BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-85-12

LARGE OBSERVATION STUDY FOR THE 1982 CENSUS OF CONSTRUCTION INDUSTRIES

by

Lawrence R. Ernst
Statistical Research Division
Bureau of the Census
Room 3524. F.O.B. #3
Washington, D.C. 20233 U.S.A.

Recommended by: Paul P. Biemer

Report completed: January 8, 1986

Revised: May 12, 1986

1. EXECUTIVE SUMMARY

This study was carried out for the 1982 Census of Construction Industries. That census did not elicit responses from all establishments in the universe, but instead, from a sample selected with probability proportional to 1981 administrative payroll. Consequently, it was subject to sampling variability. Noncertainty establishments which grew considerably from 1981 to 1982, the so-called "large growth" cases, would have had an undesirably large impact on this variability. A procedure was used in the 1982 census to reduce the effect of the large growth cases on the sampling variance. The original purpose of this study was to evaluate that procedure and to suggest any alternative procedures that might be preferable.

The procedure used for adjusting for large growth sample cases in the 1982 Construction Census was as follows. Each selected noncertainty unit with sufficiently large 1982 administrative data for either payroll, employment or receipts was classified as an "administrative large growth case"; if the 1982 reported census data were sufficiently large for any of these three items, the case was classified as a "response large growth case." Many cases were in both categories. The procedure changed the weight of each large growth case to one. In addition, all nonsample cases with sufficiently large 1982 administrative data were processed and tabulated as if they were nonrespondent sample cases with a weight of one, and census data for them were imputed. The purpose of the inclusion of these nonsample cases in the tabulations was to compensate for that

part of the downward bias resulting from the weight changes in the administrative large growth cases. No compensation was made for the downward bias for the cases that were only large growth in response data. However, it was intended that this bias would be more than offset by a reduction in variance that would result in a lower mean square error.

One critical assumption that underlies this and similar procedures for compensating for large growth is that the data used in the estimates are error-free. A preliminary examination of the data for the large growth cases revealed a considerable number of inconsistencies, particularly among the large growth response cases. It was decided to carefully review and make necessary corrections to all the response large growth data in order that the error-free assumption be satisfied as closely as possible. This review included a comparison of the data used in the tabulations to the microfilm of the original schedules. It became apparent from this review that key reasons for the data problems were the failure of the edit and imputation system used in the 1982 Construction Census to detect certain errors, and the changes to apparently correct data that this system made. It was decided to detail, as a major part of this study, those particular aspects of the edit which appeared to be less than ideal and which impacted on the key data items in the large growth response cases. However, no statistical inferences could be made from this study concerning the magnitude of the effect that any aspect of the edit had on census estimates, since the large growth response cases were, most decidedly, not a

probability sample of either the entire universe or all census sample cases.

The following were some findings of this study of the edit for the large growth response cases:

1. Errors in Total Payroll (PR) generally were not corrected by the edit because, with few exceptions, it was always assumed that PR was correct. Furthermore, uncorrected errors in PR tended to result in changes to other apparently correct data fields that were edited either directly or indirectly against PR.

2. Average Total Employment (ATE) was edited against PR by means of the ratio test PR/ATE. Because of unrealistically low, lower limits for this ratio, many errors in ATE were not detected. This problem of inappropriate limits on ratio tests occurred also for other pairs of fields. Other errors in the tabulated values of ATE occurred because of apparently incorrect editing decisions when the employment items were not consistent. Still other errors in ATE resulted from apparently correct values being changed because of uncorrected errors in PR.

3. Some errors in Total Receipt (TR) were undetected. When PR was corrected by an analyst and when the case was subsequently reedited, the portion of the edit involving TR was skipped over, even though the key edit of TR was against PR. This problem was not unique to TR. Errors in TR also occurred because TR values were changed as a result of uncorrected errors in PR.

4. Editing problems such as those just described could have been alleviated by a different type of edit, that examined

several fields simultaneously instead of generally considering only a pair of fields at a time.

5. The documentation of the edit procedures, which was excellent, greatly facilitated this study.

Because of delays in this project caused by the need to review the large growth response cases, and because of other problems, such as errors in the listing of large growth cases used in this study, which is discussed in Section 4, the original project goals were not achieved. However, the following were some of the other observations made on the large growth procedure itself and related issues.

