EIA ACTIVITIES TO IMPROVE DATA QUALITY

Douglas R. Hale, Energy Information Administration

I am pleased to participate in a public discussion on the quality of data published by statistical agencies. Judging from the dearth of scholarly and public policy papers in the area, one would suppose there are no problems. Those of us who are veterans of statistical agencies would suspect a different conclusion. The analysis of data quality is difficult, tends to result in heavy, dull reports, and makes nearly everyone mad. I hope that this session will stimulate thinking into more productive ways of assessing and reporting data quality.

My part of this discussion will be to describe how my agency, the Energy Information Administration (EIA), ensures that the data series it publishes are what we intend them to be. Given the quantities that are to be estimated, EIA's approach to quality assurance is summarized in Figure 1. We view quality assurance as a continuous process involving evaluation, reporting findings to responsible managers, and stimulating change by researching and developing better methods of doing our work, establishing standards and guidelines, and demonstrating improved methods. The starting point for quality assurance, and the one I will address today, is evaluation. Some of the hardest questions we ask involve whether a data collection is designed and operated to meet its goals for relevancy, consistency, and accuracy.

EIA relies on validation studies, quality audits, and external data comparisons to address these issues. The purpose of my talk is to describe the content of these activities, to illustrate the nature of the findings from these investigations, and to comment on their general strengths and weaknesses.

1. VALIDATION STUDIES

EIA was formed in 1977 which was a time of Congressional and public skepticism about the validity of energy data. The new agency was charged with bringing together diverse groups from across government to create a coherent National Energy Information System. To underline its concern with the integrity of the data and forecasts, the Congress took the extraordinary step of creating the Professional Audit Review Team (PART) to investigate the agency's performance. PART still reports annually to Congress. EIA's formal quality assurance activities began in an environment characterized by deep Congressional suspicion regarding the validity of the data, Congressional and public demands for policy relevant statistics, and all the problems of an agency pulled together from a dozen reluctant donors.

In response to these varied pressures, EIA developed the concept of a validation study. These ambitious studies were intended to:

- describe the history, technology, and institutional structure of the various energy industries,
- present the legal basis for data collection and to recount the history of such efforts,
- construct formal models of error in data collections with special emphasis on systematic bias (nonsampling error), and
- recommend alternative ways of collecting and verifying data based on the results of the error model,

Perhaps the most lasting contribution of this work was the building of error models of energy industries. Each error model was intended to be a formal description of the data collection together with an exhaustive error classification scheme (error taxonomy) specific to the target industry.

The starting point for these efforts was a schematic diagram of the stocks and flows, both in value and physical terms, within the industry and across related industries. Figure 2 shows one such schematic for the natural gas industry at the national level. The diagrams were the agency's formal concept of the industry's structure. The picture could be right or wrong, but it was explicit and subjected to thorough, critical review. The next step was to show explicitly the intended measurement points. An example is in Figure 3. At this stage, it was common to discover that the existing data collection made technological and institutional assumptions that were false. Typically, supply sources were missed, implied measurements were not, or could not be made, or the definitions were found to allow excessive latitude. Next, the measurement means - scales, engineering estimating techniques, grab and other samples, gauges, etc. - were examined for repeatability and reliability. Concepts such as reserves, usable inventory, minimum operating levels, capacity, and even production turned out to be elusive. The final piece of the error taxonomy was to examine the path of the data from measurement to reporting.

The difficulty came in quantitatively estimating the individual errors and adding them up. Quantifying certain errors involved extremely expensive audits of basic company records. While the direction of specific errors was stable, magnitudes varied across companies, and were highly sensitive to changes in tax law and accounting practice. Even in those cases where the errors could be quantified, no good means was found for adding up offsetting errors.

Consequently, the error models never advanced to the point of yielding an estimate of such things as the error in national gasoline production. Nevertheless, the discipline of attempting to build these models yielded worthwhile results.

