



AFRL-RH-WP-TR-2020-0075

**Strategic Policy Analysis Resource and Knowledgebase
(SPARK) Version 1.0 System Documentation**

**Johnny J. Weissmuller
Air Force Personnel Center
Strategic Research and Assessment Branch**

**Kristin A. Cazares
Osi Vision, LLC**

**July 2020
Interim Report**

DISTRIBUTION A: Approved for Public Release.

**AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in the document for any purpose other than Government procurement does not obligate the U. S. Government. The fact that the Government may formulated or supplied the drawings, specifications, or other data, does not license the holder, or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may in any relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RH-WP-TR-2020-0075 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH A SIGNED DISTRIBUTION STATEMENT.

CARRETTA.TH
OMAS.R.12289
84929

Digitally signed by
CARRETTA.THOMAS.R.12
28984929
Date: 2020.09.17 14:17:50
-04'00'

THOMAS R. CARRETTA
Work Unit Manager
Performance Optimization Branch
Airman Biosciences Division

WILLIAMS.LOG
AN.ANDREW.12
73597634

Digitally signed by
WILLIAMS.LOGAN.ANDRE
W.1273597634
Date: 2020.09.17 13:45:15
-04'00'

LOGAN A. WILLIAMS, PhD
Core Research Area Lead
Performance Optimization Branch
Airman Biosciences Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YY) 28-07-2020		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 1-05-2019 to 30-06-2020	
4. TITLE AND SUBTITLE Strategic Policy Analysis Resource and Knowledgebase (SPARK) Version 1.0 System Documentation				5a. CONTRACT NUMBER FA8650-18-F-6828	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER 62202F	
6. AUTHOR(S) *Johnny J. Weissmuller and **Kristen A. Cazares				5d. PROJECT NUMBER 5329	
				5e. TASK NUMBER 09	
				5f. WORK UNIT NUMBER H0SA (532909TC)	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **OSI Vision, LLC 100 Sandau, Ste. 210 San Antonio, TX 78216				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Airman Biosciences Division Performance Optimization Branch Wright-Patterson AFB, OH 45433				10. SPONSORING/MONITORING AGENCY ACRONYM(S) 711 HPW/RHBC	
				11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S) AFRL-RH-WP-TR-2020-0075	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution A: Approved for public release. 88ABW-2020-2866, cleared 15 September 2020					
13. SUPPLEMENTARY NOTES Report contains color.					
14. ABSTRACT This report documents the data management practices for the current Strategic Policy Analysis Resource and Knowledgebase (SPARK) Version 1.0 system, authorized by Air Force Manual 36-2664, Personnel Assessment Program, dated 16 May 2019. The purpose of SPARK is to support strategic studies by the Air Force Examining Activities Program (PE 81713F) to evaluate current or proposed personnel policies to improve "person-job-match" and general human capital management from pre-accession through retirement.					
15. SUBJECT TERMS SPARK, DRILS, DICE, FIDO					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT: SAR	18. NUMBER OF PAGES 105	19a. NAME OF RESPONSIBLE PERSON (Monitor) Thomas R. Carretta 19b. TELEPHONE NUMBER (Include Area Code) N/A
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			

Table of Contents

PREFACE	iii
1.0 SUMMARY	1
2.0 INTRODUCTION	2
2.1 The Categories of SPARK Studies	2
2.1.1. Related Scientific Professional Interactions and Studies.....	3
2.2. Overview of the SPARK System	4
2.2.1. The SPARK Databank.....	5
2.2.2. The SPARK Multi-Media Library.....	6
2.2.3. SPARK as a Software Package and Research and Development (R&D)	
Center 6	
3.0 SPARK SYSTEM AND WORKFLOW MANAGEMENT.....	8
3.1 SPARK Databank History, Management, and Design Philosophy.....	8
3.1.1. Translating the HRRD to SPARK Program Management Models	9
3.1.2. Data Classes: Operational, Psychometric, and Job Requirements Data	13
3.2 SPARK ID and Protecting Personally Identifiable Information (PII).....	14
3.2.1. SPARK ID to Support Cross-Master-File Record Matching	15
3.2.2. SPARK-ID Management Logistics	15
3.3 SPARK “Job Data Family” Management Approach	17
4.0 AFPC/DSYX’S SPARK UNIT’S MISSION, PURPOSE, & FUNCTION.....	21
4.1 Daily Functions of the SPARK Unit within DSYX.....	21
4.2 Overall Function/Goal of the DSYX SPARK Unit.....	21
5.0 NAVIGATING SPARK: DRILS & DICE	22
5.1 Data Retrieval Item Locator System (DRILS).....	22
5.2 Data Item Code Explanation (DICE)	23
5.3 DICE Index Content.....	24
5.4 DICE-AFSC.....	25
5.5 Goal of DICE	25
6.0 DSYX FILE TYPES: OVERVIEW.....	26
6.1 Master Files.....	26
6.2 Format of Master Files	27
6.3 How to produce a Master File	27
6.4 Occupational Research Databank.....	27
6.5 Project Files (Formerly, Study Files)	29
6.6 Project Request Research Support	29
7.0 PROTOCOL FOR SUBMITTING SPARK DATA PULL REQUESTS.....	31
7.1 Creating/Updating a Layout for an External Data Pull	33
7.2 Import and Export Data Tracking Log.....	34
7.3 Process for Pulling Archival Data (the importance of Date Ranges).....	35
7.4 Pulling Data from Controlled Test Material Files.....	36
8.0 MOU or DUA	38
8.1 Sample Data Use Agreement	39
8.2 Sample Memorandum of Understanding	42
9.0 SPARK’S DATA JOB FAMILIES: OVERVIEW.....	44
9.1 SPARK: Descriptions by Job Family & Refresh Rates	44
9.1.1. 00XX Infrastructure and Emerging Job Family Data Systems	44

9.1.2.	0500 Longitudinal Officer and Enlisted Personnel Data System (PDS)	44
9.1.3.	0600 Officer Accession Source Topics	44
9.1.4.	0700 Officer Personnel Topics	46
9.1.5.	0800 Enlisted Accession Source Topics.....	46
9.1.6.	0900 Enlisted Personnel Topics.....	47
9.1.7.	0950+ Special Projects	48
10.0	LAYOUT SYSTEM.....	51
10.1	Layout Versions.....	52
10.2	Data Cleaning Procedures and Auditing Process	56
10.2.1.	Auditing Process Steps	57
10.2.2.	Data Cleaning	58
11.0	RESULTS AND DISCUSSIONS	61
12.0	CONCLUSION	62
13.0	REFERENCES.....	63
	Appendix A: Instructions for Virtual (Windows XP Pro) Machine	64
	Appendix B: Access to AFPC/DSYX's Restricted Drive	66
	Appendix C: Point of Contacts (PoCs)	67
	Appendix D: Elements of Data Field Definition Lines in SPARK Layout Files	68
	Appendix E: Portal to SPARK: Layout Programs.....	75
	Appendix F: The History of SPARK Software.....	79
	Appendix G: SPARK Software Categories & Commands	81
	Appendix H: SPARK Data Pull Request Form	83
	Appendix I: SPARK Quick Reference Guide.....	84
	LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS	95

LIST OF FIGURES

Figure 1.	Examining Activities and AFCAPS	4
Figure 2.	AFHRL Divisions in 1984.....	10
Figure 3.	Manpower & Personnel Research Thrusts 1985	11
Figure 4.	Example of a DICE layout	24
Figure 5.	Example of a layout file in the Notepad program	53
Figure 6.	Sample Audit of a Variable Field	55

PREFACE

Per AFMAN 36-2664, Personnel Assessment Program, dated 16 May 2019, the Air Force Personnel Center/Strategic Research and Assessment Branch (AFPC/DSYX) is tasked with supporting the effort to optimize the United States Air Force (USAF) mission readiness and facilitating the total force capability in recruiting, selection, classification, training, promotion, utilization, and retention. AFPC/DSYX provides both broad institutional support and targeted process improvement. The “SPARK Unit” within AFPC/DSYX maintains the Strategic Policy Analysis Resource and Knowledgebase (SPARK) system that supports strategic policy analyses by the USAF Examining Activities program, located in HQ AF/AIPT.

A primary mission of the SPARK is to support the ongoing measurement and predictive analyses that ascertain the optimal fit of people, given particular attributes, into both the Air Force in general (accessions) and into specific jobs or career fields (classification). This umbrella mission has been pursued for the past 73 years in the Air Force personnel research community and is colloquially called “person-job-match.” In today’s terminology, this effort performs the analyses to evaluate Air Force personnel policies in order to shape future human capital management operations. This is policy analysis “due diligence” as envisioned in the Dixon Plan in September 1978 (USAFPP-1, The USAF Personnel Plan – Volume One – Personnel Management Objectives, 29 September 1978): Concept 7: *“Objectives, policies, practices, and procedures must reflect integrity, be authoritative, and be fully supported by sound rationale and a research and analysis capability.”*

The structure of the SPARK System is driven, in part, by the demands of these evidence-based, data-driven studies performed by the Industrial/Organizational (I/O) psychologists in the Strategic Research and Assessments Branch at the Air Force Personnel Center (HQ AFPC/DSYX). The SPARK system has three major components: the huge resource of a 50-year longitudinal databank of databases, state-of-the-art (since the 1950s) analytic methodologies supported in software, and a knowledgebase of scientific reports, historical studies, and analytical methodologies.

The cornerstone of the SPARK Databank is built with hundreds of Master and Project data files (800+ CD-ROMs) from more than three decades of personnel research at the Air Force Human Resources Laboratory (AFHRL). The 30-year legacy (1968-1998) of AFHRL’s Human Resources Research Databank ((HRRD) still drives many of the current SPARK Databank procedures (i.e., file structure and layouts, data cleaning procedures, co-signed Data Use Agreements (DUAs), data-pulls and update procedures). Details of these inherited and modified processes are provided herein as exemplars of the historical and current operational SPARK data processing practices.

This document completes Phase I of the SPARK development effort to record the existing, functional SPARK 1.0 procedures.

1.0 SUMMARY

This report documents the data management practices for the current SPARK Version 1.0 System authorized by Air Force Manual 36-2664, PERSONNEL ASSESSMENT PROGRAM, dated 16 May 2019. The SPARK databank supports AFPC/DSYX strategic studies to evaluate the effectiveness of personnel measurement, selection, and classification methods from pre-accession through retirement. Results of these studies provide support to the Strategic Personnel Assessment Program (HQ AF/A1PT Examining Activities in development of Air Force policy. The focus of this report is to document the current SPARK System as managed by AFPC/DSYX.

The SPARK databank is the successor to AFHRL's HRRD (1968-1998). SPARK procedures are driven in large part by the legacy HRRD system. Details of these inherited and modified processes are provided herein as exemplars of the historical and current operational SPARK data processing practices.

