Editor Judgment Effect: Modeling a Key Component of Nonsampling Error in Administrative Data

Kimberly Henry, Yahia Ahmed, and Ellen Legel
Statistics of Income, P.O. Box 2608, Washington, D.C. 20013-2608

Presented at the 2004 American Statistical Association Meetings

Keywords: Data Quality, Net Difference Rate, Index of Inconsistency

1. Introduction

This paper is a modest attempt to model a key component of nonsampling error in administrative data, particularly tax data. Tax data items present obstacles for statistical uses that are far outweighed by the fact that responses on tax returns are likely to be more accurate than financial-related responses to general surveys. These obstacles lead to a kind of nonsampling error that we refer to as *editor judgment error*. The IRS Statistics of Income (SOI) Division developed a processing procedure called statistical editing to abstract tax return data for statistical purposes. Statistical editing helps overcome limitations inherent in tax return statistics and achieves certain statistical definitions desired by data users. Statistical editing involves adjusting certain taxpayer entries based on supplemental information reported elsewhere on the tax return (such as attached schedules that support a reported total). It is minimal in producing SOI's individual income tax return statistics, but a major factor in producing its corporation income tax return statistics.

In Section 2, we describe the SOI corporate sample design, identify sources of nonsampling error, and define the term "editor judgment error." Section 3 describes current SOI editing and quality review processes, while Section 4 outlines the purpose of our study and its limitations. Section 5 discusses bias and variance component models, which were adapted from simple response error measurement models. Results and conclusions are summarized in Section 6.

2. Sample Design Description and Nonsampling Error Sources

The data for this study were abstracted from the 2001 SOI Corporate sample, which consisted of corporations that filed income tax returns with accounting periods ending between July 1, 2001, and June 30, 2002. The realized 2001 sample contained 147,093 returns selected from a population of 5,563,663. The sample is a stratified random sample, where stratification is based on 1120 form type. Within form type, further stratification is achieved by use of either size of assets alone, or both size of assets and a measure of income. A Bernoulli sample is selected independently from each stratum, with rates ranging from 0.25 to 100 percent. The sample is selected weekly as the Form 1120 returns are posted to the IRS Business Master File. It takes two years to select the sample due to the combination of noncalendar year filing and the six-month extension options.

Sampling errors arise from using a sample instead of a census, and SOI publishes them in the form of Coefficients of Variation (IRS, 2001, pp. 29-36). Nonsampling errors include all others, such as coverage, nonresponse, measurement, and processing errors.

Coverage errors, when a unit is not available on the sampling frame, can occur if a corporation files an extension. Imputation procedures using adjusted prior-year data are used to correct for coverage errors in large companies.

Missing data, or nonresponse errors, occur when other IRS functions have returns selected for the sample, rendering them unavailable for SOI processing. Imputation procedures and weighting adjustments are used to adjust for missing large and small companies, respectively. Noncoverage imputation and missing returns represented 0.03 percent and 0.22 percent of the 2001 sample, respectively (IRS, 2001, pp. 7-14).

Measurement errors occur when a taxpayer enters an incorrect value, for various reasons. SOI does not sample amended returns or contact taxpayers.

Finally, processing errors occur while abstracting, transcribing, and cleaning the data. Since the editors abstract administrative data from tax returns and enter them into SOI database systems for statistical purposes, editor judgment error falls into this nonsampling error category. However, it is more than transcription error because certain judgments are required from the editors due to a combination of

transcribing data collected for tax liability, which is subject to different corporate accounting practices, and study standards created for statistical purposes.

3. Current SOI Editing and Quality Review Processes

Fifty-nine editors at two IRS Service Centers abstracted approximately 1,400 corporate tax return items for the 2001 sample. This data abstraction process was complicated due to the following factors:

- The extracted items from any given return often require totals to be constructed from various other items on other parts of the return.
- There are currently ten form types, with different layouts, schedules, and attachments, so data extraction is not uniform across form type.
- There is no legal requirement that a corporation meet its tax return filing requirements by filling out, line by line, the entire U.S. tax return form. Some returns are also exempt from filling out entire sections; for example, currently, Form 1120 returns with total assets and total receipts below $250,000 do not have to report their balance sheet items.
- There is no single accepted method of corporate accounting used throughout the country. For example, different companies may report the same data item, (such as deposits, a subset of other current liabilities), on different lines of the tax form.