The fourteen validation studies completed during the period from February 1979 through September 1982 resulted in 148 discrete recommendations for EIA's consideration. Over half of

the recommendations were adopted: the others were not adopted mainly because of price deregulation, cost, and controversy over their technical feasibility. As shown in Table 1, these recommendations were distributed across the areas of definitions and categorization of industrial processes, respondent understanding, frames, editing, and data processing.

EIA no longer conducts full scale validation studies. The agency has matured; there is now a better understanding of the basic features of the energy industries. Cost is also an issue - the studies typically started out costing around half a million dollars. There is also less pressure on the agency to challenge industry data. In part, this reflects the validation studies' general finding that companies reports were generally dependable.

The idea of formally describing the data collection and accounting for the sources of error has survived. EIA's current practice is to periodically rebuild the error model in the light of technological and institutional changes. This is particularly important prior to beginning the detailed design of a new data collection. The effort to quantify each and every potential error and then aggregate up to a total has been abandoned because of high costs and apparent technical infeasibility. Instead, when an error category is believed to be significant, special studies of that particular problem are initiated.

2. QUALITY AUDITS

Once a system is designed and built, the emphasis switches to its operation. EIA uses quality audits to determine whether respondent data is being properly and efficiently processed. To a large degree, these audits are concerned with good housekeeping. The existence and accuracy of documentation; whether computer programs and edit checks work and are being used; nonresponse and item nonresponse follow-up are typical concerns. The value of these audits is that they help ensure consistency of good practice through the life of the system by bringing operational problems to the attention of management.

The audits generally proceed in three phases. In the first phase, documentation is used to construct a workflow from data receipt through publication which identified the major components of the system, their purpose, inputs, outputs, operations, and interactions.

The second phase of an audit involves the comparison of actual system operations to the workflow constructed from the documentation. This is usually where the excitement begins. The final phase of an audit is to independently replicate system results and to evaluate selected practices in the operations of the system.

In the last year and a half, we have completed about a dozen quality audits. They are generally done in a 2- to 3-month period by contractors working for a full-time government project leader, and cost on the order of $25,000 each. Table 2 illustrates some of the negative findings from four typical audits. We do find some good things; when we do, we incorporate the good ideas in standards and guidelines. We believe that these quality audits have been an effective, relatively low cost means of detecting operating problems and encouraging improved performance.

3. EXTERNAL DATA COMPARISONS

To compliment the inward-looking error models and quality audits, EIA uses data comparisons as a global check on the plausibility of its estimates. In each of the last 4 years, EIA has published studies which systematically compare given EIA data series with those from other agencies or those produced by different means.

Comparisons with other series cannot address the issue of accuracy directly: when there are differences among series, the question of which series is the more reliable remains. Nevertheless, comparisons do indicate the range of estimates from different measurement approaches. In some cases, one can validly argue that the various estimates bound the target values. Comparisons also detect changes in previously stable relationships among series that may signal trouble. Of course, if two independent series are found to correspond closely, it increases confidence in the accuracy of each.

Some of you may have heard Renee Miller describe what we found last year in a comparative assessment of EIA petroleum and natural gas data. At the risk of stealing a bit of her thunder, I will use some of her recent findings.

Renee compared coal production data from the EIA-7A, a form which is sent to all mines producing 10,000 or more short tons a year, with data from the Mine Safety and Health Administration (MSHA) and various state mining agencies. She found that annual production figures obtained from MSHA were consistently lower than those obtained from the EIA-7A survey from 1978 to 1982.

Investigation of the 1981 shortfall of 4 percent in Illinois indicated that it was the result of incomplete or inconsistent MSHA production data for a few mines. For about three-quarters of the mines reporting to both agencies, MSHA figures were within 2 percent of the EIA-7A figures. By contract, the relatively close 1981 figures for Virginia did not reflect good correspondence between the series. For one mine, the MSHA tonnage was ten times the EIA tonnage. Excluding that mine resulted in an MSHA figure that was 94 percent of the EIA-7A estimate in 1981.

Unlike the MSHA data, the estimates of production from the state agencies were generally larger than or almost equal to the EIA-7A estimates. The discrepancies generally could be traced to:

- Different definitions of coal production (for example, raw versus clean coal production),
- Different respondents to the data collections, and
- Reporting or processing errors in the state mining agency data collections.

