Journal of Data Science **1**(2003), 1-25

Data Quality Effects of Alternative Edit Parameters

Katherine Jenny Thompson and Samson A. Adeshiyan
U.S. Bureau of the Census

Abstract: This paper describes a test of two alternative sets of ratio edit and imputation procedures, both using the U.S. Census Bureau's generalized editing/imputation subsystem ("Plain Vanilla") on 1997 Economic Census data. We compare the quality of edited and imputed data — at both the macro and micro levels — from both sets of procedures and discuss how our quantitative methods allowed us to recommend changes to current procedures.

Key words: Blind testing, ratio edit, tabulation.

1. Introduction

The U.S. Census Bureau conducts an Economic Census in years ending in 2 and 7, mailing out over four million census forms to business establishments that provide commercial services to the public and other businesses. For the 1997 Economic Census, the Census Bureau developed and used a generalized editing and imputation subsystem, called Plain Vanilla (PV). The PV edit subsystem consists of three separate (generically written) edit and imputation programs: a ratio edit module; a balance edit module; and a verification module. Program areas incorporate these general programs into their existing data processing systems ("legacy systems") as described in Section 2.

For the services sectors portion of the Economic Census, the use of PV in 1997 was a significant change in editing and imputation methods from those used for their previous censuses. The main difference between the two edit systems was the ratio edit methodology. The PV ratio module tests a complete set of ratio edits simultaneously, determining the minimum number of reported fields that must be changed to satisfy all of the edits (see Section 3). Although the services sectors areas had always used ratio edits, 1997 was the first processing year in which these programs tested all ratio edits simultaneously. Misconceptions about the PV ratio edit module methodology led to some implementation problems in 1997. Consequently, at the conclusion of their 1997 production processing, we conducted a quality audit of each services sector's PV implementation. Based on the audit results, we recommended several modifications to the 1997 PV production procedures.

To evaluate the effect on data quality using these two alternative PV edit implementations[1] on 1997 Economic Census data, we conducted a test on a subset of industries and basic data items. This test used the same data and the same PV software, but differed in the implementation of the software ("the edit parameters"). After processing, we had three competing values for each edited data item in each industry: the final published value of the item in the production database (assumed "correct"); the value obtained using the 1997 production procedures; and the value obtained using the modified procedures.

This paper describes how we compared the quality of edited and imputed data — at both the macro and micro levels — from the 1997 production processing to that from the modified procedures recommended by the quality audit. Section 2 presents general information about the Economic Census data editing procedures, in particular on the Plain Vanilla (PV) subsystem. Section 3 gives detailed information on the PV ratio edit methodology. Section 4 provides background on the evaluation study. Section 5 describes the methods used to compare the macro-data (tabulations) and presents the results of these comparisons. Section 6 presents the micro-data review methodology and associated results. Section 7 discusses these results, and Section 8 provides our conclusions.

2. Economic census data editing: The plain vanilla (PV) subsystem

[1] The 1997 PV production procedures versus the audit-recommended modified PV procedures.

Economic Census data is reviewed in many different ways before publication. The first review is the micro-level review of edit-failing records. Each questionnaire is machine-edited as received (the specific set of edits for each establishment depends on the industry). Because of the volume of questionnaires received and the tight time-constraints, the analysts cannot wait to receive all cases in an industry before beginning micro-editing. Typically, selected editing-failing records are reviewed extensively at this point by subject-matter experts; remaining edit-failing records (usually small establishments) are automatically corrected. After micro-review is completed, analysts begin table cell analysis (macro-review). In many programs, selected individual records undergo a third review, reconciling reported census data to data collected from the same units in current annual surveys. Finally, there is often another stage of micro-review of data used for frame construction.

The Plain Vanilla (PV) subsystem came about as an effort to improve the efficiency of this first stage of data review. The Economic Census is administered by nine different program areas. Edit-rules - conditions for identifying "suspicious/erroneous" data items - differ by program area. However, the form of the edit rules are similar. All Economic Census programs automatically check their questionnaire data with

- Ratio edits, which compare the ratio of two correlated data items to (industry-specific) upper and lower bounds;

- Balance (additivity) edits, which compare reported total items to the sum of their associated detail items;

- Verification edits, which compare reported items to lists of legal variables.

The PV subsystem consists of three separate (generically written) FORTRAN programs, each of which performs one of these edit functions. In addition, both the Ratio and Balance edit modules provide a variety of deterministic imputation options[2]. Individual programs customize this gen-

[2] E.g., direct substitution (e.g., replace total with associated sum of details),ratio

eral purpose software in two ways: by developing program-specific scripts files and by developing industry-specific editing and imputation parameters. Examples of script file functions include defining the balance edits (describing which sets of details sum to an associated total), listing all items that are subjected to ratio edits, and providing the names and locations of edit and imputation parameter files. A separate PV script processor then creates program-specific FORTRAN code by combining the script information with the PV edit modules. The resultant object code is then linked into the program-specific data processing system[3] (Sigman 1997).