This document completes Phase I of the Strategic Policy Analysis Resource & Knowledgebase (SPARK) development effort to record the existing, functional SPARK 1.0 procedures. Note: Much of this material has been drawn from existing sources. The "HRRD-SPARK" system has been under constant revision since 2004, due to higher-level policy changes (e.g., requirements for handling of Personally Identifiable Information (PII), over and above the Privacy Act of 1974), changes in the dataset given to DSYX by data providers, and the emergence of new tests and assessments. Furthermore, a portion of the content documented herein is "in-progress", such as the replacement of all PII, including Social Security Account Number (SSAN), in both new and historical datasets managed in SPARK.

Phase II will use this document as a springboard to envision a 21st century design for the SPARK 2.0 System. It will manage (import, protect, store, retrieve, and analyze) the Air Force's historical datasets and also integrate all newly emerging datasets from tests and assessments developed by the Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX), the Air Force Research Laboratory (AFRL), their contractors, sister-services, the Department of Defense, professional colleagues, and international partners.

2.0 INTRODUCTION

The mission of the AFPC DYSX is to improve the compatibility between people and Air Force career fields (i.e., person-job-match) to improve human capital management for optimal Air Force mission capability.

2.1 The Categories of SPARK Studies

To support the AFPC/DSYX mission cited above, SPARK studies are generally performed in two primary categories:

Category 1: Program Evaluation

The first category is the exercise of due diligence in program evaluation by the Program Manager of the Air Force Examining Activities Program (EAP), (a.k.a. Chief of Air Force Testing and Survey Policy). This office budgets the funds paid to Air Force personnel in support of the United States Military Entrance Processing Command (USMEPCOM) for all pre-accession data collection on Air Force applicants as mandated in Department of Defense (DoD) Instructions and the Air Force Manual (AFMAN) 36-2664, 16 May 2019. The purpose of the program evaluation is to ensure that the tests given at the Military Entrance Processing Station (MEPS) and Military Entrance Test Sites (METS) are providing the Air Force with proper value. Longitudinal studies are performed to determine how well those tests predict success both in the Air Force at-large (i.e., a selection issue) and for classification purposes into specific Air Force training programs and subsequent career progression. Longitudinal studies are performed to evaluate the efficacy of the *test instruments and policies* for Air Force purposes, not as a measure of the success or failure of any given individual.

Category 2: Air Force Personnel Program Process Improvement

A second category of SPARK studies is for process improvement in Air Force personnel policies. These analysis projects span in scope from the broad Air Force-wide selection issues to targeted protocol development for screening candidates for classification of new recruits into specific Air Force Specialty Codes (AFSCs) or post-accession assignment decisions into special duty identifier roles, such as First Sergeant or Military Training Instructor. Examples of these studies include:

- 1) Creating large benchmark “applicant pool” datasets representing the personal attributes of the “qualified” Air Force applicants in various fiscal years;
- 2) Benchmarking “learning difficulty” for all Air Force career field training pipelines to help rationalize differential career field entrance requirements;

- 3) Realigning AFSC entrance requirements to minimize adverse impacts without a loss of operational readiness;
- 4) Recalibrating cognitive standards for combat career fields formerly closed to women;
- 5) Introducing Predictive Success Models (PSMs), (i.e., mathematical composite scores), measuring the quality of *person-job-match* by utilizing the following measures:
 - a) Cognitive (e.g., aptitude test scores);
 - b) Non-cognitive (e.g., personality, job interest, Air Force Core Value Compatibility);
- 6) Developing and deploying structured interview processes, such as the 360-protocol used for selecting Military Training Instructors and other positions.

2.1.1. Related Scientific Professional Interactions and Studies

Incidental to the above categories, reporting results in both official and professional venues is also accomplished. When proposed studies/papers are not in direct support of a DSYX mission and they plan to use Air Force data, they are vetted through official channels and abide with the relevant Institutional Review Board (IRB) and Protection of Human Subject protocols.

The Air Force Personnel Center's Strategic Research and Assessment Branch (AFPC/DSYX) serves in direct support of the USAF Examining Activities Program (EAP) per AFMAN 36-2664, 16 May 2019: *The Air Force Personnel Assessment Program*. The EAP is functionally located in AF/A1PT, Air Staff Directorate Force Management Policy for Accessions and Training. "Examining Activities" is a standard line item in the Air Force section of the DoD budget, directly adjacent to "Recruiting". In this capacity, AFPC/DSYX has inherited the 50 years of broad Air Force level personnel research missions formerly supported by the Air Force Human Resources Laboratory (AFHRL) and its predecessor organizations.

The Program Manager for the Air Force EAP is also the central manager for the Air Force Strategic Personnel Research Council, as established in AFMAN 36-2664. This function is called the Air Force Center for Applied Personnel Studies (AFCAPS) and includes input from AFPC/DSYX, the Air Force Research Laboratory, Air Force Recruiting Service, the USAF Occupational Analysis (OA) Program, the USAF Airman Advancement Program, etc. (See the USAF EAP, The AFCAPS and associated organizations in Figure 1 below. Note that some of these generic organizational labels include the other US Military Branches (Army, Navy, Marines, and Coast Guard), both directly and indirectly through the DoD level Under Secretary of Defense for Manpower Accession Policy Working Group (MAPWG). International connections include the International Military Testing Association (IMTA), which generates many individual allied interactions as well as

through various international committees, upon which Dr. Carretta at AFRL serves. Professional organizations include the Society for Industrial and Organizational Psychology (SIOP), the Performance Testing Council, and others. These contacts also generate professional studies on relevant topics of common interest.

Examining Activities: Scope and Reach for Leveraging Resources Air Force Center for Applied Personnel Studies (AFCAPS per AF/A1P, used in DTIC)

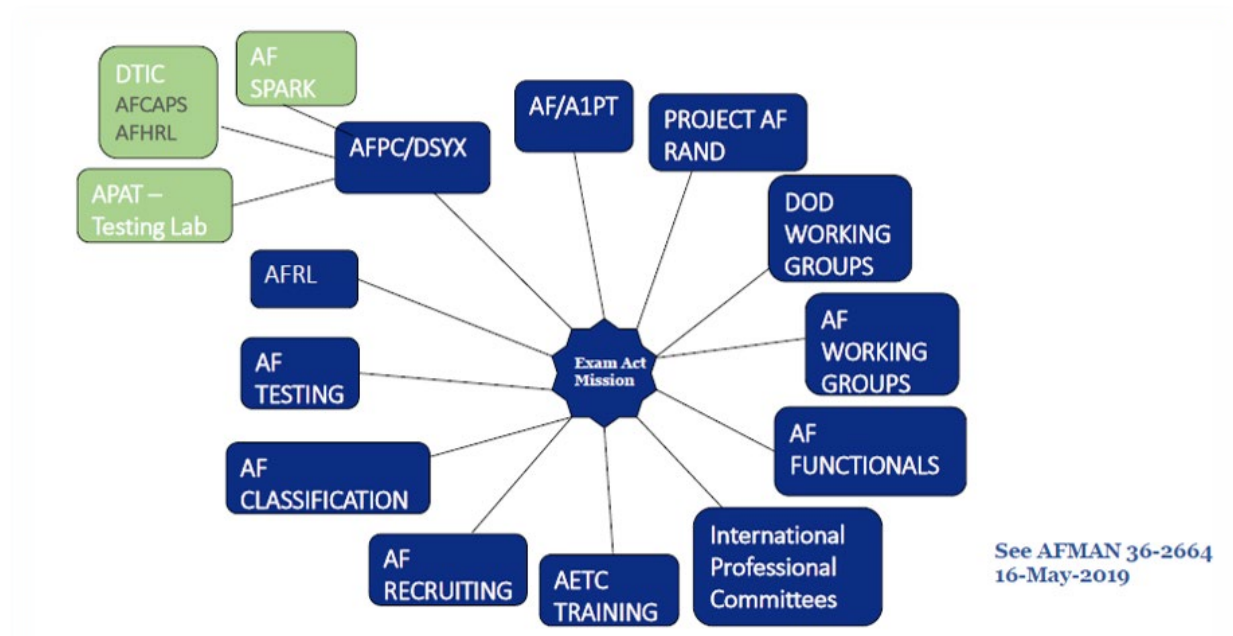


Figure 1. Examining Activities and AFCAPS

2.2. Overview of the SPARK System

AFPC/DSYX manages the SPARK databank to support the research, development, validation and monitoring of USAFwide personnel measurement, selection, classification, utilization, promotion, and proficiency evaluation processes. AFPC/DSYX manages SPARK resources to support improvements in human resources management processes while, simultaneously ensuring effectiveness, cost-benefit to the USAF and compliance with Air Force policies, professional standards, and strategic human resource management objectives.

To accomplish this mission, the SPARK system is structured into three components: the SPARK Databank, the SPARK Multi-media Library, and the SPARK Software Resource Center.

2.2.1. The SPARK Databank

In order to create a “PSM” for every Air Force Career Field (cited by Air Force Specialty Code (AFSC)), the SPARK Databank includes both person attribute data and job requirements data. “Person” attribute data includes pre-accession test scores, job interest ratings, success in training outcomes, career progression history, job satisfaction self-reports, and re-enlistment/attrition records. The SPARK Databank also includes job requirements data about each Air Force Specialty (AFS Structure, AFS job-interest profiles, literacy requirements, physical demands, and learning ability requirements).

At its central core, the SPARK Databank has integrated all Air Force historical data archives (formerly known as the Human Resources Research Databank (HRRD)), with 20 additional years of ongoing routine data-pulls from Systems of Record (SORs) from the Air Force, DoD, and other services. As the action-arm of the USAF Examining Activities Program, the PhD staff of AFPC/DSYX monitors both the current functioning of all Air Force personnel tests (i.e., program evaluation for tests in the Air Force Personnel Test (AFPT) Catalog) as well as monitoring the state-of-the-art in In I/O psychology to propose and evaluate strategic investment opportunities to improve operational Air Force personnel programs.

To support this professional analysis mission, the “SPARK Unit” within DSYX must obtain and securely manage all possible sources of available data regarding USAF applicants and current members to produce an array of studies and analyses in support of the development, revision, and validation of assessment screening tools and processes.

The SPARK Unit maintains ongoing and ad hoc data submissions from current Air Force and DoD sources, special data collection contracts, and historical data, providing access to records from data files that span over more than 50 years. The kind of data maintained by the SPARK Unit includes contractor-collected data from special projects, survey efforts, enlisted technical training data, AFSC job/task analysis data, aircrew training data, officer training data, and new data to be incorporated as required by HQ AF Force Management Division (HQ AF/A1PF) in support of future Force Management Policy analyses.

There are many sources of data, to include the following: Air Force Recruiting Service (AFRS), Basic Military Training (BMT), the USAF Air Education and Training Command’s (AETC’s) technical training, flying training records, job performance, and the emergent AETC training ecosystem. The goal of the AFPC/DSYX SPARK Unit is to obtain, audit, prepare in usable form, maintain, and document the necessary data required for analyses to support the Examining Activity’s function in the U.S. Air Force as well as the Air Force as a whole.

To that end, the SPARK Unit members maintain the SPARK System, which is a highly indexed and cross-linked databank containing current as well as historical and longitudinal databases

relevant to human resource policy topics. These policy areas include accessions, testing, personnel, performance, classification, AFSC task analyses, and training data used to support studies and analyses of issues impacting recruiting, accessions, personnel measurement, selection and classification systems, diversity, fairness, force development, force quality, utilization, and retention.