While these results are generally encouraging, previous studies have detected problems with frames and survey design and have indicated

situations when EIA data are most likely to diverge from other series. We have found these comparisons to be an invaluable means of better understanding, controlling, and explaining our data collections.

4. CONCLUSIONS

This morning, I have discussed three kinds of evaluation reports that EIA has conducted. Validation studies were helpful in educating the new agency about the specific problems of collecting reliable data in the various energy supply industries and in assuring the public that the basic data was trustworthy. Although EIA no longer conducts full scale validation studies, the notion of a formal error model has been useful for ensuring that the data collection design is appropriate to the technology and institutional structure of the target industry. Quality audits are the best single means we have found of detecting deterioration in our data processing. Comparative analyses are useful for assuring that the data collection is producing reliable information.

REFERENCE

[1] Energy Information Administration, U.S. Department of Energy, An Assessment of the Quality of Selected EIA Data Series: Coal and Electric Power Data for 1977-1982, DOE/EIA - 0292(83). Washington, D.C., April 1984.

FIGURE 1. QUALITY ASSURANCE AT EIA

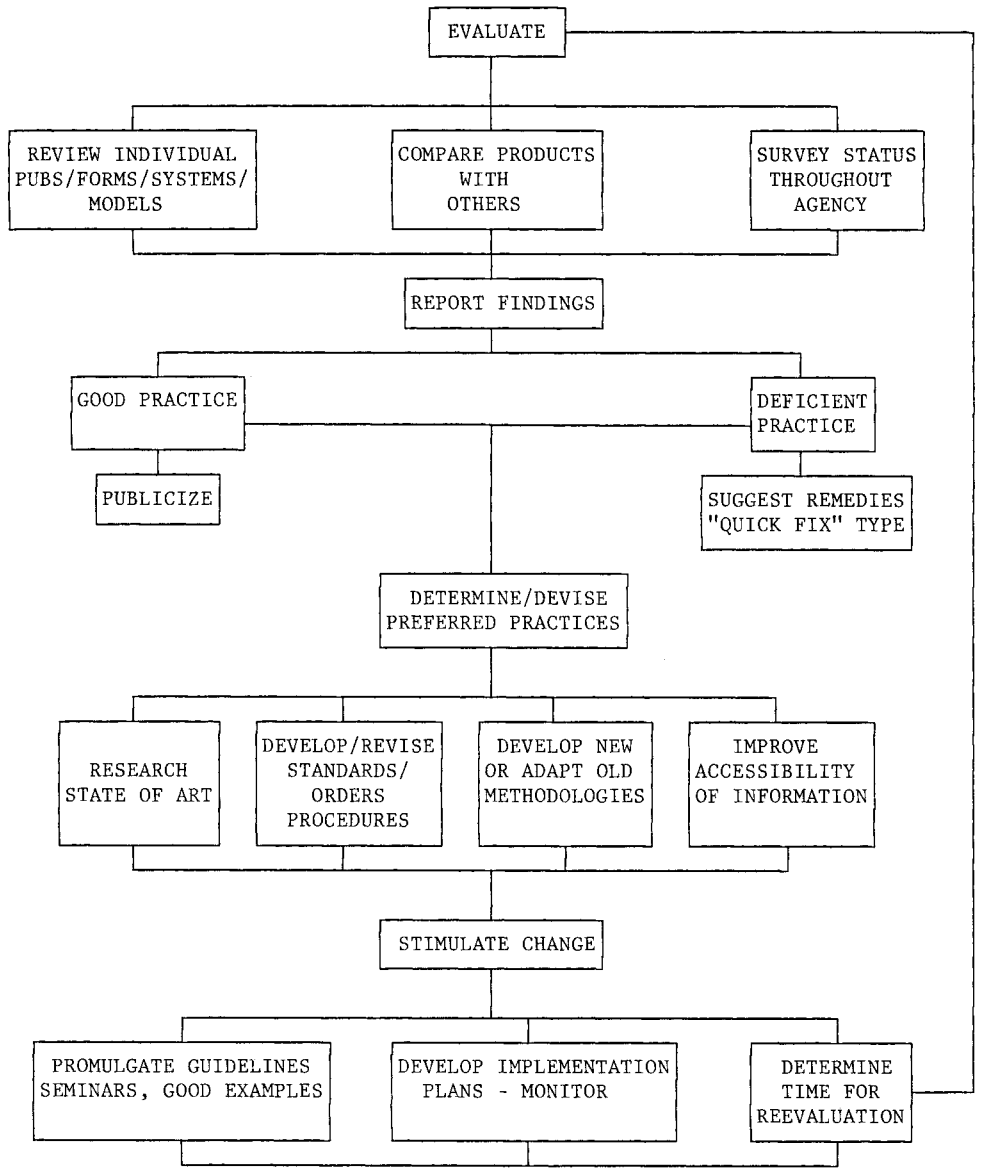

Figure 2. Simplified Natural Gas Physical Flow Chart

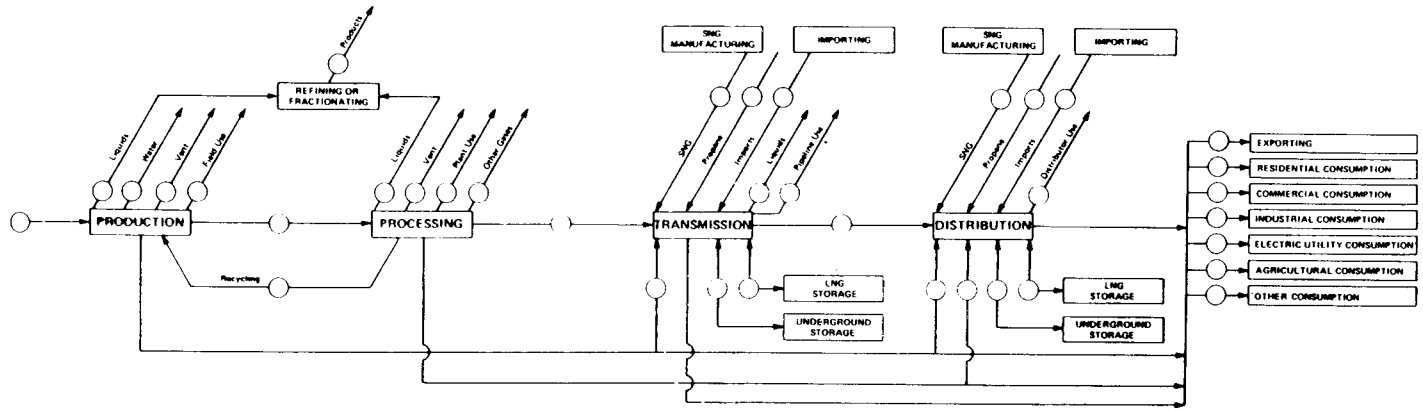

Figure 3. Measurement Points and Data Collection Forms

		BOM—6—1340—A	BOM—6—1341—A	EIA—19	EIA—23	EIA—26	EIA—50	EIA—64	EIA—81	EIA—82	EIA—87	EIA—88	EIA—146	EIA—173	EIA—174	EIA—191	FEA—P—319—M—0	FERC—16	FERC—121	FERC—122[a]	FERC—123[a]	FERC—124[a]	FPC—R0010	FPC—R0384	FPC—R0326	FPC—2	FPC/FERC—2A	FPC—8	FPC/FERC—11	FPC—14	FPC—15	FPC—108	FPC—314A	FPC—314B	FPC—334	MA—13K	MC—138	USGS—6—152	USGS—6—329	USGS—6—329—1	USGS—6—614A	NGPS
FREQUENCY OF SUBMISSION[b]		A	A	E	A	A	A	E	I	I	E	E	I	A	A	E	S	I	I	I	I	I	A	I	A	A	E	E	A	A	A	A	I	A	I	A	A	E	E	E	E