Despite complexities such as those listed above, study standards place SOI's editors in a position to make judgments during data abstraction. Errors in these judgments are the largest source of editor error in the corporate sample.

To assist the editors, SOI's National Office (NO) staff in Washington, DC implement many procedures that attempt to make the editing process consistent with the 1120 study standards and reduce editor effect. This is similar to the concept of standardized interviewing used in other survey organizations. For example:

- Detailed editing instructions are prepared every year – the 2001 manual contained more than 900 pages.
- Over 700 computerized tests are performed on abstracted data to ensure certain accounting conditions are satisfied, such as balanced totals or absence of consistent amounts between front-page items and attached schedules. All tests are reviewed and tested by NO staff the year prior to data abstraction in a process called Systems Acceptability Testing.
- The staff build utilities into the edit computer system that offer industry-specific suggestions, guidelines, and requirements for particular sections of the form.
- They review and monitor the sample throughout the program year for unusual accounting conditions and codes. During the last four months, the largest corporations within each industry are reviewed as well as the largest industry differences across asset classes.
- The NO staff conduct extensive edit training and review all items on all returns edited during certain periods of the program year to overcome inexperience due to new tax laws, edit instructions, codes, or even an entirely new program. For example, editors improving throughout the year are given more complicated returns, the first of which were completely reviewed with their supervisors.

While complete review was an excellent training tool, the editors knew in advance which returns were going to be reviewed. For the purposes of our study the returns may have been biased, so they were omitted from analysis.

During data editing, approximately fifty returns were randomly selected for each editor for quality review. Once an editor's return was selected for review, another editor on the same team independently re-edited it. After the returns were compared item-by-item and discrepancies were stored in SOI databases, the editors' supervisor determined the correct value (either the first editor's value, the second's, both, or neither). Any amounts that differed by less than $10, along with character, display, and generated item mismatches were omitted from quality review. We used only the first editor values because they are the final file values and the second editor knew which returns were for review. Assuming that a taxpayer is correct, the errors described in Table 1 are used to determine service center accuracy ratings and we included all of them:

Table 1: Types of Errors

Type of Error	Description
Amount	An incorrect amount was entered in an item.
Omitted Entry	A zero or blank item that should have a code/amount present.
Extra entry	An item with a code/amount in it should have been blank or zero.
Entry on omitted form	An item was not edited because the form or schedule was not edited.
Improper allocation	An amount that should have been allocated to another item was not moved or was moved incorrectly.

Improper allocations were the most frequent errors, so this type of error is illustrated in Table 2.

Table 2: Improper Allocation Example

Item	Edited Amount	Correct Amount	Error
A	1,000.00	0.00	1,000.00
B	0.00	1,000.00	-1,000.00
C	2,000.00	2,000.00	0.00
Total	3,000.00	3,000.00	0.00

Here, for three hypothetical items A, B, and C (which may not be located on the same page, form, or attachment), both totals match; the system will not catch the error despite errors in two of three items. An important aspect of improper allocation errors is that they often result in net error effects of zero: here, errors in items A and B cancel each other out. This is important when calculating national-level estimates for totals, but a concern for estimates of A or B.

4. Study Purpose and Limitations

The quality review system was developed to check edit manuals, measure training effectiveness, and evaluate the editors. As previously mentioned, approximately fifty returns were randomly selected for each of the fifty-nine editors for quality review. Given this pre-existing quality review system, our goal was to develop quality performance statistics and quantify the editor effect.

Table 3: Errors and Error Rates, Quality Review Study vs. Our Study

Item	QR Study	Our Study
# returns	3,080	373
# errors	9,229	760
# errors possible	33,880	4,103
error rate	.272	.185

As shown in Table 3, data used for our study were a subsample of 373 returns from the 3,080 quality review returns. All 3,080 returns were not included because returns with assets more than $250 million were only edited by a group of the most experienced editors, then reviewed by NO staff. In order to compare across all form types, service centers, teams within service, and editors within teams, we selected this subsample, which consists of all Form 1120 and Form 1120 Regulated Investment Company returns with total assets less than $250 million. Most importantly, all editors edit these returns during the program year, regardless of their experience. There were 73,115 of these returns in the corporate sample, for which NO staff relied on the editors' judgment for most of them because they were reviewed only under special circumstances. Our subsample is small compared to the SOI sample (about 0.51 percent), so the results from this relatively small sample were analyzed assuming the observations were from independent, identically distributed random variables and sample weights were not used (Brick et al., 1996).

