TREASURY INSPECTOR GENERAL FOR TAX ADMINISTRATION

Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement

September 29, 2014

Reference Number: 2014-20-063

This report has cleared the Treasury Inspector General for Tax Administration disclosure review process and information determined to be restricted from public release has been redacted from this document.

Phone Number / 202-622-6500
E-mail Address / TIGTACommunications@tigta.treas.gov
Website / http://www.treasury.gov/tigta

CUSTOMER ACCOUNT DATA ENGINE 2 DATABASE VALIDATION IS PROGRESSING; HOWEVER, DATA COVERAGE, DATA DEFECT REPORTING, AND DOCUMENTATION NEED IMPROVEMENT

Highlights

Final Report issued on September 29, 2014

Highlights of Reference Number: 2014-20-063 to the Internal Revenue Service Chief Technology Officer.

IMPACT ON TAXPAYERS

There is significant effort underway to ensure the accuracy of individual taxpayer account data on the Customer Account Data Engine 2 (CADE 2) database. This effort is an important part of its implementation because inaccurate data could delay this database from becoming the authoritative source of data, thereby increasing the cost of implementation.

WHY TIGTA DID THE AUDIT

This review was part of our Fiscal Year 2014 Annual Audit Plan and addresses the major management challenge of Modernization. The overall audit objective was to evaluate IRS efforts to ensure that the data in the CADE 2 database are accurate and complete.

The IRS requested that TIGTA evaluate the new data validation testing methodology. TIGTA performed this audit during the data validation testing process and provided the IRS with recommendations for continuous improvement.

WHAT TIGTA FOUND

Data validation efforts were efficiently performed due to adequate planning and resource coordination. For example, detailed data validation plans ensured that test activities were on track and a new process ensured that data defects were effectively managed.

The IRS identified the data fields to be verified and how each would be validated. While a large percentage of the data fields are validated with automated data compare tools, there is no documented plan to ensure that data fields validated using other means are validated periodically. The data sampling methodology for validating CADE 2 data is sound. The IRS developed a data sampling methodology to enable maximum data validation coverage by using a statistical sample, but key activities were not documented. After discussing the need to document the data sampling methodology, the IRS began development of the documentation. Several in-progress documents were provided for our review.

The IRS developed a Data Quality Scorecard to track progress in meeting data quality success criteria. However, the processes needed to effectively perform these activities were not sufficiently documented. As a result, some of the metrics were initially incorrectly reported.

WHAT TIGTA RECOMMENDED

TIGTA recommended that the Chief Technology Officer ensure that: 1) data validation test results are maintained and available for data fields not validated by automated data compare tools; 2) data validation plans include periodically validating the data fields that are not validated with automated data compare tools; 3) all data sampling processes are completely documented; 4) details needed for determining the Data Quality Scorecard metrics are completely documented; 5) all documentation needed to verify the data in the Data Quality Scorecard is stored for future reference; 6) automated data compare tools identify and report on data fields, not field identifier numbers; and 7) automated data compare tool reports clearly identify counters and align with data validation metrics.

The IRS agreed with six of the report's seven recommendations. The IRS plans to maintain results for manual data validation activities, validate changes to the data fields that are not validated with automated data compare tools, develop documentation on the procedures to collect and maintain data used to support data validation metrics and the Scorecard development process, and store Scorecard source documentation.

TREASURY INSPECTOR GENERAL
FOR TAX ADMINISTRATION

DEPARTMENT OF THE TREASURY

WASHINGTON, D.C. 20220

September 29, 2014

MEMORANDUM FOR CHIEF TECHNOLOGY OFFICER

FROM: Michael E. McKenney
Deputy Inspector General for Audit

SUBJECT: Final Audit Report – Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement
(Audit # 201320030)

This report presents the results of our review of the Customer Account Data Engine 2 data validation efforts. The overall objective of this review was to evaluate Internal Revenue Service (IRS) efforts to ensure that the data in the Customer Account Data Engine 2 (CADE 2) database are accurate and complete. This review is included in the Treasury Inspector General for Tax Administration's Fiscal Year 2014 Annual Audit Plan and addresses the major management challenge of Modernization.

While we are in general agreement with the IRS's response, one area of disagreement is whether CADE 2 Transition State 1.5 should be closed. We believe it should not be closed because, as of June 2014, only 68 percent of logic paths and 81 percent of field identifiers had been validated, and data defects were identified. There is a significant risk that additional defects will be identified as data validation continues. Therefore, we believe that CADE 2 Transition State 1.5 should remain open until several consecutive data validation cycles are completed with no new data defects identified.

Management's complete response to the draft report is included in Appendix VI.

Copies of this report are also being sent to the IRS managers affected by the report recommendations. If you have any questions, please contact me or Danny R. Verneuille, Acting Assistant Inspector General for Audit (Security and Information Technology Services).

Table of Contents

Background .... Page 1

Results of Review .... Page 4

Data Validation Efforts Were Performed Efficiently Due to Adequate Planning and Resource Coordination .... Page 4

The CADE 2 Program Management Office Identified the Data Fields to Be Verified and How Each Would Be Validated; However, All Data Fields Are Not Being Periodically Validated .... Page 5

Recommendations 1 and 2: .... Page 8

The Data Sampling Methodology for Validating CADE 2 Data Is Sound; However, Key Processes in the Implementation of the Methodology Need to Be Documented .... Page 9

Recommendation 3: .... Page 14

The Documentation and Processes for Determining the Data Quality Scorecard Metrics Need Improvement .... Page 14

Recommendations 4 and 5: .... Page 20

The Field Identifier Compare Tool Validates Data for Downstream Systems, but Data Discrepancy Reports Need Improvement .... Page 20

Recommendation 6: .... Page 22

Recommendation 7: .... Page 23

Appendices

Appendix I – Detailed Objective, Scope, and Methodology .... Page 24

Appendix II – Major Contributors to This Report .... Page 26

Appendix III – Report Distribution List .... Page 27

Appendix IV – Data Quality Scorecards .... Page 28

Appendix V – Glossary of Terms .... Page 32

Appendix VI – Management's Response to the Draft Report .... Page 35

Abbreviations

CADE 2	Customer Account Data Engine 2
EDMO	Enterprise Data Management Office
FLID	Field Identifier
IMF	Individual Master File
IRS	Internal Revenue Service
IT	Information Technology
KISAM	Knowledge, Incident/Problem, Service Asset Management
KPI	Key Performance Indicators
PMO	Program Management Office
TIGTA	Treasury Inspector General for Tax Administration

Background

In addition to standard testing procedures, several tools and methodologies have been identified and developed to validate the quality and integrity of the data.

The Customer Account Data Engine[1] 2 (CADE 2) Program is one of the top information technology modernization projects in the Internal Revenue Service (IRS). The CADE 2 mission is to provide state-of-the-art individual taxpayer account processing and data-centric technologies to improve service to taxpayers and enhance tax administration. The CADE 2 database will replace the current Individual Master File (IMF) account settlement system with a relational database processing system and become a key component in the IRS's enterprise-wide, data-centric information technology strategy. Implementation of the CADE 2 database (Database Implementation) to support this objective has introduced a greater potential for data anomalies due to a complex infrastructure, the complexity of tax processing, and the introduction of a new relational database. As such, there is a need for a comprehensive plan for ensuring the quality and integrity of the data within the CADE 2 database and the data provided to downstream systems. In addition to standard testing procedures, several tools and methodologies have been identified and developed to validate the quality and integrity of the data and to identify anomalies within the data.

In March 2013, in its definition of "authoritative source," the IRS Chief Counsel stated that if the data in CADE 2 are used as evidence of the transactions in the taxpayer's account, the information obtained from CADE 2 must be identical to the IMF at any given point in time.

On November 5, 2012, the CADE 2 Executive Steering Committee approved a conditional CADE 2 Transition State 1 Milestone 5 exit with two conditions. On April 4, 2013, the CADE 2 Executive Steering Committee closed the November 2012 Milestone 5 exit conditions and opened 2 new Exit conditions – one of which was for Data Assurance: 1) Data Assurance – "Getting the Data Right" and 2) Robust and Sustainable System Performance and Operational Readiness. These exit conditions are now being tracked by the IRS as Transition State 1.5. The criteria for closing the Data Assurance conditions are:

- Verification of a statistically sound sample (911 data fields against 270 million taxpayer accounts) of data in the CADE 2 database with no Priority 1/Priority 2 data defect tickets.
- Ability to scale data assurance tools to perform high-volume testing in time to test within filing season test windows.

[1] See Appendix V for a glossary of terms.

- Minimal (risk-based decision) code defects that could cause data defects downstream resulting in the need to use data correction tools.

The criteria for closing the Robust and Sustainable System Performance and Operational Readiness conditions are:

- Address identified system performance concerns.
- Meet organizational and operational readiness objectives.
- Meet and exceed system performance targets for database processing within budgeted time frames in production.

Over the past two years, the Treasury Inspector General for Tax Administration (TIGTA) reported on the progress of the CADE 2 Database Implementation. In September 2012, we reported that the IRS had data integrity checks in place at several levels of the CADE 2 database. Despite these controls and their data integrity testing efforts, the IRS could not ensure that the data on the CADE 2 database were consistently accurate and complete at the data field level due to the complexity of many of the data transformation rules and embedded business logic contained within IMF data fields.[2]

In September 2013, TIGTA reported that the CADE 2 database could not be used as a trusted source for downstream systems due to the 2.4 million data corrections that had to be applied to the CADE 2 database and the IRS's inability to evaluate 431 CADE 2 database columns of data for data accuracy. During the audit, the IRS was in the process of developing additional tools and implementing a new data validation testing methodology intended to achieve timeliness, accuracy, integrity, validity, reasonableness, completeness, and uniqueness.

The IRS requested that TIGTA evaluate the new data validation testing methodology. TIGTA agreed to do so[3] and performed this audit during the data validation testing process and provided the IRS with recommendations for continuous improvement. During fieldwork, the IRS took immediate steps to address concerns identified by TIGTA. Most of these actions are noted in the Management Action statements later in the report.

This review was performed at the IRS Information Technology (IT) organization's offices in Lanham, Maryland, during the period August 2013 through May 2014. We conducted this performance audit in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objective. We

[2] TIGTA, Ref. No. 2012-20-109, *The Customer Account Data Engine 2 Database Was Initialized; However, Database and Security Risks Remain, and Initial Timeframes to Provide Data to Three Downstream Systems May Not Be Met* pp. 3–4 (Sept. 2012).
[3] TIGTA, Ref. No. 2013-20-125, *Customer Account Data Engine 2 Database Deployment Is Experiencing Delays and Increased Costs* pp. 7–10 (Sept. 2013).

believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives. Detailed information on our audit objective, scope, and methodology is presented in Appendix I. Major contributors to the report are listed in Appendix II.

Results of Review

Data Validation Efforts Were Performed Efficiently Due to Adequate Planning and Resource Coordination

Detailed data validation plans were used to help ensure that test activities remain on track

The CADE 2 Database Implementation Data Validation Plan contains detailed information about the people, processes, and tools that will be leveraged to execute data validation and identify data anomalies in the Systems Acceptability Test environment and the Production Support Environment. To supplement the CADE 2 Database Implementation Data Validation Plan, the CADE 2 Program Management Office (PMO) also developed a Data Validation Execution Plan to facilitate the periodic meetings held to discuss the status of the data validation activities. The Data Validation Execution Plan included activities to be completed for each cycle of tests. Examples of activities include selecting the data samples for validation, executing the automated data compare tool, analyzing the data validation results reports, preparing problem tickets to correct defects, and assigning the problem tickets to the proper organization for resolution.

Adequate planning and resource coordination were achieved despite the Government shutdown and limited resources

The CADE 2 PMO adequately planned and coordinated the data validation testing schedule and process. Planning was accomplished despite the Government shutdown, limited testing support, and a limited testing environment during the November to December 2013 testing period. Accommodations were made to shift testing efforts from the Final Integration Testing environment to the Production Support Environment and to extend testing dates further into Calendar Year 2014. All this required a great deal of coordination among the IT and business unit organizations. Testing implementation procedures were also defined and coordinated among all involved parties.

In addition, periodic checkpoint meetings were effectively used to identify, keep all partners informed of, and resolve an issue with using the Field Identifier (FLID) Compare Tool (High Volume) (hereafter referred to as the FLID Compare Tool) in the Production Support Environment. The data validation activities for Final Integration Testing were completed on schedule in January 2014, and the data validation activities in 2014 continue to meet the target completion dates.

Data defects were effectively managed through the Knowledge, Incident/Problem, Service Asset Management (KISAM) system

Data defects identified through both automated and manual means were effectively managed through the KISAM system. Testers generated KISAM tickets when they found data discrepancies not previously identified. Triage teams then analyzed the tickets and assigned them to the appropriate groups for resolution. IRS procedures require that testers verify corrections before closing KISAM tickets. The CADE 2 PMO monitored the list of KISAM tickets generated during data validation.

Most of the data correction tools were successfully developed and deployed to enable database data defect corrections

The IRS developed three new tools to correct CADE 2 database data defects caused by loading errors, the receipt of bad data from the IMF, or software/hardware failures during daily update runs.

- The Update in Place tool executes direct updates to data on the CADE 2 database through the use of Structured Query Language update statements.
- The Account Deleter/Re-Extractor tool makes corrections by deleting erroneous data from the database, reextracting it from the IMF, and loading the corrected data into the database.
- The Taxpayer Identification Number Bypass Tool is used in conjunction with the Account Deleter/Re-Extractor tool. It allows daily update processing to proceed while temporarily bypassing updates for specific CADE 2 database accounts with known data problems until the problems can be corrected.

These tools were sufficiently tested through the combined efforts of the Enterprise Services Enterprise Systems Testing and the Applications Development organizations (both a part of the IT organization) and were successfully deployed into production in Calendar Year 2014. The last data correction tool, the FLID Specific Update Tool, is scheduled for deployment on June 27, 2014.

The CADE 2 Program Management Office Identified the Data Fields to Be Verified and How Each Would Be Validated; However, All Data Fields Are Not Being Periodically Validated

The Government Accountability Office's *Standards for Internal Control in the Federal Government* state that control activities include verifications and accurate and timely recording of transactions and events.[4] Transactions should be promptly recorded to maintain their relevance and value to management in controlling operations and making decisions.

According to information technology industry standards, data quality assurance can be achieved only when the following criteria are met:

- Accuracy: Data must be correct and consistent.
- Completeness: All related data must be linked from all possible sources.
- Availability: Data must be available upon demand.
- Timeliness: Current data must be available.

Data quality for the CADE 2 database is dependent on the database matching corresponding IMF data. The CADE 2 Database Implementation Data Validation Plan for 2013/2014 documents the activities that need to be performed in order to validate the CADE 2 database. This encompasses validation of all CADE 2 data fields that are derived from the IMF. In addition, data quality ensures that the CADE 2 data records match the corresponding data records from the IMF. This encompasses validation of all data fields that are fed downstream from the IMF currently and that will be fed to downstream systems by the CADE 2 database.

For the 2014 database format, the CADE 2 PMO prepared a data coverage matrix that identified 1,018 verifiable IMF data fields that would be validated. Figure 1 provides the distribution of the validation methods.

[4] Government Accountability Office (formerly known as the General Accounting Office), GAO/AIMD-00-21.3.1, *Internal Control: Standards for Internal Control in the Federal Government* (Nov. 1999).

Figure 1: Data Fields Grouped by Validation Methods

2014 Data Field Count		
Number of Fields to Be Validated		1,018
Fields Validated by FLIDs		(911)
Fields Validated by Other Methods		107
Other Validation Method Details		
No Need to Validate		3
Database Integrity Check	20	
Systems Acceptability Testing Cases	41	
General Transcript Report Test	2	
Manual Compare	41	
Total Fields Validated by Other Methods		104
Total		107

Source: CADE 2 Database Data Field Coverage v2.4.2 11222013_Final. Figures in parentheses are negative (subtractions).

The FLID Compare Tool will validate 911 data fields that will be fed to downstream systems. The Data Quality Scorecard metrics used to monitor and report the status of data validation efforts will focus on only the data fields fed to downstream systems. Therefore, there will be no status reporting on the remaining 107 data fields.

We requested test documentation for each category to review the validation of the 104 data fields needing validation (three fields required no validation; see Figure 1). While the test documentation was not readily available, by May 9, 2014, we received sufficient testing documentation for 100 of the 104 data fields supporting that the data fields were initially validated.

In addition, the CADE 2 PMO determined how often the data fields derived from the IMF will be validated during production. The data validation execution schedule dated May 8, 2014, details data validation activities planned for production cycles 5 through 22. The data validation activities are concentrated on the data fields that will be fed to downstream systems. While we obtained test documentation supporting the initial validation of 100 of 104 data fields currently not fed to downstream systems, all 107 data fields not validated by the FLID Compare Tool are derived from the IMF; therefore, they should be periodically validated if the CADE 2 database is to become the authoritative source of data.

Without periodically validating all data derived from the IMF and maintaining adequate documentation of the validation results, management will not have full assurance that the data are complete and accurate.

On April 29, 2014, the CADE 2 Executive Steering Committee approved a proposal to close the Transition State 1.5 Data Assurance exit condition by June 27, 2014, after testing transmission of data to selected downstream systems. However, a Data Quality Scorecard reported that as of June 27, 2014, there were five open Priority 2 data defect tickets. Three of the five were from the data validation activities that were recently completed on June 27, 2014. Therefore, the exit condition that requires verification of a statistically sound sample (911 data fields against 270 million taxpayer accounts) of data in the CADE 2 database with no Priority 1 or 2 data defect tickets was not successfully met. We believe that Transition State 1.5 should not be closed until several consecutive cycles of data validation results show that no Priority 1 or 2 data defect tickets remain open. The IRS indicated that data validation is a dynamic process and when reviewing problem tickets the nature of the ticket needs to be considered. In this case, the open tickets were of low impact and minimal risk.

The IRS closed the Data Assurance exit condition on June 17, 2014. With this closure, IRS management indicated acceptance of the risk of data defects occurring as data validation proceeds through the remainder of the processing year.

Recommendations

The Chief Technology Officer should:

Recommendation 1: For data fields not validated through automated data compare tools, ensure that data validation test results are maintained and available.

> ***Management's Response:*** The IRS agreed with this recommendation and asserts that processes are in place. These test results are an integral part of maintaining transparency with CADE 2 stakeholders and delivery partners. The business organization data validation results and testing results are maintained based on the organization's official procedures. The IRS affirms that it will continue to maintain results for manual data validation activities in accordance with standard procedures, on an ongoing basis.

Recommendation 2: Ensure that data validation plans include periodically validating the data fields that are not validated with automated data compare tools.

> ***Management's Response:*** The IRS agreed with this recommendation. Any changes to the data fields that are not validated with automated data compare tools, such as annual filing season updates, will be validated through standard testing procedures. The IRS has updated the data validation plan to reflect the frequency and process of manually validating data fields not fed to downstream systems.

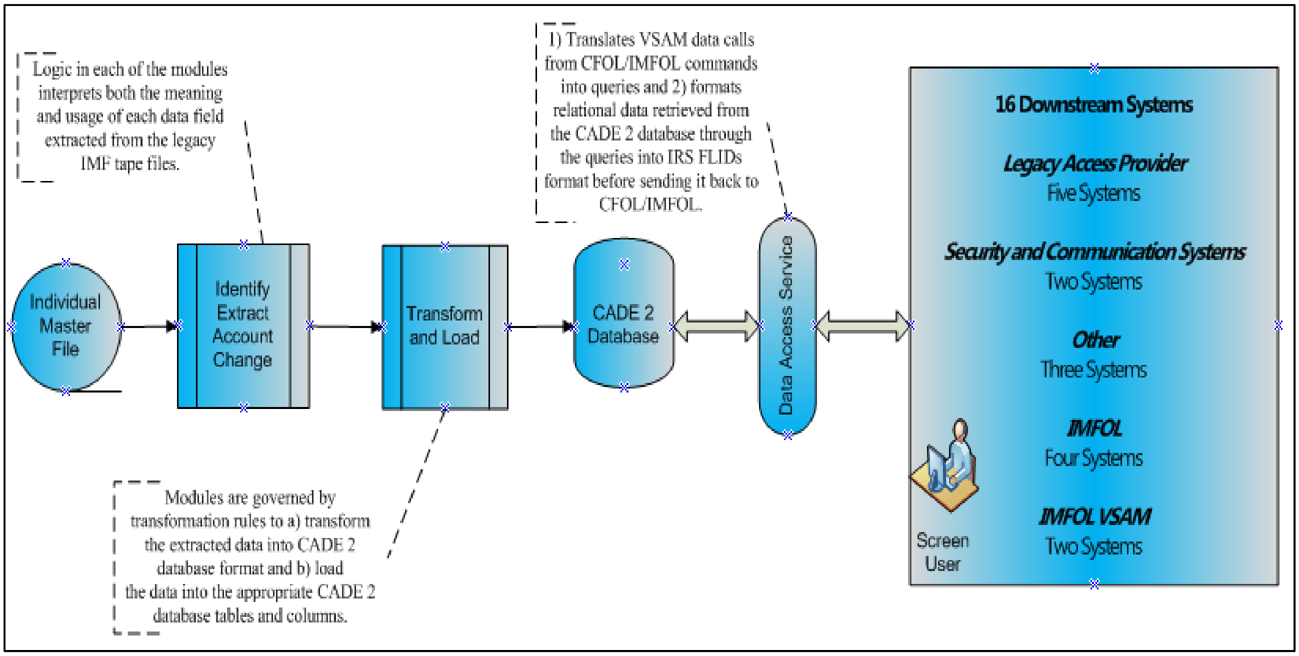
The Data Sampling Methodology for Validating CADE 2 Data Is Sound; However, Key Processes in the Implementation of the Methodology Need to Be Documented

The Government Accountability Office's *Standards for Internal Control in the Federal Government* state that control activities include verifications and accurate and timely recording of transactions and events. Transactions should be promptly recorded to maintain their relevance and value to management in controlling operations and making decisions. According to industry standards, data quality assurance can be achieved only when the following criteria are met: 1) accuracy; 2) completeness; 3) availability; and 4) timeliness.