A1	GAS FROM WELLS		•		•	•												C							•	•		•		C										•
A2	LIQUID HYDROCARBONS				•	•					•																								•	•	•	•		
A3	WATER																																				•	•	•		
A4	VENTED OR FLARED GAS		•																																		•	•	•		•
A5	FIELD USE		•																						•		•										•	•	•		
A6	LIQUID PRODUCTS					•		•			•														•		•								•					
B1	PRODUCTION SALES TO PROCESSING		•			•		•					•							•	•	•			•		•			•	•	•		•	•	•			•
B2	LIQUID HYDROCARBONS			•				•			•														•		•								•					
B3	VENTED OR FLARED GAS		•					•																																	
B4	PLANT USE		•	•				•					•														•								•					
B5	OTHER GASES																																								
B6	RECYCLED GAS		•					•																														•	•	•		•
C1	PROCESSING SALES TO TRANSMISSION		•					•					•					•							•	•	•			C					•					
C2	SNG			•														•								•	•			C										
C3	PROPANE AIR INJECTION														•			•								•	•														
C4	IMPORTS					•												•											•	C										
C5	LIQUID HYDROCARBONS																									•			•												
C6	PIPELINE USE		•										•					•								•			•											
C7	PRODUCTION SALES TO TRANSMISSION		•			•												•		•	•	•			•	•	•			C	•	•	•		•		•	•		•
C8	UNDERGROUND STORAGE		•													•		•								•		•	•											
C9	LNG STORAGE																									•															
D1	TRANSMISSION SALES TO DISTRIBUTION	•	•				•		•	•			•					•								•	•		•											
D2	SNG			•			•		•	•			•																											
D3	PROPANE AIR INJECTION						•		•	•			•		•																										
D4	IMPORTS						•		•	•			•																	•										
D5	DISTRIBUTOR USE	•	•										•																												
D6	PRODUCTION SALES TO DISTRIBUTION	•	•			•	•		•	•			•							•	•	•									•	•	•				•	•	•	•
D7	PROCESSING SALES TO DISTRIBUTION		•				•	•	•	•			•																											
D8	UNDERGROUND STORAGE	•	•						•	•			•			•																								
D9	LNG STORAGE								•				•																											