6. The processing system did not always take the intended action in terms of which cases should be subject to the large growth procedure. Any case with census data above a large growth cutoff at the end of any pass through the edit was treated as a large growth case even if the final data used in the tabulations were all below the cutoffs, while many true administrative large growth cases did not have their weights changed to one.

7. The structure of the employment questions on the schedules, particularly those that were arithmetical combinations of other items, appeared to contribute to the response errors on those items for the large growth response cases. This is detailed in Section 2.3.

8. Weighted data, instead of unweighted data, should be used in the determination of which cases should be subject to a large growth procedure. This is explained in Section 2.4.

2. DETAILED FINDINGS OF THE STUDY

2.1 Introduction

Section 2.2 discusses the impact of the edit system used in the 1982 Construction Census on the data errors in the large growth response cases. Section 2.3 considers the relationship between the structure of the employment questions on the schedules and the response errors on these questions. Section 2.4 contains several observations about the large growth procedure itself.

2.2 The Edit System and the Large Growth Response Cases

In this subsection the effect of the edit on the large growth response cases will be detailed. The emphasis of this study will be on the items PR, ATE and TR since, in addition to their general importance in the census, these were the items that determined large growth response status. Other data items, such as detail and administrative data for payroll, employment and receipts, and Total Numbers of Construction Hours Worked (THOURS), that either were or, perhaps, should have been an integral part of the edit of PR, ATE and TR will also be discussed.

There is one point that must be strongly emphasized concerning the findings to be presented on the effects of the edit on the large growth response cases. These cases do not, in any way, represent a probability sample of either the construction universe or all census sample cases. (The sampling frame for the census consisted of approximately 485,000 establishments, of which approximately 180,000 were selected.) In fact, a highly

disproportionate number of errors would be expected among these cases, since it was these errors that caused some of these cases to be declared large growth cases. Consequently, no statistical inferences could be drawn from this study concerning the number or magnitude of uncorrected errors in the construction census. However, the problems with the edit that were discovered for the large growth response cases should also apply to other edited cases for which there were similar errors.

This study covered all 570 of the large growth response cases on a listing obtained from Construction Statistics Division (CSD). (It was later discovered that additional cases should have been included, while the tabulated values for 33 cases that were included did not meet the large growth response criteria. The reasons for these errors in the listing are discussed in Section 4.) Of these 570 cases, 187 were also large growth in administrative data. All 570 cases were carefully reviewed and corrections were made where applicable. The review included a comparison of the final edited data for each case to a microfilm of the original report.

After changes to the data for PR, ATE and TR resulting from this review, 109 cases no longer met the large growth response criteria. (This does not include the 33 cases that did not meet the criteria even before these changes.) Interestingly, all of these cases were among the 383 cases that were not large growth in administrative data. This is not particularly surprising; if a case was large growth only in response data, there was data inconsistency. There were also errors detected in cases other

than these 109 that did not effect the large growth status, but these other cases will not be discussed in this report. Note that focusing only on cases with errors in the tabulated data further precludes a balanced evaluation of the edit system. Such an evaluation was not a purpose of this study, only a documentation of the role of the edit in the errors detected in the review of the large growth response cases.

These 109 cases were divided into two broad categories. For 23 of these cases, it was the judgment of this author that their incorrect large growth statuses might not have been detected by any computerized edit. For example, in some of the cases the errors resulted from keying errors of apparently correct data that were not large enough to make the data inputted to the edit appear inconsistent. Such errors could only have been detected by comparison with the data on the schedules. For other cases among the 23, there were so many errors in the reported data that it would have been difficult for any computerized edit to have made proper corrections.

For the other 86 cases among these 109 cases, it was the judgment of this author that they would not have been classified as large growth response cases if a sufficiently improved edit had been used. There may be disagreement on this point for some of these cases, since without a specific alternative to test, this judgement is, by necessity, subjective. However, an attempt was made to be conservative in classifying a case in the latter group.

Table 1 indicates for each of these two groups, and for each of the six possible combinations of error for the fields PR, ATE and TR, the number of cases in the tabulated data with that combination of errors.