The CADE 2 PMO developed a data sampling methodology to identify datasets (random and Smart samples) to cover all transformation logic paths and define appropriate Taxpayer Identification Numbers and modules for each validation method. Implementation of this methodology is ongoing and being refined.

The data sampling methodology was used throughout Systems Acceptability Testing and Final Integration Testing of the 2013 and 2014 version of the data and continues to be used for production validation in order to maximize coverage of data transformation logic between the IMF and the CADE 2 database. Figure 2 illustrates the data flow and transformation process between the IMF and the CADE 2 database and from the CADE 2 database to downstream systems. The methodology identifies the probability of certain transformation logic paths occurring and pinpoints specific Taxpayer Identification Numbers that can be used for data validation that meet specific business conditions.

Figure 2: The CADE 2 Database Corporate Files Online/ IMF Online/Data Access Service Interface Data Flow

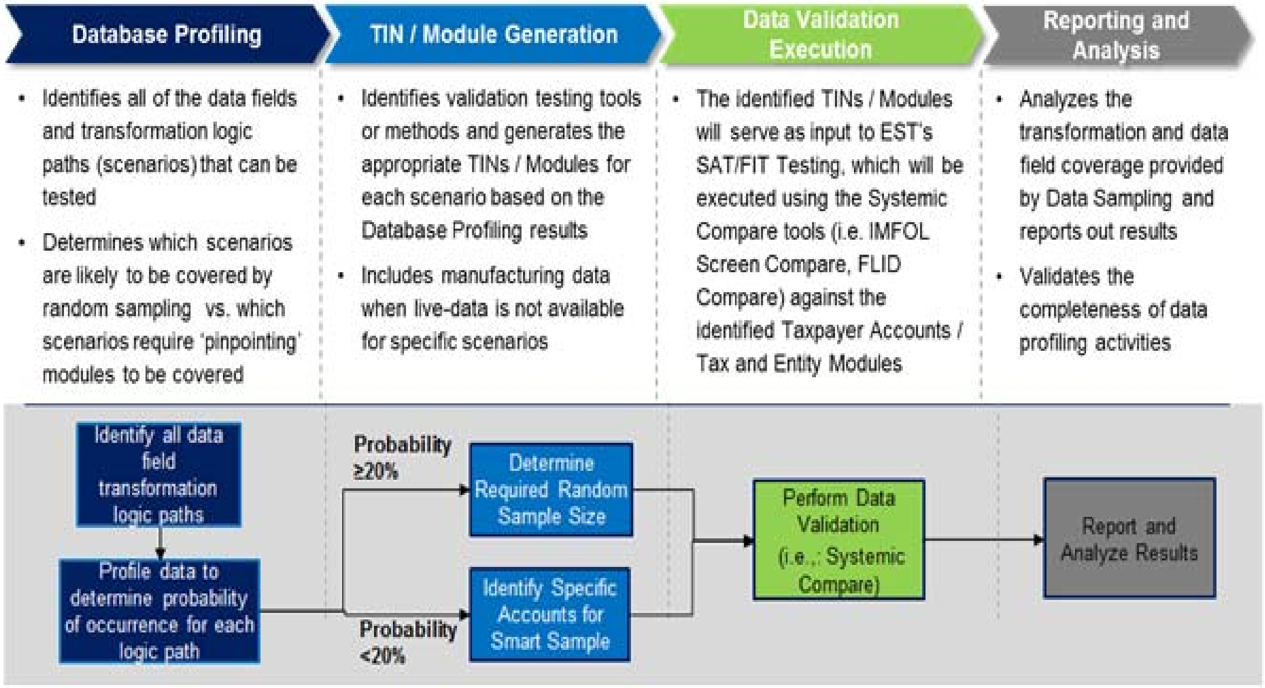

Logic in each of the modules interprets both the meaning and usage of each data field extracted from the legacy IMF tape files.
1) Translates VSAM data calls from CFOL/IMFOL commands into queries and 2) formats relational data retrieved from the CADE 2 database through the queries into IRS FLIDs format before sending it back to CFOL/IMFOL.
Individual Master File
Identify Extract Account Change
Transform and Load
CADE 2 Database
Data Access Service
16 Downstream Systems
Legacy Access Provider
Five Systems
Security and Communication Systems
Two Systems
Other
Three Systems
IMFOL
Four Systems
IMFOL VSAM
Two Systems
Screen User
Modules are governed by transformation rules to a) transform the extracted data into CADE 2 database format and b) load the data into the appropriate CADE 2 database tables and columns.

Source: TIGTA, Ref. No. 2013-20-125, Customer Account Data Engine 2 Database Deployment Is Experiencing Delays and Increased Costs p. 8 (Sept. 2013), and a presentation for the CADE 2 Executive Steering Committee Meeting held on April 29, 2014, slide 16. VSAM – Virtual Storage Access Method. CFOL – Corporate Files Online. IMFOL – Individual Master File Online.

The data sampling process to maximize coverage of transformation logic during data validation execution consists of the following activities:

- *Database Profiling identifies all of the data fields and transformation logic paths that can be tested as well as the probability of each transformation occurring in the data for that processing cycle* – Because some business transactions occur infrequently or are unique, production data may not be available to validate those transformation rules until later in the processing year. Figure 3 outlines the high-level approach to the data sampling methodology, which will provide test cases as inputs to the Automated Compare Data Validation tool. Transformation logic paths that have a 20 percent or greater probability of occurring in the data will be included in a random sample; those with less than 20 percent probability will be included in a Smart sample.

Figure 3: Data Sampling Methodology – High-Level Approach



Database Profiling

- Identifies all of the data fields and transformation logic paths (scenarios) that can be tested
- Determines which scenarios are likely to be covered by random sampling vs. which scenarios require 'pinpointing' modules to be covered

TIN / Module Generation

- Identifies validation testing tools or methods and generates the appropriate TINs / Modules for each scenario based on the Database Profiling results
- Includes manufacturing data when live-data is not available for specific scenarios

Data Validation Execution

- The identified TINs / Modules will serve as input to EST's SAT/FIT Testing, which will be executed using the Systemic Compare tools (i.e. IMFOL Screen Compare, FLID Compare) against the identified Taxpayer Accounts / Tax and Entity Modules

Reporting and Analysis

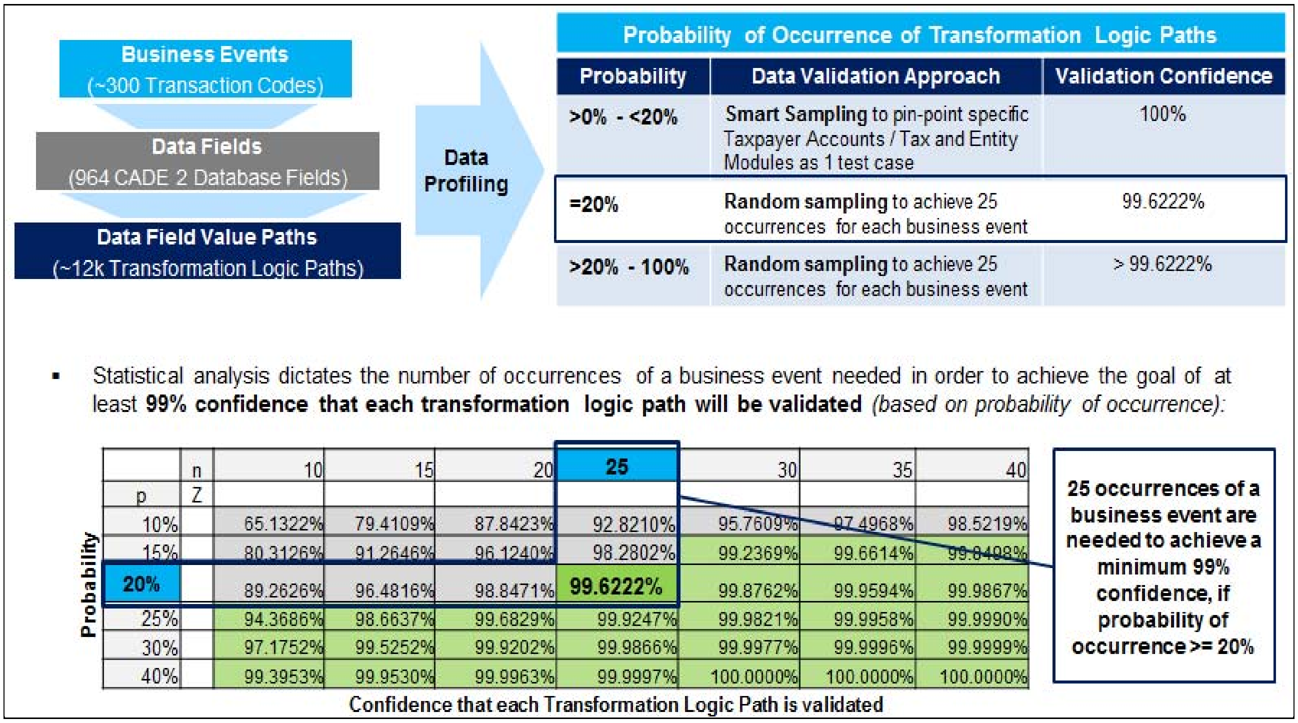
- Analyzes the transformation and data field coverage provided by Data Sampling and reports out results
- Validates the completeness of data profiling activities

Identify all data field transformation logic paths
Profile data to determine probability of occurrence for each logic path
Probability ≥20%
Determine Required Random Sample Size
Probability <20%
Identify Specific Accounts for Smart Sample
Perform Data Validation (i.e.,: Systemic Compare)
Report and Analyze Results

Source: CADE 2 Database Implementation Data Validation Plan, Version 2.0, p. 34, dated February 3, 2014. TIN – Taxpayer Identification Number. EST – Enterprise Systems Testing. SAT – Systems Acceptability Testing. FIT – Final Integration Testing. IMFOL – Individual Master File Online.

Figure 4 provides the data sampling methodology that applies a statistical approach to determine the validation confidence. It determines the probability of each transformation logic path occurring through Database Profiling. For example, a business event with at least a 20 percent probability of occurring must occur 25 times to achieve a confidence level of 99.6222 percent.

Figure 4: Data Sampling Methodology – Statistical Approach



Probability of Occurrence of Transformation Logic Paths

Probability	Data Validation Approach	Validation Confidence
>0% - <20%	**Smart Sampling** to pin-point specific Taxpayer Accounts / Tax and Entity Modules as 1 test case	100%
=20%	**Random sampling** to achieve 25 occurrences for each business event	99.6222%
>20% - 100%	**Random sampling** to achieve 25 occurrences for each business event	> 99.6222%

- Statistical analysis dictates the number of occurrences of a business event needed in order to achieve the goal of at least **99% confidence that each transformation logic path will be validated** *(based on probability of occurrence):*

	n	10	15	20	25	30	35	40
p	Z							
10%		65.1322%	79.4109%	87.8423%	92.8210%	95.7609%	97.4968%	98.5219%
15%		80.3126%	91.2646%	96.1240%	98.2802%	99.2369%	99.6614%	99.8408%
20%		89.2626%	96.4816%	98.8471%	99.6222%	99.8762%	99.9594%	99.9867%
25%		94.3686%	98.6637%	99.6829%	99.9247%	99.9821%	99.9958%	99.9990%
30%		97.1752%	99.5252%	99.9202%	99.9866%	99.9977%	99.9996%	99.9999%
40%		99.3953%	99.9530%	99.9963%	99.9997%	100.0000%	100.0000%	100.0000%

25 occurrences of a business event are needed to achieve a minimum 99% confidence, if probability of occurrence >= 20%

Source: *Data Integrity Validation Smart Sampling Deep Dive Draft, dated April 18, 2013.*

- *Taxpayer Identification Numbers/Module Generation includes identifying specific data (Taxpayer Accounts or Tax and Entity Modules) that can be tested by the data validation tools, which cover specific business conditions (that are unlikely to occur in a random sample of data)* – We met several times with the Smart Sampling subject matter expert to discuss how this activity and the data profiling activities were performed. We were provided a spreadsheet that contained information such as transaction codes and the profiling analysis used for identifying the data and business conditions that can be tested. However, neither the identification process nor an explanation of the spreadsheet data was documented. Thus, we were unable to evaluate the process. The CADE 2 PMO stated it had not yet documented the processes because executing data quality activities (*e.g.*, preparing random and Smart samples in time for data validation) had priority over the documentation.

- *Data Validation Execution includes testing the sampled Taxpayer Identification Numbers/Modules using the identified data validation methods* – The Data Validation Execution Plans and FLID Compare reports show that random and Smart samples were used in the data validation tests. Validation of completeness is reported on the Data Quality Scorecard under the Data Coverage Section. This section was first populated for

Production cycles 5 and 6, which reported on only the percentage of transformation logic paths covered. The methodology for validating completeness had not been documented.

On March 11, 2014, a Fast Smart sampling process was tested in cycle 5. It reuses the regular Smart sampling process but can be applied to production on a weekly basis, while the regular Smart sampling process requires at least four weeks. The results indicate that the Fast Smart sampling method added five times more coverage than the regular random sampling method and helped to identify new defects. As a result, it was officially implemented for cycles 9 and 10, in addition to using random sampling. We received two results spreadsheets that summarized the results used to conclude that Fast Smart sampling provided more coverage with fewer cases. We received seven of the eight source documents to support the summary spreadsheets; therefore, we were unable to completely confirm the numbers.

- *Reporting and Analysis* – The following activities are associated with this step:

  a. *Analyzes the transformation and data field coverage provided by data sampling and reports out results.* Transformation Logic Paths coverage and data field coverage were included on the Data Quality Scorecard beginning with cycles 5/6 and 9/10, respectively.

  b. *Validates the completeness of data profiling activities.* We have not seen any documentation on the status of this activity.

Our statistician determined that the concept and process of using the data sampling methodology to ensure that infrequently used data fields will be included in data validation testing and to provide a statistical basis for deciding how many instances of a particular data field or business event are to be sampled, based on the probability of occurrence and target confidence level, is sound. While the process used to implement the data sampling methodology was verbally described by IRS personnel in meetings, these processes had not been documented and were not available for review.

In addition, the process for measuring the effectiveness and success of the data sampling methodology in providing the expected coverage had not been documented. For example, the process for determining the percentage of transformation logic paths covered was not documented. This information is needed to ensure that the percentage of transformation logic paths, FLIDs, and data fields covered are accurately identified for the Data Quality Scorecard.

Due to the significant time pressure and limited resources faced by the CADE 2 PMO to ensure that the CADE 2 data validation activities stay on course, conducting the data sampling activities had priority over fully documenting the processes for profiling the data and evaluating the effectiveness of the data sampling methodology. In addition, the CADE 2 PMO explained that although the methodology has been implemented, they are still in the process of refining it.

Until data validation processes are formally documented, IRS management cannot have full confidence that the correct data validation procedures are performed consistently. This may also reduce the assurance that CADE 2 data are effectively and completely tested. These processes should be documented as soon as possible to avoid the risk of losing the knowledge that only the subject matter experts have and to provide a reference for current and future use.

Management Action: After discussing the need to document the data sampling methodology with CADE 2 PMO management, they recognized the urgency of the need and began development of the documentation. Several in-progress documents were provided for our review, including the Defect Verification Process used by Smart Sampling and the CADE 2 Data Validation Smart Sample Process Overview documents.

Recommendation

Recommendation 3: The Chief Technology Officer should ensure that all data sampling methodology processes such as data profiling and calculating data field and transformation logic coverage are completely documented and that the documents are readily available for review. Where applicable, the documentation should include procedures to collect and maintain source data used to support data validation metrics.

> ***Management's Response:*** The IRS agreed with this recommendation. The IRS is developing documentation on the procedures to collect and maintain source data used to support data validation metrics.

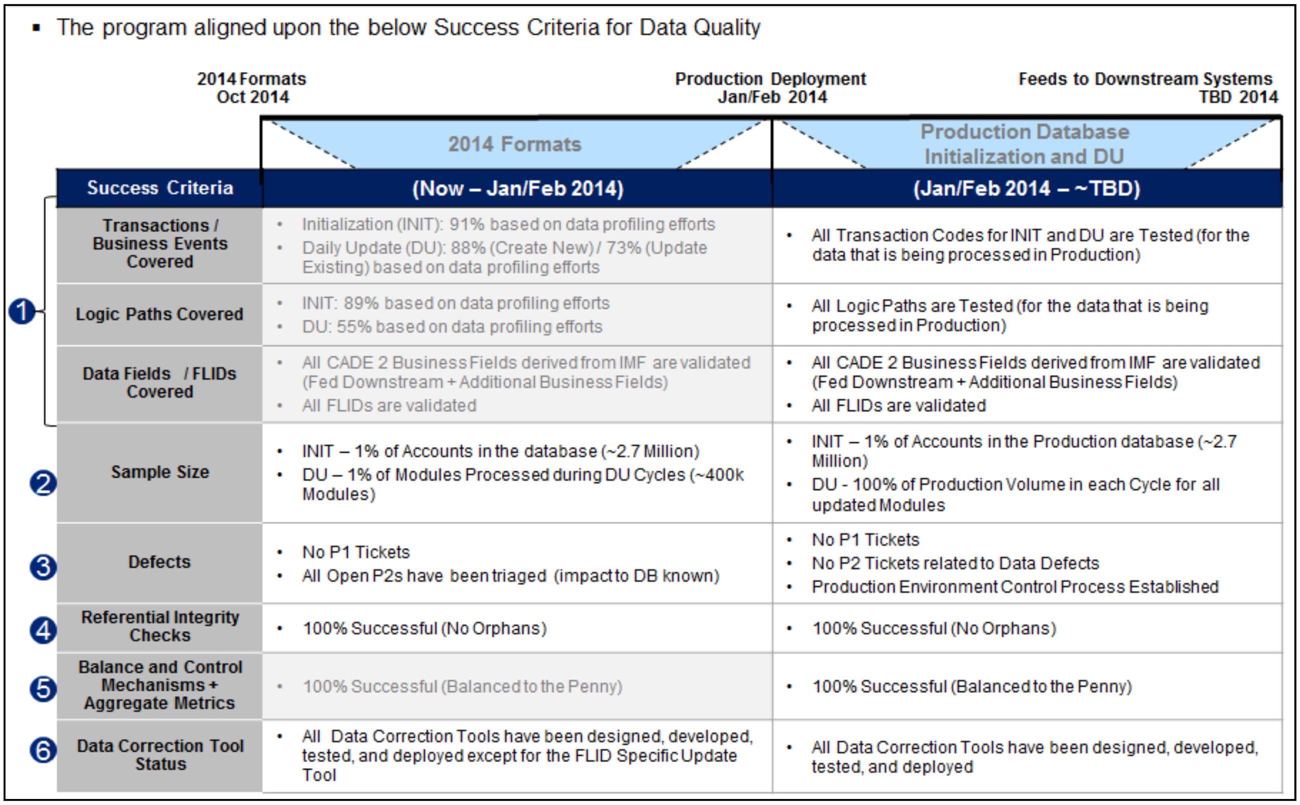
The Documentation and Processes for Determining the Data Quality Scorecard Metrics Need Improvement

The Government Accountability Office's *Standards for Internal Control in the Federal Government* state that control activities include verifications and accurate and timely recording of transactions and events. Transactions should be promptly recorded to maintain their relevance and value to management in controlling operations and making decisions.

According to the Data Quality Team Charter v 0.3 dated July 26, 2013, the team's mission is to ensure the quality and integrity of the data within the CADE 2 database and the data fed to downstream systems by providing execution support for defect management activities and establishing a comprehensive Data Quality Scorecard to measure the progress towards data quality goals.

The Data Quality Team developed a Data Quality Scorecard that includes six key performance areas with success criteria: 1) Data Coverage; 2) Sample Size; 3) Data Validation Defect Summary; 4) Referential Integrity Checks; 5) Balance and Control Mechanisms Plus Aggregate Metrics; and 6) Data Correction Tool Status. Figure 5 provides the defined key performance indicators (KPI) and success criteria for each area. The KPIs that are grayed were not included in the initial Data Quality Scorecard because the information was not available.

Figure 5: Key Performance Indicators and Success Criteria



- The program aligned upon the below Success Criteria for Data Quality

2014 Formats
Oct 2014

Production Deployment
Jan/Feb 2014

Feeds to Downstream Systems
TBD 2014

		2014 Formats	Production Database Initialization and DU
	Success Criteria	(Now – Jan/Feb 2014)	(Jan/Feb 2014 – ~TBD)
1	Transactions / Business Events Covered	• Initialization (INIT): 91% based on data profiling efforts • Daily Update (DU): 88% (Create New) / 73% (Update Existing) based on data profiling efforts	• All Transaction Codes for INIT and DU are Tested (for the data that is being processed in Production)
	Logic Paths Covered	• INIT: 89% based on data profiling efforts • DU: 55% based on data profiling efforts	• All Logic Paths are Tested (for the data that is being processed in Production)
	Data Fields / FLIDs Covered	• All CADE 2 Business Fields derived from IMF are validated (Fed Downstream + Additional Business Fields) • All FLIDs are validated	• All CADE 2 Business Fields derived from IMF are validated (Fed Downstream + Additional Business Fields) • All FLIDs are validated
2	Sample Size	• INIT – 1% of Accounts in the database (~2.7 Million) • DU – 1% of Modules Processed during DU Cycles (~400k Modules)	• INIT – 1% of Accounts in the Production database (~2.7 Million) • DU - 100% of Production Volume in each Cycle for all updated Modules
3	Defects	• No P1 Tickets • All Open P2s have been triaged (impact to DB known)	• No P1 Tickets • No P2 Tickets related to Data Defects • Production Environment Control Process Established
4	Referential Integrity Checks	• 100% Successful (No Orphans)	• 100% Successful (No Orphans)
5	Balance and Control Mechanisms + Aggregate Metrics	• 100% Successful (Balanced to the Penny)	• 100% Successful (Balanced to the Penny)
6	Data Correction Tool Status	• All Data Correction Tools have been designed, developed, tested, and deployed except for the FLID Specific Update Tool	• All Data Correction Tools have been designed, developed, tested, and deployed

Source: CADE 2 Data Quality Scorecard for the 2014 Version of the Data as of December 16, 2013. TBD – To Be Determined. P1, P2 – Priority 1 or 2.