2.2.2. The SPARK Multi-Media Library.

The SPARK “knowledgebase” contains over 2,000 technical reports (scanned in their entirety and key-word indexed) that document the personnel research methods and studies performed by the AFHRL with its predecessors and successors. SPARK also includes documents from other services and the DoD, including multiple reports from Air Staff archives as well as their associated datasets (e.g., assessments across all AFSCs of Strength-Aptitude Requirements, Literacy Requirements, and other individual occupational analysis (OA) studies within each Job Family).

Other historical resources include hundreds of microfiche with archived Occupational Reports on Air Force Specialties, several hundred CD-ROMS indexing and archiving past research studies of the AFHRL from 1957—1998, (with limited hard-copy study-folders), and the library of professional personnel research books donated by the Institute for Job and Occupational Analysis (IJOA).

In addition to the historical archives, The SPARK Multi-Media Library includes current AFPC/DSYX contracted and in-house research published in DTIC, under our imprint (Air Force Center for Applied Personnel Studies, (AFCAPS)).

SPARK Multi-Media Library resources also include a video library with military personnel programs (e.g., 1957 AF Video on Specialty Knowledge Test Development and 1961 AF Occupational Analysis/Personnel Research), as well as a growing number of realistic job previews and animation-based Situational Judgment Test (SJT) items, circa 2018. The aforementioned resources are available on the AGPC/DSYX Restricted Network Drive or on the following sets of DVDs: AFHRL_Technical_Reports, briefings and other agency reports, Occupational_Research_Data_Bank (ORDB), USAF_officer_research, and the remotely-piloted-aircraft-and-pilot_selection_development (limited distribution).

2.2.3. SPARK as a Software Package and Research and Development (R&D) Center

The Applied Performance Assessment and Testing (APAT) Center is a psychometric laboratory established by AFPC/DSYX for the controlled development and use of experimental (i.e., *psychometric*) data. AFPC/DSYX collects data through the APAT center in order to create and validate replacement assessments as well as to develop new assessment instruments. Once the data flows to a designated psychometric SPARK Job Family, the dataset is deemed “experimental”

and per the Informed Consent Form established with APAT, never becomes part of a subject's personnel record or used in any official Air Force action. To reiterate what was emphasized previously in this document: AFPC/DSYX only releases *psychometric* data for policy evaluation purposes and model development, in accordance with established DUAs, or in the case of APAT, the Informed Consent Form.

The *operational* "personnel records" imported into SPARK are maintained solely for the statistical research or program evaluation purposes of the Program Manager of the Strategic Personnel Assessment Program (HQ AF/A1PT Examining Activities and the SPAP per AFMAN 36-2664). As a responsible steward of Air Force personnel data, AFPC/DSYX's data management plan ensures program effectiveness and compliance with all selection, classification, and professional testing standards as well as protection of Personally Identifiable Information (PII). An initial step in importing personnel data into SPARK is to remove PII and assign each *statistical record* a SPARK-ID capable of tracking and matching records across decades, regardless of original ID tracking (i.e. SSAN, DoD ID/EDIPI, Air Force Serial Number, Air Force Applicant ID, etc.). Once incorporated into the SPARK Databank from any external System of Record (SOR), the data becomes part of the databank of statistical observations, subject to modification only for statistical analysis purposes, and can no longer be released or used to make decisions on the rights, benefits, or entitlements of any individual. This is in accordance with the AFI 33-332, para 2.14.4.5 Privacy Act (k)(4) exemption which "...applies to records maintained solely for the statistical research or program evaluation purposes and which are not used to make decisions on the rights, benefits, or entitlements of an individual"

The primary focus of the current SPARK Development Project is the set of software and procedures used to maintain the SPARK Databank. The goal is to upgrade all these processes (anchored in the 1960s) to 21st century methods, while at the same time, maintaining the flexibility of a "databank" philosophy to index and retrieve data sets from diverse sources, each with unique challenges.

The portal to SPARK is the set of algorithms used for formatting and running SPARK programs, which will also facilitate the transition from the Turbo Pascal programming language to the more topical programming language that will be chosen to support AFPC/DSYX's psychometric and operational requirements in the near and more distant future. The foundation of the SPARK databank is the Data Job Families (included herein) and the Macros (utility and special programs included in the Appendix) to run them, which together serve as the basis for all analyses by the AFPC/DSYX.

3.0 SPARK SYSTEM AND WORKFLOW MANAGEMENT

This documentation describes the SPARK system, including History, Management, Design Philosophy as well as technical descriptions for external data imports, existing historical (Master File/Project File) data sources, and data capture procedures for internally or experimentally generated datasets. Also included are the indices of the archive of historical technical reports and documentation of the Job Family method for file cleaning and integration processes, file storage, timelines and frequencies of pulls and updates as well as the file structure for each data source.

3.1 SPARK Databank History, Management, and Design Philosophy

The SPARK Databank has incorporated Air Force historical databases from over 50 years ago to the present day, and continues to integrate new data streams as new sources and assessment instruments are developed. Some “people” datasets are deemed to be “predictor” data, such as personnel pre-accession test scores, job interest profiles, ability, personality assessments, etc. Other “people” datasets are “criterion” data, such as success in training, job satisfaction, and retention decisions. The other kind of data (i.e., “job” data), includes data collected to measure AFSC requirements, such as literacy (e.g., reading grade levels and need for specialized domain jargon), physical demands (e.g., strength and stamina), and learning difficulty demands from each training course pipeline. The SPARK resource services data requests from a diversity of organizations and individuals: DSYX staff I/O scientists, Career Field Managers (CFMs), the Air Force Institute of Technology (AFIT), the RAND Corporation’s Project Air Force (PAF), and from a variety of other Air Force Functionals and multi-level Working Groups. All of the aforementioned types of datasets must be in a format that facilitates rapid extraction into high priority projects, with a minimum “data cleaning” load for any requesting project analyst.

In accordance with the historical data management principles and Findable, Accessible, Interoperable, and Re-usable (FAIR) guidelines, DSYX developed both a Project Charter Protocol to scope emerging projects and a SPARK Data Pull Request Form to instantiate the data to be used in the project. The Project Charter documents the objective for a given project along with the contact information for all the Stakeholders and staff members working on the project. The SPARK Data Pull Request Form is completed by the DSYX project’s Principal Investigator (a.k.a., Task Scientist) and defines the kinds of data and date ranges to be covered. The SPARK technician will translate this into the actual datasets required as well as the dataset merge required to produce a de-identified project data file. The actual datasets used by the SPARK technician will provide the Project Analyst/Principal Investigator with references to the Job Family (with explanatory information), and layout (data dictionary) with the variable identification, format specification, and meanings for each value of coded variables. This data management practice facilitates the “finding” and “accessing” of SPARK data by research clientele.

3.1.1. Translating the HRRD to SPARK Program Management Models

Because SPARK incorporates the 30 years (1968-1998) of data and procedures from the HRRD that was managed by the AFHRL, any indices of these older resources use the then current role titles for personnel accomplishing the research program. Note that this SPARK Unit is a subset of the Strategic Research and Assessments Branch of the Air Force Personnel Center (AFPC/DSYX). In function, however, AFPC/DSYX is attempting to replicate the functionality of two divisions of AFHRL, specifically, the Manpower and Personnel Research Division and the Technical Services Division. The Technical Services Division aggregated decades of data into longitudinal databases and provided analysts to execute research projects requested by Task Scientists in the Manpower and Personnel Research Division.

Figure 2 shows the organizational breakout within each of these two aforementioned divisions. The SPARK “onit” within AFPC/DSYX provides the technical services indicated in the figure, except for the actual main frame computer (Computer Operations) and other hardware management (Information Resources). Not necessarily clear from the figure, the Computer Programming Branch included the units dedicated for collecting and archiving Master File data from all external sources as well as managing the layouts and code structures as they evolved over the decades. The work associated with layouts and code-tracking (File Item Data Overview, FIDO) will be discussed in depth later in this SPARK Phase I Report. The Computer Programming Branch also included several units dedicated to doing data analysis per task scientist requests. These project descriptions may have required the assigned analyst in the branch to translate notional data sources into exact available data sources (using a Generalized Data Index Table, (GDIT)), layouts, and FIDO code structures in order to create the dataset appropriate for the projects as described in the project request. Translation of these legacy systems (GDIT, FIDO, etc.) will be addressed in detail in sections that follow.

MANPOWER AND PERSONNEL DIVISION Col R L Kerchner Ext 2244 UF (MO)	TECHNICAL SERVICES DIVISION Dr R Bottenberg Ext 3841 GP (TS)
FORCE ACQUISITION BRANCH -- Dr L Valentine Ext 3256 UFA (MOA)	COMPUTER OPERATIONS BRANCH -- Mr R Martinez Ext 3931 GPP (TSC)
FORCE UTILIZATION BRANCH -- Dr R B Gould Ext 3648 UFB (MOD)	INFORMATION RESOURCE MANAGEMENT OFFICE -- Mr A Johnson Ext 3677 GPL (TSM)
TEST AND TRAINING RESEARCH BRANCH -- Vacant Ext 3570 UFD (MOE)	COMPUTER PROGRAMMING BRANCH -- Mr J Souter Ext 3928 GPO (TSO)
MANPOWER AND FORCE MGT SYSTEMS BRANCH -- Lt Col R Rue Ext 2912 UFC (MOM)	SCIENTIFIC & TECHNICAL INFORMATION OFFICE -- Ms N Perrigo Ext 3877 GPN (TSR)
PROGRAM SUPPORT OFFICE -- Maj S Trubshaw Ext 2244 UFF (MOS)	

Figure 2. AFHRL Divisions in 1984

Within AFPC/DSYX, the I/O Psychologist staff provides the functionality of the AFHRL Manpower and Personnel Research Division. Figure 3 provides the 1985 Thrusts for this AFHRL Division:

MANPOWER AND FORCE MANAGEMENT THRUST

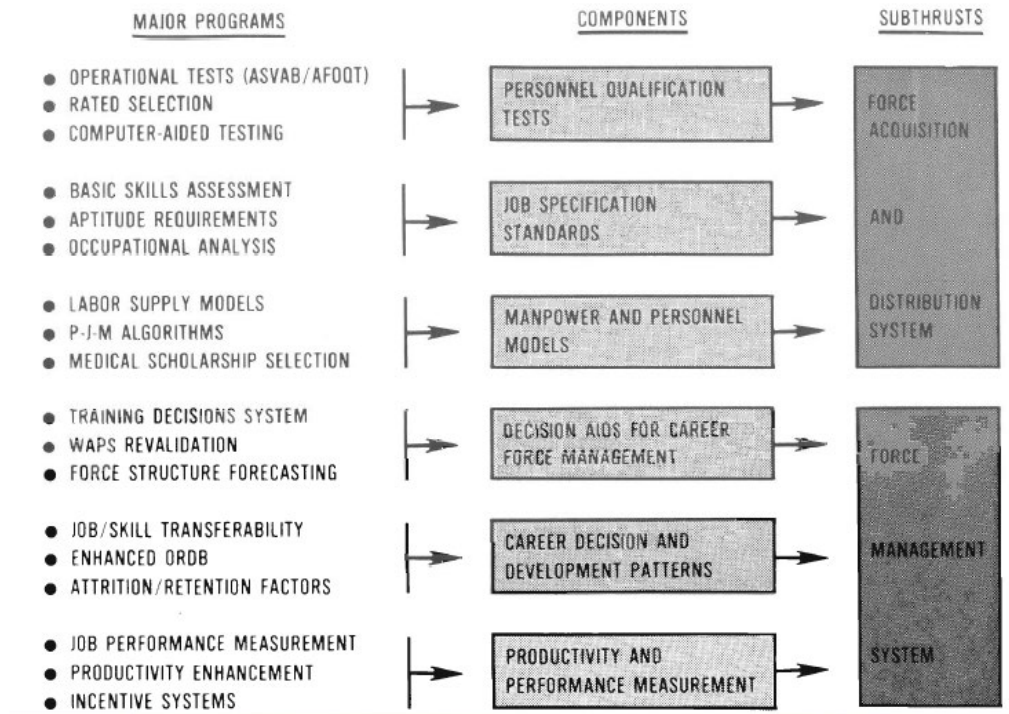


Figure 3. Manpower & Personnel Research Thrusts 1985

Below is a list of the new SPARK role titles that correspond to each identified AFHRL-period's equivalent job position:

- 1) SPARK Program Manager – AFHRL Computer Programming Branch Chief in the Technical Services Division. This person manages interactions with external agencies for establishing Data Use Agreements (or Data Sharing Agreements, or Test Use Agreements), as well as Privacy Impact Assessments (PIAs), System of Record Notices (SORNs), and any other Air Force, DoD, or federal requirements for disclosures of data maintained by the Government. All such documents are prepared for signature by AFHRL Division management (currently, AFPC/DSY Division Chief). The SPARK Program Manager reviews proposed projects/studies, clarifies requirements and suggested analysis needs, and then assigns projects to the appropriate analyst(s). This Project Review is comparable to the Project Charter Review currently conducted by the AFPC/DSYX Branch Chief. The focus here, however, is only on the SPARK Unit and therefore only involves review of the “SPARK Request Form”, discussed later in this report.
- 2) SPARK Data Importer – AFHRL Member of the Computer Programming Branch who requires specialized training and certifications to directly (or indirectly) access data from

our external sources as authorized in a Data Use Agreement (DUA). Currently, when these data are pulled from their external source, the new file is stored on the AFPC/DSYX Restricted Drive (DATA-TRANSER directory) and the SPARK Data Importer alerts the SPARK Data Manager (described below) to the new submission. In today's AFPC this is done by AFPC/DSYR (Data Retrievals) for *operational* data sources and AFPC/DSYX (Strategic Research and Assessments) for a combination of *operational* and *psychometric* data.

- 3) SPARK Data Manager – AFHRL Unit Team Leader in the Data Management Unit. Due to the new PII-handling requirements for SPARK, this person may be the POC for all first-level data importing, up through the removal of PII from imported datasets.
- 4) SPARK Job Data Family Specialist – Member of the Data Management Branch with responsibility for importing, normalizing, sanitizing, and updating Master Files within their Job Family areas. This person's focus is to update a Master File for subsequent use by other analysts responding to Project Charter needs.
- 5) SPARK Data Analyst – AFHRL System-Analyst/Programmer in the Computer Programming Branch running "study" projects as assigned in a Study Package written by the AFHRL Task Scientist (now aa AFPC/DSYX Staff Psychologist, typically a PhD I/O Psychologist) and approved by the AFHRL Branch Chief. This person translates the Project Charter and SPARK Data Request into actionable analyses. Note: Due to the analytical expertise and computational prowess of today's staff psychologists, often times the SPARK data analyst simply turns over the extracted custom SPARK project dataset to the requestor.
- 6) AFPC/DSYX staff psychologist (I/O psychologist) – the AFHRL task scientist working in the Manpower and Personnel Research Division. Many years ago, very few AFHRL task scientists did their own data analyses; however, in modern times, most I/O psychologists perform their own.
- 7) High-Level Sponsor – The USAF office or organization with a need for personnel research that will be translated into a research study package (currently called a project charter) by the AFHRL task scientist (now, AFPC/DSYX staff psychologist). In the past, this sponsor would write a Request for Personnel Research (RPR). A review panel consisting of the Technical Services Division Chief, the Computer Programming Branch Chief, and the USAF Examining Activities program manager would meet to establish the priorities for these RPRs over the next 18 months.

Currently, the SPARK program manager (PM) has inherited and adapted an ad hoc data management workflow with an adherence to FAIR principles of data management. In practice, AFPC/DSYX's SPARK databank is managed to support applied research purposes and outputs that can be evaluated and reused: SPARK data is managed by AFPC/DSYX to be found, accessed, reformatted, and reused by the USAF managers and researchers (cited above), who often request that the SPARK PM and/or staff psychologists develop PSMs for them, rather than receiving raw data extracts. By performing the requisite analysis and subsequent synthesis of a working model, DSYX provides process-improvement in support of AFMAN 36-2664, Personnel Assessment.

3.1.2. Data Classes: Operational, Psychometric, and Job Requirements Data

In concert with the principles of FAIR-ness, the current SPARK data management plan is guided by the Department of Defense Instruction (DoDI) 3200.12: DoD Scientific and Technical Information Program (STIP), 22 AUG 2013, yet with qualified exemptions. There are three classes of data that comprise the SPARK databank: *operational* data drawn from a System of Record (SOR) and *psychometric* data generated for development of new assessment instruments or data generated for revision of existing ones. As for the operational data, AFPC/DSYX is covered by the following exemption: DoDI 3200.12 does NOT apply to "DoD programs involving day-to-day operations the warfighter uses unless required for scientific and technical analysis." If dissemination is required of operational data, AFPC/DSYX refers the requesting agency to the originator of the dataset. As for psychometric data, AFPC/DSYX follows professional guidelines tailored to each work product by disseminating data to the highest degree possible. For example, in the case of a report that contains Controlled Test Material (CTM), AFPC/DSYX adheres to professional guidelines by publishing a Restricted Distribution document. AFPC/DSYX may follow that publication with a redacted version (sans CTM) that is published with Unlimited Distribution. AFPC/DSYX only releases psychometric data for policy evaluation purposes and model development, in accordance with established DUAs.

SPARK data is managed for "interoperability" and re-use, since information-requestors are able to reformat the data in accordance with their particular purpose, e.g., in the following possible formats: .CSV (text file using a special delimiter character), .TXT (text file with fixed field widths), .XLS (Excel spreadsheet), or .SAS (Statistical Analysis System dataset). SPARK procedures/macros exist to generate load scripts for other analysis tools such as the Statistical Package for the Social Sciences (SPSS), which is used at the Air Force Research Laboratory (AFRL) and by several contracting organizations used by AFPC/DSYX.

In summary, all documentation and dissemination of information from the SPARK databank is *informed* by DoDI 3200.12, as well as each unique DUA or Memorandum of Understanding (MoU) between AFPC/DSYX and partnering agencies. AFPC/DSYX endeavors to abide by best practices in data management at all times and as infrastructure and instruments (hardware, software, applications) evolve, the SPARK data management plan is modified in accordance with

the aforementioned guidelines, to support the new, ever-changing model. Though the data management procedures evolve with new requirements, the AFPC/DSYX mission remains steadfast: To facilitate USAF total force capability and to optimize mission readiness via support of the strategic policy analyses by the HQ AF/A1PT Examining Activities program.

3.2 SPARK ID and Protecting Personally Identifiable Information (PII)

Like SPARK's predecessor, the Air Force HRRD (1968-1998), every "personnel" SPARK Master File datum is stored in a flat ASCII format, (American Standard Code for Information Interchange). This means any data file can be inspected and easily read by humans with a NotePad or Text-based Editor. However, unlike the HRRD (which used a Social Security Account Number (SSAN) as the master key for all personnel records), SPARK uses its own uniquely generated identifier called the SPARK-ID. For maximum backwards compatibility, a SPARK-ID is a 9-character alpha-numeric value (the same size field as a Social Security Number), yet this unique identifier is not a simple encryption of the subject's SSAN. Moreover, unlike the SSAN, the SPARK-ID is not considered PII.

Upon receipt by the SPARK Data Manager, all data records from external organizations are moved to the secure off-network SPARK clean-room. The external SOURCE files are put through the appropriate import process. When the import is completed, the original SOURCE data file is archived in the secure off-network SPARK clean-room, in case reprocessing is required. Preparation of a releasable SPARK data file (i.e., a file which may leave the SPARK clean-room) has all SSANs, DoD IDs (formerly called Electronic Data Interchange Personal Identifiers (EDIPIs)) or any other PII (as well as other highly unique fields) removed, leaving only a single 9-character unique SPARK-ID. Standard PII removed (by label and known content) include: SSAN, DoD ID, Air Force Serial Number, and Recruiting Service Applicant ID. Other variables discovered in the standard import protocol with a high percent of unique values are also removed. These non-standard variables have included such items as Student ID, Unit Manpower Document (UMD) Position, and a variable labeled "ASSIGNMENT", which actually contains the 11-digit SSAN with dashes.

Once processed into the SPARK Databank, these records are only a statistical snapshot in time, not a personal record to be kept current or updated or replaced by newer data. Since these data are statistical observations, it is prohibited to pull individual records for the purposes of using that data within for any personnel action or to determine the rights, benefits, or entitlements of any individual. In SPARK operations, records are never individually pulled by SSAN, DoD ID or any other form of PII. Furthermore, PII and highly unique data elements have already been removed during the "import" protocol, long before release from the clean-room for any analysis. In operation, SPARK records are pulled from a SPARK Master file because of characteristics (variables) relevant in the aggregate to the Air Force policy being evaluated, such as "success of anyone who took a specified training course in 2017-2019."

3.2.1. SPARK ID to Support Cross-Master-File Record Matching

To be useful for policy analysis, every record in a releasable SPARK file must have a “SPARK-System-wide” unique SPARK-ID. For example, in any statistical analysis project, this requirement ensures that the predictor variables (such as pre-accession test scores) from one SPARK Job Family (0843) can be matched to their corresponding criterion value, e.g., Final School Grade (several years later) for the training course in a different SPARK Job Family (0801). These paired records are merged (based on a common SPARK-ID) into a single record for each person in the entire set of people in a specified training course. If some records do not find a match between the two datasets, they are simply dropped from the analysis. This analysis is not about individual performance, but instead, a search for a statistical relationship such as that between pre-accession test scores (a policy set prerequisite) and success in a training course (pay-off to the Air Force.) The set of matched and merged scores are run in a regression analysis to create an evidenced-based PSM. PSMs with a minimum passing score will then become codified in the Air Force Enlisted Classification Directory as a new additional requirement. This new classification requirement will screen future applicants for entry into that Air Force Specialty with that training course. This de-identified dataset used to create the PSM will be archived to document the basis for that new personnel policy for that particular Air Force Specialty. No product from this developmental analysis will ever go back into any record of the unknown individual’s, whose de-identified records were used in an evaluation that culminated in policy. In most cases, as this is evidenced-based policy analysis from the toolbox of Industrial/Organizational Psychologists, in order to be pulled for inclusion in this study, each subject/student must have already been selected for and finished the target course, and has therefore, passed the stage of their career for potentially being affected by any new resultant policy change.