The remainder of this section deals solely with the 86 cases that were classified as having an incorrect large growth response classification that could have been corrected by an adequate computerized edit. Before studying these cases in detail, the actions taken by the edit are summarized in Table 2. In this table, for each field, each of these cases is classified according to its status on input to the edit and its final tabulated data. For these 86 cases, the edit system corrected only two errors among the three fields and failed to detect or incorrectly changed 89 input errors. Among these 89 errors, only 3 for TR and 1 for ATE were changed at all by the edit. Furthermore, the edit system actually created 35 errors in apparently correct data. And, for the 5 cases where one of these data items was not reported, the edit imputed a clearly incorrect value for 4 of them. Of the 35 errors created by the edit, 34 resulted from values of ATE and TR failing an edit test because of an incorrect value of PR. The remaining error created by the edit resulted from ATE being changed because of a keying error in one of the employment detail items. Table 2 includes the final result of changes made by both the computerized edit and analyst review. The two cases where errors in PR were corrected could only have been detected as a result of an analyst's review. Note, however, that the only data sets used in this review were

the completed report form and the data used in the tabulations. Consequently, any keying errors that were corrected by an analyst's review would be unknown to this study and not included as corrections in the table.

A key aspect of the edit process was the use of ratios formed from pairs of data items. If a ratio was not within acceptable limits, then one of these items would commonly be imputed from the other by multiplication by an average ratio. In the portion of the edit that is of interest in this study, the first such ratio computed was PR/ATE. However, prior to this, the payroll and employment items were edited separately for the purpose of ensuring that items that were defined to be arithmetical combinations of other items actually satisfied the required relationships. For the employment items, this was where the first editing problems arose, affecting 16 cases. For 12 of these cases, this occurred as follows. On the report form the respondent was instructed to compute ATE by summing Average Number of Construction Workers (ACW) and Average Number of Other Employees (AOE). (For a complete description of the employment items see Section 2.3.) For each case where the values of ATE, ACW and AOE inputted to the edit did not satisfy ATE = ACW + AOE, the edit forced equality. For some conditions this was accomplished by changing ATE to equal the sum of ACW and AOE. However, when ATE was greater than the sum and either ACW or AOE (but not both) was zero, the edit left ATE unchanged and, instead, changed the zero entry to obtain additivity. This situation occurred in these 12 cases and it was always AOE that

was initially zero. With the possible exception of one case, it appeared that the zero value for AOE was correct and the action that should have been taken was to change ATE to equal ACW. This editing problem could have been avoided if the following two observations had been used in making the editing decision. First, since ATE, unlike PR for example, was not a figure that respondents would typically have had in their files, a value of ATE other than ACW + AOE should always have been suspect. Secondly, the zero entry in AOE could have been checked against Other Employees Salaries (OES), with a zero value for OES tending to confirm the AOE value. In fact, such a check was made later in the edit process and for 11 of these 12 cases AOE was changed back to zero because OES was indeed zero. However, by this stage in the edit process ATE was not permitted to be changed, and instead ACW was set equal to ATE. The net result for each of these 11 cases was that AOE was correctly tabulated as zero, but both ATE and ACW were tabulated as the incorrectly imputed value of ATE rather than the apparently correct value of ACW.

For these 12 cases, the most common reason for error in the data inputted to the edit, which occurred in 9 cases, was that the respondent entered the Total Number of Construction Workers (TCW) (i.e., the sum of the 4 quarters of construction employment) in both its correct location and where ATE should have been entered. This particular problem appeared to result partly from the design of the employment questions, which is discussed in Section 2.3. As a result of the edit procedures, when this error

occurred the tabulated values for both ACW and ATE were generally 4 times too large since ACW was defined as TCW/4.

In this edit of the employment detail a closely related problem occurred for three of the four other cases where the employment items inputted to the edit were not in balance. The input in these three cases satisfied the relationships ATE = ACW > TCW/4, that resulted from an error in ACW which, as would have been expected, carried over to ATE. The edit first correctly changed ACW to equal TCW/4, since this was how the respondent had been instructed to compute ACW. However, the change in ACW was not carried over to ATE as it should have been; consequently, the same conditions were created, namely ATE > ACW, AOE = OES = 0, that were just considered, and the edit, of course, treated the data in the same manner. The tabulated values for both ATE and ACW became their input values, which were in error. The edit was then forced to change the apparently correct value of TCW to preserve the relationship ACW = TCW/4. Thus, in cases such as these three, an obvious arithmetical error would not be corrected by the edit.