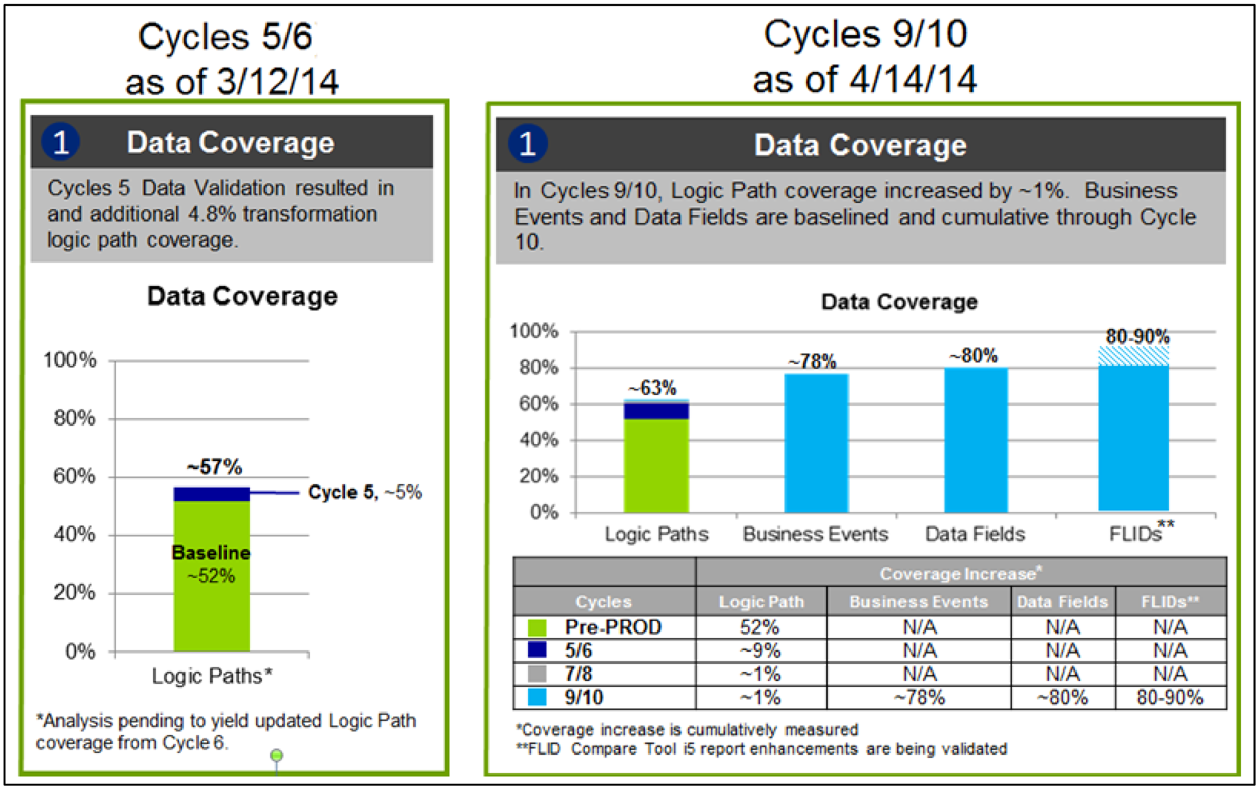
The first published Data Quality Scorecard, dated December 16, 2013, reported on pre-production data and was distributed to stakeholders on December 20, 2013. The Scorecard is presented in Appendix IV, Figure 1. The IRS initially planned to prepare a Scorecard every two weeks for distribution to stakeholders. On March 21, 2014, we received information that a Scorecard will be produced for each data validation cycle.

We attempted to fully assess the accuracy of the entire Data Quality Scorecard for a specific cycle. However, due to the lack of supporting source documentation we were unable to complete the assessment. Alternatively, we validated the individual sections of the Scorecard when sufficient source information was made available.

<u>The results of our review follow</u>:

Section 1 – Data Coverage: This section includes the Transactions/Business Events, the Logic Paths, and the Data Fields and FLIDs covered. The IRS relies on summary spreadsheets to report the data validation results for the first three KPIs. The Data Quality Scorecard for cycles 9/10 as of April 14, 2014, reported metrics for Logic Paths, Business Events, Data Fields, and FLIDs. We received summary spreadsheets for the first three metrics. We also received source documentation supporting the summary spreadsheet for the Logic Paths KPI but not for the Business Events and the Data Fields KPIs. Although the Scorecard reported 80 – 90 percent coverage of the FLIDs, we did not receive any documentation to support that metric. Figure 6 displays the Data Coverage portion of the Data Quality Scorecards.

Figure 6: Data Coverage



Cycles 5/6
as of 3/12/14

1 Data Coverage

Cycles 5 Data Validation resulted in and additional 4.8% transformation logic path coverage.

*Analysis pending to yield updated Logic Path coverage from Cycle 6.

Cycles 9/10
as of 4/14/14

1 Data Coverage

In Cycles 9/10, Logic Path coverage increased by ~1%. Business Events and Data Fields are baselined and cumulative through Cycle 10.

	Coverage Increase*			
Cycles	Logic Path	Business Events	Data Fields	FLIDs**
🟩 Pre-PROD	52%	N/A	N/A	N/A
🟦 5/6	~9%	N/A	N/A	N/A
⬜ 7/8	~1%	N/A	N/A	N/A
🟦 9/10	~1%	~78%	~80%	80-90%

*Coverage increase is cumulatively measured
**FLID Compare Tool i5 report enhancements are being validated

Source: Excerpts of the Data Quality Scorecards provided by the CADE 2 PMO. Pre-PROD – Pre-production. i5 – Iteration 5.

Section 2 – Sample Size: This section includes the targeted number of Taxpayer Identification Numbers and/or modules expected to be compared and the actual number of Taxpayer Identification Numbers and/or modules compared for data validations performed prior to production cycles 5/6. Beginning with production cycles 5/6, the objective was to compare and report on modules. The source for the number of actual modules compared during production should have been documented in an FLID report. Until the end of April 2014, the number of actual modules compared was incorrectly reported because the IRS did not base the numbers on the FLID report. Instead, they used the targeted volumes for the random and Smart samples as the basis for reporting the actual modules compared. The IRS was not referring to the FLID reports for the actual number of modules compared because the FLID reports did not clearly indicate the actual number of modules compared. In addition, the process for determining the actual number was not documented. Also, the Wage and Investment Division Business Modernization Office (hereafter referred to as Business Modernization Office) stated that they were in the process of learning how to read the FLID reports and verify the contents. As a result, the incorrect numbers were included in presentations submitted to CADE 2 executives and the Chief Technology Officer for their discussions.

Figure 7 shows the incorrect and correct number of modules actually compared. The Business Modernization Office personnel stated that after learning more about the data captured in the FLID report (how to read them and verify the contents), they updated the Scorecards from cycles 5/6 through the present accordingly to accurately reflect the actual number of modules compared. Prior to that, the numbers were based on the targeted volumes for the random and Smart samples. It appears that the IRS learned of the need to make the corrections after our repeated requests for documented source information.

Figure 7 – The Incorrect and Correct Number of Actual Modules Compared As Reported on Various Iterations of the Data Quality Scorecard

	Actual Modules Compared	
Cycles	**Incorrect Number**	**Correct Number**
5/6	500,000/500,000 577,794 / 576,618	590,229/588,630
7/8	591,302/589,264	591,652/590,308
9/10	623,372/500,000	500,042/611,374

Source: Data Quality Scorecards provided by the CADE 2 PMO.

Section 3 – Data Validation Defect Summary: This section reports the number of new data defect tickets open and, of those, the number that remain open for that cycle as of the Scorecard date. It does not report the cumulative number of open unresolved tickets from other cycles as of that date. For example, the Data Quality Scorecard for cycles 15/16 as of May 12, 2014, reported that all of the new tickets opened during that time were closed because they were later determined not to be data issues. Because the Scorecard showed no open tickets, it might appear that all of the data are correct. However, this is not the case because the Scorecard does not carry over the unresolved data defect tickets from prior cycles that remain in open status. For this reporting period, another management report shows seven open data defect tickets. All were estimated to be resolved and closed by May 28, 2014. IRS management indicated that initial

Scorecards did not report cumulative open unresolved data defect tickets because each Scorecard covered only a two week period. As of May 19, 2014, the IRS began producing an Aggregate Scorecard that includes all open data defect tickets.

Also, as of April 3, 2014, there are 10 open known data defects on the Known Defect List. These are data defects that have occurred on more than one occasion and need to be corrected. These, along with the new data defects that are identified during the data validation process, must be corrected before the CADE 2 database can replace the current IMF account settlement system with a relational database processing system and become a key component in the IRS's enterprise-wide, data-centric information technology strategy.

Although the information is available, the Data Quality Scorecard does not show the impact of the data defects. For example, the Scorecard does not show the number of tax and/or entity modules or taxpayers affected. When resources are limited, knowing the impact of the data defects could help prioritize the order in which data defects are resolved.

We also found a discrepancy between the Data Quality Scorecard for cycles 15/16 dated May 12, 2014, and the CADE 2 Data Implementation Health Report dated May 19, 2014 (hereafter referred to as the Health Report). The Scorecard, which was also embedded in the Health Report, reported "Eight new data defect tickets were initially opened for cycle 15/16 production Data Validation, but after further analysis, these tickets were determined to not be data issues and were closed." The Health Report reported that eight data defect tickets opened as a result of cycles 15 and 16 data validation; however, seven of them were deemed to be "no trouble found." The remaining ticket was scheduled to be closed upon the delivery of FLID Compare Tool Iteration 6 in early June 2014.

The Data Quality Scorecard for Production Cycles 5/6 dated March 12, 2014, correctly reported that 12 new data defects were open and one of the 12 was subsequently closed. However, we found two discrepancies in this section. The first is in the bar graph, which shows eight open tickets for cycle 5 and three for cycle 6. The spreadsheet with the source information shows seven open tickets for cycle 5 and four for cycle 6.

The second discrepancy is with the percentages in the pie chart. The chart shows that 41 percent and 17 percent of the Defect Origin/Source were from Solutions Engineering–Data Engineering and Identify and Extract Account Changes, respectively. However, based on the source spreadsheet, Solutions Engineering–Data Engineering had six (50 percent) of the 12 and Identify and Extract Account Changes had one (8 percent) of the 12.

Section 4 – Referential Integrity Checks: Referential Integrity Checks are run against the database to ensure that tax account information that is spread over many tables can be reassembled into a coherent tax account (*i.e.*, prevent orphan data in the database). Identified issues should be resolved according to standard operating procedures. As of April 24, 2014, all Data Quality Scorecards reported that all checks for the cycles passed. We obtained and reviewed 14 source reports for cycles 201250 through 201310 but none corresponded to the Data

Quality Scorecards we received. Therefore, we were unable to confirm that all Referential Integrity checks passed.

Section 5 – Balance and Control Mechanisms + Aggregated Metrics: This section reports results from two sources:

- Simplified Financial Balance Reports: These are financial integrity checks to ensure that amounts from the IMF equal the CADE 2 database amounts. Chief Financial Officer requirements include balancing the sum of certain financial fields. Specialized financial reports are generated and provided to the Chief Financial Officer for manual comparison and verification. For the Data Quality Scorecard for Cycles 5/6 as of March 12, 2014, we received and compared the nine IMF reports to the nine CADE 2 database reports and found that all nine balanced to the penny.

- CADE 2/IMF Analytical Report Business Objects Enterprise Comparisons: This activity validates that CADE 2 data match IMF data by comparing data from certain IMF and CADE 2 database reports. As planned, these metrics were first reported on the Data Quality Scorecard for cycles 9/10. The April 14, 2014, version shows that the data fields in nine of the 10 reports matched. The remaining report has an 87 percent match rate, but the CADE 2 PMO is expecting results from another test report. We received a summary report that supported all the data in the Business Objects Enterprise Report Execution Analysis section except for the data in the CADE Fields Used column. However, we did not receive documents supporting the statistics in the summary report.

Section 6 – Data Correction Tools: We received documentation which confirms that six of the seven tools were implemented into production. Therefore, this section correctly reported the status of the tools.

Because the IT organization and the Business Modernization Office worked together to develop the Scorecard and the KPIs, the Scorecard should meet the stakeholders' needs. In addition, the processes used to ascertain the actual statistical data need to be documented to ensure that they are correctly and accurately determined. This will help stakeholders fully understand what the statistics represent if they request an explanation for the basis of the statistics. When processes are not sufficiently documented, there is a risk that they are not correctly performed. For example, because the FLID report does not clearly state the total number of actual modules compared and there were no documented instructions for identifying this, the number of actual modules compared were incorrectly determined and incorrectly reported on the Scorecards through April 2014 and incorrectly reported in presentations to management.

Management Action: After meeting with the CADE 2 PMO regarding the lack of sufficient supporting documentation needed to validate the metrics on the Data Quality Scorecard, it began collecting and providing us with the documentation. For example and as stated above, we received source documentation that confirmed the logic path KPI metric.

Recommendations

The Chief Technology Officer should:

Recommendation 4: Ensure that all processes for determining the metrics needed to populate the Data Quality Scorecard are completely documented and that the documents are readily available for review.

> **Management's Response:** The IRS agreed with this recommendation. The IRS has developed and will be publishing documentation of the Scorecard development process. The IRS will continue to update, maintain, and develop documentation around the Data Quality Scorecard to ensure that its inputs and processes are transparent to CADE 2 stakeholders.

Recommendation 5: Ensure that all documentation needed to verify the data in the Data Quality Scorecard is stored for future reference and to provide the information needed for oversight activities, such as spot checks to confirm the accuracy of the Scorecard.

> **Management's Response:** The IRS agreed with this recommendation. The IRS has documented procedures for developing the Scorecard, a checklist to verify the contents, and has begun storing all Scorecard sources in a SharePoint repository. The IRS will ensure that the repository remains organized and easily accessible.

The Field Identifier Compare Tool Validates Data for Downstream Systems, but Data Discrepancy Reports Need Improvement

The IRS data strategy requires that data fields be uniquely and consistently identified across systems. The validation of data on the CADE 2 database is critical to the database becoming a trusted source of data for downstream systems and ultimately the file of record for IMF data.

The FLID Compare Tool was developed as an automated way to compare high volumes of IMF data to CADE 2 data during the data validation process. It was the main tool used for automated data validation during the 2014 Filing Season. The tool leverages the existing IRS process of using field identifiers (*i.e.*, FLIDs) to help identify IMF data. Currently, Corporate Files Online processing builds FLIDs for IMF data from the IMF Virtual Storage Access Method files. The new CADE 2 Data Access Service builds these same FLIDs for data from the CADE 2 database. The FLID Compare Tool compares FLIDs from both sources and identifies any discrepancies in their data values.

Current IMF processing sends IMF data to downstream systems in files using FLIDs. By comparing FLIDs built from IMF Virtual Storage Access Method files to FLIDs built from the CADE 2 database, the FLID Compare Tool can cover all the data consumed by downstream systems. Therefore, 911 (89 percent) of the 1,018 verifiable data fields on the CADE 2 database can be identified through the use of FLIDs; the remaining 107 (11 percent) of the data populated into the CADE 2 database from the IMF are not related to an FLID number. Other validation methods are used to ensure coverage of the data fields not covered by the FLID Compare Tool. (This information is summarized in Figure 1 of this report.)

The FLID Compare Tool produces several reports on the results of its comparisons. One of them, the Discrepancy Detail Report, lists all data discrepancies by FLID number, FLID name, IMF data field name, and CADE 2 database table and column. The Business Modernization Office used this report to review and analyze details on data discrepancies found during the data validation process.

The FLID Coverage Count Report, added for the 2014 Filing Season, provides metrics on FLID coverage during execution of the FLID Compare Tool. It provides a complete list of all unique FLID numbers, whether or not the FLID was compared, and the match/no-match count for each compared FLID.

The Enterprise Data Management Office (EDMO) maintains the list of FLIDs. We compared the EDMO FLID list to the one in the FLID Coverage Count Report and found discrepancies.

- 10 FLID numbers on the EDMO list were missing from the FLID Coverage Count Report.
- 23 FLID numbers on the FLID Coverage Count Report did not have FLID names.
- 36 FLID numbers in the FLID Coverage Count Report were listed as “reserved,” compared to 37 in the EDMO list.

These discrepancies raise questions as to whether the FLID Compare Tool is accurately comparing all data at the FLID level.

After we alerted the IRS to the 10 missing FLID numbers, the IRS researched the issue and found that the missing FLIDs should have been included in the FLID Coverage Count Report and compared by the FLID Compare Tool. The IRS plans to add the missing FLIDs to the next iteration of the FLID Compare Tool scheduled for implementation in the summer of 2014. In the interim, the IRS is using another automated tool to review the 10 missing FLIDs.

While FLID numbers are currently used by the IMF to pass data to downstream systems, FLID numbers do not uniquely identify data on the IMF. They are used in conjunction with their position on the IMF data record to provide uniqueness. There are 805 FLID numbers[5] and 911 FLID data fields on the CADE 2 database. This indicates that some FLID numbers are used more than once for data field coverage. For example, the last name in the IMF data field Taxpayer Nameline is represented by FLID 0733. However, FLID 0733 is mapped to three separate data fields on the CADE 2 database. Specifically:

[5] FLID number sequence count (842) minus reserved FLID numbers (37) = 805 FLID numbers used in 2014.

- Taxpayer_Nameline.Joint_Last_Nm.
- Taxpayer_Nameline.Primary_Last_Nm.
- Taxpayer_Nameline.Secondary_Last_Nm.

The FLID Coverage Count Report counts by unique FLID number only; it does not trace back to unique data fields on the database. Without this traceability, it is impossible to verify that all database fields are validated by the FLID Compare Tool without additional analysis. After we raised this issue to the IRS, the IRS responded that it will explore ways to address the one-to-many relationship of FLIDs to data fields in future iterations of the FLID Compare Tool.

The FLID Compare Tool is used to gather metrics for data validation reporting. The Extended Discrepancy Counts Report is used to provide sample size counts for the Data Quality Scorecard; however, the report takes counts by program name, and documentation does not indicate how these program names translate to sample size counts. Therefore, data in this report may be misinterpreted and lead to incorrect information reported to management. In addition, if the FLID list used in the FLID Compare Tool does not match the FLID list maintained by the EDMO, the IRS cannot be assured that it is accurately and completely validating all FLIDs that are intended to be fed to downstream systems. Finally, if the FLID Compare Tool cannot trace back to the 911 data fields on the CADE 2 database that it is tasked with validating, the IRS cannot guarantee the accuracy or the completeness of those fields.

Recommendations

The Chief Technology Officer should:

Recommendation 6: Ensure that automated data compare tools identify and report on data fields, not FLID numbers, to align CADE 2 data validation efforts with the IRS's data strategy goal of uniquely identifying data fields across systems.

> **Management's Response:** The IRS disagreed with this recommendation. Data defects are identified at the FLID level; the output from the FLID Compare Tool provides counts by FLID number. Traceability to unique data fields is established through the use of transformation rules analyzed during the defect triage process. This provides the acceptable level of traceability to unique data fields. The IRS's data strategy goal for uniquely identifying data fields across systems is considered a guiding principal; however, programs are given discretion for when identifying at the data field level is necessary.
>
> **Office of Audit Comment:** TIGTA maintains its position that CADE 2 data validation efforts should identify and report on individual data fields. The IRS Data Strategy and Roadmap (dated August 27, 2012) stresses that information should be consistently represented across systems, available at the same level of granularity, and have summary levels so that meaningful comparisons can be made. The Data Strategy does not mention

that discretion is given to programs to determine when this principle would or would not apply.

Recommendation 7: Ensure that automated data compare tool reports clearly identify counters and align with data validation metrics.

Management's Response: The IRS agreed with this recommendation. The High Volume FLID Compare Tool Design Document will be updated to explain the source of the numbers that are populated for those program names in Report 4, which will provide the actual input record count. This will allow for accurate reporting of actual sample size on the Scorecard.

Appendix I

Detailed Objective, Scope, and Methodology

Our overall audit objective was to evaluate IRS efforts to ensure that the data in the CADE 2 database[1] are accurate and complete. To accomplish our objective, we:

I. Assessed the effectiveness of the CADE 2 Data Validation methodology.

   A. Reviewed the CADE 2 Database Implementation Data Validation Plan.

   B. Evaluated the data sampling methodology.

II. Evaluated the implementation and effectiveness of automated compare tools in the CADE 2 data validation process.

   A. Reviewed documentation to determine if formal planning and resource coordination occurred for the implementation of the automated compare tools in the CADE 2 data validation process.

   B. Interviewed subject matter experts to determine how each automated compare tool is used in the data validation process.

   C. Reviewed testing results generated from each tool to determine the effectiveness of the tool in the data validation process.

III. Evaluated the effectiveness of the CADE 2 Data Quality Team.

   A. Reviewed the CADE 2 Data Quality Team Charter.

   B. Determined what metrics (if any) currently exist for CADE 2 data validation activities and how these metrics are being used to measure data quality.

   C. Evaluated KPIs developed by the team to ensure that they adequately measure CADE 2 data quality.

   D. Evaluated the monitoring and reporting processes for KPIs.

   E. Evaluated the effectiveness of the data defect management process.

[1] See Appendix V for a glossary of terms.