3.2.2. SPARK-ID Management Logistics

Although the SPARK-ID is a 9-character alphanumeric string, the unique identifier is not simply an encoded form of the associated person’s SSAN. The first encounter of any new dataset can cause the generation of the “base” SPARK-ID for people in that dataset, regardless of the PII available in the dataset. Hence, if the DOD ID is available on that dataset, the SPARK Identify Master File (SIMF) will be searched for a matching DoD ID. If no match is found, the next available randomly generated SPARK-ID will be used, which also encrypts the flag that the source was a DoD ID. The possibility exists that an individual may have multiple SPARK-IDs until a dataset enters SPARK, which actually necessitates two or more PII elements to permit the “cross-walk”. Every Master File import is processed via the SIMF, at which time the SPARK-IDs are compared and the “original” SPARK-ID is propagated to the new Master Data File stored and indexed in the SIMF.

The SPARK Databank manages (oversees, indexes, imports, and integrates) a large collection of

diverse datasets from many outside sources, each with its own requirements, quirks, and history of changes in operation. (This process is explained later in the discussion on Job Families and their unique scripts for importing.) Regardless of its original database host system, data imported into SPARK generates a Master File stored as a flat, ASCII text file. In most cases, the external dataset is exported from its original format in order to provide SPARK with a CSV (Comma Separated Value) format file with the variable labels stored in row one of the data file. In some cases, the external dataset is provided in a fixed length field for each data item. Fixed length record files also come with a separate file/document called a data dictionary (or file layout), which delineates specifications for each data field in the file. From either format, data are tentatively imported into SPARK and become a fixed length record with a fixed width for each data field, and then is documented by a data dictionary called a “layout.” The SPARK version of a layout file is also a text file with one or more rows of specifications for each data field in the imported file. SPARK routines exist to accept the SPARK layout file, automatically produce a run-stream (batch file) to perform a data distribution on the imported file for each data field, inspect that field, and then recommend “reformat” actions and flag potential PII (and/or highly unique items) for an analyst’s review.

As each imported dataset is extracted from SPARK, its indexing key (usually SSAN or DoD ID) is matched against the current SIMF to ensure the first or “base” SPARK-ID is used to maximize match rates between other SPARK statistical datasets in the current analyses. Aside from not being PII, another advantage of the SPARK-ID is the following: Before SPARK, if two databases did not contain exactly the same PII elements for matching, those records could not be combined for statistical analysis. In contrast, with the SPARK-ID, as soon as another dataset is imported with two or more varieties of PII elements, a bridge is forged, and now the two datasets with only unmatching “unique” PII can be matched and used in a statistical analysis.

For an example, consider the pre-accession job interest values yielded from the Air Force Work Interest Navigator (AF-WIN). AF-WIN values are indexed by a person’s unique USAF Recruiting Service Applicant ID, and there is no other index PII key available on that dataset. By itself, then, that file cannot be used for evaluation if the AF-WIN scores result in better job satisfaction (and is worth the cost of “job preference” matching to the Air Force). When the USAF Recruiting Service swears-in a person and then ships that person to Basic Military Training (BMT), this action generates a record in the dataset (specifically, Job Family 0850-Recruit-RDB), which includes both the USAF Recruiting Service Applicant ID and the person’s SSAN. At the time of import, the SIMF is triggered to incorporate the SSAN into the existing SIMF record for this person/SPARK-ID. In this way, an applicant’s responses to Job Interest (pre-accession) can now be linked to a reported job satisfaction cited in an occupational survey taken after four years on the Air Force job. Another benefit to the SPARK-ID is that all people who took the AF-WIN yet who did not come in the Air Force can be “identified”, because those candidates have a SPARK-ID with only an Air Force Recruiting Service ID and no other PII. This differentiation allows the

I/O Psychologists to identify the interest profiles of people who do not come into the Air Force. The subsequent AFPC/DSYX analyses can then be shared with the Recruiting Service Market Analysis shop.

The master SPARK Identity Master File (SIMF, 0999-SIMF.DAT) is maintained in the SPARK clean-room. Each SPARK-ID record in the SIMF lists the previously encountered personal identifiers linked to this SPARK-ID (i.e., SSAN, DoD ID, Air Force Serial Number, and Recruiting Service Applicant ID). The original version of the SIMF was generated from the Personnel Data System (PDS, Job Family 0501) using the DoD ID as the preferred foundation for the “base” SPARK-ID for each person. PDS is an extract maintained by the Air Force Personnel Center and is a cumulative archive of both enlisted and officers in the current force, as well as those who have served in the past. Other than the SMIF, only the submission from external data providers maintain their complete PII in the SPARK clean-room; however, these “AS SUBMITTED” files are not releasable to anyone other than the SPARK Data Manager.

3.3 SPARK “Job Data Family” Management Approach

To best support a diversity of projects, the data handling portion of the SPARK system is structured as a “databank”, as opposed to a single “database.” To illustrate via an analogy, assume each data source (by organization) and “kind” of data is routinely collected and then stored within its own safety deposit box with its own access requirements. In SPARK, the safety deposit box is actually a computer directory labeled with a Job Data Family Number (typically called a “Job Family”, for brevity). Because of differing rules of engagement, there must be differentiation between the three primary classes of Job Family data that are imported into SPARK. *Operational* data are data record “snapshots” drawn from an official System of Record (SOR) belonging to an outside organization. *Psychometric* data are data generated incidentally in the development process for new assessment instruments and/or for revision of existing instruments by I/O psychologists within DSYX. (In other words, the generation of psychometric data for analysis by psychologists is akin to a “work product” for lawyers.) *Job Requirements/Attributes* data are records that document attributes of Air Force jobs or occupations, as opposed to data about people. These measures for Air Force Specialty Codes cover topics such as strength-aptitude requirements, literacy reading-grade level, as well as task and job learning loads.

Each SPARK class has its own rules for the “release” of data. For example, DSYX does NOT release *operational* data. Instead, AFPC/DSYX refers any requesting agency to the originator of the dataset to inquire about its release. AFPC/DSYX only releases *psychometric* data for policy evaluation purposes to qualified research teams in accordance with established Data Use Agreements. This data may also be released (de-identified) to duly authorized working-groups or action teams for statistical model development. *Job Requirements* data are rarely requested by users outside of the AF/C/DSYX professional staff.

The historical SPARK databank spans over 50 years and is comprised of Master Files, layout files, and other data files, all labeled for easy identification. The core SPARK databank of historical records is composed of records drawn from the official System of Record from the Air Force or from DoD sources. This databank includes several “Job Families” such as Accession Test Scores (Armed Services Vocational Aptitude Battery (ASVAB) (0843) from the Defense Manpower Data Center (DMCD), Tailored Adaptive Personality Assessment System (TAPAS) (0950) from MEPS, Cyber Test (CT) (0960) from MEPS, Air Force Recruiting Service “As Shipped to Basic Training” (0850) Data, Occupational Survey Data (0060) from Air Education and Training Command (AETC)-OA, Personnel Data System (PDS) Personnel (0501) Data from AFPC, Flying Training (0665) from AETC, Enlisted Technical Training (0801-ADSS) from AETC, Weighted Airmen Promotion System (0851) promotion scores, and Air Force Officer Qualifying Test (AFOQT) scores covering Job Families 0641, 0986, 0987, 0991, 0992, and 0993).

Prior to any data flow to a designated SPARK Job Family, the source agency establishes their SYSTEM OF RECORD repository system and then a Data DUA is negotiated establishing the data use (or sharing) relationship. All of these prior “records” are accrued to the aggregate databank of “statistical observations” in the designated SPARK Job Family, meaning that they are no longer deemed to be an official record on a person but rather an “example” statistical observation from a given point in time. As new observations are added, a transmission loss rate of 5% or less of incoming records is considered acceptable, i.e., within statistical limits. No attempt is made to replace earlier records with newer records for the same SSAN within any data family. AFPC/DSYX also analyzes “test—retest” dynamics, which means data is needed from all attempts at a given test to determine the stability of the test over time. Furthermore, these statistical observations may be changed to improve the quality of the data for a given analysis. AFPC/DSYX is also responsible for doing adverse impact studies, so if a record is missing demographic information such as gender, race, or ethnicity, it may be modified based on matching that data from other data sources which have that same data element.

These records are maintained solely for the statistical research or program evaluation purposes of the Program Manager of the Strategic Personnel Assessment Program (SPAP). To build on the analogy offered previously, each Job Family can be considered a different safety deposit box in the data “bank”, with a compartmentalized set of rules and regulations for access and use of the data within. This compartmentalization of the data is necessary because each imported dataset has its own unique set of protocols (and challenges) both for importing and later extraction for use in model building or other statistical analyses. Moreover, SPARK collects then imports, de-identifies, normalizes, and merges longitudinal data flowing in from a highly diverse set of operational and research organizations, each of which are consolidating data from numerous remote offices. Some data sources have known issues, like data originating on mark-sense answer sheets which reads the black marks in bubbled areas. Responding on bubble sheets is subject to

human error in key fields like SSAN, date of test, or the person's name being bubbled and scanned as "RQY" instead of "ROY".

SPARK is intended to maintain these large scale, longitudinal datasets in an ever ready, interoperable state to minimize any delay in responding to data requests from AFPC/DSYX staff psychologists who, in turn, respond to analysis requests from Air Force managers. This means that one group of people in the SPARK Unit deal with external agencies to secure and integrate snapshots of these operational datasets on a recurring basis. The goal is to minimize the workload and schedule delays for the AFPC/DSYX Project Analyst, who must use data from one or more Job Families to complete their project. Datasets are extracted from their Job Family (computer directory) then "matched and merged" into a de-identified Project dataset (a.k.a. a database) only when the requirements of a given project are clearly defined. Because the Project Analyst is not the person managing this data on a daily basis, the Job Family structure approach allows the SPARK Data Manager to document all the quirks of each dataset (in the Layout for that Job Family) to ensure the using analyst has access to all the rules of engagement with that data source.

Because SPARK Job Families cover datasets from both long-running operational programs and quick changing psychometric data from research programs, the SPARK Data Manager must vet each and every dataset turned over for import into the SPARK system. Some Job Families are easy to maintain, such as Job Family 0664 – TBAS and PCSM (Test of Basic Aviation Skills and Pilot Candidate Selection Method). The primary score extracted from this Job Family is the PCSM score (soon to be renamed the PSM-11x score) for screening pilot and other aircrew applicants. What makes this SPARK import "easy" is that the PCSM contractor always provides a complete copy of their database, which only goes back to 2007. (This database has less than 43,000 cases as of 01 Jul 2020.) In other words, there is no pre-existing Master File already in SPARK to which the new data needs to be integrated. All data provided (after PII removal) and conversion of dates to four-digit year formats will be made available upon the next release of the SPARK Master File for this Job Family (0664-20E.DAT).