E1	EXPORTS					•							•																	•										
E2	RESIDENTIAL CONSUMPTION	•	•				•		•	•			•																											
E3	COMMERCIAL CONSUMPTION	•	•				•		•	•			•													•	•													
E4	INDUSTRIAL CONSUMPTION	•	•				•		•	•			•													•	•													
E5	ELECTRIC UTILITY CONSUMPTION	•	•				•		•	•			•													•	•		•											
E6	AGRICULTURAL CONSUMPTION												•																											
E7	OTHER CONSUMPTION	•	•				•		•	•			•													•	•		•											

[a] On March 14, 1980, the April 1, 1980, filing date for these forms was extended until further notice (Federal Register, vol 45, no 60, March 26, 1980, p. 19546).

[b] A - annually, S - semiannually, M - monthly or twice monthly, I - irregularly.

[c] Indicates estimated or projected value only

Table 1. Validation Study Recommendations by Functional Categories

Categories	Number of Recommendations Implemented	Number of Recommendations Not Implemented EIA Systems
Changes to Form Involving:		
• Definitions and Categorization of Industrial Processes ...	21	20
• Respondent Understanding, Instructions, Procedures	16	17
• Frames (list of potential and actual survey participants)	13	9
Changes to System Involving:		
• Editing Procedures	16	5
• Data Processing	14	17
Totals	80	68

Table 2. Results of Four Audits

Problem	Survey Audited			
	1	2	3	4
1. Publication tables were created manually, even though data processing was automated.	X	X	X	
2. Components of the data system, such as source codes and JCL streams, were not organized consistent with EIA standards.	X		X	X
3. Statistical editing procedures were not implemented correctly and did not serve the intended purpose.				X
4. Performance statistics were not maintained for every survey cycle.				X
5. Software to update basic survey frame information was nonexistent.			X	X
6. Software and file structures did not reflect changes in survey forms; for example, gaps were found in each data record causing waste in computer resource utilization.	X			X
7. Survey documentation was not up to date.	X		X	X