Internal controls methodology

Internal controls relate to management's plans, methods, and procedures used to meet their mission, goals, and objectives. Internal controls include the processes and procedures for planning, organizing, directing, and controlling program operations. They include the systems for measuring, reporting, and monitoring program performance. We determined that the following internal controls were relevant to our audit objective: the Government Accountability Office's *Standards for Internal Control in the Federal Government*,[2] the CADE 2 Database Implementation Data Validation Plan, various meetings such as the CADE 2 Weekly Executive Status Meetings and periodic data validation execution checkpoint meetings, design documents, and data validation policies and procedures. We evaluated these controls by conducting interviews with IRS management and staff; attending CADE 2 meetings; and reviewing and evaluating documents such as the CADE 2 Data Quality Team Charter, the CADE 2 Database Implementation Data Validation Plan and Data Validation Execution Plans, the FLID Compare Tool design documents, and related FLID reports.

[2] Government Accountability Office (formerly known as the General Accounting Office), GAO/AIMD-00-21.3.1, *Internal Control: Standards for Internal Control in the Federal Government* (Nov. 1999).

Appendix II

Major Contributors to This Report

Alan R. Duncan, Assistant Inspector General for Audit (Security and Information Technology Services)
Danny Verneuille, Director
Myron Gulley, Audit Manager
Tina Wong, Lead Auditor
Richard Borst, Senior Auditor
Arlene Feskanich, Information Technology Specialist
Erika D. Axelson, Ph.D., Statistician

Appendix III

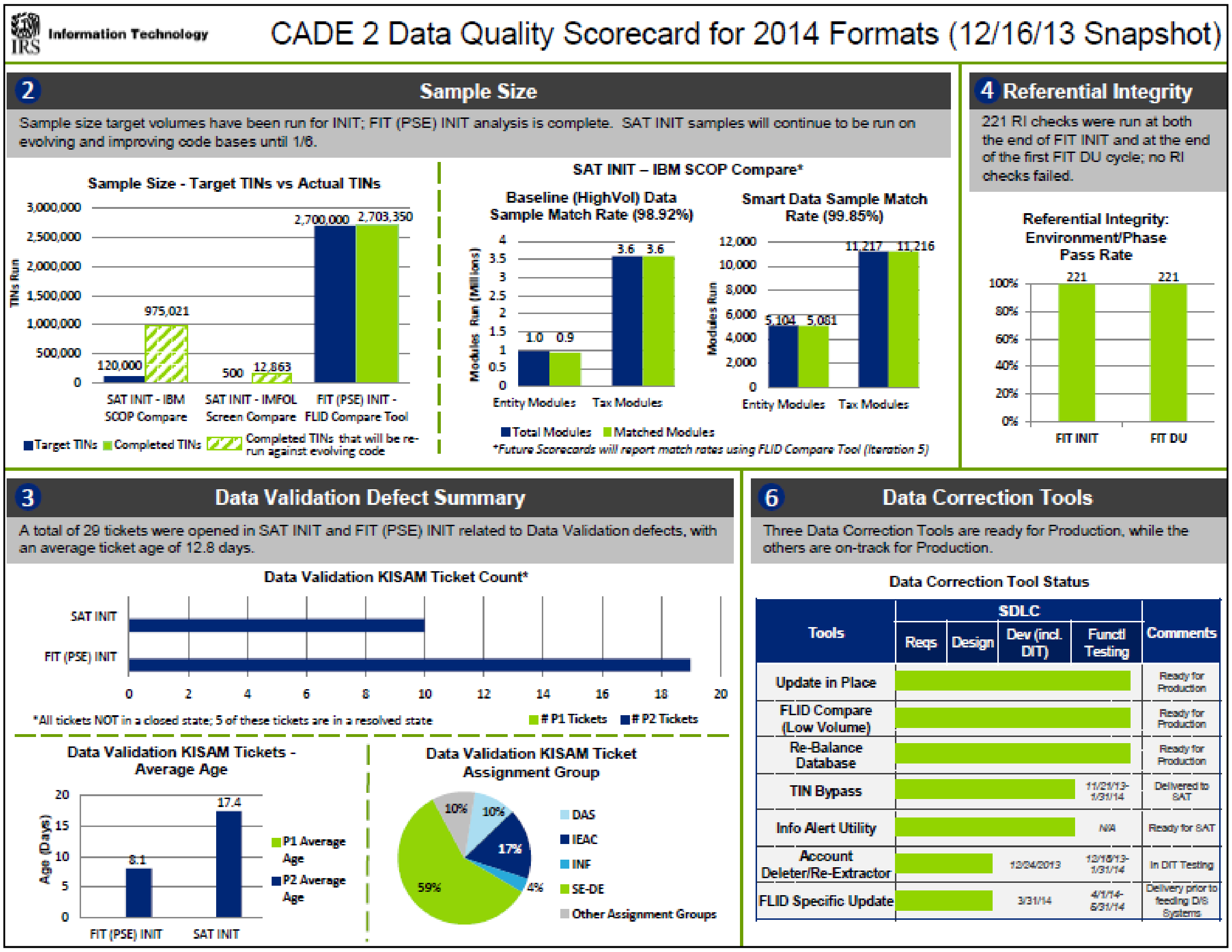
Report Distribution List

Commissioner C
Office of the Commissioner – Attn: Chief of Staff C
Deputy Commissioner for Operations Support OS
Deputy Commissioner for Services and Enforcement SE
Commissioner, Wage and Investment Division SE:W
Deputy Chief Information Officer for Operations OS:CTO
Associate Chief Information Officer, Applications Development OS:CTO:AD
Associate Chief Information Officer, Enterprise Information Technology – Program Management Office OS:CTO:EIT
Director, Enterprise Systems Testing OS:CTO:AD:EST
Chief Counsel CC
National Taxpayer Advocate TA
Director, Office of Legislative Affairs CL:LA
Director, Office of Program Evaluation and Risk Analysis RAS:O
Office of Internal Control OS:CFO:CPIC:IC
Audit Liaison: Director, Risk Management Division OS:CTO:SP:RM

Appendix IV

Data Quality Scorecards

**Figure 1: First Published Data Quality Scorecard
Snapshot as of December 16, 2013**



IRS Information Technology

CADE 2 Data Quality Scorecard for 2014 Formats (12/16/13 Snapshot)

2 Sample Size

Sample size target volumes have been run for INIT; FIT (PSE) INIT analysis is complete. SAT INIT samples will continue to be run on evolving and improving code bases until 1/6.

*Future Scorecards will report match rates using FLID Compare Tool (Iteration 5)

4 Referential Integrity

221 RI checks were run at both the end of FIT INIT and at the end of the first FIT DU cycle; no RI checks failed.

3 Data Validation Defect Summary

A total of 29 tickets were opened in SAT INIT and FIT (PSE) INIT related to Data Validation defects, with an average ticket age of 12.8 days.

*All tickets NOT in a closed state; 5 of these tickets are in a resolved state

6 Data Correction Tools

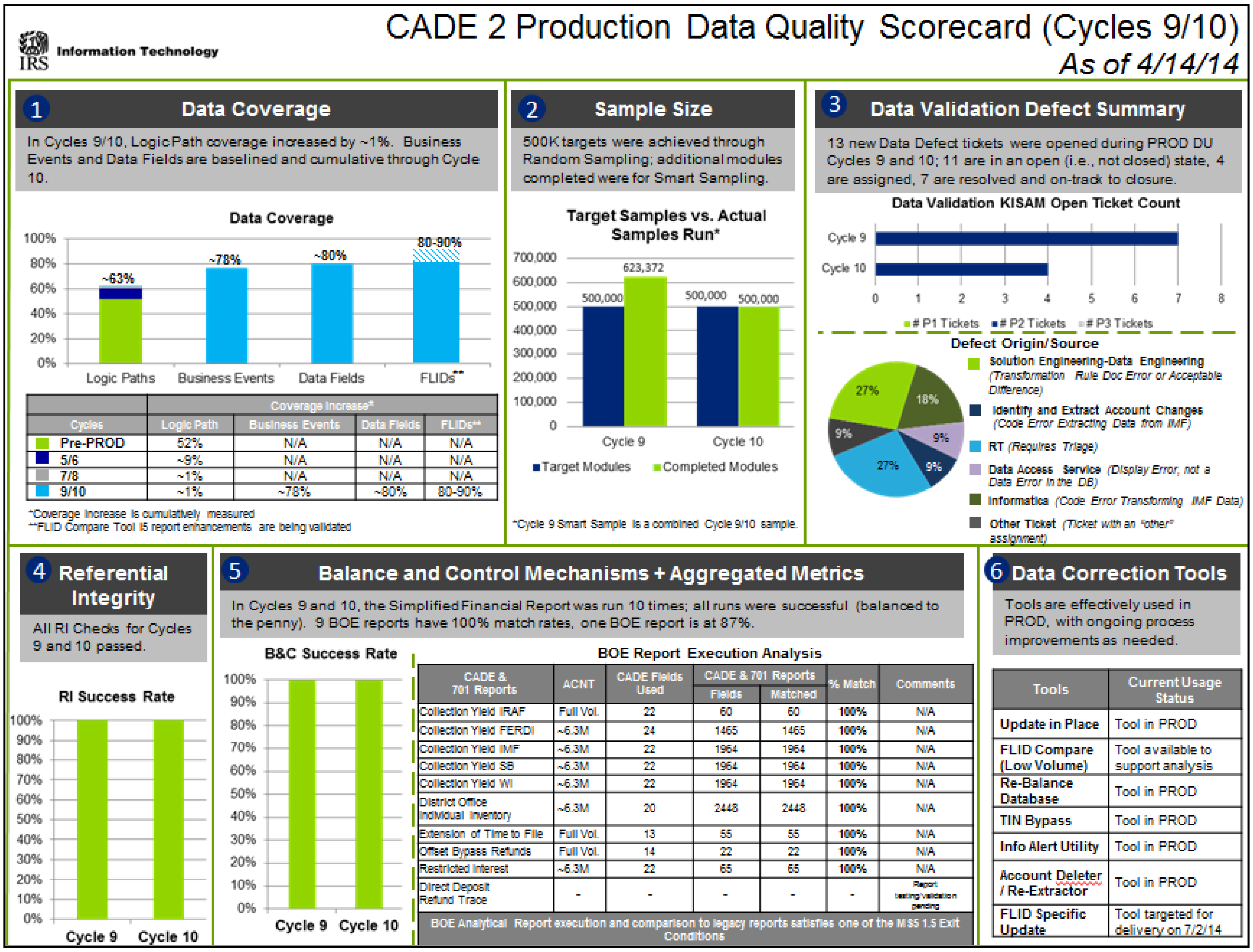
Three Data Correction Tools are ready for Production, while the others are on-track for Production.

Data Correction Tool Status

Tools	SDLC				Comments
	Reqs	Design	Dev (incl. DIT)	Functl Testing	
Update in Place	🟩	🟩	🟩	🟩	Ready for Production
FLID Compare (Low Volume)	🟩	🟩	🟩	🟩	Ready for Production
Re-Balance Database	🟩	🟩	🟩	🟩	Ready for Production
TIN Bypass	🟩	🟩	🟩	11/21/13-1/31/14	Delivered to SAT
Info Alert Utility	🟩	🟩	🟩	N/A	Ready for SAT
Account Deleter/Re-Extractor	🟩	🟩	12/24/2013	12/16/13-1/31/14	In DIT Testing
FLID Specific Update	🟩	🟩	3/31/14	4/1/14-5/31/14	Delivery prior to feeding D/S Systems

Source: CADE 2 PMO. INIT – Initialization. FIT – Final Integration Testing. PSE – Production Support Environment. SAT – Systems Acceptability Testing. RI – Referential Integrity. DU – Daily Update. TIN – Taxpayer Identification Number. IBM – International Business Machines. IMFOL – Individual Master File Online. SCOP – Standard Corporate Files On Line Overnight Processing. Vol – Volume. P1, P2, P3 – Priority 1, 2, or 3. DAS – Data Access Service. IEAC – Identify and Extract Account Changes. INF – Informatica. SE-DE – Solutions Engineering – Date Engineering. SDLC – Systems Development Life Cycle. Reqs – Requirements. Dev – Development. DIT – Development, Integration, and Testing. Functl – Functional.

Figure 2: Data Quality Scorecard for Production Cycle 9/10 as of April 14, 2014



CADE 2 Production Data Quality Scorecard (Cycles 9/10)

As of 4/14/14

1 Data Coverage

In Cycles 9/10, Logic Path coverage increased by ~1%. Business Events and Data Fields are baselined and cumulative through Cycle 10.

	Coverage Increase*			
Cycles	Logic Path	Business Events	Data Fields	FLIDs**
🟩 Pre-PROD	52%	N/A	N/A	N/A
■ 5/6	~9%	N/A	N/A	N/A
■ 7/8	~1%	N/A	N/A	N/A
■ 9/10	~1%	~78%	~80%	80-90%

*Coverage Increase is cumulatively measured

**FLID Compare Tool I5 report enhancements are being validated

2 Sample Size

500K targets were achieved through Random Sampling; additional modules completed were for Smart Sampling.

*Cycle 9 Smart Sample is a combined Cycle 9/10 sample.

3 Data Validation Defect Summary

13 new Data Defect tickets were opened during PROD DU Cycles 9 and 10; 11 are in an open (i.e., not closed) state, 4 are assigned, 7 are resolved and on-track to closure.

- 🟩 Solution Engineering-Data Engineering *(Transformation Rule Doc Error or Acceptable Difference)*
- ■ Identify and Extract Account Changes *(Code Error Extracting Data from IMF)*
- ■ RT *(Requires Triage)*
- ■ Data Access Service *(Display Error, not a Data Error in the DB)*
- ■ Informatica *(Code Error Transforming IMF Data)*
- ■ Other Ticket *(Ticket with an "other" assignment)*

4 Referential Integrity

All RI Checks for Cycles 9 and 10 passed.

5 Balance and Control Mechanisms + Aggregated Metrics

In Cycles 9 and 10, the Simplified Financial Report was run 10 times; all runs were successful (balanced to the penny). 9 BOE reports have 100% match rates, one BOE report is at 87%.

BOE Report Execution Analysis

CADE & 701 Reports	ACNT	CADE Fields Used	CADE & 701 Reports		% Match	Comments
			Fields	Matched		
Collection Yield IRAF	Full Vol.	22	60	60	100%	N/A
Collection Yield FERDI	~6.3M	24	1465	1465	100%	N/A
Collection Yield IMF	~6.3M	22	1964	1964	100%	N/A
Collection Yield SB	~6.3M	22	1964	1964	100%	N/A
Collection Yield WI	~6.3M	22	1964	1964	100%	N/A
District Office Individual Inventory	~6.3M	20	2448	2448	100%	N/A
Extension of Time to File	Full Vol.	13	55	55	100%	N/A
Offset Bypass Refunds	Full Vol.	14	22	22	100%	N/A
Restricted Interest	~6.3M	22	65	65	100%	N/A
Direct Deposit Refund Trace	-	-	-	-	-	Report testing/validation pending
BOE Analytical Report execution and comparison to legacy reports satisfies one of the MS5 1.5 Exit Conditions						

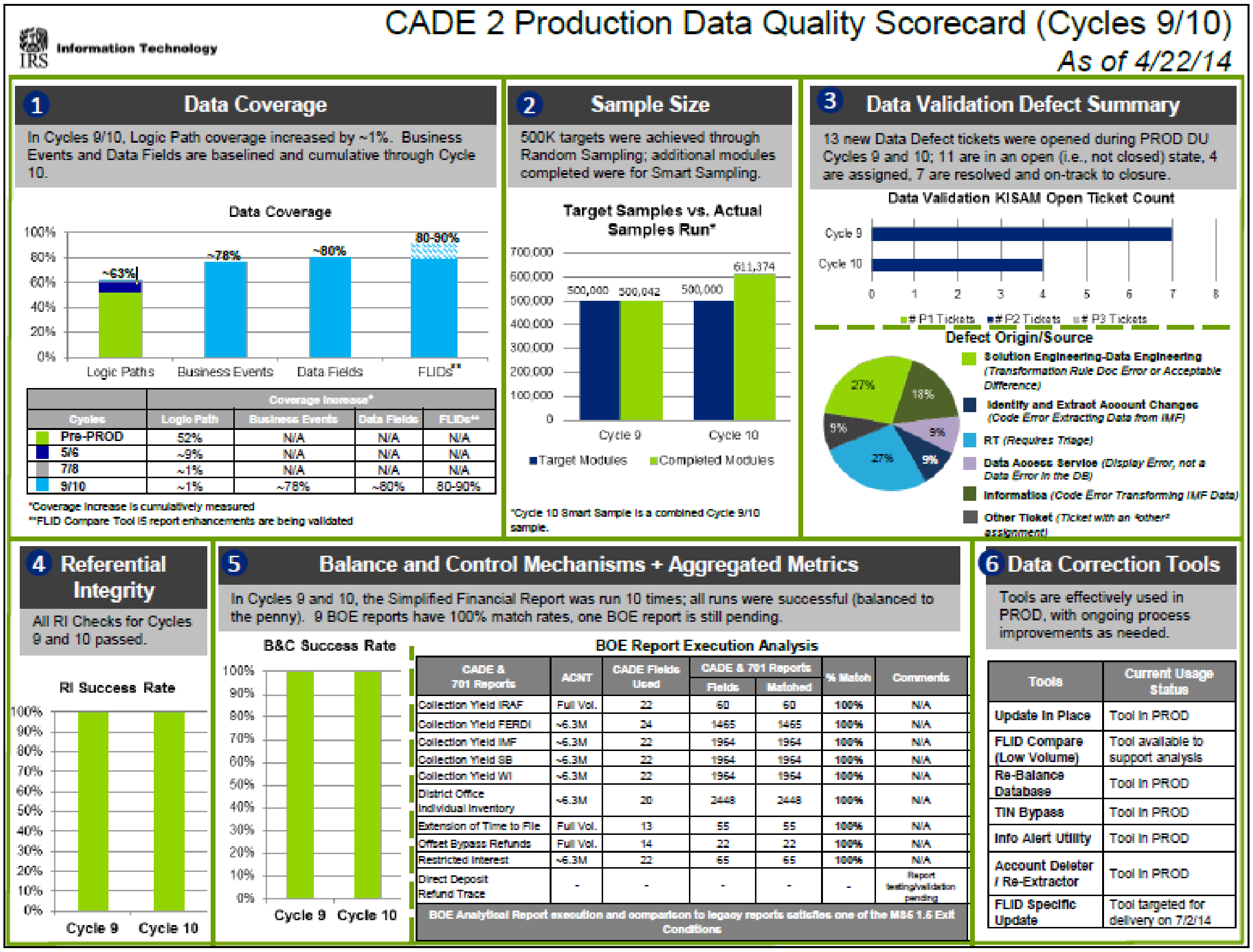
6 Data Correction Tools

Tools are effectively used in PROD, with ongoing process improvements as needed.

Tools	Current Usage Status
Update in Place	Tool in PROD
FLID Compare (Low Volume)	Tool available to support analysis
Re-Balance Database	Tool in PROD
TIN Bypass	Tool in PROD
Info Alert Utility	Tool in PROD
Account Deleter / Re-Extractor	Tool in PROD
FLID Specific Update	Tool targeted for delivery on 7/2/14

Source: CADE 2 PMO. Pre-PROD – Pre-production. PROD DU – Production Daily Update. P1, P2, P3 – Priority 1, 2, or 3. DB – Database. RI – Referential Integrity. B&C – Balance and Control. BOE – Business Objects Enterprise. ACNT – Accounts. IRAF – Individual Retirement Account File. FERDI – Federal Employee/Retiree Delinquency Initiative. SB – Small Business and Self-Employed. WI – Wage and Investment. MS – Milestone. TIN – Taxpayer Identification Number.

Figure 3: Revised Data Quality Scorecard for Production Cycle 9/10 as of April 22, 2014



Information Technology

CADE 2 Production Data Quality Scorecard (Cycles 9/10)

As of 4/22/14

1 Data Coverage

In Cycles 9/10, Logic Path coverage increased by ~1%. Business Events and Data Fields are baselined and cumulative through Cycle 10.

	Coverage Increase*			
Cycles	Logic Path	Business Events	Data Fields	FLIDs**
🟩 Pre-PROD	52%	N/A	N/A	N/A
■ 5/6	~9%	N/A	N/A	N/A
■ 7/8	~1%	N/A	N/A	N/A
■ 9/10	~1%	~78%	~80%	80-90%

*Coverage Increase is cumulatively measured

**FLID Compare Tool I5 report enhancements are being validated

2 Sample Size

500K targets were achieved through Random Sampling; additional modules completed were for Smart Sampling.

*Cycle 10 Smart Sample is a combined Cycle 9/10 sample.

3 Data Validation Defect Summary

13 new Data Defect tickets were opened during PROD DU Cycles 9 and 10; 11 are in an open (i.e., not closed) state, 4 are assigned, 7 are resolved and on-track to closure.

- 🟩 **Solution Engineering-Data Engineering** *(Transformation Rule Doc Error or Acceptable Difference)*
- ■ **Identify and Extract Account Changes** *(Code Error Extracting Data from IMF)*
- ■ **RT** *(Requires Triage)*
- ■ **Data Access Service** *(Display Error, not a Data Error in the DB)*
- ■ **Informatica** *(Code Error Transforming IMF Data)*
- ■ **Other Ticket** *(Ticket with an "other" assignment)*

4 Referential Integrity

All RI Checks for Cycles 9 and 10 passed.

5 Balance and Control Mechanisms + Aggregated Metrics

In Cycles 9 and 10, the Simplified Financial Report was run 10 times; all runs were successful (balanced to the penny). 9 BOE reports have 100% match rates, one BOE report is still pending.