Many SPARK source files are from long-running operational programs for which there are existing Master Files, going back decades. Operational organizations sometimes upgrade their mission and procedures (i.e., computer systems or software packages), which causes change in the data they collect and hence provide to the SPARK Unit. Rarely will a source agency provide a "heads up" summary on the nature or scope of changes. Often times, the source agency just processes its standing order (i.e. re-runs the existing script) for the SPARK data-pull. A common scenario is for the source agency to just "give you what we got", which requires the SPARK Data Manager to detect differences between the old standard and the new data submitted. Once the differences are identified, the SPARK Data Manager needs to decide what it means to add the new dataset to the existing longitudinal Master File system. Sometimes, the Master Layout is adjusted to include data fields from both the old and new standards with corresponding annotations. Some variables

may no longer be provided (like Age, for example, noted “no longer available since Jan 2020”) and others may have their coding standards changed (such as RACE from 6 categories to 31 categories, generating two variables “RACEOLD” and “RACENEW” with the appropriate annotations in Layout comments as well as the explicit codes listed). Whatever the SPARK Data Manager decides will be documented in the COMMENTS and NOTES in the new Master File Layout. In some cases, the changes may be so dramatic that the existing Job Family designation will be declared to be a Closed Archive, and an entire new Job Family will be created to accommodate very different import protocols from this point forward. An example of this major disconnect is the conversion from AFOQT Form S (Job Family 0987, Aug 2005-Feb 2015) into AFOQT Form T, with three new Job Families (0991, 0992, and 0993, Feb 2015-present).

At times, there are separate Job Families dedicated to IMPORTING highly unique raw data formats, yet that produce common scores that can be combined into a single database and EXPORTED for operational use from a different Job Family, which is the case for the AFOQT. This is an exceptional case, since AFPC/DSYX is responsible for the development and deployment of the AFOQT. DSYX developed all the pre-processing software to scan and score the AFOQTs, and then gave the software to AFPC/DP1 for operational usage under a different Job Family. The AFOQT test forms (Form S and Form T) later had unique demands, since the answer sheets were processed by the mark-sense scanner: Even though the tests appear different during scanning (i.e. one S answer sheet Air Force Personnel Test (AFPT) 987 versus two T answer sheets, AFPT 991 and AFPT 993), there are a small set of final AFOQT Composite Scores that are common among these tests, yet the ultimate use by the analyst can be delivered in an integrated format.

4.0 AFPC/DSYX'S SPARK UNIT'S MISSION, PURPOSE, & FUNCTION

The SPARK Unit is tasked with updating historical and current files on various data sources and DUAs. Further, the common functions of DSYX consist of auditing the data files pulled/received, creating layouts for data files when needed, creating/updating Master Files for Job Families as applicable, maintaining the Data Item Code Explanations (DICES)—(formerly, FIDO, which is the code file on all variables)—and creating/updating the Data Retrieval Item Locator System (DRILS), (formerly, the GDIT) Master File as well as the DRILS Roster. In addition to these routine functions, AFPC/DSYX is responsible for pulling data on an as needed basis for projects and analyses to support the Air Force Research and Examining Activity's requirements.

4.1 Daily Functions of the SPARK Unit within DSYX

- Continuously pulling/receiving data based on the agreements with internal and external agencies.
- Auditing new data to make sure it is usable and can be incorporated with the previous submissions of similar data within AFPC/DSYX.
- Updating/creating appropriate files to incorporate new data pulled/received.
- Updating and maintaining appropriate Master Files within AFPC/DSYX.
- Copying active AFPC/DSYX Master Files onto DVD/CD-ROMs for internal and offsite storage and backup of the internal AFPC/DSYX server.
- Negotiate new Memorandums of Understandings (MOUs) with outside agencies as needed to receive/pull necessary datasets in support of AF research.
- Pull data and appropriate layouts of data as needed for AFPC/DYSX's internal/external use.

4.2 Overall Function/Goal of the DSYX SPARK Unit

The SPARK Unit's function is to provide both quick and long-term internal and external/contracted research and analysis data support. This means AFPC/DSYX will either maintain within SPARK or have access to all data sources required to support force management analyses and needs.

5.0 NAVIGATING SPARK: DRILS & DICE

The DRILS and DICE files are used as indices of the longitudinal data stored by AFPC/DSYX. The DRILS is used to summarize and index the datasets maintained by DSYX, while the DICE file is used to index/explain the numerous variables used throughout the history of the AF. If either file is not accurately maintained and updated, then longitudinal studies would not be easily performed and AFPC/DSYX would not function properly.

5.1 Data Retrieval Item Locator System (DRILS)

The DRILS file (i.e., spreadsheet with tabs for each of the Job Families) can be utilized to navigate, identify, open, and save various parts of the SPARK databank. This file contains all historical and current Job Families, Master Files, Master File layouts, and DICE codes. There is an abbreviated list of current and historical datasets contained in SPARK, along with a brief description of what the dataset contains. The DRILS Roster summarizes the Job Families (datasets) maintained by DSYX. The Roster provides the applicable Job Family code that is used to identify datasets along with a brief description of what type of variables are contained in the dataset.

The DRILS index maintains details on all the Master Files for data that DSYX has established DUAs)or MOUs with various internal/external agencies. The DRILS index is within the SPARK databank and lists all datasets and Master Files that have currently and historically been maintained by DSYX as a longitudinal databank. DRILS also contains information about the appropriate layout of the data elements, description of the content of each Master File, a POC for each data source, the frequency of pulling/receiving the data, and Job Codes for the SPARK Master File families that are maintained by AFPC/DSYX.

The DRILS file must identify all Master File submissions within a Job Family Code (with corresponding date ranges) along with the appropriate layout for that Master File submission. A Job Family code is used to identify what the data represents or where it comes from. For example, the Job Code 0641 refers to AFOQT. Both CTM and regular extract files are referenced in the DRILS file, so when pulling data, it is important to take note of this difference as it effects how/where the data is pulled from.

The DRILS index is designed to explain each SPARK dataset (Job Family Code) and lists out every individual Master File within each Job Family Code. A DRILS entry or update needs to be done anytime a new Master File is created or appended to an existing Master File for any Job Family in SPARK.

- 1) The Job Family Code Name (Ex., AFOQT, Weighted Airman Promotion System (WAPS), etc.)
- 2) Job Number that corresponds to the Family Name (e.g., AFOQT is 0641).
- 3) Brief Summary of the type of data contained under that Job Family.
- 4) Detailed account of each Master File contained within that Job Family.
 - a. Contained in the each DRILS Master File listing is:

- i. FILE-NAME; TITLE; COVERAGE; SEQUENCE; LAYOUT; REC COUNT; REELS (if it is a historical file); BLOCK; LRL (REC LENGTH/6); NOTES.
- b. This includes:
 - i. **Master File Name** (standard format).
 - ii. **Title**
 - 1. Provides explanation of data contained in that Master File.
 - iii. **Coverage dates**
 - 1. Range of dates that Master File includes.
 - iv. Sequence (used for historical purposes)
 - v. **Layout file name**
 - 1. This is the master layout that goes with the Master File.
 - vi. REC Count (Record Count)
 - vii. LRL (Record Length of the file plus two for character linefeed)
 - viii. Notes/Remarks

(*NOTE: The items in **bold** above are those that are referenced when trying to pull a specific data file from SPARK. These allow for the proper file and layout to be identified and matched up. In addition, the number of reels corresponds to the number of times that file will be on a CD/DVD spindle within the historical archives of SPARK. So, for example, a DRILS Master File with four historical reels will have four separate files on an historical CD and each needs to be pulled, because they all contain the needed information. Reels basically determine how many pieces a file is saved in on the CD spindles.)

In addition to the DRILS Master File, a DRILS Roster needs to be maintained. The Roster is a condensed version of the DRILS Master that describes all SPARK Master Files, yet does not list in detail each Master File submission that is contained in a Job Family, as the DRILS Master does. Instead, the DRILS Roster will have a listing of all Job Family Code names followed by a brief summary of the type of data contained in that file/family. (Note: Most of this information will be exactly the same as that referenced in the DRILS Master File, but in a condensed version.) The difference between the Roster and the Master DRILS is that the Roster is easier to navigate and displays the type of data SPARK contains, because every individual Master File submission for each Job Family Code is not listed on the Roster as it is in the DRILS Master.

5.2 Data Item Code Explanation (DICE)

The DICE file is a current version of what used to be called the data dictionary. The file contains a list of all data elements that are found within the Job Family Master Files, both historically and currently maintained by AFPC/DSYX. This file maintains records of variable names and codes throughout the history of the Air Force, to allow quick reference and interpretation of otherwise meaningless numbers/alpha values listed within the data files.

The DICE code is a flat ASCII text file (Notepad) that contains all the appropriate codes for a particular variable. For each data element, there is the title of the variable/element (i.e., the DICE

code), the name of the variable (may be a brief description), the date the code started being used, and the ending date that the code stopped being used, if applicable.

Each file is saved as a name similar to the variable ID (must be 8 characters or less) in a format like XX-000. When looking through a layout variable, for example, “Academic Education Level (Highest) Applicant Education”, the format would have a specific variable ID and then a corresponding DICE code shown (Ex. AC-025). The layout indicates that to interpret what the dataset shows for this variable, the DICE code (AC-025) needs to be searched in the DICE code index to determine the meaning. Otherwise, the data just has values for that variable column that are untranslatable, or the coding may change from year to year and could cause confusion without standardization, especially when doing a longitudinal study.

The DICE code file supports virtually all facets of research and development for military personnel testing, because many efforts in this area involve the use of current and historical personnel databases. Maintaining a proper DICE code file allows for these efforts to continue in a timely manner by having one uniform document that references variable codes and values. Without this file, numerous hours would be spent trying to locate code values for variables within a dataset.

5.3 DICE Index Content

The DICE code file must contain the following information:

- 1) The code value that is shown in the data file/layout.
 - a. This code is identified only in the corresponding layout (Figure 4) of a data file and identifies which DICE code needs to be looked up in the DICE index in order to interpret the dataset for that variable.
- 2) Date that the code became active (YYMMDD).
- 3) Date that the code ended (YYMMDD) or 999999 if the code is still active.
- 4) Asterisk to denote the code was deleted or changed meaning on a certain date.
- 5) Asterisk to denote that more lines of description follow. (A blank in this column means this is the last line; an asterisk means the description takes up more than one line on the page.)
- 6) Spelled-out meaning for the code, e.g., “Clear Text”, description of the variables meaning.

DICE layout – AC-205		
1	890405999999	Currently in an AFIT BA Completion Program.
1	751101890405 *	Deleted this date 730401.
2	801001999999 *	Does not hold at least a BA degree and is currently in a USAFIT Sponsored non-degree program (See number 3 above).

Figure 4. Example of a DICE layout

Note: Notice how one column is for code value, one column is for start and end date, one column is for a first asterisk (if needed), one column is for a second asterisk (if needed), and the last column is for the spelled-out meaning/description of the code.