We selected eleven variables from the balance sheet and income statement sections of the returns in our study that were of interest to our subject-matter specialist; it is obvious from their names that many are ambiguous. Table 4 displays the number of errors and error rates for the eleven selected variables.

Table 4: Number of Errors and Error Rate, by Item

Item	# Errors	Error Rate
Gross Receipts	58	0.014
Other Assets	68	0.017
Other Costs	72	0.018
Other Current Assets	57	0.014
Other Current Liabilities	58	0.014
Other Deductions	110	0.027
Other Income	81	0.020
Other Investments	76	0.019
Total Deductions	62	0.015
Total Income	63	0.015
Trade Notes/Accounts Receivable	55	0.013

Error rate is equal to number of errors out of the 4,103 errors possible. Other Deductions has the highest error rate of 2.7 percent because Deduction item editing tasks are more complicated due to complex and varying accounting rules.

5. Bias Estimation and Variance Decomposition

Measurement error modeling was first proposed by Hansen et al. (1952) and Seth and Sukhatme (1952). Their model specified that a single observation y_i from a randomly selected respondent i is the sum of two terms: a true value, m_i, and an error term, e_i. Mathematically, this is written as

$$y_i = m_i + e_i \qquad (5.1)$$

While we did not measure response error, we adopted these models to our data to measure editor judgment

error. In model (5.1), m_i, the true value, is a random variable whose distribution depends on the sample design. The distribution of the editor error variable e_i is conceptual; it could be viewed as sampling from a hypothetical population of errors. Thus, the assumptions for model (5.1) are

$$E[e_i|i] = B_i \neq 0$$
$$Var[e_i|i] = s_i^2$$
$$E[s_i^2] = s^2$$
$$Cov[e_i e_j] = 0,\ i \neq j$$

In words, a systematic bias exists because the mean of the errors is not zero and the variances are not equal. Also, errors are uncorrelated: the errors for a first or second edited return do not affect other returns in the same edit period and errors across edit periods for the same return are uncorrelated.

Assuming unrestricted simple random sampling,

$$E[m_i] = \bar{m}$$
$$V[m_i] = s_m^2$$
$$Cov[m_i, m_j] = 0,\ i \neq j$$

In our study, the observed value is the first editor's value on the file, while the true value is either the first or second editor's value (whichever was determined to be correct by their supervisor), and i denotes unit. It deserves mention that model (5.1) has potential weaknesses, particularly if the first and second editor's values are correlated, but it can provide a useful approximation for the editor's contribution of error. The model also allows for calculating statistics to measure editor accuracy further than number of errors out of number of errors possible.

Under model (5.1), we assume that the first editor's error term no longer averages to zero, possibly due to editor bias, defined as

$$B = \sum_{i=1}^{N} (y_i - m_i) \qquad (5.2)$$

The bias can be estimated by the *Net Difference Rate* (NDR), which is given by

$$NDR = \bar{y} - \bar{m} \qquad (5.3)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and n is the sample size.

It can be shown that if m_i is the true value, then the expected value of the NDR is the bias and its variance exists (Biemer and Atkinson, 1992). Table 6 shows the estimated NDR and BR values for our eleven items, where the *Bias Ratio* (BR) measures the relative magnitude of bias to the standard error of the NDR. Negative bias values should be interpreted as editors underestimating variables and positive NDR estimates indicate overestimates.

Table 6: Net Difference Rate, by Item

Item	NDR	BR
Gross Receipts	-749,441	0.16809
Other Assets	293,125	0.23662
Other Costs	7,847	0.00683
Other Current Assets	361,062	0.19090
Other Current Liabilities	1,989,871	0.26820
Other Deductions	-958,930	0.26017
Other Income	-662,720	0.27392
Other Investments	-59,372	0.03116
Total Deductions	543,972	0.21601
Total Income	500,441	0.16296
Trade Notes	32,635	0.01395

At first, the NDR estimates look very large in both directions. Since most errors are improper allocations, an entire amount is determined to be in error. The BR estimates, however, are all quite small, which implies that editor judgment appears to be a random error, not a systematic error as first assumed. Since all bias ratios are less than 1, confidence interval probabilities for SOI sample estimates from these particular returns are almost unaffected (Cochran, 1977). Therefore, we can assume that $E[e_i|i] = b_i = 0$, i.e., the editor error averages to zero because it is a random error.