In the one remaining case of unbalanced employment items, the value of TCW inputted to the edit was greater than the sum of the 4 quarters of construction employment due to a keying error in TCW. Instead of changing TCW to equal this sum, the 4 quarters of construction employment, ACW and ATE were all changed to be consistent with TCW. ATE was changed from 3 to 53 as a result of this action.

After the preliminary separate edits of the payroll and employment items were completed, the ratio PR/ATE was computed. Two editing problems arose. First, if the ratio was not within the limits for this ratio, PR was assumed correct, with two exceptions, and ATE was imputed from PR. (One exception was that if PR/(1000xATE) was within the limits, it was assumed that PR had been reported in dollars instead of thousands of dollars. The other exception was that if detail had been provided for employment but not payroll then ATE was assumed correct and PR was imputed from ATE.) However, the assumption that PR was correct proved false in 27 cases. Furthermore, 23 of the 24 cases in Table 2 for which apparently correct values of ATE were changed by the edit system were among these 27 cases; these 23 changes were caused by the PR/ATE ratio being above the upper limits because of the error in PR. (The only other case where an apparently correct value of ATE was changed was the case just discussed where ATE was changed as a result of a keying error in TCW.)

The situation just described is typical of the editing errors that can arise in a system such as this one and some others currently used at the Census Bureau, which are sequential and typically only compare two fields at a time. Distinguishing between correct and incorrect fields can be difficult, or even impossible, with two-way comparisons. Furthermore, to begin a sequential edit, one field, PR for this edit, is commonly assumed correct. For cases where that field is in error, not only is it tabulated incorrectly, but any other field that is directly or

indirectly edited against it may also be tabulated incorrectly. More generally, an editing error of any field in a sequence can affect the edit of all fields that are subsequently edited directly or indirectly against it. Further illustration of these problems will be provided later in this subsection.

For this census, these problems perhaps could have been alleviated by substitution of an edit that examined a number of related fields simultaneously in attempting to determine the fields in error, such as an analyst would do when reviewing a case. For example, a simultaneous edit of PR, ATE, TR, THOURS, and administrative data for payroll, employment and receipts should readily be able to identify any large errors in PR only. In this author's opinion, a particular shortcoming of the current edit is that it makes almost no use of administrative data.

The second problem associated with the PR/ATE ratio test is that the lower limits on this ratio were simply too low, with final limits as low as 1.5 for some SIC's. (All cash values in this report are in thousands of dollars.) As a result there were 49 cases where the following all held: PR was apparently correct; the ATE value inputted to the edit was too large; the ratio test passed and hence ATE was not changed; the ATE error could have been detected by the use of an edit system of the type just described. This includes all 15 cases previously discussed where the ATE error resulted in an imbalance among the employment items. Furthermore, for many of these 49 cases, the ATE error could have been detected by the PR/ATE test alone if more realistic limits had been used. (Among the other 7 cases of the

total of 56 cases of uncorrected ATE errors, PR was also in error for 1 case, while for the remaining 6 cases, which will be discussed later, the ATE error occurred indirectly because of an error in AOE.) In fact, unrealistic ratio limits were a common occurrence with the current edit. This was primarily because the upper and the lower limits for each ratio in each SIC were defined to be the fourth highest and the fourth lowest value respectively for that ratio among all sample cases in that SIC, including part-year reports. Unfortunately, there often appeared to be more than 3 serious errors in the same direction for a field in a SIC, but the excess errors could not be detected by the appropriate ratio test.

One alternative approach to setting the edit limits would be to review a sample of cases from the previous census to obtain estimates of the proportion of the cases that were truly in error among those failing the edit for various limits, and then to use this information in determining the limits. For at least some ratio tests, consideration should also be given to having the limits vary not only with the SIC but also with the size of the establishment. For example, a small establishment with mostly part-time employees working second jobs might have a genuinely low value for PR/ATE, but it would appear less likely that a larger establishment would function in this fashion.