BOE Report Execution Analysis

CADE & 701 Reports	ACNT	CADE Fields Used	CADE & 701 Reports		% Match	Comments
			Fields	Matched		
Collection Yield IRAF	Full Vol.	22	60	60	100%	N/A
Collection Yield FERDI	~6.3M	24	1465	1465	100%	N/A
Collection Yield IMF	~6.3M	22	1964	1964	100%	N/A
Collection Yield SB	~6.3M	22	1964	1964	100%	N/A
Collection Yield WI	~6.3M	22	1964	1964	100%	N/A
District Office Individual Inventory	~6.3M	20	2448	2448	100%	N/A
Extension of Time to File	Full Vol.	13	55	55	100%	N/A
Offset Bypass Refunds	Full Vol.	14	22	22	100%	N/A
Restricted Interest	~6.3M	22	65	65	100%	N/A
Direct Deposit Refund Trace	-	-	-	-	-	Report testing/validation pending
BOE Analytical Report execution and comparison to legacy reports satisfies one of the MS5 1.5 Exit Conditions						

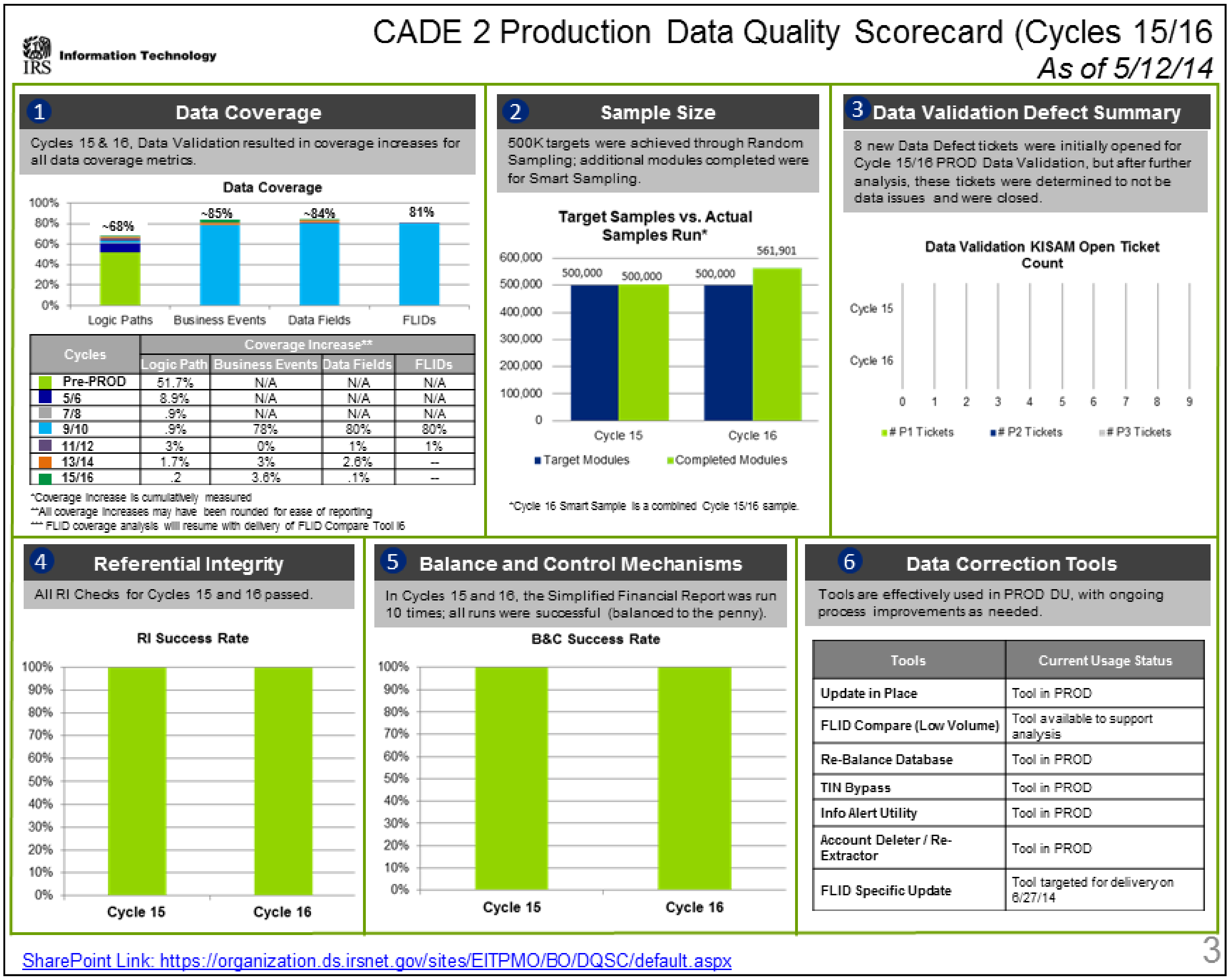
6 Data Correction Tools

Tools are effectively used in PROD, with ongoing process improvements as needed.

Tools	Current Usage Status
Update In Place	Tool In PROD
FLID Compare (Low Volume)	Tool available to support analysis
Re-Balance Database	Tool In PROD
TIN Bypass	Tool In PROD
Info Alert Utility	Tool In PROD
Account Deleter / Re-Extractor	Tool In PROD
FLID Specific Update	Tool targeted for delivery on 7/2/14

Source: CADE 2 PMO. Pre-PROD – Pre-production. PROD DU – Production Daily Update. P1, P2, P3 – Priority 1, 2, or 3. DB – Database. RI – Referential Integrity. B&C – Balance and Control. BOE – Business Objects Enterprise. ACNT – Accounts. IRAF – Individual Retirement Account File. FERDI – Federal Employee/Retiree Delinquency Initiative. SB – Small Business and Self-Employed. WI – Wage and Investment. MS – Milestone. TIN – Taxpayer Identification Number.

Figure 4: Data Quality Scorecard for Production Cycle 15/16 as of May 12, 2014



IRS Information Technology

CADE 2 Production Data Quality Scorecard (Cycles 15/16

As of 5/12/14

1 Data Coverage

Cycles 15 & 16, Data Validation resulted in coverage increases for all data coverage metrics.

Cycles	Coverage Increase**			
	Logic Path	Business Events	Data Fields	FLIDs
🟩 Pre-PROD	51.7%	N/A	N/A	N/A
■ 5/6	8.9%	N/A	N/A	N/A
■ 7/8	.9%	N/A	N/A	N/A
■ 9/10	.9%	78%	80%	80%
■ 11/12	3%	0%	1%	1%
🟧 13/14	1.7%	3%	2.6%	--
🟩 15/16	.2	3.6%	.1%	--

*Coverage increase is cumulatively measured
**All coverage increases may have been rounded for ease of reporting
*** FLID coverage analysis will resume with delivery of FLID Compare Tool I6

2 Sample Size

500K targets were achieved through Random Sampling; additional modules completed were for Smart Sampling.

*Cycle 16 Smart Sample is a combined Cycle 15/16 sample.

3 Data Validation Defect Summary

8 new Data Defect tickets were initially opened for Cycle 15/16 PROD Data Validation, but after further analysis, these tickets were determined to not be data issues and were closed.

4 Referential Integrity

All RI Checks for Cycles 15 and 16 passed.

5 Balance and Control Mechanisms

In Cycles 15 and 16, the Simplified Financial Report was run 10 times; all runs were successful (balanced to the penny).

6 Data Correction Tools

Tools are effectively used in PROD DU, with ongoing process improvements as needed.

Tools	Current Usage Status
Update in Place	Tool in PROD
FLID Compare (Low Volume)	Tool available to support analysis
Re-Balance Database	Tool in PROD
TIN Bypass	Tool in PROD
Info Alert Utility	Tool in PROD
Account Deleter / Re-Extractor	Tool in PROD
FLID Specific Update	Tool targeted for delivery on 6/27/14

SharePoint Link: https://organization.ds.irsnet.gov/sites/EITPMO/BO/DQSC/default.aspx

3

Source: CADE 2 PMO. Pre-PROD – Pre-production. P1, P2, P3 – Priority 1, 2, or 3. RI – Referential Integrity. B&C – Balance and Control. PROD DU – Production Daily Update. TIN – Taxpayer Identification Number.

Appendix V

Glossary of Terms

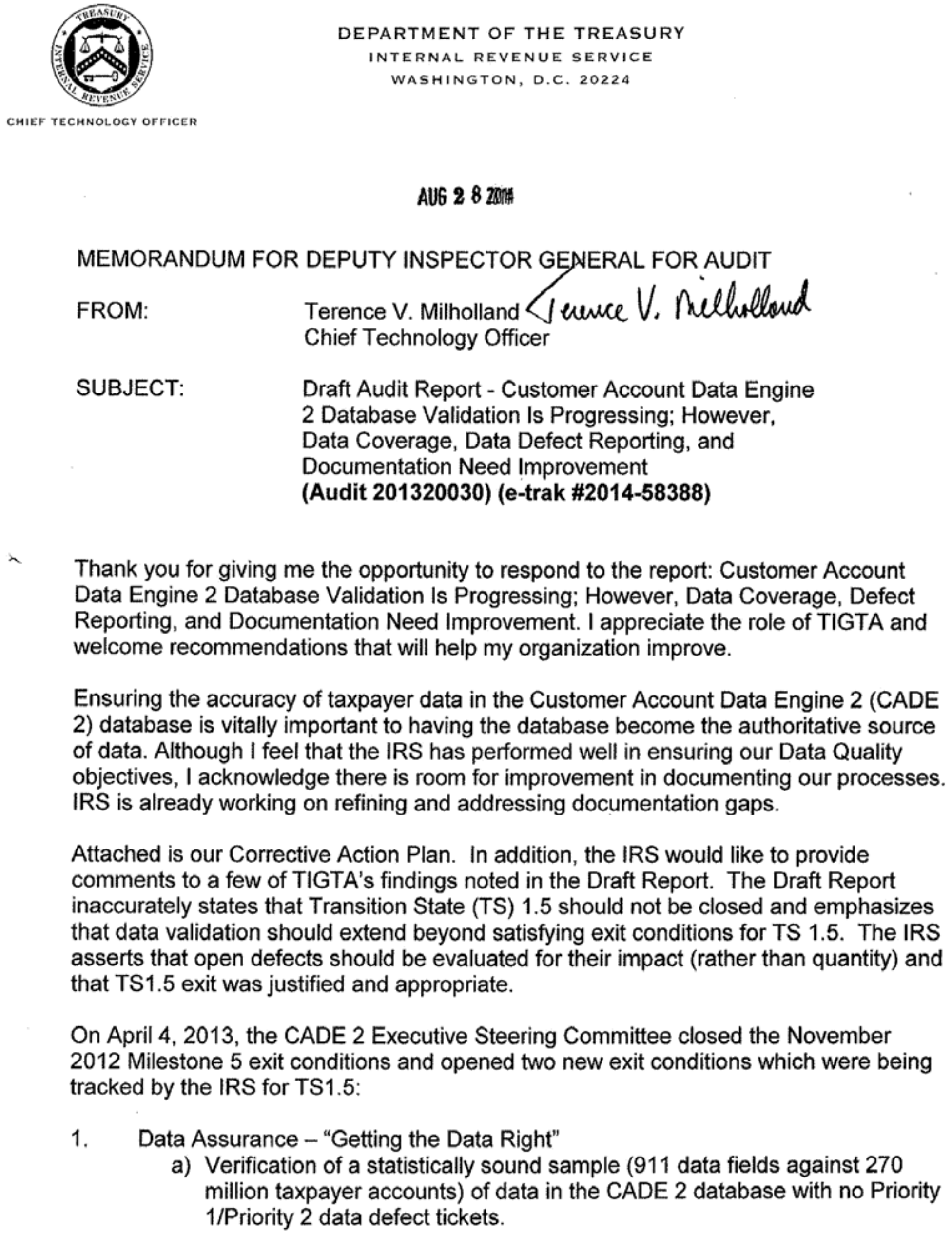
Term	Definition
Applications Development	The development organization for systems that manage taxpayer accounts from the initial filing of a tax return to interactions with the taxpayers and potential audit and collection activities. It also provides enterprise-wide administrative systems related to workforce support, human capital, financial, and facilities.
Business Event	Consists of transactions and nontransactions. A transaction is a business event. An example of a transaction is the posting of a tax return to the taxpayer's account. A nontransaction is usually generated by a transaction. An example of a nontransaction is the balance section of the taxpayer's account.
Corporate Files Online	A collection of "read only" files extracted from the Master Files and maintained at the Enterprise Computing Centers in Memphis, Tennessee, and Martinsburg, West Virginia.
Customer Account Data Engine (CADE)	The foundation for managing taxpayer accounts in the IRS modernization plan. It will consist of databases and related applications that will replace the existing IRS Master File processing systems and will include applications for daily posting, settlement, maintenance, refund processing, and issue detection for taxpayer tax account and return data.
Cycle	A week, which is usually designated by a cycle number when referring to IRS processing activities.
Data Access Service	A set of common capabilities that mediate relationships between applications throughout the enterprise and the external community. In general, the Data Access Service layer supports inter-application integration and sharing of data and functions that are maintained in separate application systems.
Database	A collection of information that is organized so that it can easily be accessed, managed, and updated.
Data-Centric	Refers to a focus on the specific data relevant to a given task.

Term	Definition
Field Identifier (FLID)	An IRS file format that uses a numeric field (*i.e.*, FLIDs) to identify a data field.
FLID Compare Tool (High Volume)	An automated tool that compares a high volume of taxpayer accounts (the business requirement is to compare 1 million tax modules in 40 hours). The tool is intended to compare data in the IMF and CADE 2.
Filing Season	The period from January through mid-April when most individual income tax returns are filed.
Final Integration Testing	A system test consisting of integrated end-to-end testing of mainline tax processing systems to verify that new releases of interrelated systems and hardware platforms can collectively support the IRS business functions allocated to them.
General Transcript Report	A report used by the Chief Financial Officer and Business Modernization Office during data validation to compare the corresponding data fields to ensure identical data.
Individual Master File	The IRS files that maintain transactions or records of individual tax accounts.
Individual Master File Online	This provides online access to individual taxpayer returns.
Knowledge, Incident/Problem, Service Asset Management	An IRS application that maintains the complete inventory of information technology and non–information technology assets, including computer hardware and software. It is also the reporting tool for problem management with all IRS developed applications, and shares information with the Enterprise Service Desk.
Milestone	Provides for "go/no-go" decision points in a project and are sometimes associated with funding approval to proceed.
Priority 1 Defect Ticket	An incident ticket issue exhibiting the following characteristics: 1) resulting in severe mission-critical work stoppage or any issue relating to safety or health (*e.g.*, fire, electrical shock); 2) affecting vital IRS customer commitments of national or area-wide scope; 3) affecting multiple internal or external customers and service to taxpayers; and 4) requiring immediate action.

Term	Definition
Priority 2 Defect Ticket	An incident ticket issue with the potential to result in a work stoppage and/or to lead to severe mission-critical work stoppage if actions are not taken to resolve the incident.
Production Support Environment	A close replica of the IRS production environment used for various activities such as performance testing and data validation.
Requirement	A statement of capability or condition that a system, subsystem, or system component must have or meet to satisfy a contract, standard, or specification.
Risk	A potential event that could have an unwanted impact on the cost, schedule, business, or technical performance of an information technology program, project, or organization.
Smart Sample	A sample of modules selected as a result of the Smart sampling process, which is part of the CADE 2 data validation data sampling methodology. The Smart sampling process will ensure that infrequently seen data fields will be included in data validation testing. It will also provide a statistical basis for deciding how many instances of a particular data field or business event are to be sampled based on the probability of occurrence and target confidence level.
Structured Query Language	A standard interactive and programming language for getting information from and updating a database.
Systems Acceptability Testing	Testing conducted to verify a system satisfies application requirements.
Transformation Logic Path	This is the value of a data field based on the transformation rule conditions it meets.
Transformation Rule	A rule to set the value in a field in the CADE 2 database. It may contain multiple conditions to decide the value of that field. Each condition defines a logic path for the transformation.

Appendix VI

Management's Response to the Draft Report



DEPARTMENT OF THE TREASURY
INTERNAL REVENUE SERVICE
WASHINGTON, D.C. 20224

CHIEF TECHNOLOGY OFFICER

AUG 2 8 2014

MEMORANDUM FOR DEPUTY INSPECTOR GENERAL FOR AUDIT

FROM: Terence V. Milholland Terence V. Milholland
Chief Technology Officer

SUBJECT: Draft Audit Report - Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement
(Audit 201320030) (e-trak #2014-58388)

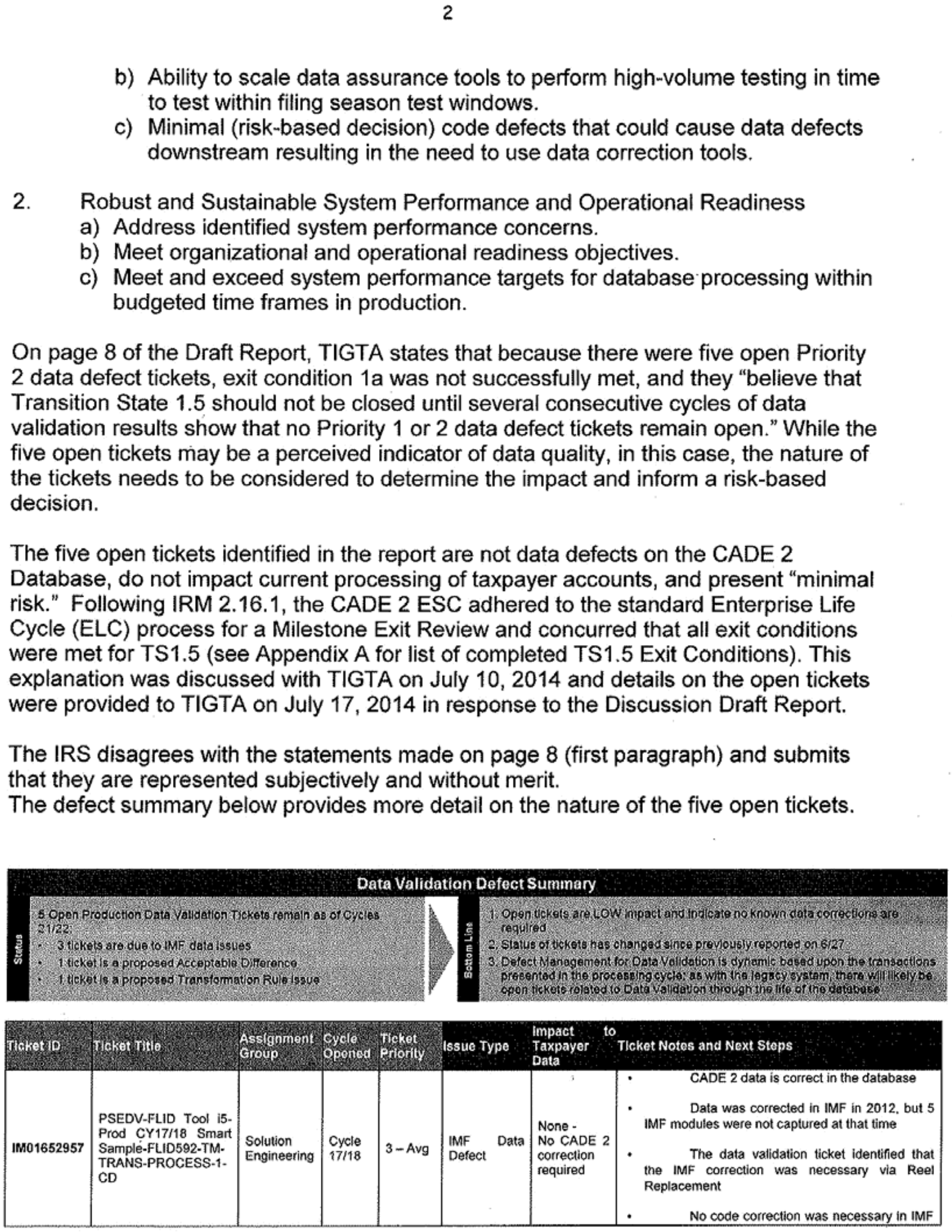
Thank you for giving me the opportunity to respond to the report: Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Defect Reporting, and Documentation Need Improvement. I appreciate the role of TIGTA and welcome recommendations that will help my organization improve.

Ensuring the accuracy of taxpayer data in the Customer Account Data Engine 2 (CADE 2) database is vitally important to having the database become the authoritative source of data. Although I feel that the IRS has performed well in ensuring our Data Quality objectives, I acknowledge there is room for improvement in documenting our processes. IRS is already working on refining and addressing documentation gaps.

Attached is our Corrective Action Plan. In addition, the IRS would like to provide comments to a few of TIGTA's findings noted in the Draft Report. The Draft Report inaccurately states that Transition State (TS) 1.5 should not be closed and emphasizes that data validation should extend beyond satisfying exit conditions for TS 1.5. The IRS asserts that open defects should be evaluated for their impact (rather than quantity) and that TS1.5 exit was justified and appropriate.

On April 4, 2013, the CADE 2 Executive Steering Committee closed the November 2012 Milestone 5 exit conditions and opened two new exit conditions which were being tracked by the IRS for TS1.5:

1. Data Assurance – "Getting the Data Right"
   a) Verification of a statistically sound sample (911 data fields against 270 million taxpayer accounts) of data in the CADE 2 database with no Priority 1/Priority 2 data defect tickets.



2

b) Ability to scale data assurance tools to perform high-volume testing in time to test within filing season test windows.
c) Minimal (risk-based decision) code defects that could cause data defects downstream resulting in the need to use data correction tools.

2. Robust and Sustainable System Performance and Operational Readiness
   a) Address identified system performance concerns.
   b) Meet organizational and operational readiness objectives.
   c) Meet and exceed system performance targets for database processing within budgeted time frames in production.

On page 8 of the Draft Report, TIGTA states that because there were five open Priority 2 data defect tickets, exit condition 1a was not successfully met, and they "believe that Transition State 1.5 should not be closed until several consecutive cycles of data validation results show that no Priority 1 or 2 data defect tickets remain open." While the five open tickets may be a perceived indicator of data quality, in this case, the nature of the tickets needs to be considered to determine the impact and inform a risk-based decision.

The five open tickets identified in the report are not data defects on the CADE 2 Database, do not impact current processing of taxpayer accounts, and present "minimal risk." Following IRM 2.16.1, the CADE 2 ESC adhered to the standard Enterprise Life Cycle (ELC) process for a Milestone Exit Review and concurred that all exit conditions were met for TS1.5 (see Appendix A for list of completed TS1.5 Exit Conditions). This explanation was discussed with TIGTA on July 10, 2014 and details on the open tickets were provided to TIGTA on July 17, 2014 in response to the Discussion Draft Report.