The PV Balance and Verification modules were developed "from scratch" for the 1997 Economic Census. The PV Ratio module described below modified pre-existing code that had been used successfully at the Census Bureau by other economic programs since the early 1980s (Greenberg, Draper and Petkunas 1990).

3. PV ratio edit module methodology

A ratio edit compares the ratio of two highly correlated items to upper and lower bounds, called tolerances. Reported items that fall outside of the tolerances are considered edit failures, and one or both of the items in an edit-failing ratio are either imputed or flagged for analyst review. From a subject-matter analyst perspective, ratio edits are useful because it is often difficult to evaluate the "reasonableness" of a data item's value by itself. By comparing an item to other related values in the questionnaire, one can determine if a response appears valid (e.g., Annual Payroll divided by 1st Quarter Payroll should be approximately four). Ratio edits are equally appealing from a mathematical perspective. Greenberg (1986) proves that by augmenting the analyst-provided explicit edits with implied ratio edits, set-covering procedures can be used to determine the minimum number of edit-failing reported data fields that need to be changed to *simultaneously*

imputation, or regression imputation.

[3] The ratio edit and balance edit portion of the Economic Census editing process is often referred to as the "complex edit." Some program-specific processing is required before invoking the PV edit modules (for example, all establishments must be classified into a NAICS industry before subsequent editing). Legacy systems perform this program-specific data preparation.

satisfy the complete set of edits[4]. This approach — developing the complete set of edits, finding the minimal number of fields to impute, and requiring imputed values to satisfy all edits — is known as the Felligi-Holt model of editing (Felligi and Holt 1976).

The Imputation and Tolerance (I&T) Ratio Edit software used to edit the services sector portion of the 1992 Economic Census employed a very different methodology to resolve ratio edit failures. First, ratio tests were performed separately; there were no joint requirements. Second, the system required a minimum of two edit failures per item before marking the item for imputation. Misconceptions about the differences between the I&T and PV ratio edit methodology led to some PV implementation problems for this portion of the 1997 Economic Census, illustrated by the following two examples:

Example 1: Three (Explicit) Ratio Edits, No Additional Implied Edits

Edit 1: $2 \leq$ Annual Payroll/1st Quarter Payroll ≤ 8
Edit 2: $30 \leq$ Annual Payroll/Employment ≤ 60
Edit 3: $7.5 \leq$ 1st Quarter Payroll/Employment ≤ 15

Establishment's reported data: Annual Payroll = 1000, 1st Quarter Payroll = 145, Employment = 30

Edit 3 fails (1st Quarter Payroll/Employment ≈ 4.8)

[4] The complete set of edits is defined as the user-specified edits provided in the script (the explicit edits), plus the other ratio tests implied by the explicit set. [Note: any pair of ratio edits with a common data item implies another ratio edit]. For example, each of the services-sectors collects data on annual payroll (APR), 1st quarter payroll (QPR), and employment (EMP). To guarantee that the imputed value of APR is never smaller than the imputed value of QPR and that the ratio of APR to QPR is never "far from" the industry average of four, the edit-developer specifies that $1 \leq$APR/QPR ≤ 6. Since employment is usually a good predictor of annual payroll, an edit developer defines an explicit test between those two variables with industry-specific tolerance limits ($30 \leq$APR/EMP ≤ 60). These two tests imply a third relationship, namely 1st quarter payroll to employment tested by $30/6=5 \leq$ QPR/EMP ≤ 60. The PV edited/imputed record from this industry must satisfy all three edits.

- With the 1992 I&T edit: No reported data is changed (two edit failures/item required);

- With the 1997 PV ratio edit: either 1st Quarter Payroll or Employment is changed.

Example 2: Four Explicit Ratio Edits, Two Implied Ratio Edits

Explicit (Analyst-Supplied) Edit 1: $2 \leq$ Annual Payroll/1st Quarter Payroll ≤ 8
Explicit (Analyst-Supplied) Edit 2: $30 \leq$ Annual Payroll/Employment ≤ 60
Explicit (Analyst-Supplied) Edit 3: $7.5 \leq$ 1st Quarter Payroll/Employment ≤ 15
Explicit (Analyst-Supplied) Edit 4: $7.5 \leq$ Sales/Annual Payroll ≤ 90
Implied Edit 5: $15 \leq$ Sales/1st Quarter Payroll ≤ 720
Implied Edit 6: $112.5 \leq$ Sales/Employment ≤ 5400

Establishment's reported data: Annual Payroll = 1000, 1st Quarter Payroll = 145, Employment = 30, Sales = 580

- The 1992 I&T edit system would use explicit edits 1 through 4. Of these, edit 3 fails (1st Quarter Payroll/Employment ≈ 4.8) and edit 4 fails (Sales/Annual Payroll ≈ 0.58). Since each ratio-edit failing item is in at most one edit, no data are changed;