Once the edit of PR against ATE was completed, neither data item could be changed at subsequent stages of the edit. Consequently, any errors in these fields that remained could only be corrected by an analyst's review, which, for these 86 cases,

appeared to happen only twice for PR and not at all for ATE. Furthermore, as will be illustrated, subsequent ratio tests that paired either PR or ATE with another field sometimes resulted in the other field being incorrectly changed because of an error in PR or ATE.

After the edit of ATE against PR, payroll and employment detail were edited further. In this portion of the edit the following problem occurred for 13 cases. The value of AOE inputted to the edit was too high and resulted in the ATE being incorrectly above the large growth cutoff. In none of these cases had ATE been changed by the edit as a result of the PR/ATE ratio test, and in six cases, even with an improved edit, the error could only have been detected by examination of the employment detail. The edit of the employment detail in this portion of the edit did result in the ratio test ACW/ATE being below the lower limit for 3 of these 13 cases. However, both ACW and AOE were then imputed from ATE for these cases, resulting in tabulated values for both ACW and ATE that were too large. No employment data were changed for the other 10 cases. Another ratio, OES/AOE, that could have detected the error in AOE was later computed. However, the final lower limits on this test were as low as .5 for some SIC's, and only 4 of the 13 cases would have failed this test with the final limits. (Some cases were edited with preliminary limits that differed from the final limits so it is not known which cases actually failed the test.) Furthermore, when an establishment failed this test, no data were changed as a result. Instead the case was merely

flagged for analyst review. Since the edit generated 600,000 such flags, it is not surprising that the analysts were not able to review all such cases, and AOE did not appear to have been changed by an analyst for any of these 13 cases. Again, an edit that simultaneously examined, AOE, ACW, ATE, OES, Construction Worker Wages (CWW), PR, ATE, THOURS, and administrative payroll and employment, would have ascertained that AOE was in error in these cases. It would have changed AOE and recomputed ATE, which would have corrected the original error in editing employment.

Also note that since payroll and employment detail were edited to be consistent with PR and ATE respectively in this portion of the edit, errors in either PR or ATE carried over to their detail.

An imputation procedure that could be improved was also observed in this portion of the edit. If CWW and OES were positive, and ACW and AOE were to be imputed, they were imputed from ATE only. As a result the ratios CWW/ACW and OES/AOE, particularly the latter, were not always reasonable. As an alternative, values for ACW and AOE could be imputed from CWW and OES respectively, and then the employment detail could be "raked" to obtain agreement with ATE. Similarly, if ACW and AOE were positive while CWW and OES were to be imputed, they could be imputed from ACW and AOE respectively, and then the payroll detail could be "raked" to obtain agreement with PR. In the current edit, PR alone was used to impute payroll detail in this situation.

The next step in the edit process was the edit of the Quarterly Construction Hours Worked the i-th Quarter (HOURSi), i=1,2,3,4. HOURSi was edited by means of the ratio HOURSi/CWi where CWi was the Number of Construction Workers, i-th Quarter. For each quarter where this ratio was not within limits, HOURSi was imputed from CWi. THOURS then became the sum of the edited values of HOURSi, i=1,2,3,4. However, for most of the cases in which the tabulated values of ATE were incorrect, the error was caused, at least in part by (or resulted in errors in), the CWi's. Consequently, in many of the cases in which ATE was in error, apparently correct values for THOURS and HOURSi were changed by the edit.

If an edit which compared ACW, THOURS and CWW simultaneously had been used, and if CWW/THOURS was acceptable but THOURS/ACW was not, this would have pointed to ACW as the field likely in error. In fact, a particular shortcoming of the current edit was that the ratio CWW/THOURS was not originally used in the edit process at all, since this should have been a particularly stable ratio, with a legally prescribed minimum. This shortcoming was discovered during the processing, and a supplementary edit was performed which reimputed THOURS against CWW. This gave a generally better imputation due, in part, to the closer relationship between these two fields than between THOURS and ACW, but more importantly, because there were more errors in the edited values of ACW than CWW. However, since the employment fields were not reimputed and since the payroll fields were sometimes in error, the tabulated values of THOURS and CWW

resulted in values of THOURS/CWW that were sometimes unreasonable, including two cases below 100 and two cases above 6000. Being forced to choose which pair of fields would be allowed to be inconsistent is another consequence of a sequential edit. Furthermore, the supplementary edit imputed THOURS only if at least one of the HOURSi/CWi ratios had failed. Some establishments passed these ratio tests even though they had values for CWW/THOURS below the minimum hourly wage. No data were changed by this supplementary edit for these cases.