The IRS disagrees with the statements made on page 8 (first paragraph) and submits that they are represented subjectively and without merit.
The defect summary below provides more detail on the nature of the five open tickets.

Data Validation Defect Summary

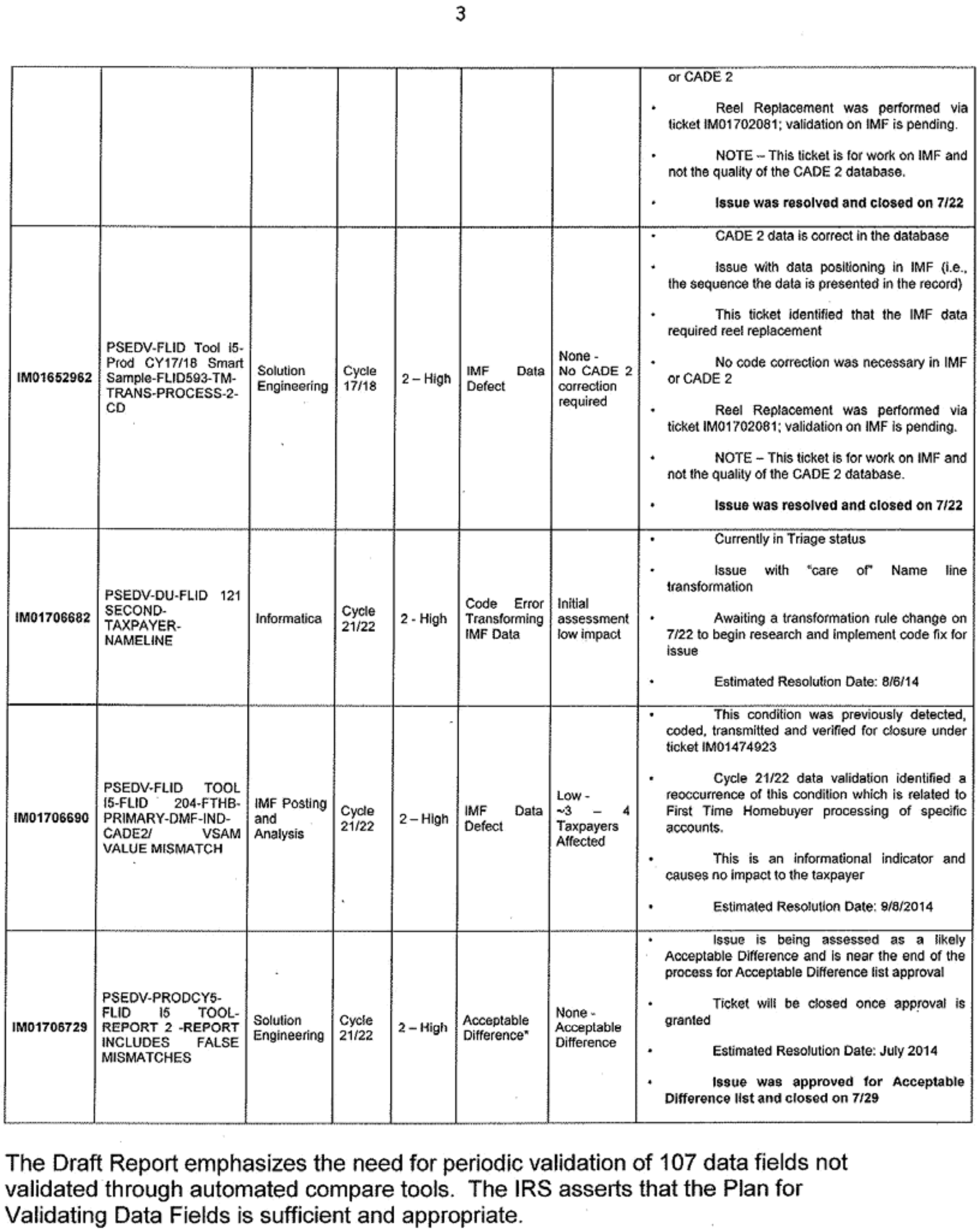
Status

5 Open Production Data Validation Tickets remain as of Cycles 21/22:
- 3 tickets are due to IMF data issues
- 1 ticket is a proposed Acceptable Difference
- 1 ticket is a proposed Transformation Rule issue

Bottom Line

1. Open tickets are LOW impact and indicate no known data corrections are required
2. Status of tickets has changed since previously reported on 6/27
3. Defect Management for Data Validation is dynamic based upon the transactions presented in the processing cycle; as with the legacy system, there will likely be open tickets related to Data Validation through the life of the database

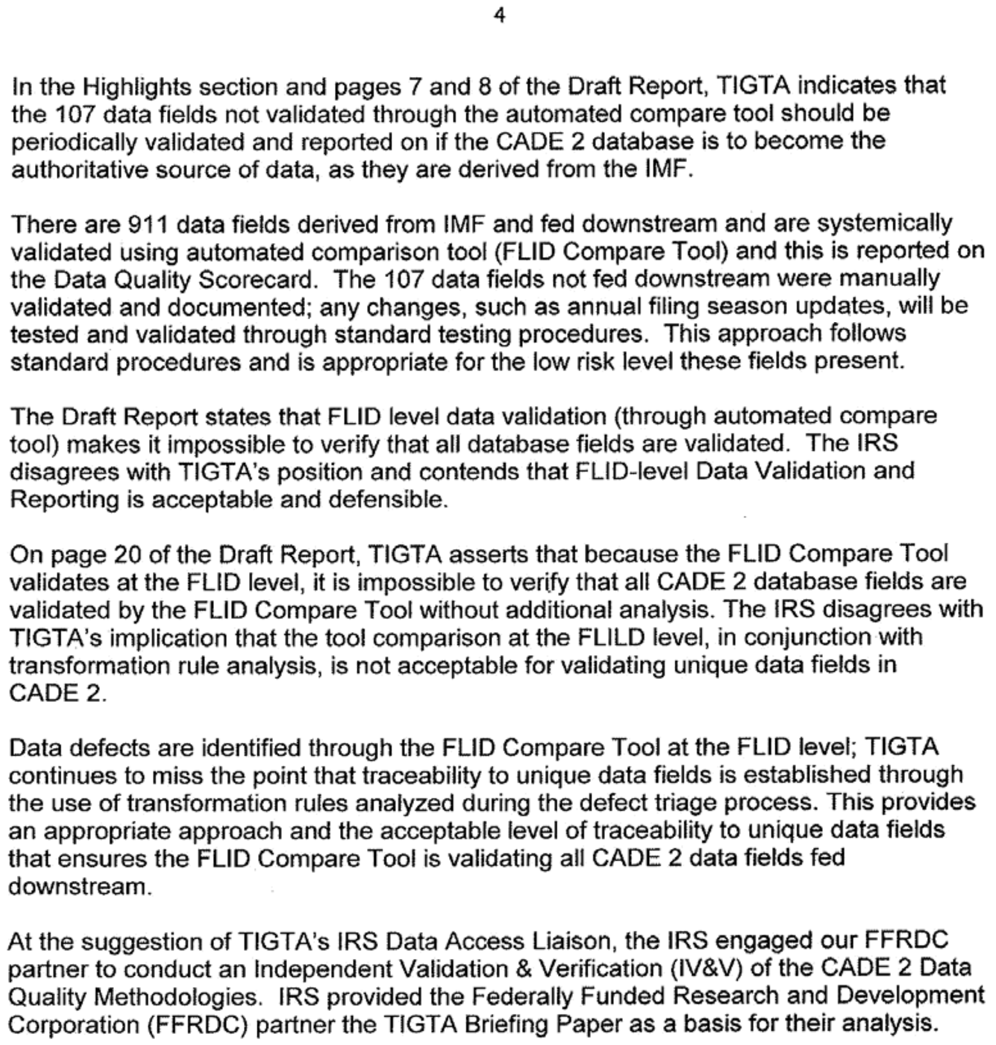
Ticket ID	Ticket Title	Assignment Group	Cycle Opened	Ticket Priority	Issue Type	Impact to Taxpayer Data	Ticket Notes and Next Steps
IM01652957	PSEDV-FLID Tool i5-Prod CY17/18 Smart Sample-FLID592-TM-TRANS-PROCESS-1-CD	Solution Engineering	Cycle 17/18	3 – Avg	IMF Data Defect	None - No CADE 2 correction required	• CADE 2 data is correct in the database • Data was corrected in IMF in 2012, but 5 IMF modules were not captured at that time • The data validation ticket identified that the IMF correction was necessary via Reel Replacement • No code correction was necessary in IMF



3

							or CADE 2 • Reel Replacement was performed via ticket IM01702081; validation on IMF is pending. • NOTE – This ticket is for work on IMF and not the quality of the CADE 2 database. • **Issue was resolved and closed on 7/22**
IM01652962	PSEDV-FLID Tool I5-Prod CY17/18 Smart Sample-FLID593-TM-TRANS-PROCESS-2-CD	Solution Engineering	Cycle 17/18	2 – High	IMF Data Defect	None - No CADE 2 correction required	• CADE 2 data is correct in the database • Issue with data positioning in IMF (i.e., the sequence the data is presented in the record) • This ticket identified that the IMF data required reel replacement • No code correction was necessary in IMF or CADE 2 • Reel Replacement was performed via ticket IM01702081; validation on IMF is pending. • NOTE – This ticket is for work on IMF and not the quality of the CADE 2 database. • **Issue was resolved and closed on 7/22**
IM01706682	PSEDV-DU-FLID 121 SECOND-TAXPAYER-NAMELINE	Informatica	Cycle 21/22	2 - High	Code Error Transforming IMF Data	Initial assessment low impact	• Currently in Triage status • Issue with "care of" Name line transformation • Awaiting a transformation rule change on 7/22 to begin research and implement code fix for issue • Estimated Resolution Date: 8/6/14
IM01706690	PSEDV-FLID TOOL I5-FLID 204-FTHB-PRIMARY-DMF-IND-CADE2/ VSAM VALUE MISMATCH	IMF Posting and Analysis	Cycle 21/22	2 – High	IMF Data Defect	Low - ~3 – 4 Taxpayers Affected	• This condition was previously detected, coded, transmitted and verified for closure under ticket IM01474923 • Cycle 21/22 data validation identified a reoccurrence of this condition which is related to First Time Homebuyer processing of specific accounts. • This is an informational indicator and causes no impact to the taxpayer • Estimated Resolution Date: 9/8/2014
IM01706729	PSEDV-PRODCY5-FLID I5 TOOL-REPORT 2 -REPORT INCLUDES FALSE MISMATCHES	Solution Engineering	Cycle 21/22	2 – High	Acceptable Difference*	None - Acceptable Difference	• Issue is being assessed as a likely Acceptable Difference and is near the end of the process for Acceptable Difference list approval • Ticket will be closed once approval is granted • Estimated Resolution Date: July 2014 • **Issue was approved for Acceptable Difference list and closed on 7/29**

The Draft Report emphasizes the need for periodic validation of 107 data fields not validated through automated compare tools. The IRS asserts that the Plan for Validating Data Fields is sufficient and appropriate.



In the Highlights section and pages 7 and 8 of the Draft Report, TIGTA indicates that the 107 data fields not validated through the automated compare tool should be periodically validated and reported on if the CADE 2 database is to become the authoritative source of data, as they are derived from the IMF.

There are 911 data fields derived from IMF and fed downstream and are systemically validated using automated comparison tool (FLID Compare Tool) and this is reported on the Data Quality Scorecard. The 107 data fields not fed downstream were manually validated and documented; any changes, such as annual filing season updates, will be tested and validated through standard testing procedures. This approach follows standard procedures and is appropriate for the low risk level these fields present.

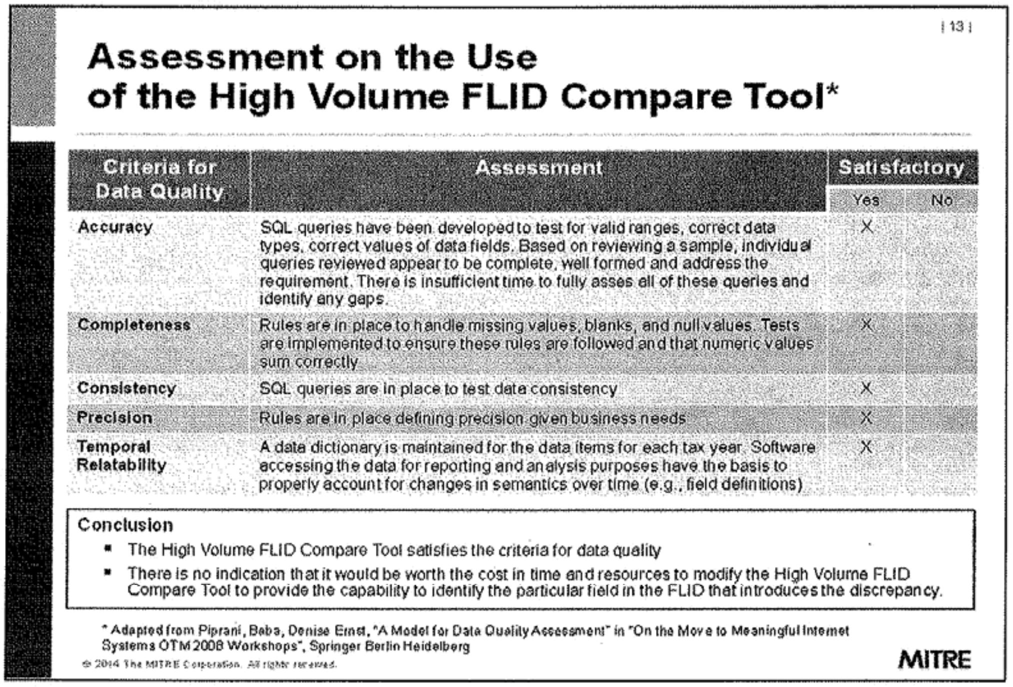
The Draft Report states that FLID level data validation (through automated compare tool) makes it impossible to verify that all database fields are validated. The IRS disagrees with TIGTA's position and contends that FLID-level Data Validation and Reporting is acceptable and defensible.

On page 20 of the Draft Report, TIGTA asserts that because the FLID Compare Tool validates at the FLID level, it is impossible to verify that all CADE 2 database fields are validated by the FLID Compare Tool without additional analysis. The IRS disagrees with TIGTA's implication that the tool comparison at the FLILD level, in conjunction with transformation rule analysis, is not acceptable for validating unique data fields in CADE 2.

Data defects are identified through the FLID Compare Tool at the FLID level; TIGTA continues to miss the point that traceability to unique data fields is established through the use of transformation rules analyzed during the defect triage process. This provides an appropriate approach and the acceptable level of traceability to unique data fields that ensures the FLID Compare Tool is validating all CADE 2 data fields fed downstream.

At the suggestion of TIGTA's IRS Data Access Liaison, the IRS engaged our FFRDC partner to conduct an Independent Validation & Verification (IV&V) of the CADE 2 Data Quality Methodologies. IRS provided the Federally Funded Research and Development Corporation (FFRDC) partner the TIGTA Briefing Paper as a basis for their analysis.

Figure 2: IV&V Assessment of the Use of the High Volume FLID Compare Tool



| 13 |

Assessment on the Use of the High Volume FLID Compare Tool*

Criteria for Data Quality	Assessment	Satisfactory	
		Yes	No
Accuracy	SQL queries have been developed to test for valid ranges, correct data types, correct values of data fields. Based on reviewing a sample, individual queries reviewed appear to be complete, well formed and address the requirement. There is insufficient time to fully asses all of these queries and identify any gaps.	X	
Completeness	Rules are in place to handle missing values, blanks, and null values. Tests are implemented to ensure these rules are followed and that numeric values sum correctly	X	
Consistency	SQL queries are in place to test data consistency	X	
Precision	Rules are in place defining precision given business needs	X	
Temporal Relatability	A data dictionary is maintained for the data items for each tax year. Software accessing the data for reporting and analysis purposes have the basis to properly account for changes in semantics over time (e.g., field definitions)	X	

Conclusion

- The High Volume FLID Compare Tool satisfies the criteria for data quality
- There is no indication that it would be worth the cost in time and resources to modify the High Volume FLID Compare Tool to provide the capability to identify the particular field in the FLID that introduces the discrepancy.

* Adapted from Piprani, Baba, Denise Ernst, "A Model for Data Quality Assessment" in "On the Move to Meaningful Internet Systems OTM 2008 Workshops", Springer Berlin Heidelberg

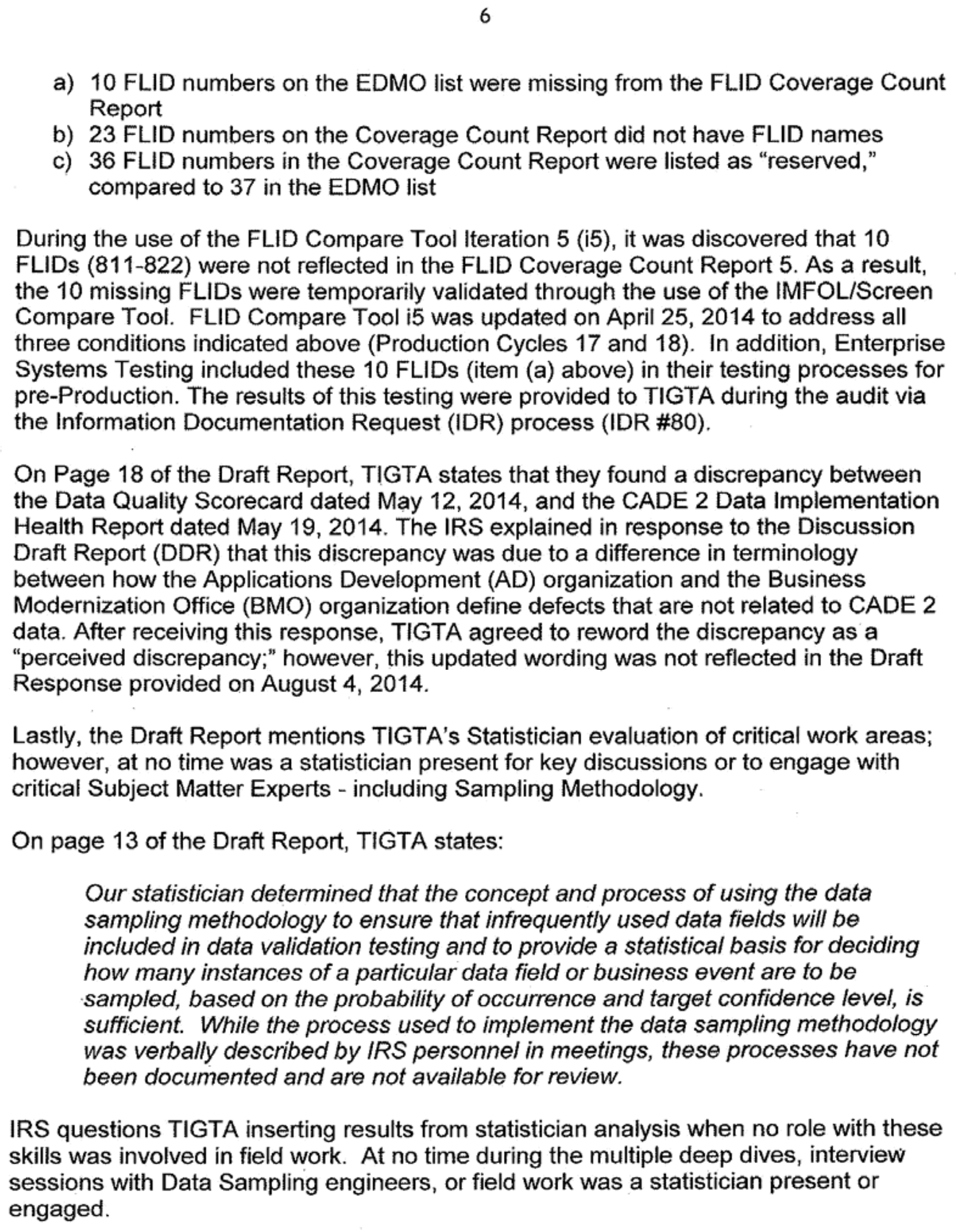
MITRE

The FFRDC partner's assessment indicated satisfactory indicators for the key criteria for Data Quality including: Accuracy, Completeness, Consistency, Precision and Temporal Relatability. Their conclusion on the Assessment of the Use of the High Volume FLID Compare Tool* is below:

> *There is no indication that it would be worth the cost in time and resources to modify the High Volume FLID Compare Tool to provide the capability to identify the particular field in the FLID that introduces the discrepancy.*
> ** Adapted from Piprani, Baba, Denise Ernst, "A Model for Data Quality Assessment" in "On the Move to Meaningful Internet Systems OTM 2008 Workshops", Springer Berlin Heidelberg*

TIGTA maintains that discrepancies found in the FLID Coverage Count Report raise questions as to whether the FLID compare tool is accurately comparing all data at the FLID Level. The IRS disagrees with the severity of this statement since alternative means were temporarily leveraged to perform data validation.

On page 20 of the Draft Report, TIGTA noted the following discrepancies between the EDMO FLID list and the FLIDs in the FLID Coverage Report:



6

a) 10 FLID numbers on the EDMO list were missing from the FLID Coverage Count Report
b) 23 FLID numbers on the Coverage Count Report did not have FLID names
c) 36 FLID numbers in the Coverage Count Report were listed as "reserved," compared to 37 in the EDMO list

During the use of the FLID Compare Tool Iteration 5 (i5), it was discovered that 10 FLIDs (811-822) were not reflected in the FLID Coverage Count Report 5. As a result, the 10 missing FLIDs were temporarily validated through the use of the IMFOL/Screen Compare Tool. FLID Compare Tool i5 was updated on April 25, 2014 to address all three conditions indicated above (Production Cycles 17 and 18). In addition, Enterprise Systems Testing included these 10 FLIDs (item (a) above) in their testing processes for pre-Production. The results of this testing were provided to TIGTA during the audit via the Information Documentation Request (IDR) process (IDR #80).