- The 1997 PV ratio edit module would use edits 1 through 6. In addition to the two edit-failures mentioned above, edit 6 fails (Sales/Employment ≈ 19.33). Since both Sales and Employment are each involved in *two* edit failures, and Annual Payroll and 1st Quarter Payroll each fail *one* edit, Sales and Employment are replaced by imputed values. The imputation regions for Sales and Employment are obtained as follows:

Imputation region for sales

(from edit 4) $7.5 \leq$ Sales/Annual Payroll ≤ 90
$(7.5)(1000) \leq$ Sales $\leq (90)(1000)$
$7500 \leq$ Sales ≤ 90000

(from edit 5) $4 \leq$ Sales/1st Quarter Payroll ≤ 720
$(15)(145) \leq$ Sales $\leq (720)(145)$
$2175 \leq$ Sales ≤ 104400

Therefore, the imputation region for Sales that will satisfy all edits given the reported values of Annual Payroll and 1st Quarter Payroll is

$$7500 \leq \text{ Sales } \leq 90000$$

Imputation region for employment
(from edit 2) $30 \leq$ Annual Payroll/Employment ≤ 60
$(1/60)(1000) \leq$ Employment $\leq (1/30)(1000)$
$16.67 \leq$ Employment ≤ 33.33

(from edit 3) $7.5 \leq$ 1st Quarter Payroll/Employment ≤ 15
$(1/15)(145) \leq$ Employment $\leq (1/7.5)(145)$
$9.67 \leq$ Employment ≤ 19.33

Therefore, the imputation region for Employment that will satisfy all edits given the reported values of Annual Payroll and 1st Quarter Payroll is

$16.67 \leq$ Employment ≤ 19.33

Note that if PV considered only the *explicit* edits 1 through 4 (like the I & T Edit System above), then all four items would be involved in one edit failure. In this case, PV would attempt to impute 1st Quarter Payroll or Employment (from failed edit 3) and Sales or Annual Payroll (from failed edit 4). If both 1st Quarter Payroll and Annual Payroll are selected for imputation, there is no imputation region for Annual Payroll given the reported values of Sales and Employment: to satisfy edit 2, Annual Payroll must fall between 900 and 1800, and to satisfy edit 4, Annual Payroll must fall between 6.44 and 77.33. Greenberg (1986) proves *why* PV requires the complete set of ratio edits to find a minimum deletion set that will satisfy all of the edits. This example illustrates this requirement.

The PV Ratio module also allows the user to specify reliability weights for each item to influence the probability of deleting/imputing a given data field, with lower weights indicating *higher* reliability [Reliability weights are listed in the PV script next to the ratio edit items]. Failure counts for each edit-failing item are multiplied by its reliability weight, so that minimizing the number of items to be deleted is equivalent to maximizing the weighted failure count (number of edit failures for the item multiplied by the item reliability weight). For example, suppose that the user assigned the following reliability weights to the three items tested in Example 1 above: 1 for Annual Payroll, 1.5 for 1st Quarter Payroll, 2 for Employment. In Example 1, edit 3 fails, and either 1st Quarter Payroll or Employment must be imputed. Without reliability weights, each item is involved in one edit failure, and the PV ratio module would randomly pick one of these two items for imputation. However, the weighted failure count for 1st Quarter Payroll is (1.5)(1) = 1.5 and the weighted failure count for Employment is (2)(1)=2, so Employment would be selected for imputation.

Missing reported data items are automatically imputed by the PV ratio edit module. If only one data item is reported, it is not edited (two non-missing data items are required for a ratio edit). However, a complete record is imputed from the single (unedited) reported item.

4. Evaluation study background

The services sectors portion of the Economic Census is a mail-out/mail back census that comprises five trade areas: Retail Trade; Wholesale Trade; Service Industries; Transportation, Communication, and Utility Industries (Utilities); and Finance, Insurance, and Real Estate (FIRE). Data are collected on approximately one hundred fifty different industry-specific questionnaires. Some trade areas further classify the establishments within industry by legal form of organization, type of operation, and tax status. We used these editing-processing classifications for our evaluation, but refer to each classification as an industry.

Trade area subject-matter-experts provided the industries used for this test. These industries were selected because they were particularly problematic in 1997 and were not meant to be representative of the trade area as a whole. A "side effect" of this criterion was that some of these industries could be very intractable in terms of edit and imputation parameter development. We had a small number of industries per trade area: four in Retail; 14 in Wholesale; seven in Services; four in Utilities; and four in FIRE. We performed our evaluation by industry within trade area. Our test data consisted of active full-year reporter records. The data items used in this study varied slightly by trade area. Besides Annual Payroll (APR), 1st Quarter Payroll (QPR), and Number of Employees (EMP), all trade areas collect Sales/Receipts (SLS). In addition, Wholesale collects Operating Expenses, Purchases, Beginning Inventories, and Ending Inventories, and Services collects Operating Expenses in tax-exempt industries.