The last step in the edit that is of concern in this study was the edit of TR through the ratio TR/PR. If this ratio was not within the specified tolerance limits, TR was generally imputed from the presumed correct value of PR. However, in 11 of the 12 cases where the tabulated value of PR was in error, an apparently correct value of TR was changed by the edit. The remaining case in which the edit created an error in an apparently correct value of TR resulted from the method used to reedit a record after an analyst's corrections. Only those portions of the edit that involved fields changed by an analyst were reedited. Since PR and TR were in separate portions of the edit, a correction to PR would not cause TR to be reedited, despite the fact that the key edit of TR is against PR. In the particular case in question, the values PR=10,000, TR=350 were reported. The edit changed TR to 49,560. Apparently, an analyst in reviewing the case realized that actually PR had been reported incorrectly in dollars instead of thousands of dollars and changed PR to 10, while leaving TR at 49,560. Since TR was not

reedited, this obviously grossly inflated value became the tabulated value for TR. This problem could have been avoided if, whenever a field was corrected, a reedit was performed for every field that was originaly directly or indirectly edited against the corrected field. Similarly, in a second case, in which the input values of both PR and TR were in error, neither field was corrected by the edit, but PR alone was corrected by an analyst. TR was not corrected on reedit, again.

There was a possible situation in which TR would not be imputed even though TR/PR was above the upper limit. If SO $\geq$ .75 TR, where SO is the amount paid by the establishment for subcontracted work to other construction companies, then TR was accepted but flagged. However, at a later stage in the edit, if SO/TR > .75 then for about half the SIC's SO was rejected and imputed from TR. Thus for these SIC's, TR could be accepted because of a value of SO that itself was not believed to be correct. Although this situation did not occur for any of the cases in this study, it did not appear to be logical.

2.3 The Employment Questions

As described in Section 2.2, nine of the response errors in employment resulted from the respondent copying the TCW entry onto the ATE line in the schedule, and the edit failing to correct these errors. However, this type of error and some of the other errors in employment items might not have been made if a different set of employment questions had been asked. For 1982 the respondent was first asked to provide the number of construction workers employed during a specified pay period in each of

the four quarters (CW1 - CW4), to sum these four numbers to obtain TCW, and to divide by four to obtain ACW. Then the respondent was to record the number of other employees during a specified pay period in the first quarter (AOE), and finally sum ACW and AOE to obtain ATE.

Three of these fields, TCW, ACW and ATE, were arithmetical combinations of other fields. From an operational point of view, having the respondent perform the arithmetic, as opposed to having a computer alone perform the operations, was inefficient. It created an extra burden for both the respondent and the keyer, and created additional opportunities for respondent confusion and error. Furthermore, adding one quarter of other employment to an average of four quarters of construction worker employment was not a natural operation, which may have resulted in such errors as copying the TCW entry onto the ATE line, and recording incorrect values on the AOE line. Also, the expression "Total number of employees", which appeared on the ATE line, was an imprecise expression that might also have contributed to errors on this line.

In addition, there did not appear to be any particularly strong offsetting advantages in having the respondent perform the arithmetic on the employment questions as there were, for example, for the payroll inquiry, where the respondent was asked to sum CWW and OES to obtain PR. In the case of payroll, an advantage in having a PR line was that a total payroll figure was generally available in the respondent's records. If used by the respondent as a check against the sum of CWW and OES, entering

this figure might have enabled the respondent to detect errors in payroll detail. Also, if the respondent would not or could not provide payroll detail, at least a field was available to record PR. However, it was unlikely the respondent would have figures for TCW, ACW or ATE in their files, and any value given for these fields without the corresponding detail should have been considered suspect.

2.4 General Observations on the Large Growth Procedure

Although the original purpose of this study, the evaluation of the large growth procedure and suggested alternatives, was not completed, several observations concerning the procedure and its operation will be presented here.

1. Irrespective of the large growth procedure used, the procedure implemented should have been the procedure that was intended to have been implemented. However, in two situations, one for response large growth cases, and one for administrative large growth cases, the tabulations were affected because the weights used were not the intended weights.