Since simple random sampling is assumed and the bias is zero, it can be shown that the variance of a mean over all possible editing review samples and all possible editing trials can be decomposed into

$$Var[\bar{y}] = Var[\bar{m}] + \frac{s^2}{n} \qquad (5.5)$$
$$= SV + EV$$

The *sampling variance*, SV, is the ordinary variance with no editor error. The *editor variance*, EV, is the

variability of returns averaged over conceptual repetitions of the editing under the same conditions.

Hansen et al. (1964) define the *Index of Inconsistency* (IOI) as

$$IOI = \frac{EV}{SV + EV} \qquad (5.6)$$

which we use to estimate the proportion of random errors associated with editor judgment error in total variance. Estimated IOI values are shown in Table 5.

Table 5: Index of Inconsistency, by Item

Item	IOI
Gross Receipts	0.0155
Other Assets	0.3084
Other Costs	0.0140
Other Current Assets	0.1526
Other Current Liabilities	0.1829
Other Deductions	0.2091
Other Income	0.1365
Other Investments	0.0464
Total Deductions	0.0247
Total Income	0.0336
Trade Notes	0.0370

Other Assets (0.3084) and Other Deductions (0.2091) are the items with the greatest proportion of editor judgment error. All other IOI estimates were less than 0.2, which is a small proportion compared to other surveys (Lessler and Kalsbeek, Ch. 11).

6. Conclusions

To summarize, despite large NDR values in both directions due to editor judgment errors, particularly improper allocations, the expected value of the bias for all items is zero. Further analysis of the NDR yielded different results by edit team. Internal examinations of NDR comparison graphs by team, item, and editor were useful in identifying strengths and areas of editing improvement that can be addressed through training. Third, the BR values are also small, much less than the upper-bound of 1.1 stated by Cochran (1977).

Most importantly, editor judgment error for these returns is a variable error, not a systematic error. Variance decomposition for our eleven items showed editor variance is a small component of total variance. Variable errors tend to cancel each other out. Overall, our measure demonstrate high quality editing, so reliance on their judgment is justified when every possible error scenario cannot be programmed, foreseen, or identified by National Office Staff.

This study is a first attempt, and a modest one, to quantify the effect of SOI's editors on data quality. Our encouraging results are a strong argument of the necessity for more research. We examined the simplest tax returns in order to compare the editors, returns whose errors have the smallest impact on overall quality of national estimates. The largest errors associated with the largest tax returns require a separate error measurement study because they are sampled with certainty and therefore do not contribute to sampling error. Further, the validity of taxpayer values, which are assumed to be correct when corporate returns reach SOI, is another area deserving examination.

Resources

Biemer, P. and Atkinson, D., "Estimation of Measurement Bias Using a Model Prediction Approach," *1992 Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 64-73.

Biemer, P.P. and Fesco, R.S. (1995), "Evaluating and Controlling Measurement Error in Business Surveys," *Business Survey Methods*, John Wiley & Sons, New York, pp. 257-281.

Biemer, P. and Stokes, L. (1991), *Measurement Errors in Surveys*, John Wiley & Sons, Ch. 24, pp. 487-516.

Brick, M., Kim, K., Nolin, M.J., and Collins, M. (1996), "Estimation of Response Bias in the NHES:95 Adult Education Survey," *Working Paper Series*, National Center for Education Statistics, Washington, DC.

Cochran, W. (1977), *Sampling Techniques*, John Wiley & Sons, New York, p. 380.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R., (2004), *Survey Methodology*, John Wiley & Sons, New York, p. 276.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1952), *Sample Survey Methods and Theory*, John Wiley & Sons, New York, Volume II.

Hansen, M.H., Hurwitz, W.N, and Pritzker, L. (1964), "The estimation and interpenetration of gross differences and the simple response variance," *Contribution to statistics*, in C.R. Rao (ed.) Pergamon Press, Oxford, and Statistical Publishing Society, Calcutta, pp. 111-136.

Internal Revenue Service, Statistics of Income–2001, *Corporation Income Tax Returns*, Washington, DC

Lessler, J.T. and Kalsbeek, W.D. (1992), *Nonsampling Errors in Surveys*. John Wiley & Sons, New York

Sukhatme, P.V. and Seth, G.R. (1952), "Non-sampling errors in surveys," *Journal of Indian Society of Agricultural Statistics*, pp. 5-51.