There are two key differences between the sets of ratio edits employed by the new and old scripts. First, the old scripts contained more ratio-tested variables, including trailer data and administrative data tests. Second, the old scripts provided tolerances for the complete sets of ratio edits (explicit and implicit), resulting in very tight edit-acceptance spaces. The new scripts dropped all trailer data tests (often, these data items were poorly correlated with the basic data items, causing edit failures/imputations based on tenuous relationships) and specified a very limited set of explicit ratio edits (APR/QPR, APR/EMP, and SLS/APR in all trade areas, with a few additional tests in one Services industry and in Wholesale Trade). Although we recommended including administrative data ratio tests in the new scripts, they were not included in our test scripts because of operational concerns. Different item reliability weights were used (for the same items) in each script. Finally, there was some discrepancy in edit and imputation parameter quality. The old-script parameters had undergone several revisions; while our parameters were reviewed once. Consequently, we viewed equally good results from the old and new scripts as evidence of improved methodology in the new scripts.

Thompson *et al.* (2001) provides details on the new script development process. There are two key differences between the sets of ratio edits employed by the new and old scripts. First, the old scripts contained more ratio-tested variables, including trailer data and administrative data tests. Second, the old scripts provided tolerances for the complete sets of ratio edits (explicit and implicit), resulting in very tight edit-acceptance spaces. The new scripts dropped all trailer data tests (often, these data items were poorly correlated with the basic data items, causing edit failures/imputations based on tenuous relationships) and specified a very limited set of explicit ratio edits (APR/QPR, APR/EMP, and SLS/APR in all trade areas, with a few additional tests in one Services industry and in Wholesale Trade). Although we recommended including administrative data ratio tests in the new scripts, they were not included in our test scripts because of operational concerns. Different item reliability weights were used (for the same items) in each script. Finally, there was some discrepancy in edit and imputation parameter quality. The old-script parameters had undergone several revisions; while our parameters were reviewed once. Consequently, we viewed equally good results from the old and new scripts as evidence of improved methodology in the new scripts.

5. Macro-level evaluation (Tabulation comparisons)

Our first set of analyses compares data item tabulations from the old and new edit results to the tabulations based on final 1997 publication data (our “gold standard”). Table 1 presents the format of our data item tabulations. Each tabulation is cross-classified within industry by establishment size category (small and large).

Using the ratios of old/final and new/final displayed in Table 1, we first examined which script yielded better results for each data item. In each industry, we compared the two alternative tabulations for each item (columns 5 and 6 of Table 1) to the final data tabulation (column 4 in Table 1) and selected the “better” tabulation as the one with the ratio (in columns 7 and 8) closer to 1. When both ratios were within five-percent of the final value, then the two scripts tied. Table 2 summarizes our data item comparisons for total establishments summed over industries within a trade area. The small and large establishment tabulations showed similar patterns.

Table 1: Table Shell for Comparison of Original and New Script Edited Tabulated Data with Final 1997 Tabulated Data

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Small					
		Large					
		Total					

(1)= Industry code, (2)= Data item, (3)= Establishment size, (4)= 1997 published data total (final), (5)= Original PV-script edited data total (old), (6)= New PV-script edited data total (new), (7)= Ratio of old/final, (8)= Ratio of new/final

Except for Retail, the new edit script tabulations are generally closer to the final published values than the corresponding old scripts tabulations for all data items, often by a wide margin. The two inventory items in Wholesale Trade are notable exceptions: both scripts performed equally poorly. In the six industries that collected inventories data, the average ratio of new script to final beginning inventories and new script to final ending inventories were 13.3 (min = 1.8, max = 35.8) and 15.4 (min = 1.9, max = 54.9) respectively, and the average ratio of old script to final beginning inventories and old script to final ending inventories was 12.03 (min = 2.03, max = 35.9) and 13.4 (min = 1.9, max = 54.9) respectively. Inventories items are difficult to ratio edit; although they have high pairwise correlation, they are poorly correlated with the other basic data items.

Next, we compared total results by industry within trade area. In each industry, we summed up our ratio classifications (new better/old better/tied) for each data item to get an industry-level determination. Table 3 summarizes the industry level comparisons for total establishments.