For the response large growth cases, the determination of which cases were subject to the large growth procedure should have been on the basis of the final data used in the tabulations. However, as actually implemented any noncertainty case which at the end of any pass through the edit had census data above a large growth cutoff was treated as a large growth response case and, hence, tabulated with weight one, even if the final data were all below the cutoffs.

Many of the cases that were large growth in administrative data only were tabulated incorrectly with their original weights, due to a programming error.

2. The term "large growth" may be a misnomer for some cases. The sampling weight was based on the 1981 administrative payroll alone, with a certainty cutoff of 280. However, the upper limit for the ratio TR/PR in the edit is either 15 or 25, depending on the SIC. Assuming that these limits were reasonable, an establishment with 1981 payroll slightly below 280 might easily have had 1981 receipts above the 2,500 large growth cutoff for receipts and, thus, be classified as a large growth case even if there was no growth at all from 1981 to 1982 in any of the key data items. It would appear that this situation could only have been avoided if the sampling weight had been based on 1981 administrative payroll, receipts and employment.

3. Weighted data should be used to determine which cases should be subject to a large growth procedure. For example, the contribution to the estimates of an establishment with data slightly above the cutoffs and a weight of 1.5 is less than the contribution from an establishment with data slightly below the cutoffs and a weight of 5, yet it is only the former case which is subject to the large growth procedure. An alternative procedure for large growth response cases only, would be to reduce the weight to the point where none of the weighted PR, ATE, and TR exceed a cutoff value. Some theoretical justification for such a procedure is given in Ernst (1980),

although the assumptions in that paper do not exactly match the conditions in the construction census.

3. SAMPLE DESIGN AND SELECTION

The sample for this study was all large growth cases in the 1982 Construction Census. This census was actually a probability sample for which establishments with 1981 administrative payroll less than 280 were generally selected with probability less than 1. (Duke (1983) provides complete details on the census sampling procedures.) All selected noncertainty cases for which edited PR $\geq$ 500, ATE $\geq$ 50, or TR $\geq$ 2,500 were considered large growth response cases. All select cases for which the 1982 administrative payroll, employment, or receipts was at or above the corresponding cutoffs were considered "select" administrative large growth cases. Cases could be in both of these categories. Furthermore, all nonselect cases for which the 1982 administrative payroll, employments or receipts were at or above the corresponding cutoff were considered "nonselect" large growth cases.

The sample for the edit study portion of this report was all large growth response cases, whether or not they were also administrative large growth cases.

4. METHODOLOGY

CSD, as part of the census processing procedures, maintained listings of the response large growth cases, select administrative large growth cases and nonselect large growth cases. CSD combined these listings to obtain a single listing containing the Identification Number of each large growth case together with a

code indicating in which category or categories of large growth the case was included. This file was used to identify the sample cases for this study. At the time this file was created, CSD removed approximately 400 cases which had been processed as large growth response cases even though the data used in the tabulations were all below the large growth cutoffs. This problem was discussed in Section 2.4.

The Statistical Research Division (SRD) matched this listing to the 1982 Construction Census Data Register and created a record for each matched case, which contained an abbreviated form of the data register record together with the large growth status code. Soon after this file was created, CSD informed SRD that some cases should be deleted. This was principally due to the fact that in the processing some certainty cases had improperly been keyed with weights greater than one (an error that was corrected before the data were tabulated) and, as a result, had incorrectly been classified as large growth cases. After removing these cases the resulting file, according to the large growth status codes, contained 383 cases of only response large growth, 208 select cases of only administrative large growth , 187 cases of both response and administrative large growth, and 733 nonselect large growth cases. This was the file used in this study.

However, it was soon discovered that, due to a number of problems, this SRD file was not complete, nor were all the codes correct. The missing cases mainly resulted because some ID's on the CSD listing of large growth cases did not match the data

register. Part of the reason for this was that the listing was obtained through a keying operation. By a manual check of the listing, it was discovered that there were 30 cases that did not match due to keying errors. These cases were later added to the file, but too late to be used in this study.

