of the level of unemployment for example, but reduce the sampling variance for estimates of monthly change or estimates of gross flows between employment states.
- The level and pattern of rotation will have different effects on the three uses.
- If we consider measurement error on the classification of employment status (employed, unemployed or economically inactive) we know that small levels of measurement error will not have a major impact on the estimates of level or even

change. But those same small measurement errors can have an extremely severe effect on the estimates of gross flows. So much so that many National Statistics Offices will not publish the estimates. There is a further issue here since measurement error (assuming zero mean and nonzero variance) will add to the variance for a simple estimate of population total or mean. But that same measurement error applied to the same variable will lead to bias if the variable is used as an explanatory variable in a regression analysis, for example (which is essentially the gross flow situation), or as the duration variable in a survival model. Hence the fundamental nature of the impact of measurement error depends upon the way that the variable is used.

- Nonresponse that arises from a relatively constant process over time may have an important impact on the estimates of level but perhaps less so on the estimates of change.
- Overall nonresponse rates may give us confidence that the survey has been well conducted but may mask severe nonresponse rates for particular sections of the community and hence will have a much greater potential for impact for some domain estimates.

These examples simply illustrate that measures of quality or estimates of the size of sources of error cannot be treated as characteristics of the data source but need to be linked to the use. It is this question of use-related measures of quality that is of most relevance to users. But a National Statistics Office supports a wide variety of uses and much of the data are used for secondary analyses or input into economic or other models. The modern trend is toward user-defined outputs rather than producer-determined outputs and this increases the distance between the National Statistics Office and the end use.

This does not imply that measures of quality applied to data sources are uninformative. Basic measures of quality build confidence for users since they can recognise that the process for producing the data source has been well conducted. However, additional measures of quality that reflect the use to which the data is put can be produced only when that use is identified. When users carry out secondary analyses on a micro-data source (such as the individual responses to a household survey) then some quality measures such as variance estimates for the new analyses can, in principle, be produced. But official statisticians have done little to try to anticipate other measures that might be useful. Should we, for example, produce response rates for often-used sub-groups of the data, identify subsets of the respondents for which extensive editing and imputation has been required, or characteristics of the population for which the sampling frame is deficient? Of course, it can be argued that the sub-groups of interest cannot be anticipated and that the effort cannot be justified, but we may cumulate much more understanding about the quality of the survey process if we systematically disaggregate some of the existing quality measures in this way.

5. Quality of Administrative Sources

The article focuses on surveys as the basic data source and there is very little reference to quality issues concerning administrative sources specifically. The implication is that the same issues apply to both surveys and administrative sources but there are some important differences. The main difference is that the questions used in surveys can be designed to

measure as closely as possible the concepts that are required. For administrative sources the main purpose is the administrative one, and very often the data contained in the source do not measure the required concept. For example, the ''registered unemployed:'' those who claim unemployment-related state benefits are not the same as those who would be classified as unemployed using the ILO definitions. Often, too, an administrative source can be deficient for important sections of society. The quality issue that dominates the use of such administrative sources for statistical purposes is not accuracy but relevance. And this requires a different approach from that of the official statistician. Investigations are needed to understand the nature of the difference between the available information and the concept that we seek to measure. The size and direction of differences needs to be quantified. If the relationship between the administrative source and the required statistic changes over time then this too needs to be monitored and reported on. Such work is not usually a by-product of the statistical compilation process from the administrative source. It often requires separate, independent studies that can be resource intensive. But when administrative sources are used for high profile statistical purposes this sort of quality assurance is essential. The official statisticians have a duty to inform users about the differences and to provide an interpretation that links the administrative source to the required statistical measure.

Another issue is that often the administrative sources are very large and comprehensive for their own purpose but may be deficient for the population as a whole. For example, personal tax records are a valuable source of information on personal income but omit low earners whose income is below the tax threshold. For some purposes this is a serious deficiency.

Again, issues of this kind require a different approach to quality issues than those applied to surveys although in this case the issue is analogous to the problem of a frame deficiency for a survey. To investigate this fully requires an assessment of the undercoverage and the impact that this may have on the statistical outputs. Often other sources, such as household surveys, can throw light on the issue but these too may be imperfect. In the case of personal income, for example, the measurement in a household survey may be less accurate than from tax records or may omit some sources. Also survey nonresponse may be higher for some types of household who are more likely to have lower incomes. These issues are not easy to resolve and call for detailed investigation and reconciliation that is very different from the sort of quality assurance and quality measures associated with standard survey analysis.

There is a strong case that quality assessment related to the use of administrative sources is different from that used for surveys and merits separate consideration. The growing demand for sub-national estimates is increasing the use of administrative sources and a framework for quality assessment is needed.

6. Data Coherence and User Confidence

We know less than we should about the way that users develop confidence in the quality of Official Statistics. But what we do know is that ''users'' are not homogeneous. At the extremes, some are very knowledgeable, statistically capable and have a detailed understanding of their field of interest. Others are statistical novices and depend entirely on

the National Statistical Office for the statistics and their interpretation. It follows that users will develop confidence in the quality of the outputs in very different ways. It could be argued that there are two fundamentally different approaches to confidence building. One route, the process route, is through the quality measures that are produced by official statisticians and the other route, the outcome route, is based upon the face value validity that the outputs appear to have.

Even for the most advanced users, it is not clear to what extent the quality measures that official statisticians produce are actively used. One needs to separate the fact that such measures are produced from their actual content. A National Statistics Office that produces quality measures in a systematic way is publicly declaring its commitment to quality and demonstrating its professional competence. This itself will build confidence in the professionalism of the Office and through this it will build confidence in the quality of the statistics produced. And capable users who develop confidence through this process route will consciously or unconsciously convey that confidence to less informed users.

The second route, based upon the face value validity of the outputs, depends upon the coherence and consistency that the statistical outputs display. Estimates of employment, for example, derived from surveys of households or employers need to be coherent, or discrepancies understood. The coherence between output, employment and productivity brings reassurance as, more generally, does the coherence of National Accounts and the various economic aggregates and indexes that are essential components of the system. Time series that display erratic and unpredictable characteristics will undermine confidence even if the process quality measures are produced and reported on. The approach to producing this type of quality assurance is much closer to the sort of ex-post analysis and investigation that is needed when assuring the quality of statistics derived from administrative sources. There is a clear need for greater emphasis on quality assurance of this kind and for a framework of quality assessment to be developed.

7. Conclusion

In summary, like Platek and Särndal, I have placed more emphasis on the development of a more comprehensive framework for quality assessment than for a single integrated theory to be developed. We have to recognise that professional judgements are called for in all aspects of the official statistician's work. No theoretical development will alter this and we should welcome the challenge of providing adequate justification for the judgements that we make. However, a stronger framework of quality assessment is needed. In particular the use of administrative sources and the need to demonstrate the coherence of related statistical outputs should be more emphasised.

Received December 2000