Again except for Retail, the new scripts generally produce results closer to the final tabulated results than the old scripts. Unfortunately, in the two Retail industries where the old edit was better, the new script-imputed tabulations were between two and three times as large as the final tabulations. Initially, we believed that the different scripts' results were due to

Table 2: Data Item Comparisons (Tie Indicates That Both Tabulations Are Within 5% of Final Value)

Trade Area	Data Item	New Better	Old Better	Tie
FIRE	Sales	3	1	0
	Annual Payroll	3	1	0
	1st Quarter Payroll	3	1	0
	Employment	4	0	0
Retail	Sales	2	2	0
	Annual Payroll	2	2	0
	1st Quarter Payroll	2	2	0
	Employment	2	2	0
Services	Sales	2	2	3
	Annual Payroll	2	1	4
	1st Quarter Payroll	3	1	3
	Employment	3	1	3
	Operating Expenses	0	1	0
Utilities	Sales	2	1	1
	Annual Payroll	1	1	2
	1st Quarter Payroll	1	1	2
	Employment	2	0	2
Wholesale	Sales	9	5	0
	Annual Payroll	8	4	2
	1st Quarter Payroll	7	2	5
	Employment	9	1	4
	Operating Expenses	6	5	3
	Purchases	6	0	0
	Beginning Inventories	2	4	0
	Ending Inventories	3	3	0

Table 3: Number of Industries Where New Script-imputed Data Items Are as Good or Better Than Old Script- imputed Data Items

Trade Area	New Script Better	Old Script Better	Tied
FIRE	3	1	0
Retail	2	2	0
Services	3	2	2
Utilities	2	1	1
Wholesale	8	4	2

overly-wide tolerances, so we tightened the limits on the ratio edits in those industries and re-edited. The second set of new script-imputed tabulations was not noticeably different from the previous set. Clearly, this was not a parameter problem.

To characterize the cases that were poorly edited with the new scripts, we examined the records with the largest new-script-imputed values for each data item. In both industries, the difference between the old-script and new-script imputed tabulations were caused by a few establishments having the following reporting problems:

- All dollar value items were reported in the wrong units (reported in units instead of thousands); or

- Only one basic data item was reported, and it was obviously incorrect (unreasonably large or small).

In general, ratio edits are poor at correcting the first type of error ("rounding"). The new scripts attempted to identify rounding errors by placing a relatively low reliability weight on employment, the only non-dollar basic data item. This strategy failed when number of employees was missing or was not reported correctly and also had the undesirable side effect of almost never changing the reported value of employment. In the two problem Retail industries, only dollar value items were reported in the rounding-error-establishments, and the new scripts imputed values of employment consistent with the *unrounded* (unit) values. The second case — only one item reported — cannot be ratio edited. Instead, the module imputes a complete record from the one reported item. In these industries, the reported item was annual payroll or sales, reported in units instead of thousands.

Doubting that these two reporting problems were unique to Retail, we performed a macro-comparison in all trade areas. Based on the reported values of the four common basic data items (APR, QPR, SLS, EMP), we re-classified each establishment into the following seven categories:

- Full-impute (delinquents)
- One item reported, bad reported value, administrative data available
- One item reported, bad reported value, no administrative data
- One item reported, good reported value, administrative data available
- One item reported, good reported value, no administrative data
- Two or more items reported, no rounding errors
- Two or more items reported, at least one rounding error

The "one item reported" category was subdivided into good/bad reported value categories to examine the effect of the new imputation parameters: the script with better imputation parameters should yield consistently closer tabulations to the final tabulated results. We further subdivided the "one item reported" category by administrative data availability. The old scripts included administrative data tests, so cases with administrative data were not truly "one item reported" cases in both scripts. After reclassifying the establishments, we produced the same tabulations and ratios as in Table 1 with the Establishment Size column replaced by the "record characteristic" category and performed similar analyses to those presented in Tables 2 and 3.

Both scripts performed equally well for full-impute establishments. As expected, the old scripts imputed better records when only one (good or bad) data item was reported as long as at least one administrative data item was available. The old scripts also did a better job of correcting rounding errors than the new scripts (probably due to the old scripts' inclusion of tests to administrative data). Otherwise, the new scripts generally imputed better records. The two situations where the old script was consistently preferable accounted for a very small percentage of the tested establishments (less than two percent of the establishments in any trade area).

6. Micro-level evaluation (Blind testing)

6.1. Description of test

Comparing tabulations is a fairly objective way to determine if there is a systematic difference between the two sets of edits in terms of effect on the tabulations. However, large establishments are very influential in this type of comparison. Although all Economic Census forms are machine-edited, analyst review of edit-failing cases is generally restricted to large establishments because of time-constraints. Subsequent stages of data review usually focus on large establishments because they most impact the tabulations. So, a limitation of the macro-level evaluation is that while it does well on the whole and for large establishments in particular, it does not necessarily do well for detecting edit problems with small establishments.

Given these limitations, we did not want to use the published microdata for our micro-level comparison. Moreover, the subject-matter-expert analysts wanted to review the microdata from both edit scripts. So, we conducted blind testing in all of our industries. The original motivation for the blind testing was to (hopefully) make the analysts comfortable with the revised procedures on a case-by-case basis. We also planned to use blind test results to either confirm the macro-level results presented in Section 4 or to uncover systematic edit/imputation problems for certain classes of establishments (e.g., small establishments).