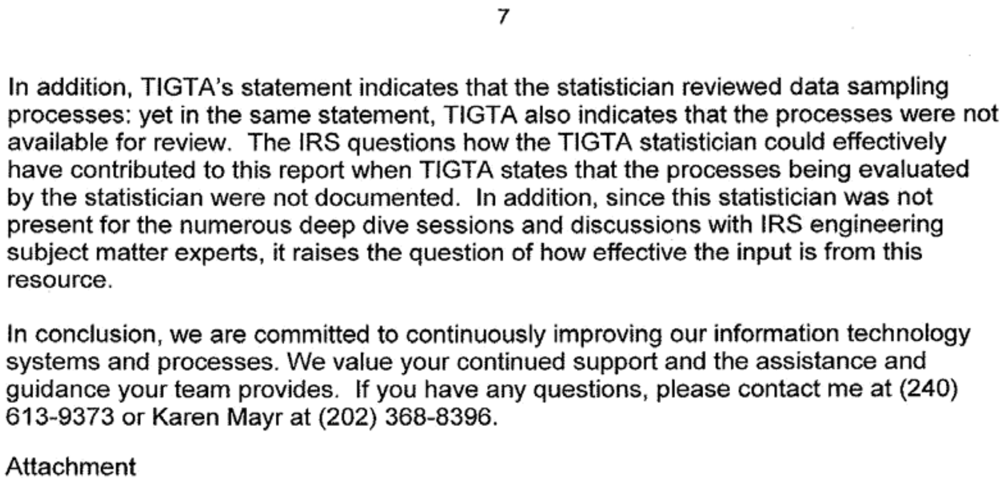
On Page 18 of the Draft Report, TIGTA states that they found a discrepancy between the Data Quality Scorecard dated May 12, 2014, and the CADE 2 Data Implementation Health Report dated May 19, 2014. The IRS explained in response to the Discussion Draft Report (DDR) that this discrepancy was due to a difference in terminology between how the Applications Development (AD) organization and the Business Modernization Office (BMO) organization define defects that are not related to CADE 2 data. After receiving this response, TIGTA agreed to reword the discrepancy as a "perceived discrepancy;" however, this updated wording was not reflected in the Draft Response provided on August 4, 2014.

Lastly, the Draft Report mentions TIGTA's Statistician evaluation of critical work areas; however, at no time was a statistician present for key discussions or to engage with critical Subject Matter Experts - including Sampling Methodology.

On page 13 of the Draft Report, TIGTA states:

> *Our statistician determined that the concept and process of using the data sampling methodology to ensure that infrequently used data fields will be included in data validation testing and to provide a statistical basis for deciding how many instances of a particular data field or business event are to be sampled, based on the probability of occurrence and target confidence level, is sufficient. While the process used to implement the data sampling methodology was verbally described by IRS personnel in meetings, these processes have not been documented and are not available for review.*

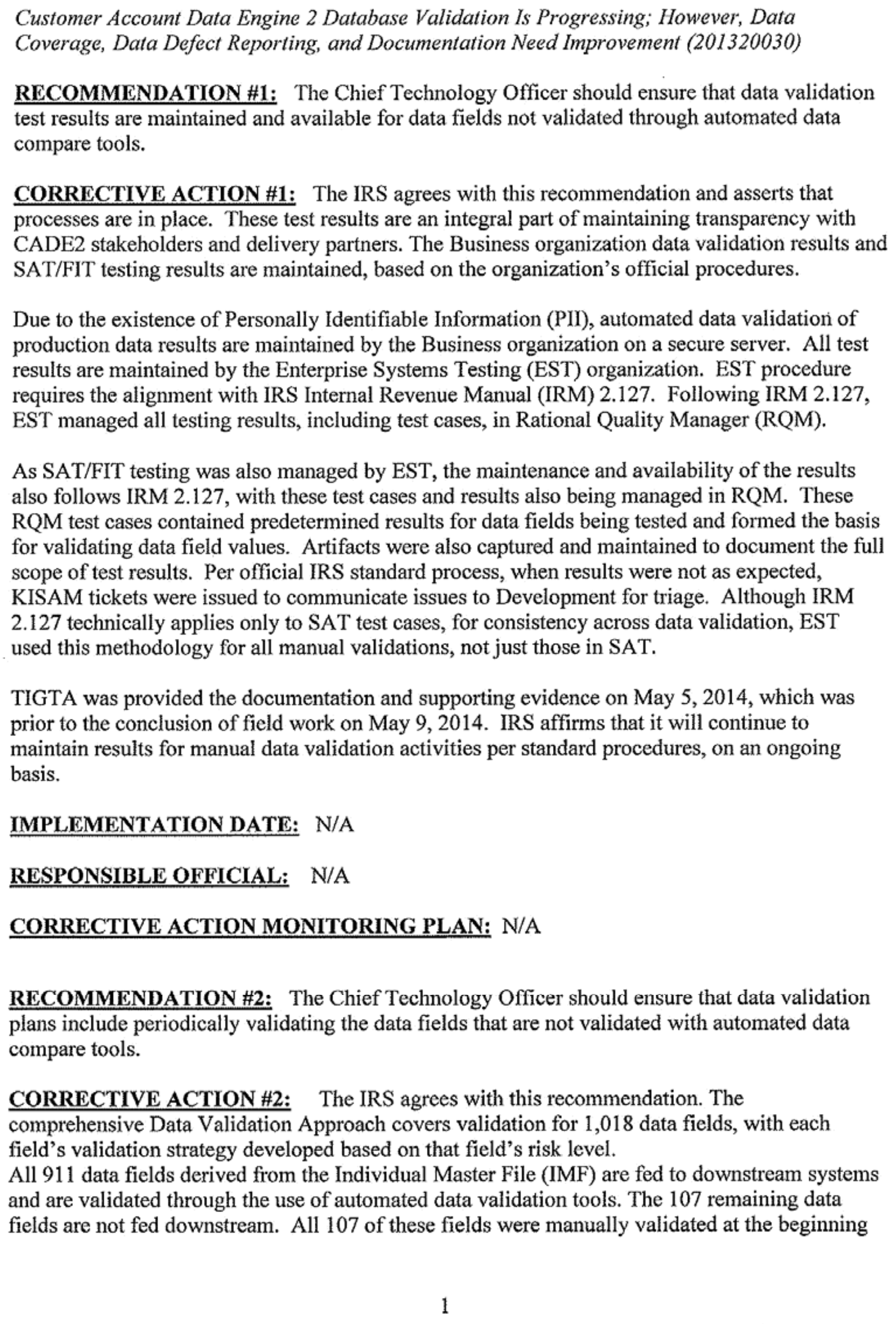
IRS questions TIGTA inserting results from statistician analysis when no role with these skills was involved in field work. At no time during the multiple deep dives, interview sessions with Data Sampling engineers, or field work was a statistician present or engaged.



In addition, TIGTA's statement indicates that the statistician reviewed data sampling processes: yet in the same statement, TIGTA also indicates that the processes were not available for review. The IRS questions how the TIGTA statistician could effectively have contributed to this report when TIGTA states that the processes being evaluated by the statistician were not documented. In addition, since this statistician was not present for the numerous deep dive sessions and discussions with IRS engineering subject matter experts, it raises the question of how effective the input is from this resource.

In conclusion, we are committed to continuously improving our information technology systems and processes. We value your continued support and the assistance and guidance your team provides. If you have any questions, please contact me at (240) 613-9373 or Karen Mayr at (202) 368-8396.

Attachment



Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement (201320030)

RECOMMENDATION #1: The Chief Technology Officer should ensure that data validation test results are maintained and available for data fields not validated through automated data compare tools.

CORRECTIVE ACTION #1: The IRS agrees with this recommendation and asserts that processes are in place. These test results are an integral part of maintaining transparency with CADE2 stakeholders and delivery partners. The Business organization data validation results and SAT/FIT testing results are maintained, based on the organization's official procedures.

Due to the existence of Personally Identifiable Information (PII), automated data validation of production data results are maintained by the Business organization on a secure server. All test results are maintained by the Enterprise Systems Testing (EST) organization. EST procedure requires the alignment with IRS Internal Revenue Manual (IRM) 2.127. Following IRM 2.127, EST managed all testing results, including test cases, in Rational Quality Manager (RQM).

As SAT/FIT testing was also managed by EST, the maintenance and availability of the results also follows IRM 2.127, with these test cases and results also being managed in RQM. These RQM test cases contained predetermined results for data fields being tested and formed the basis for validating data field values. Artifacts were also captured and maintained to document the full scope of test results. Per official IRS standard process, when results were not as expected, KISAM tickets were issued to communicate issues to Development for triage. Although IRM 2.127 technically applies only to SAT test cases, for consistency across data validation, EST used this methodology for all manual validations, not just those in SAT.

TIGTA was provided the documentation and supporting evidence on May 5, 2014, which was prior to the conclusion of field work on May 9, 2014. IRS affirms that it will continue to maintain results for manual data validation activities per standard procedures, on an ongoing basis.

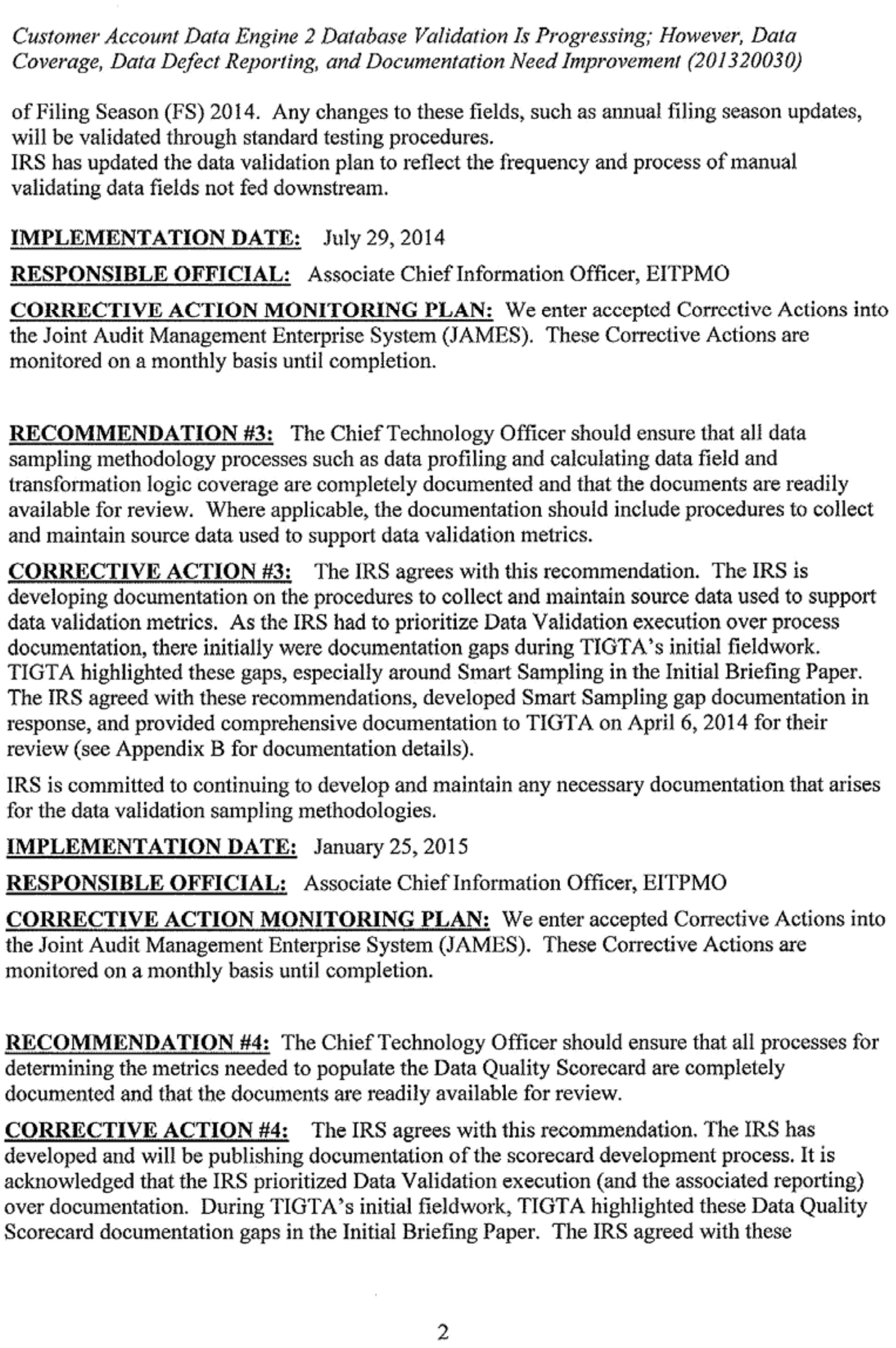
IMPLEMENTATION DATE: N/A

RESPONSIBLE OFFICIAL: N/A

CORRECTIVE ACTION MONITORING PLAN: N/A

RECOMMENDATION #2: The Chief Technology Officer should ensure that data validation plans include periodically validating the data fields that are not validated with automated data compare tools.

CORRECTIVE ACTION #2: The IRS agrees with this recommendation. The comprehensive Data Validation Approach covers validation for 1,018 data fields, with each field's validation strategy developed based on that field's risk level.
All 911 data fields derived from the Individual Master File (IMF) are fed to downstream systems and are validated through the use of automated data validation tools. The 107 remaining data fields are not fed downstream. All 107 of these fields were manually validated at the beginning



Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement (201320030)

of Filing Season (FS) 2014. Any changes to these fields, such as annual filing season updates, will be validated through standard testing procedures.
IRS has updated the data validation plan to reflect the frequency and process of manual validating data fields not fed downstream.

IMPLEMENTATION DATE: July 29, 2014

RESPONSIBLE OFFICIAL: Associate Chief Information Officer, EITPMO

CORRECTIVE ACTION MONITORING PLAN: We enter accepted Corrective Actions into the Joint Audit Management Enterprise System (JAMES). These Corrective Actions are monitored on a monthly basis until completion.

RECOMMENDATION #3: The Chief Technology Officer should ensure that all data sampling methodology processes such as data profiling and calculating data field and transformation logic coverage are completely documented and that the documents are readily available for review. Where applicable, the documentation should include procedures to collect and maintain source data used to support data validation metrics.

CORRECTIVE ACTION #3: The IRS agrees with this recommendation. The IRS is developing documentation on the procedures to collect and maintain source data used to support data validation metrics. As the IRS had to prioritize Data Validation execution over process documentation, there initially were documentation gaps during TIGTA's initial fieldwork. TIGTA highlighted these gaps, especially around Smart Sampling in the Initial Briefing Paper. The IRS agreed with these recommendations, developed Smart Sampling gap documentation in response, and provided comprehensive documentation to TIGTA on April 6, 2014 for their review (see Appendix B for documentation details).

IRS is committed to continuing to develop and maintain any necessary documentation that arises for the data validation sampling methodologies.

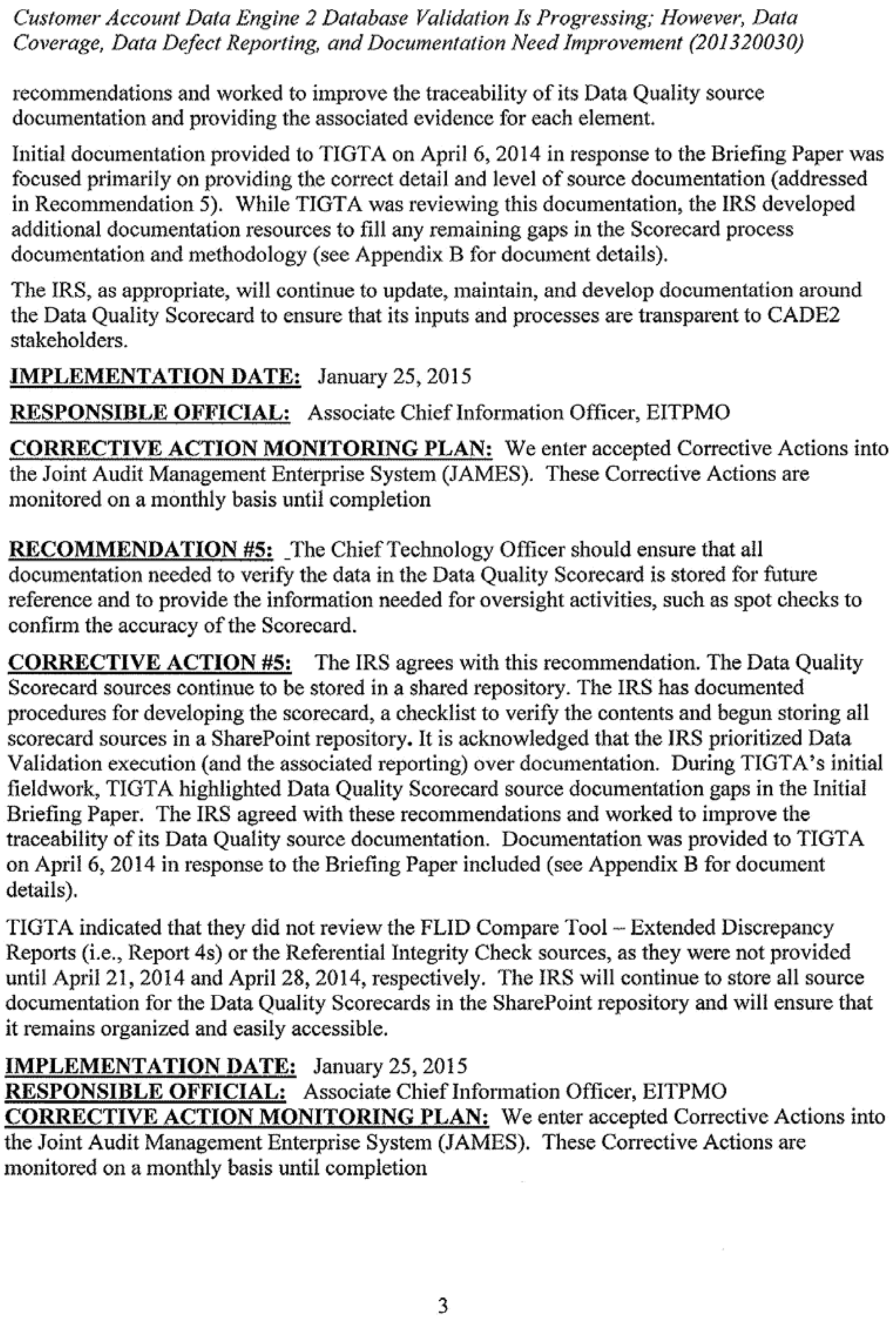
IMPLEMENTATION DATE: January 25, 2015

RESPONSIBLE OFFICIAL: Associate Chief Information Officer, EITPMO

CORRECTIVE ACTION MONITORING PLAN: We enter accepted Corrective Actions into the Joint Audit Management Enterprise System (JAMES). These Corrective Actions are monitored on a monthly basis until completion.

RECOMMENDATION #4: The Chief Technology Officer should ensure that all processes for determining the metrics needed to populate the Data Quality Scorecard are completely documented and that the documents are readily available for review.

CORRECTIVE ACTION #4: The IRS agrees with this recommendation. The IRS has developed and will be publishing documentation of the scorecard development process. It is acknowledged that the IRS prioritized Data Validation execution (and the associated reporting) over documentation. During TIGTA's initial fieldwork, TIGTA highlighted these Data Quality Scorecard documentation gaps in the Initial Briefing Paper. The IRS agreed with these



recommendations and worked to improve the traceability of its Data Quality source documentation and providing the associated evidence for each element.

Initial documentation provided to TIGTA on April 6, 2014 in response to the Briefing Paper was focused primarily on providing the correct detail and level of source documentation (addressed in Recommendation 5). While TIGTA was reviewing this documentation, the IRS developed additional documentation resources to fill any remaining gaps in the Scorecard process documentation and methodology (see Appendix B for document details).

The IRS, as appropriate, will continue to update, maintain, and develop documentation around the Data Quality Scorecard to ensure that its inputs and processes are transparent to CADE2 stakeholders.