The errors in the large growth status codes were detected by comparing the code for each record in the SRD file against the data. Among the discrepancies discovered were 33 cases that were coded as response large growth only and 101 cases that were listed as select administrative large growth only, but should have been coded as both according to the data. (Note however, that among the 101 cases, 67 were total nonrespondents, with census data imputed.) This was due to various processing complications and errors. Since these cases were all listed as large growth in one category, the misclassifications alone had no effect on the census estimates. These processing problems would not have resulted in a case being excluded from the large growth listing. It is not known if there were other errors that resulted in some cases being missed entirely from the SRD file. In addition, 33 cases were listed as response large growth according to the codes, but the data used in the tabulations were all below the large growth cutoffs. These appeared to be cases that CSD missed when they removed the approximately 400 cases from their listing.

As an additional test, the weight of each record in the SRD file was checked to ascertain if it was indeed "1". For 93 cases, it was not. This problem was discussed in Section 2.4.

Because all these misclassifications had not been corrected at the time the edit study was done, the cases studied were the 570 with status codes indicating they were response large growth cases. Edward Ricketts of CSD reviewed all these cases. His review included a comparison of the data used in the tabulations to a microfilm of each original schedule. All cases for which data were changed as a result of the review were forwarded to this author for further study. The results were presented in Section 2.2.

5. RELATIONSHIP OF THIS STUDY TO SIMILAR STUDIES

There is some relationship between this study and three of the other studies of the 1982 censuses.

The extensive data collected in the Processing Study was of use in our study. For a few of the large growth response cases that were also in the Processing Study sample, the Processing Study data were used to confirm the type of actions taken by the edit system. Another area of relationship between these two studies arose from the explanations given in this study of some of the actions taken by the edit system, which could also be used to explain some of the observations noted in the Processing Study. For example, it was noted in the Processing Study that the THOURS field for an establishment frequently was changed several times; at least a partial explanation was provided in this report.

This study is obviously related to the Evaluation of the Edit and Imputation Procedures in the Business Censuses, since both studies concentrated on edit systems for economic

censuses. However, the focus of the two studies was different. The objective of the business edit was to conduct an analytic evaluation of the methodologies and procedures incorporated into the automated routines within the edit.

The Content Evaluation Pilot Study investigated rental payments, inventories, assets, capital expenditures and depreciation for construction and other economic censuses, and, among other findings, suggested that some changes in the questionnaires might improve the quality of the reporting. As noted in Section 2.3 of this report, this also might be true for the employment questions.

References

Construction Statistics Division (1982), "Census Edit," Bureau of the Census report.

Duke, Dennis (1983), "Documentation of Statistical Methodology for the 1982 Census of Construction Industries," Bureau of the Census memorandum to Jesse Pollock, dated February 28.

Dyke, T. Christopher (1985), "1982 Economic Census Processing Study," Bureau of the Census, 1982 Economic Censuses Evaluation Task Force report.

Ernst, Lawrence R. (1980), "Comparison of Estimators of the Mean Which Adjust for Large Observations," Sankya, Series C, 42, 1-16.

Greenberg, Brian and Petkunas, Thomas (1985), "An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses," Bureau of the Census, Economic Censuses Evaluation Task Force report.

Van Nest, Gary (1985), "Content Evaluation Pilot Study," Bureau of the Census, Economic Censuses Evaluation Task Force report.

Table 1. Error Combinations in Cases where Review Removed Large Growth Response Status

Error Combinations	All Errors Correctable by a Computer Edit	One or More Errors Not Correctable by a Computer Edit
PR	1	5
ATE	55	6
TR	3	4
PR and ATE	10	4
PR and TR	2	3
ATE and TR	0	1
PR, ATE and TR	15	0
Total	86	23

Table 2. Status of Data Inputted to Edit and Final Tabulated Data for Cases Where All Errors Correctable by a Compute Edit

Input Data	Tabulated Data	Data items PR	ATE	TR	Total
Apparently Correct	Apparently correct	56	6	65	127
	In error, > cutoff	0	20	9	29
	In error, ≤ cutoff	0	4	2	6
In error, > cutoff	Apparently correct	2	0	0	2
	In error, > cutoff	27	56	6	89
Not reported	Apparently correct	0	0	1	1
	In error, ≤ cutoff	1	0	3	4
Total		86	86	86	258

Note: Correct data never exceed cutoffs for these cases. Also, there are no cases for input data, tabulated data combinations not listed in table.