For the blind test, analysts from each trade area were provided with the basic information displayed in Figure 1 for 200 randomly selected cases

(100 cases per size category) per trade area and were asked to select which — if either — edit outcome (edit A or edit B) was acceptable. The label for edit A and edit B was randomly assigned, so that neither the analysts nor the evaluators knew which script was used to obtain either outcome. To avoid potentially biasing the outcome, analysts were not to be able to identify a particular establishment. They also were not given any edit flags. In effect their data review tools were more limited than they would be in a production system. Also, the tabulated results described in Section 4 were not provided until blind testing was completed. Analysts were asked to review at least 50 of the 100 cases per size category.

We were interested in gaining insight to the following questions:

1. At the micro-level, did the analysts have a preference for one script outcome over another?

2. Were the results from the micro-evaluation (analyst preference) consistent with those from the macro evaluation? If not, can we explain the reasons for the differences and what can we do to correct these differences in the future?

We addressed the first question using the standard categorical analyses presented in Section 5.2 and examined the second set of questions by conducting an exploratory review of the records in which the analysts clearly preferred the old script results.

Our ability to analyze the blind test results was greatly handicapped by the sample design. Ideally, we would have selected a sample that was stratified within industry and establishment size cell by "magnitude of edit difference." Also, our population would have been the set of all establishments whose edit results (new script/old script) contained at least one basic data item changed by either script. Because of processing timing concerns, we were not involved in the sampling plans. Instead, the analysts were provided with a stratified random sample of establishments selected within trade area (not industry) and size category (large, small) that had at least one imputed data item from either script that differed from the published value. Consequently, the blind test data contained a high percentage of

Case:

ID: [Multi or single unit and size class]
Bridge code:
Kind of business:
Tax status:
Type of operation:
Legal form of operation:

Data:

Variable	Reported Value	Admin Value	92 Census Value	92 Census Flag	Edit A Value	Edit B Value
Sales						
Annual Payroll						
1-st Quarte Payroll						
Employment						
Operating Expenses						
Beginning Inventories						
Ending Inventories						
Purchases						

Figure 1: Format for Analyst Evaluation of Blind Test Cases

Table 4: Tabulation of blind testing data new PV edit

Old PV \ New PV	Acceptable	Not Acceptable	Total
Acceptable	N_{11}	N_{12}	N_{1+}
Not Acceptable	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

N_{11} = both acceptable, N_{12} = only old acceptable, N_{21} = only new acceptable, N_{22} = neither acceptable.

cases with nearly identical edit outcomes.

6.2. Statistical analysis tools: Tests for association and analyst preference

Test 1. Test for Association

We first tested whether the analysts have a preference between the two edit scripts with

H_0 : Choice of the new or old edit is independent of each other (i.e., no preference between edits).

H_1 : Choice of the new or old edit is dependent on each other (i.e., one type of edit is preferred over the other)

using the standard Pearson chi-squared test (Agresti 1990) with the count data shown in Table 4.

Rejecting the hypothesis of independence allows us to conclude that the analysts tend to prefer one edit over the other, but it does not tell us *which* edit is preferred. To determine the preferred edit, we focused on the highlighted cells in Table 4, where the analyst made a clear choice: “Only New Acceptable” (N_{12}) or “Only Old Acceptable” (N_{21}). Formally, we tested

Test 2. Test for Analyst Preference

$H_0:$ $p_{21} \leq p_{12}$ (or $p_{21} - p_{12} \leq 0$, i.e., the proportion of "only new acceptable" is less than or equal to the proportion of "only old acceptable.")

$H_1:$ $p_{21} > p_{12}$ (or $p_{21} - p_{12} > 0$, i.e., the proportion of "only new acceptable" is greater than the proportion of "only old acceptable.")

where proportions $p_{12} = N_{12}/N_{++}$ and $p_{21} = N_{21}/N_{++}$. The difference in proportions is estimated by $(p_{21} - p_{12})$, and the standard error (se) is estimated by

$$\text{se}(p_{21} - p_{12}) = \sqrt{\frac{p_{21}(1 - p_{21})}{N_{++}} + \frac{p_{12}(1 - p_{12})}{N_{++}} + \frac{2p_{21}p_{12}}{N_{++}}} \tag{6.1}$$

yielding the test statistic distributed as $t_{\alpha,N_{++}-1}$. Rejecting H_0 provides evidences that the analysts prefer the new edit script over the old edit script. Since this is a one-tailed test, any t-statistic with negative value implies that we cannot reject H_0.

6.3. Results

Table 5 presents the counts of the blind test data by trade area. As expected, "Both Acceptable" was the most common choice. Wholesale had a high percentage of "either Acceptable" cases, likely attributable to the poor inventory edit results (see Section 4).