IMPLEMENTATION DATE: January 25, 2015

RESPONSIBLE OFFICIAL: Associate Chief Information Officer, EITPMO

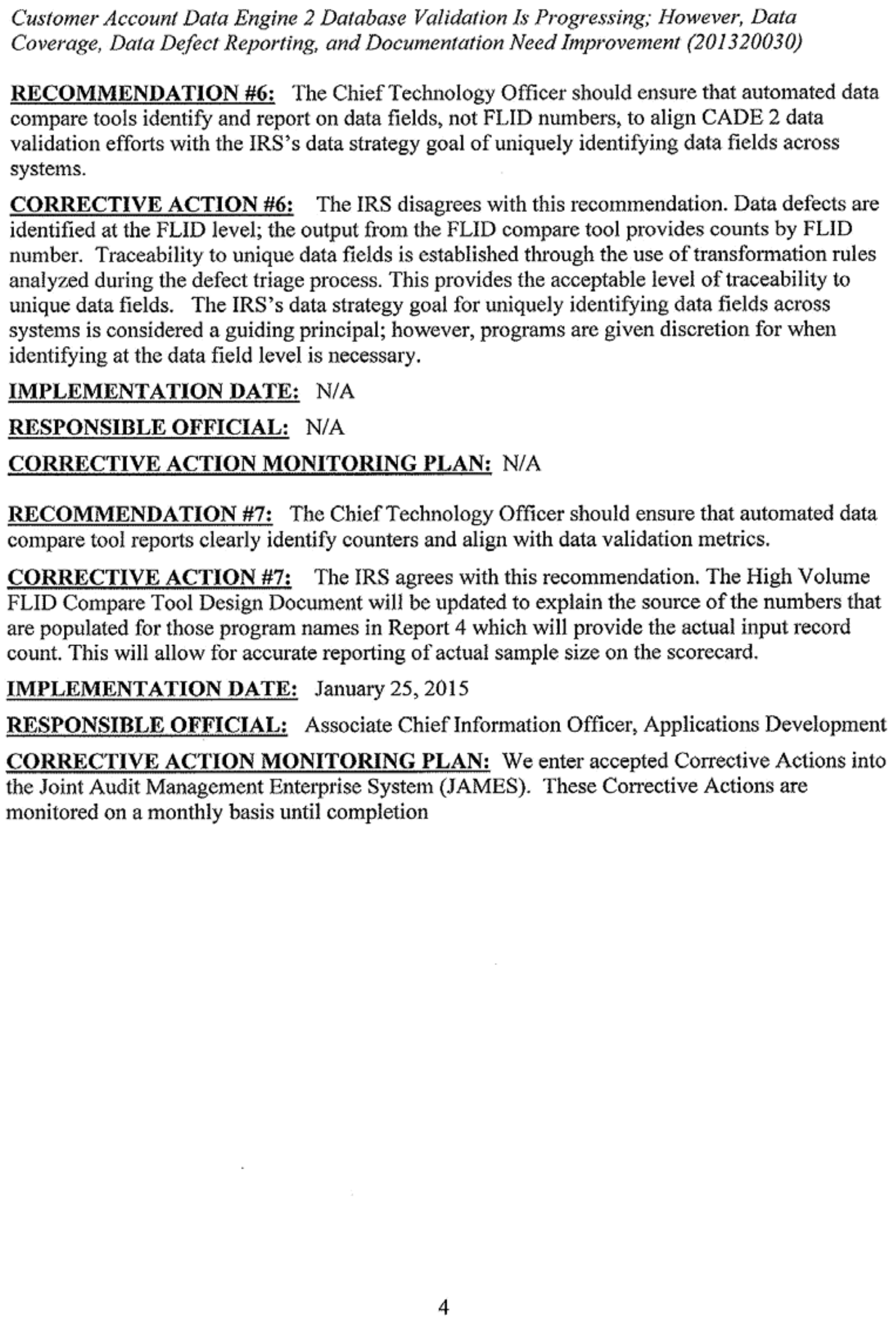
CORRECTIVE ACTION MONITORING PLAN: We enter accepted Corrective Actions into the Joint Audit Management Enterprise System (JAMES). These Corrective Actions are monitored on a monthly basis until completion

RECOMMENDATION #5: _The Chief Technology Officer should ensure that all documentation needed to verify the data in the Data Quality Scorecard is stored for future reference and to provide the information needed for oversight activities, such as spot checks to confirm the accuracy of the Scorecard.

CORRECTIVE ACTION #5: The IRS agrees with this recommendation. The Data Quality Scorecard sources continue to be stored in a shared repository. The IRS has documented procedures for developing the scorecard, a checklist to verify the contents and begun storing all scorecard sources in a SharePoint repository. It is acknowledged that the IRS prioritized Data Validation execution (and the associated reporting) over documentation. During TIGTA's initial fieldwork, TIGTA highlighted Data Quality Scorecard source documentation gaps in the Initial Briefing Paper. The IRS agreed with these recommendations and worked to improve the traceability of its Data Quality source documentation. Documentation was provided to TIGTA on April 6, 2014 in response to the Briefing Paper included (see Appendix B for document details).

TIGTA indicated that they did not review the FLID Compare Tool – Extended Discrepancy Reports (i.e., Report 4s) or the Referential Integrity Check sources, as they were not provided until April 21, 2014 and April 28, 2014, respectively. The IRS will continue to store all source documentation for the Data Quality Scorecards in the SharePoint repository and will ensure that it remains organized and easily accessible.

IMPLEMENTATION DATE: January 25, 2015
RESPONSIBLE OFFICIAL: Associate Chief Information Officer, EITPMO
CORRECTIVE ACTION MONITORING PLAN: We enter accepted Corrective Actions into the Joint Audit Management Enterprise System (JAMES). These Corrective Actions are monitored on a monthly basis until completion



Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement (201320030)

RECOMMENDATION #6: The Chief Technology Officer should ensure that automated data compare tools identify and report on data fields, not FLID numbers, to align CADE 2 data validation efforts with the IRS's data strategy goal of uniquely identifying data fields across systems.

CORRECTIVE ACTION #6: The IRS disagrees with this recommendation. Data defects are identified at the FLID level; the output from the FLID compare tool provides counts by FLID number. Traceability to unique data fields is established through the use of transformation rules analyzed during the defect triage process. This provides the acceptable level of traceability to unique data fields. The IRS's data strategy goal for uniquely identifying data fields across systems is considered a guiding principal; however, programs are given discretion for when identifying at the data field level is necessary.

IMPLEMENTATION DATE: N/A

RESPONSIBLE OFFICIAL: N/A

CORRECTIVE ACTION MONITORING PLAN: N/A

RECOMMENDATION #7: The Chief Technology Officer should ensure that automated data compare tool reports clearly identify counters and align with data validation metrics.

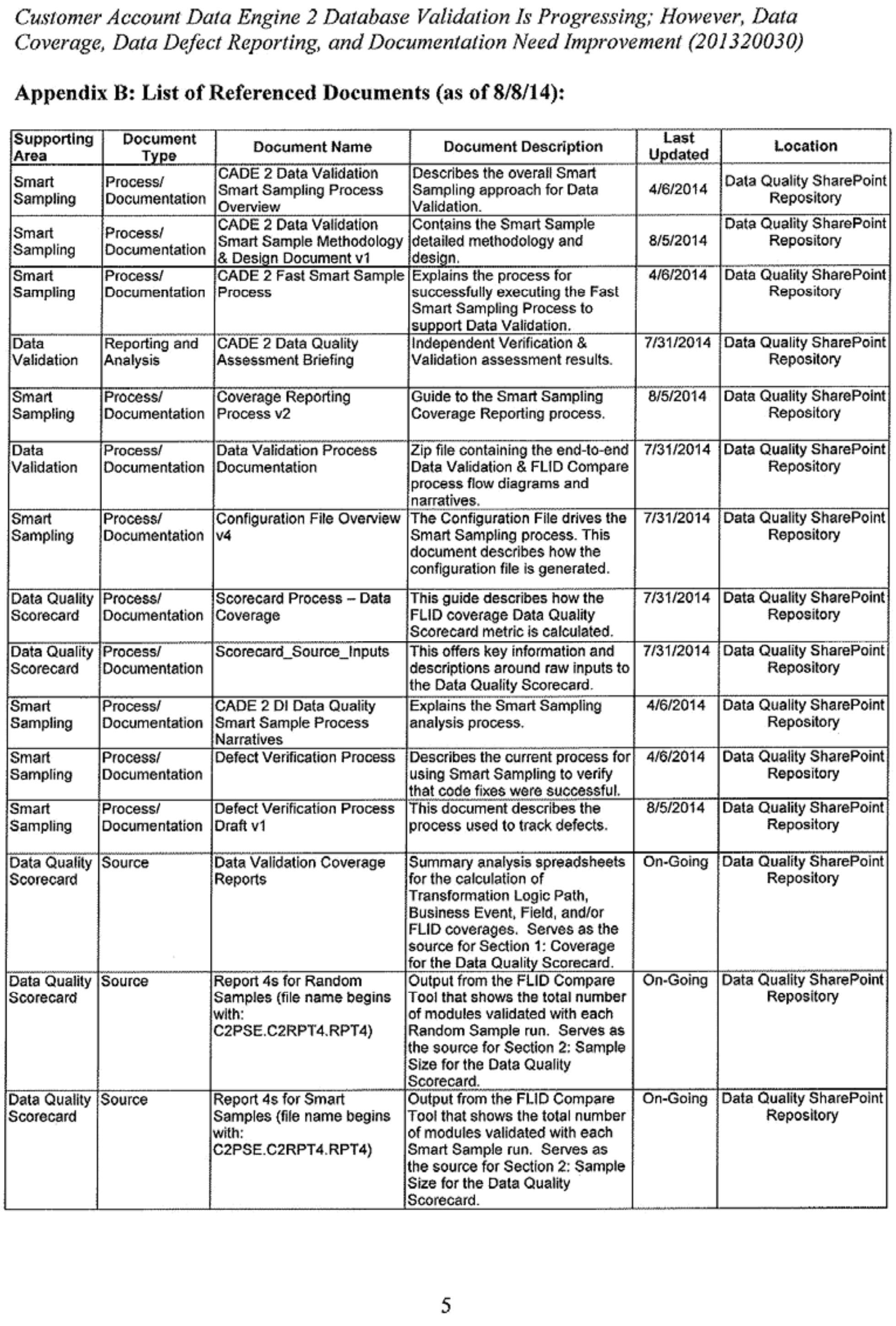
CORRECTIVE ACTION #7: The IRS agrees with this recommendation. The High Volume FLID Compare Tool Design Document will be updated to explain the source of the numbers that are populated for those program names in Report 4 which will provide the actual input record count. This will allow for accurate reporting of actual sample size on the scorecard.

IMPLEMENTATION DATE: January 25, 2015

RESPONSIBLE OFFICIAL: Associate Chief Information Officer, Applications Development

CORRECTIVE ACTION MONITORING PLAN: We enter accepted Corrective Actions into the Joint Audit Management Enterprise System (JAMES). These Corrective Actions are monitored on a monthly basis until completion

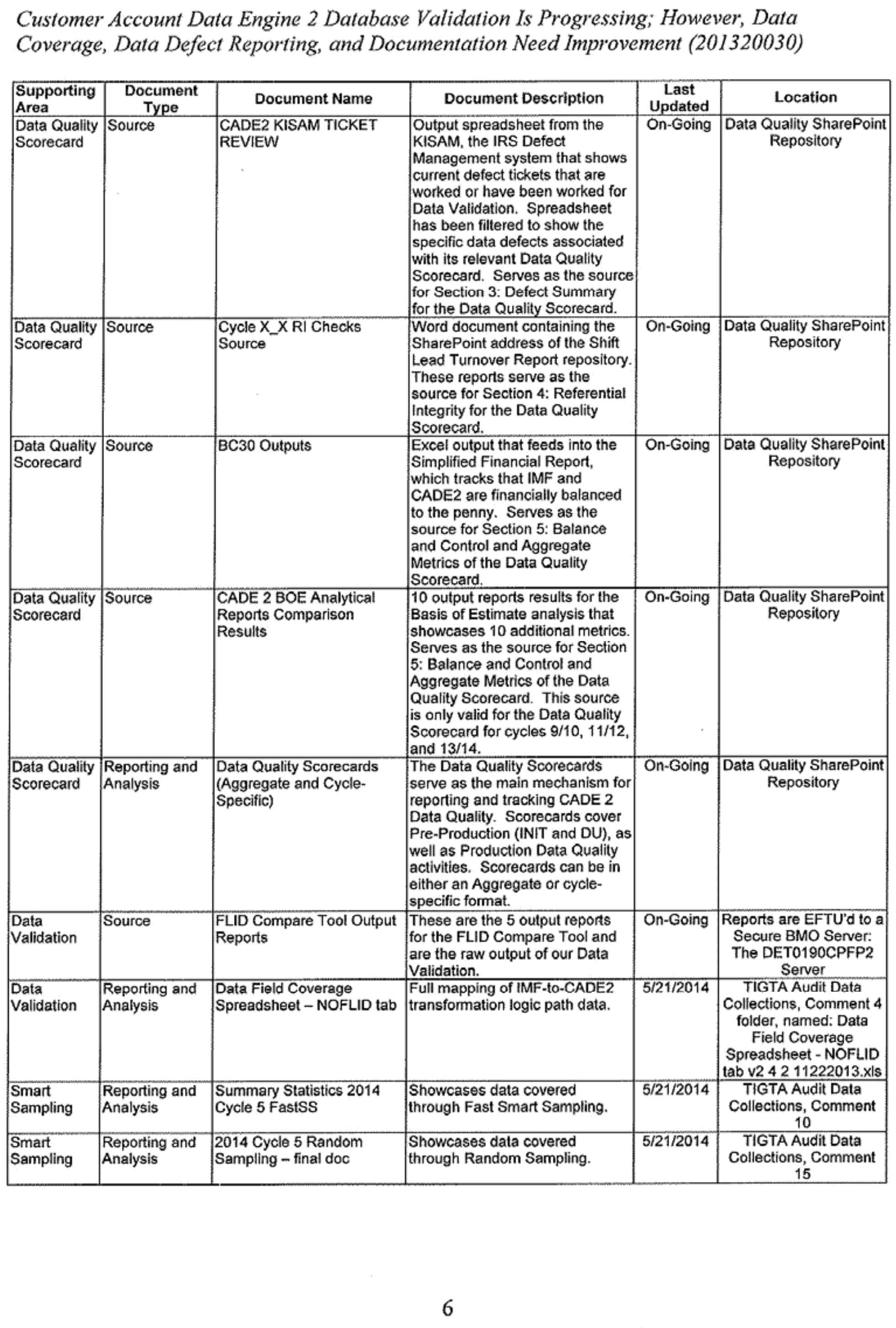
4



Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement (201320030)

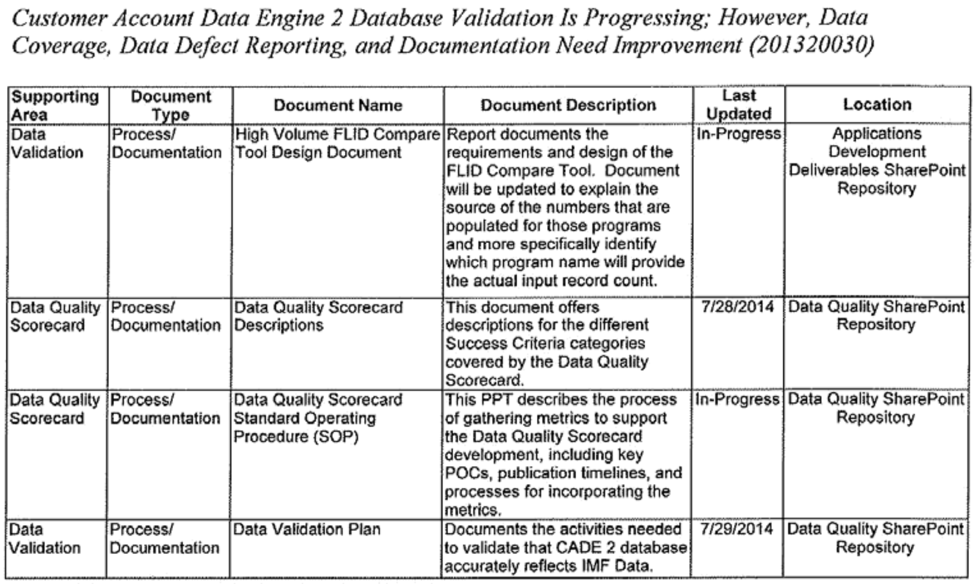
Appendix B: List of Referenced Documents (as of 8/8/14):

Supporting Area	Document Type	Document Name	Document Description	Last Updated	Location
Smart Sampling	Process/ Documentation	CADE 2 Data Validation Smart Sampling Process Overview	Describes the overall Smart Sampling approach for Data Validation.	4/6/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	CADE 2 Data Validation Smart Sample Methodology & Design Document v1	Contains the Smart Sample detailed methodology and design.	8/5/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	CADE 2 Fast Smart Sample Process	Explains the process for successfully executing the Fast Smart Sampling Process to support Data Validation.	4/6/2014	Data Quality SharePoint Repository
Data Validation	Reporting and Analysis	CADE 2 Data Quality Assessment Briefing	Independent Verification & Validation assessment results.	7/31/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	Coverage Reporting Process v2	Guide to the Smart Sampling Coverage Reporting process.	8/5/2014	Data Quality SharePoint Repository
Data Validation	Process/ Documentation	Data Validation Process Documentation	Zip file containing the end-to-end Data Validation & FLID Compare process flow diagrams and narratives.	7/31/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	Configuration File Overview v4	The Configuration File drives the Smart Sampling process. This document describes how the configuration file is generated.	7/31/2014	Data Quality SharePoint Repository
Data Quality Scorecard	Process/ Documentation	Scorecard Process – Data Coverage	This guide describes how the FLID coverage Data Quality Scorecard metric is calculated.	7/31/2014	Data Quality SharePoint Repository
Data Quality Scorecard	Process/ Documentation	Scorecard_Source_Inputs	This offers key information and descriptions around raw inputs to the Data Quality Scorecard.	7/31/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	CADE 2 DI Data Quality Smart Sample Process Narratives	Explains the Smart Sampling analysis process.	4/6/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	Defect Verification Process	Describes the current process for using Smart Sampling to verify that code fixes were successful.	4/6/2014	Data Quality SharePoint Repository
Smart Sampling	Process/ Documentation	Defect Verification Process Draft v1	This document describes the process used to track defects.	8/5/2014	Data Quality SharePoint Repository
Data Quality Scorecard	Source	Data Validation Coverage Reports	Summary analysis spreadsheets for the calculation of Transformation Logic Path, Business Event, Field, and/or FLID coverages. Serves as the source for Section 1: Coverage for the Data Quality Scorecard.	On-Going	Data Quality SharePoint Repository
Data Quality Scorecard	Source	Report 4s for Random Samples (file name begins with: C2PSE.C2RPT4.RPT4)	Output from the FLID Compare Tool that shows the total number of modules validated with each Random Sample run. Serves as the source for Section 2: Sample Size for the Data Quality Scorecard.	On-Going	Data Quality SharePoint Repository
Data Quality Scorecard	Source	Report 4s for Smart Samples (file name begins with: C2PSE.C2RPT4.RPT4)	Output from the FLID Compare Tool that shows the total number of modules validated with each Smart Sample run. Serves as the source for Section 2: Sample Size for the Data Quality Scorecard.	On-Going	Data Quality SharePoint Repository



Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement (201320030)

Supporting Area	Document Type	Document Name	Document Description	Last Updated	Location
Data Quality Scorecard	Source	CADE2 KISAM TICKET REVIEW	Output spreadsheet from the KISAM, the IRS Defect Management system that shows current defect tickets that are worked or have been worked for Data Validation. Spreadsheet has been filtered to show the specific data defects associated with its relevant Data Quality Scorecard. Serves as the source for Section 3: Defect Summary for the Data Quality Scorecard.	On-Going	Data Quality SharePoint Repository
Data Quality Scorecard	Source	Cycle X_X RI Checks Source	Word document containing the SharePoint address of the Shift Lead Turnover Report repository. These reports serve as the source for Section 4: Referential Integrity for the Data Quality Scorecard.	On-Going	Data Quality SharePoint Repository
Data Quality Scorecard	Source	BC30 Outputs	Excel output that feeds into the Simplified Financial Report, which tracks that IMF and CADE2 are financially balanced to the penny. Serves as the source for Section 5: Balance and Control and Aggregate Metrics of the Data Quality Scorecard.	On-Going	Data Quality SharePoint Repository
Data Quality Scorecard	Source	CADE 2 BOE Analytical Reports Comparison Results	10 output reports results for the Basis of Estimate analysis that showcases 10 additional metrics. Serves as the source for Section 5: Balance and Control and Aggregate Metrics of the Data Quality Scorecard. This source is only valid for the Data Quality Scorecard for cycles 9/10, 11/12, and 13/14.	On-Going	Data Quality SharePoint Repository
Data Quality Scorecard	Reporting and Analysis	Data Quality Scorecards (Aggregate and Cycle-Specific)	The Data Quality Scorecards serve as the main mechanism for reporting and tracking CADE 2 Data Quality. Scorecards cover Pre-Production (INIT and DU), as well as Production Data Quality activities. Scorecards can be in either an Aggregate or cycle-specific format.	On-Going	Data Quality SharePoint Repository
Data Validation	Source	FLID Compare Tool Output Reports	These are the 5 output reports for the FLID Compare Tool and are the raw output of our Data Validation.	On-Going	Reports are EFTU'd to a Secure BMO Server: The DET0190CPFP2 Server
Data Validation	Reporting and Analysis	Data Field Coverage Spreadsheet – NOFLID tab	Full mapping of IMF-to-CADE2 transformation logic path data.	5/21/2014	TIGTA Audit Data Collections, Comment 4 folder, named: Data Field Coverage Spreadsheet - NOFLID tab v2 4 2 11222013.xls
Smart Sampling	Reporting and Analysis	Summary Statistics 2014 Cycle 5 FastSS	Showcases data covered through Fast Smart Sampling.	5/21/2014	TIGTA Audit Data Collections, Comment 10
Smart Sampling	Reporting and Analysis	2014 Cycle 5 Random Sampling – final doc	Showcases data covered through Random Sampling.	5/21/2014	TIGTA Audit Data Collections, Comment 15



Customer Account Data Engine 2 Database Validation Is Progressing; However, Data Coverage, Data Defect Reporting, and Documentation Need Improvement (201320030)

Supporting Area	Document Type	Document Name	Document Description	Last Updated	Location
Data Validation	Process/ Documentation	High Volume FLID Compare Tool Design Document	Report documents the requirements and design of the FLID Compare Tool. Document will be updated to explain the source of the numbers that are populated for those programs and more specifically identify which program name will provide the actual input record count.	In-Progress	Applications Development Deliverables SharePoint Repository
Data Quality Scorecard	Process/ Documentation	Data Quality Scorecard Descriptions	This document offers descriptions for the different Success Criteria categories covered by the Data Quality Scorecard.	7/28/2014	Data Quality SharePoint Repository
Data Quality Scorecard	Process/ Documentation	Data Quality Scorecard Standard Operating Procedure (SOP)	This PPT describes the process of gathering metrics to support the Data Quality Scorecard development, including key POCs, publication timelines, and processes for incorporating the metrics.	In-Progress	Data Quality SharePoint Repository
Data Validation	Process/ Documentation	Data Validation Plan	Documents the activities needed to validate that CADE 2 database accurately reflects IMF Data.	7/29/2014	Data Quality SharePoint Repository