5.4 DICE-AFSC

The DICE-Air Force SpecialtyCode (AFSC; formerly known as the FIDO-AFSC) file is a specific type of DICE index that references only AFSC changes that have occurred throughout the history of the Air Force.

AFSC changes can occur 2-4 times per year now and this file maintains a record of any changes and the corresponding AFSC names that go along with the changes. This file is important to keep accurate and current, because of the need to tie historical data/AFSCs to the current data/AFSCs used for internal and external longitudinal studies.

There is a single DICE-AFSC file for all Enlisted members of the AF and a single DICE-AFSC file for all Officers of the Air Force. Each one needs to be accurately updated and maintained.

Each DICE-AFSC file (one for officers and one for enlisted) needs to contain the name of the AFSC, the AFSC number/code (as assigned by the Air Force), the date at which the AFSC started, the date at which the AFSC ended, and the previous AFSC name and code that it changed from (if applicable). In addition to maintaining separate files for officers and enlisted personnel, these two files need to be merged into one final DICE-AFSC file on a yearly basis. This allows for one overall historical DICE-AFSC to be utilized in the future for AFPC/DSYX purposes.

DSYX must combine the Officer and Enlisted AFSC changes into one Master DICE File each year.

5.5 Goal of DICE

The idea behind the DICE Code File is that every variable (that does not have a literal meaning) ever stored within SPARK has a corresponding DICE code and that code has an appropriate code structure to explain the element's meaning and dates of use.

Anytime data is pulled/received it needs to be properly saved, audited, and incorporated into appropriate files and layouts and stored in the appropriate DSYX directories as specified within this manual. The necessary updates to the DICE, DICE-AFSC, and DRILS files need to be made to reflect the current status of files, saved in the appropriate format and appropriate directory in SPARK, as specified in this document. CD-ROM copies of Master Files need to be created as needed and stored accordingly within AFPC and shipped appropriately to the offsite AFPC/DSYX storage facility.

AFPC/DSYX updates and changes are critical for being able to quickly locate sources of needed data and allow for proper identification of schema with the inclusive dates of specific elements, so data can be efficiently and accurately analyzed to respond to Air Force personnel research issues. The goal of AFPC/DSYX is to maintain and provide to contractors validated and easily usable data with sufficient documentation to permit immediate follow-on analyses. The proper documentation and files necessary to maintain and provide are outlined in the sections that follow.

6.0 DSYX FILE TYPES: OVERVIEW

Maintaining SPARK files as a longitudinal databank is pertinent to the overall functioning of AFPC/DSYX. If the DRILS, DICE, and master layouts are not properly maintained, updated, and saved appropriately, then AFPC/DSYX is greatly inhibited. The importance of these files cannot be over-emphasized, so it is necessary to distinguish between them and understand what their purposes are within AFPC/DSYX and the larger Air Force research function.

There are two types of files within SPARK, Project (formerly, “Study”) Files and Master Files. A Project File is a subset of a Master File, with particular parameters. Project Files are specifically generated for the person who requested the file. Once a Project File is requested and completed, the files are sent to the SPARK Program Manager’s drive on the network. He sends an e-mail to DYSX with the layout and the file. Master Files are more generalized and are disseminated to everyone. The Personally Identifiable Information (PII) is encrypted on all files, which minimizes the risk associated with the data. A Random User Identification (RU ID) is generated to work with encryption programs. All master files have a four-digit number, starting with zero (e.g., 0060, 0843, etc.). Project File numbers begin from 1,000 and are suffixed by a five-digit phase number (e.g., 1022-xxxxx). These numbers are used for the tracking of data within SPARK. Since the 1960’s, the first nine numbers in a layout file are the member’s social security number. Files are reformatted for the purpose of conducting analyses quickly. The source layout is the input and the target layout is the output (after the conversion program is applied to the change format). The conversion program for SPARK is written in Pascal. The input and output layouts are written in Notepad (text) documents. The eventual goal is to transition from the current Turbo Pascal language to a more topical programming language that serves the AFPC/DSYX administrators and technicians, as well as the customers they serve.

6.1 Master Files

A Master File is a CD-ROM version of a data file that encompasses all data for a specific Job Family within a specified date range. A Master File is a standard formatted data pull (either on a periodic or event-driven basis) from a data provider whose purpose is to augment the Master File section of SPARK. Data must be part of a recognized family in the DRILS Master Index, which is based on historical and current DUAs and MOUs. (See Section 7 for each Job Family of data.)

A Master File contains all data elements within a specified time frame for a certain Job Family, meaning that for every MOU/DUA that AFPC/DSYX currently and historically has had in place there needs to be a corresponding Master File that contains every individual record/entry within that file dataset during the specified time frame. A Master File’s purpose is to have a single consistent record for each individual/case that is compiled by record ID and by grouping down to the most recent/appropriate record. There are some cases where multiple records will exist for a single individual/case (for example, training courses taken, re-test scores, etc.) but the main idea is to have a single record of data for every individual within each data set.

6.2 Format of Master Files

A Master File is a “flat” ASCII file that is copied to a CD-ROM and is assigned and saved with a name consistent with the AFPC/DSYX file naming procedures. In addition to creating/updating Master Files as needed, each one needs to be copied and sent to the off-site storage facility per the DSYX contract and Air Force Instruction (AFI). This means every historical DSYX Master File needs to be copied onto a CD/DVD and sent to the off-site facility. This process should be done by using any default CD/DVD burn program that comes with a standard computer setup.

Every Master File within the SPARK databank, both historical and future ones, needs to be sent to the offsite storage facility on a quarterly basis and the files that are sent need to be documented in the Data Tracking Log under the export tab.

For every Master File there needs to be the following: An electronic version on the DSYX Restricted Drive (allows for quick access to data for Contractor/Project pulls), one CD-ROM (copy) stored internally with AFPC/DSYX, and one CD-ROM (copy) stored at the off-site storage facility for safety. The offsite storage facility will be at the Defense Manpower Data Center (DMDC),.

A Master File will contain PII and other unique identifiers to permit longitudinal cross file matches.

6.3 How to produce a Master File

A Master File is produced by combining the previous Master File submission of the Job Family with the current submission, after the appropriate auditing has been accomplished. If no previous submission exists for that Job Family or a new Job Family needs to be created, then audit the data to make sure it is usable, create a layout for that submission, and update the DICE and DRILS, as needed. The current submission will then need to be saved in the Master File format and indicated in the DRILS, layout, etc. that this is the now a Master File for the Job Family covering a specific date range.

A Master File is ultimately stored as a CD-ROM (one locally at AFPC/DSYX and one at the offsite storage facility) in addition to being saved on the internal server in the SPARK databank. Therefore, once the appropriate data file, Master File, and Master layout have been created, a CD-ROM needs to be burned that contains the Master file and Master layout for the Job Family.

The following link shows the spreadsheet used to track/document which Master Files have been created/copied to CD/DVDs. These include both newly created masters and archived Master Files that need to be copied and sent to the off-site AFPC/DSYX storage facility.

6.4 Occupational Research Databank

The Occupational Research Databank (ORDB) is part of the Master Files currently maintained by DSYX. As of October 1, 2009, ORDB became a recognizable Job Family (0999) within the DRILS and is updated on an annual basis (depending upon the specific ORDB data).

The ORDB consists of a 3-tier system of data that covers occupational research issues:

The first tier of the ORDB consists of classification data on both officers and enlisted airmen. This tier of AFPC/DSYX data is updated on a quarterly basis (after each classification directory is published) and simply updates the officer and enlisted classification directory (AFECD and AFOCD) that is published internally by AFPC and is available for download through the internal AFPC website.

The second tier of the ORDB is the Career Field Education and Training Plans (CFETPs). These documents are updated semi-annually and are available to download through E-pubs. All that AFPC/DSYX programmers need to do is go to that website and request access to the appropriate folder (through e-mail) and then download the necessary files for every AFSC. There are specific AFSCs/CFETPs that are not available for direct download. For these specific files, AFPC/DSYX must request additional access from the Air Force Occupational Measurement Squadron (AFOMS) before being able to download.

The third tier of the ORDB consists of AFOMS files. This tier is updated quarterly and contains all briefings, survey data, reports, etc. available from the AFOMS Community of Practice (CoP) webpage for direct download. DSYX's programmers can go to the website and request access to the appropriate folder to have direct download capabilities of the files. There are specific AFSCs that do not have any files available without a special request beyond general access. For those specific AFSCs, AFPC/DSYX programmers need to contact AFOMS and request access in order to obtain the proper files.

The 3-tier system of ORDB data contains the following: AFOMS occupational analysis reports, CFETPs, and Officer and Enlisted Classification information. The data is pulled from AFOMS, E-Pubs, and AFPC internal sites. In order to properly maintain ORDB, the three tiers need to be updated as follows: Air Force Classification Directories (AFCDs) = Quarterly, CFETPs = semi-annually, AFOMS data = quarterly. When at all possible, appending new updates to the existing files is preferred; however, maintaining records of each iteration/update of the data as separate files may be more feasible.

The ORDB needs to be maintained on AFPC/DSYX's internal servers and a copy of the ORDB data needs to be burned to a CD/DVD as a Master File for internal and off-site storage of AFPC/DSYX materials. ORDB files need to be maintained as For Official Use Only (FOUO). The amount of data contained in the ORDB makes it difficult to monitor each document within the databank. Therefore, FOUO is applied to the entire databank and so also to any document within the databank. The intent is to make sure that the ORDB data distribution is limited and viewed only by those who have proper permissions.

ORDB files that are restricted access need to be stored separately from the rest of the ORDB files and labeled as such. These restricted documents should be treated as CTM and direct access to the information needs to be blocked through security/password protection, stored in a safe that is compliant with security standards, or saved to a secure server that meets DoD security standards.

6.5 Project Files (Formerly, Study Files)

Project Files are an integration of one or more Master Files, which have been subject to auditing (in terms of consistency with Master File specifications) and data cleaning procedures prior to dissemination to the Project Analyst. A Project File is created anytime a contracted study/project through AFPC/DSYX is completed with the use of its data. A Project File is produced by burning a CD/DVD to include the original source file of the data from AFPC/DSYX, the final data after manipulation/cleaning, etc. the file by the contractor/customer, and the report/results of the project rendered by the contractor. The Project File is then indexed accordingly within the Project Log and kept within the AFPC/DSYX safe for storage.

A “project” is actually a “sync for time”; in other words, a project could be for discussing policy issues and in that instance, would not necessarily translate to a data file. Yet, every SPARK data pull (discussed below) should be tied to a project and so accompanied by a DUA or MOU. Furthermore, AFPC/DSYX is currently in the process of creating Project Charters (e.g., for Space Command, etc.), which will be shared on the Restricted Drive.

Project Files, which may result from merging Master Files and new data sources, will be purged of all PII, leaving only a SPARK ID in the final file delivered.