Table 6 provides the results of our tests for association (Test 1) and analyst preference (Test 2) using the count data shown Table 5. Except for Services, Table 6 provides evidence of association in analysts' old and new edit choices at the 5% significance level. For FIRE and Utilities, the old edit is preferred to the new edit; for Wholesale and for Retail, the new edit is preferred to the old edit; and we were unable to make a conclusion about direction of preference for Services.

These results are very interesting because they appear to conflict with the macro-level results discussed in Section 4. However, before interpreting

Table 5: Counts from Analyst Review by Trade Area

(a) \ (b)	FIRE	Retail	Services	Utilities	Wholesale
Both Acceptable (N_{11})	137	294	509	176	765
Only New Acceptable (N_{21})	73	115	183	73	263
Only Old Acceptable (N_{12})	149	55	276	92	116
Neither Acceptable (N_{22})	45	41	82	13	744

(a)= analysts' choice, (b)= trade area.

Table 6: Results of Tests for Association and Analyst Preference

	Test 1	Test 2	
Trade area	Reject H_0	t-statistic	Reject H_0
FIRE	yes	−4.75	no
Retail	yes	4.00	yes
Services	no	−4.50	—
Utilities	yes	−1.25	no
Wholesale	yes	7.00	yes

these results, we wanted to confirm that these test results were truly indicative of the analysts' preferences and not merely a function of a poor sample. For example, Table 6 provides evidence that the analysts preferred the new Retail script. In two Retail industries, the new script was clearly better than the old script at the macro-level (Table 3). If the majority of Retail "only New Acceptable" (N_{21}) cases were sampled from those two industries, then the blind test results would actually be consistent with the macro-level results. Further examination of the industry distribution of sample data was required.

Table 7 post-stratifies the blind test cases where analysts clearly preferred one script over the other (i.e., the N_{12} and N_{21} cases) by the Table 3 industry-level classifications. Similar distributions of N_{12} and N_{21} counts in Table 5 and Table 7 distributions would provide evidence that the analysts could predict the macro-level results from their micro-review (i.e., that the

Table 7: Distribution of N_{12} and N_{21} Cases by Macro-evaluation Industry Classification

Trade Area	New Script Better from Macro Evaluation	Old Script Better from Macro Evaluation	Both Old and New Scripts Tied
FIRE	66.22%	0.00%	33.78%
Retail	39.41%	61.59%	0.00%
Services	39.00%	30.50%	30.50%
Utilities	23.64%	17.58%	58.79%
Wholesale	58.84%	20.58%	20.58%

consistently preferred script would have the better tabulated results). Dissimilar results are more difficult to interpret. For example, if the difference in both sets of macro-level results quality was caused by a few establishments (none of which were included in the blind test), then the sampled cases in an industry characterized as "Old Script Better" in Section 4 could actually have the same or even more consistent results with the new script.

For Retail, only 39.41% of the cases where analysts clearly preferred one script (N_{21} and N_{12}) were selected from industries where the new script did better overall. However, the analysts preferred the new script over the old script in approximately 68% of the N_{12} and N_{21} cases (115/(115+55)). In the other trade areas, a high percentage of these N_{21} and N_{12} cases were sampled from industries where both the new and old scripts tied. In these cases, it is tricky to draw parallels between analyst preference and macro-level results: for example, analysts might have preferred the edit that preserved more reported data or might have a preference for changing the value of one data item over another. With the Utilities data, the majority of N_{21} and N_{12} cases are from industries where both scripts tied, making it impossible to draw any parallels between macro- and micro-level results. For both FIRE and Wholesale, most of the cases came from industries where the new scripts did better. The majority of FIRE's test cases (66.22%) were selected from industries where the new scripts did better, but the analysts tended to prefer the old script. The majority of Wholesale's test cases

(58.84%) were selected from industries where the new scripts did better, and the analysts also tended to prefer the new script.

We found the apparent contradiction between macro-level and the micro-level results perplexing. Tabulations from section 4 showed marked improvements in edit outcome with the new scripts, but the blind test results showed that analysts tended to prefer the old script results at the micro-level. To understand this preference, we conducted an exploratory review of the records where the analysts clearly preferred the old script results (N_{12} cases). This led to two major findings (both confirmed by the analysts). First, the analysts usually preferred the script that changed fewer reported values or used administrative data for imputation, even when final edited data contained unusual ratios (e.g., an Annual Payroll to 1st Quarter payroll ratio of 12, far from the industry average of four). Second, analysts did not always provide complete requirements for tolerance limits (ratio edit bounds). For example, by design, our new script's tolerances guaranteed that Sales had to be greater than or equal to Annual Payroll. After reviewing the blind test results, we learned that the FIRE analysts, in addition, preferred that Sales should not exceed five times the Annual Payroll (this requirement was explicitly accounted for in the old script's parameters).

