Resolving Duplicate Detection Issues in Data Validation

Background. The DV software contains criteria for all populations except Tax 5 to detect duplicate records. The software first checks all records for syntax and other errors (e.g., records with dates out of range) and puts them into the error table. All remaining records, which meet the conditions for assignment into subpopulations—records that "parse"—are then examined using criteria that vary by population to see whether two or more records are duplicates. All records considered duplicates are sent to the general errors table and also to a separate duplicates table. Validators examine each set of duplicate records, determine which record is the countable one, retain that record in the extract file and remove the "true duplicate" (or duplicates) from the extract file before reloading the file.

The DV software occasionally rejects as duplicates legitimate, countable transactions. False or apparent duplicates are not a common problem, and do not occur in all populations. But when they do occur, they present a challenge to the validator and programmer staff. This guide indicates the steps that may be taken to work around the issue of apparent duplicates.

Duplicate Detection Criteria by Population. The following table gives the duplicate detection criteria for each population. The third column gives the basic Reporting Rule, e.g., for Population 1, that each week can only be claimed once. The second column, Duplicate Detection Fields, gives the fields DV uses to determine whether the records in the extract file conform to or violate that rule. During the load process, the DV software examines every record that has no syntax or logic errors; if multiple records have identical values for those fields, at least one of those records is presumed to be a duplicate and all are set aside as errors for the validator to examine. If the validator determines that all of the records are reportable, they must be rendered unique by modifying values in one of the Duplicate Detection fields. The "Modifiable Field" column identifies that field--if the population record has one--and explains why some Population records cannot be modified.

Duplicate Detection Criteria by Benefits and Tax Populations			
Benefits			
Population	*Duplicate Detection Fields*	*Reporting Rule*	*Modifiable Field*
1	SSN, Claim Week-end Date	Claim each week once	None. There can only be 1 record for a week.
2	SSN, Mail Date, Check #/ Unique ID	Normally one Final Payment per Benefit year.	Unique ID
3	• All claim types: SSN, Claim-Filed Date, Claim Type • UI New & Transitional Claims: SSN, claim type, Claim-Filed Date, Claim Sufficient/Insufficient	Only one of each claim type filed on a given day; multiple new and transitional UI claims are OK if those with insufficient monetary precede sufficient monetaries, and there is only one sufficient claim establishing a benefit year.	None. The extensive criteria cover all countable transactions and do not allow for false duplicates.
3a	SSN, Claim-Filed Date, Separation Date	Add claim must involve ≥ one week break in claim series due to employment	None. There can only be one additional claim for each separation.

4	SSN, Intra/Inter, Week-End Date, Mail Date	Only one week compensated for a given week	None. At present this population does not have a modifiable field. See discussion below.
5	SSN, Unique ID, Issue Type, Notice Date	Count every nonmonetary determination once.	Unique ID
6	SSN, Docket #/Unique ID	Count every appeal once	Unique ID
7	SSN, Docket #/Unique ID	Count every appeal once	Unique ID
8	SSN, Docket #/Unique ID	Count every appeal once	Unique ID
9	SSN, Docket #/Unique ID	Count every appeal once	Unique ID
10	SSN, Docket #/Unique ID	Count every appeal once	Unique ID
11	SSN, Docket #/Unique ID	Count every appeal once	Unique ID
12	SSN, Date Established, Unique ID	Count each overpayment once	Unique ID
13	SSN, Unique ID, Activity Type, Date of Activity	Count each transaction once	Unique ID
14	SSN, Unique ID	Count each overpayment once	Unique ID
15	SSN, Date Established, Unique ID	Count each overpayment once	Unique ID
Tax			
1	EAN	Count each employer once	None. Each employing unit is assigned one EAN, even for multiple business locations..
2	EAN, Employer Report Quarter (ERQ)	Each employer owes only one report for each ERQ	None. One report from each EAN due per each ERQ.
3	EAN, Status Type, Status Date, Predecessor Account Number	Count each status determination once; may be > 1 determination for an EAN	EAN
4	• Established: EAN, trans date, estab date, ERQ or Due Date, amt estab • Liquidated or Uncollectible: EAN, trans date, ERQ or Due date, trans type, trans amount • Removed: EAN, ERQ or Due Date, amount removed • Balance: EAN, ERQ or Due Date, balance amount	Report each transaction once	EAN
5	EAN, Audit ID number	Count each audit once	No duplicate criteria.

Where Duplicate Detection Failures Occur. The software can erroneously send cases of legitimate, reportable, multiple transactions to the error file as duplicates. The two main causes seem to be:

1. The State cannot populate a Duplicate Detection Field) . The typical field is the Unique ID field, which is conditionally required in Benefits Populations 5 and 12-15. Some state systems do not create a unique ID for nonmonetary determinations and overpayments. When the ID is missing,

the software determines duplicates on the basis of the other fields, e.g., SSN, Issue Type, and Notice Date for Population 5. A number of apparent duplicates can thus appear in the error table for the validator to sort out.

2. State Practice Conflicts with DV Duplicate Criteria. Some states generate multiple instances of certain transactions that fail the duplicate criteria but are countable. Even in states that have unique ID numbers for various transactions, apparent duplicates can be generated as follows:
    a. *Tax Population 4,* when multiple transactions of the same transaction type and amount are posted to the system on the same day for the same account;
    b. *Benefits Population 2*, when a claimant receives two legitimate final payments on the same day (additional funds would have to be returned to the account and given the same payment ID.)
    c. *Benefits Population 4*, when multiple partial payments for the same week compensated are made on the same day.

Handling Erroneous Duplicate Records. In all the cases except 2 (c) above, the remedy is the same: validator/programmer must modify one of the fields in the false duplicate records to make them unique. The field must be able to be modified, so that changing it will not cause it to fail Data Element validation. As the table above indicates, this is the unique ID field in Benefits and the EAN field in Tax. If the problem is the lack of a Unique ID in the state, the validator can simply add a 1 or 2 or 3 to the blank field. (EAN is a required field for all Tax records so the case of a missing EAN does not occur in Tax.) Both fields are very large character fields that the existing EAN or Unique ID does not exhaust, so there is plenty of room to add an additional character to differentiate the record and allow it to be parsed and counted.

Case 2 (c), involving Benefits Population 4, is a different matter. At least one state has reported that more than one partial payment can be legitimately posted on the same day. Unfortunately, Population 4 does not have a modifiable field that is used for duplicate detection. Although all population 4 records contain a check number or payment number that is unique to the payment, it is not presently used for duplicate detection, so modifying it for one or more of the records will not affect their duplicate status. At present there are no plans to modify the criteria for population 4, so a state encountering false duplicates should first of all see whether their number is sufficient to cause the population to fail because the rejection of false duplicates as errors causes a group validation count to be understated. If this is the case, the validator should document this fact in the Comments field and inform the National Office by a follow-up e-mail message as a reminder to review the submission and change the population score.