7. Discussion

When we originally planned this evaluation study, our goal was to prove that the PV ratio module — if properly implemented — could achieve excellent edit results for the services-sectors portion of the census with little or no human intervention. For the most part, we succeeded. We are not finished, however. This evaluation revealed two major systematic problems — common to all trade areas — with the new edit scripts: poor detection of rounding errors and an increased probability (compared to the old script) of imputing a complete record from a single unedited data item. We will reduce occurrences of these edit problems in the 2002 Economic Census by including ratio tests to administrative data in our edit scripts. We will also safeguard against these situations by data-filling blank items with other reported sums of details (and/or administrative data) and correcting rounding errors prior to ratio-editing.

We were initially disappointed by apparent contradictions between the macro-level and blind test results. We knew that the new procedures improved overall data quality, but the analysts reviewing the blind test data concluded differently. The deficiencies in the sample itself made it difficult for us to understand the implications of the blind test results. Even so, the blind testing was a useful analysis tool. First, it revealed a "disconnect" between analyst preference and industry-level ratio requirements (e.g. preferring edit outcomes that retained more reported or administrative data, even if results contradicted industry level tolerances). And, by reviewing the cases where the analysts accepted only the old script results (N_{12}) cases, we found some systematic problems with edit parameters.

We view the blind test as a "dress rehearsal" for the production micro-review. The analysts are clearly knowledgeable about their subject-area. They are not, however, as knowledgeable about ratio edit implementation. We must address this by developing training that conveys the connection between ratio edit tolerances and edit outcomes so that the analysts can build all program requirements into their 2002 edit scripts.

The macro-level analyses described in Section 4 were quite effective at both evaluating the edit results overall and indicating systematic edit implementation problems (especially when combined with a micro-review of "problem" records). We do not really expect the analysts' micro-review to be as revealing. Micro-review checks individual records for internal consistency, not for conformance with the industry norms. Usually outliers cannot be detected in the absence of the full distribution. Recognizing this, the 2002 PV edit implementation should include ongoing summary audits of ratio edit failures within industry to reveal problem ratio edits or tolerances, rather than relying on the analysts ability to recognize and articulate such problems.

8. Conclusion

This paper describes a test of two alternative sets of ratio edit and imputation procedures on data from the service-sectors portion of the 1997 Economic Census: the production processing procedures and a modified set of procedures resulting from a quality audit. We showed that the modified procedures resulted in improved edited data quality. Moreover, the evaluation process revealed further necessary enhancements to the modified procedures, which will be implemented in the 2002 census.

A less measurable - but equally important - deliverable from this evaluation study was the dialog between edit-implementor (methodologists) and subject-matter experts. Discussing both the macro-level results and the blind test results began a long-overdue edit-implementation training process. As a consequence of this study, we are establishing workgroups that consist of methodologists, production specialists, and subject-matter-experts to develop edit parameters and scripts for the 2002 census. Also with the work groups in place, we can develop training courses for the analysts that address the deficiencies revealed by this evaluation.

Acknowledgements

The authors would like to thank Scot Dahl, Ruth Detlefsen, Nash Monsour, Kenneth Sausman, Richard Sigman, and the editor for their valuable contributions to earlier versions of manuscript. This paper presents the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties and to encourage discussion of work in progress.

References

Agresti, A.(1990). *Categorical Data Analysis,* New York: Wiley.

Felligi, I. P. and Holt, D. (1976). Systematic Approach to Automatic Editing and Imputation. *Journal of the American Statistical Association,* **71**, 17-35.

Greenberg, B. (1986). The Use of Implied Edits and Set Covering in Automated Data Editing. Unpublished report, Washington, DC: U.S. Bureau of the Census.

Greenberg, B., Draper, L., and Petkunas, T. (1990). On-Line Capabilities of SPEER. *Proceedings of the Statistics Canada Symposium*, Statistics Canada, 235-243.

Sigman, R. S. (1997). Development of a "Plain Vanilla" System for Editing Economic Census Data. *Proceedings of the Conference of European Statisticians, Section on Statistical Data Editing.* United Nations Statistical Commission and Economic Commission for Europe, Working Paper 24.

Thompson, K.J., Sausman, K., Walkup, M., Dahl, S., King, C., Adeshiyan, S. (2001). Developing Ratio Edits And Imputation Parameters For the Services Sector Censuses (SSSD) Plain Vanilla Ratio Edit Module Test, unpublished report, Washington, DC: U.S. Bureau of the Census.

Received September 8, 2001; accepted April 19, 2002

Katherine Jenny Thompson
Room 3108-FOB 4
U.S. Bureau of the Census
Washingon, D.C. 20233
U.S.A.
katherine.j.thompson@census.gov

Samson A. Adeshiyan
Room 2754-FOB 3
U.S. Bureau of the Census
Washingon, D.C. 20233
U.S.A.
samson.a.adeshiyan@census.gov