6.6 Project Request Research Support

This section applies to ad-hoc pulls and data requests commonly done for AF/A1, AFPC, and outside contractors and agencies as requested/needed. These data-pulls and projects are entirely different from the routine master file processing that is the foundation of the DSYX longitudinal data. AFPC/DSYX maintains DUAs only for data that is needed to maintain the longitudinal aspect of DSYX’s projects and allows current and future AF data to be dovetailed back into the historical files. There are multiple systems that AFPC/DSYX has access to that contain various data points, but DUAs are not required to maintain access to the data. Outside agencies may request additional data that AFPC/DSYX has access to (yet not part of AFPC/DSYX files) that needs to be incorporated into the longitudinal data that *is* part of AFPC/DSYX. The process is described below:

1. Step 1 – Initial request (from whomever receives it) needs to be given to AFPC/DSYX personnel so a project number can be assigned to it. The request needs to be logged into the internal AFPC/DSYX files that are used for tracking/records purposes.
2. Step 2 – AF/C/DSYX needs to negotiate with requesting agency to determine exactly what data elements are available and what is being requested, determine what exactly DSYX will provide (what Master Files/data type/Job Family, specific time periods, specific variables, etc.), and determine a relative timeframe in which to expect the data from AFPC/DSYX.
3. Step 3 – AFPC/DSYX will pull appropriate data elements from all necessary sources (internal AFPC/DSYX files that may be historical or current, external databases DSYX has

access to, our other outside agencies that DUAs exist for) based upon the Customer Requirement Data Form. Upon pulling all appropriate data elements the data needs to be merged into a single file.

4. Step 4 – An appropriate layout needs to be created that summarizes the merged file from Step 3. This layout should follow the same AFPC/DSYX standards as outlined in Section 8 of this manual.
5. Step 5 – An audit needs to be run on the newly created file for review purposes. Any sensitive Privacy Act information needs to be removed accordingly.
6. Step 6 – Error checking and validation of the data and layout needs to be done based upon the audit results. If any corrections are warranted or manipulation to the data/layout is needed, make appropriate changes and re-run audit accordingly.
7. Step 7 – Once the final data file and layout are created, the data needs to be saved in an appropriate format consistent with AFPC/DSYX standards and sent to AFPC/DSYX personnel for review.
8. Step 8 – Once reviewed by AFPC/DSYX staff the project tracking log can be appended to reflect completion of the project and the file can be mailed to appropriate parties. Upon mailing the internal DSYX data tracking log should be updated.
9. Step 9 – Any queries or issues involving the data file from the contractor/requesting agency will be handled by AFPC/DSYX staff.
10. Step 10 – AFPC/DSYX should receive the project results (data results and reports) back from the requesting agency and these should be logged into the Project File and given to AFPC/DSYX staff.
11. Step 11 – Once AFPC/DSYX receives the data and project results, the files need to be merged with the original Project File (source data) that AFPC/DSYX created in Steps 2-8. These files should be merged into a Project CD with all the data and the findings from the project. This should be properly saved and labeled accordingly in the Project Files index (which is separate from the Master File indices).
12. Step 12 – The PM then provides the project CD to for review and editing by AFPC/DSYX staff in order to publish the project and results through DTIC.

7.0 PROTOCOL FOR SUBMITTING SPARK DATA PULL REQUESTS

When pulling data from SPARK for internal AFPC/DSYX projects or external agencies/contractor requests, the key is to make sure exactly what the agency is requesting is understood in order to pull the proper data to meet the customer's needs. To make this process as simple as possible, the customer should have a clear understanding of what SPARK contains.

The first step will be for the customer to examine the DRILS Roster to get an understanding of what data is contained in SPARK (may not be necessary by some customers who are familiar with the process). The tracking log (discussed below) is contingent upon the data requests AFPC/DSYX receives. In order to process those requests efficiently, there is a SPARK Data Pull Request form, which lists the basic parameters for a data-pull. This requirements document allows the customer to supply the information AFPC/DSYX needs to pull the appropriate data. Once the outside agency examines the DRILS Roster, the second step is to fill out the SPARK Data Pull Request form, specifying the types of records the agency is requesting to be pulled, the date range of the pull, the type of data sources it needs to be pulled from, the type of data file that needs to be created, along with other pertinent information. This procedure allows the external agency to specify in detail the type of information they desire and affords AFPC/DSYX the opportunity to tailor the data pull to meet the agency's needs. If all this is done properly it will save time and frustration by making sure the customer receives what they were looking for and AFPC/DSYX does not have to waste time pulling unnecessary data. AFPC/DSYX will save the SPARK Data Pull Request form along with a copy of the data sent to the customer, to maintain proper internal documentation.

There is a separate log for data pulls (on the same spreadsheet as the Project Log) with an important distinction: There are two kinds of data pulls: psychometric (research) and operational (scores). For the research pulls, the date range should be wide, going back a long way into the historical records. The operational pulls, on the other hand, the date should be as current as possible or the date range as narrow as possible. Though each data pull is unique, the following are general requirements for initiating a SPARK Data Pull Request Form:

- The SPARK Data Pull Request should be issued a week (5 business days) prior to the "needed-by" date.
- Each iteration will require 24 hours.
- The DUA or MOU must be current.
- If the request is for scores (operational), for example, by a Board, a list of SSANs must be submitted with the SPARK Data Pull Request Form for DSYX to pull operational scores for the Board. *Note of clarification: Requests from Boards are not actually serviced by any SPARK database* – in fact, the most common data request is by the OTS board reviewing candidates for accession as Rated or Non-Rated. Although SPARK does maintain these types of *statistical records*, any request by a board is only serviced by pulling data from the Official USAF system of record – to which we have direct access because developed the tests for turn over to an operational agency for actual use in Air Force decision-making.
- If the request is for research (psychometric), a definition of the pool of interest (for example, which career field, course ID, etc.) so that AFPC/DSYX can pull the SSAN's

from the training file within the date range required; another option is to supply the grade range of personnel in a particular career field within a specified date range, so that AFPC/DSYX can pull data from the Personnel File (e.g., AAE (enlisted) or BAE (officer), etc.) to identify a cohort from a training class (active duty sample).

- Parameters for the project must be submitted.

Most data pull requests invoke more than one pull (e.g., there is one set of predictor variables and one set of criterion variables), and so typically there is more than once source for a data pull. Once the SPARK Data Pull Request Form has been received back from the customer, AFPC/DSYX must approve the request before initiating the data-pull. Below is an overview of the steps AFPC/DSYX takes to pull data to fulfill requests:

1. Based upon customer request, extract the appropriate tables/data files from the DYSX DRILS Master Files and databanks. Examples include DMDC, AFOQT, Oracle Training Application (OTA), TTMS, etc. Depending on the customer's knowledge of AFPC/DSYX, their request may indicate exactly what Master Files to pull from or based upon the description of the data, DSYX may need to determine which is appropriate.
2. Make sure the requested years are defined when pulling the data to make sure the customer's needs are met (may be FYs, dates, graduation weeks, etc.).
3. Ensure that the appropriate data file types are pulled from the proper DRILS Master Files. Examples within a master file include but are not limited to: Active Duty, Officers, Airmen, Guard, Reserve, applicants, etc. This should be based upon the SPARK Data Pull Request Form.
4. Pull appropriate layouts that go with the Master Files used and locate necessary DICE codes for data descriptions.
5. Merge all requested/needed file types from whatever source into one file within the appropriate program. Record the steps taken to accomplish this, sources/files used, and variables cleaned/manipulated (and how) for tracking and re-creation purposes. Need to remove or flag duplicate files/records as appropriate using the group by key based upon customer request.
6. Retrieve or create the appropriate layout—(see layout Section 8)—(only if necessary, since some pulls may match an existing Master File layout) to match the data layout and provide proper DICE codes to explain variables within the data file. This layout should include variable names, codes, and record lengths for all variables needed to meet the customer's needs.
7. Save the file into the customer requested style (.TXT, flat ASCII, .CSV, Excel, etc.) and export the file to the personal computer (PC).
8. A quick audit needs to be run to examine the data in comparison to the layout. The audit should consist of running distributions on all variables in the layout. This is done to compare the output of the audit to the layout and DICE codes.
 - a. Examine the audit for the following:
 - i. Total size/number of cases – should match layout record length.

- ii. DICE codes match – make sure the DICE codes that are expected match what the data contains.
 - iii. Anomalies – make sure all cases contain data (as appropriate) and that date ranges, score values, etc., within the data match what would be expected based upon the layout and DICE codes.
- 9. Once the audit is complete, review the results. If there are problems, check the merged Master Files, layouts, etc. and re-run the audit. Once the audit reveals that no problems arise, then proceed to the next step.
- 10. Once audit is completed, ensure that the layout and DICE codes are in the appropriate format. Save the audit, layout, and DICE codes to an appropriate data file folder for this pull (will be utilized in the next step).
- 11. Create two copies of the CD/DVD (one for AFPC/DSYX's storage and one for the customer). Each CD should be labeled accurately, and the customer copy should be encrypted as necessary (via Privacy Act or CTM material guidelines). The CD/DVD should include the following:
 - a. Final layout with DICE codes
 - b. Final data file
 - i. File should be in requested form by customer or flat ASCII text file if no preference was given.
 - c. Audit Trail/Data – this should show the distributions and N sizes for each variable.
- 12. The customer's copy of the CD/DVD/file should be sent to AFPC/DSYX for final approval. Once approved, log the file into the import/export tracking log.
- 13. AFPC/DSYX will handle logging the file into their internal data tracking system and then provide/send the data to the customer, as appropriate.

The process of pulling the data needs to be properly TRACKED and AUDITED to ensure two things: 1. The appropriate data is pulled to satisfy customer needs; and 2. There is a clear audit trail to track how the data was manipulated, in case any questions/errors present in the future.

7.1 Creating/Updating a Layout for an External Data Pull

Initially, the AFPC/DSYX technician needs to send out the SPARK Data Pull Request form to the customer, so they can indicate what specifically needs to be pulled from the SPARK databank. Once the requested data pull is approved, then complete the following steps:

1. Extract the appropriate tables from the SPARK databank based on the customer's SPARK Data Pull Request form. This includes the data sets that DSYX maintains (e.g., DMDC, AFOQT, Oracle Training Assistant (OTA), Technical Training Management System (TTMS), etc.).
2. Identify the date range the customer needs and filter the data to include just those dates, which can be identified through the electronic DRILS Roster.

3. Identify that the type of data the customer wants pulled (e.g., officers, airmen, accessions, etc.) and filter the data set to include the content requested.
4. Join all necessary files together into single access file and remove duplicates using the group by key, IF APPROPRIATE.
 - Make sure these are actual duplicates and not multiple records for an individual. (For example, test dates/scores may appear multiple times for an individual but are not duplicates since they are separate cases/records for a single individual).
5. Re-verify that all the customer's needs were satisfied, and the appropriate data variables have been captured.
6. If all data is accounted for, save as a text file in the appropriate folder.
7. Once the correct data has been pulled, make sure a layout is created to reflect all variables pulled into the dataset.
8. Once completed, make a copy of the text file and layout and post to the DSYX Network's Restricted Drive and make a CD/DVD copy and send to the DSYX PM for approval. Once the data pull is complete the DSYX data export tracking log sheet needs to be updated to reflect the work that was done.

The steps above provide a general guideline and specific steps/elements may differ depending on the specific pull.

7.2 Import and Export Data Tracking Log

DSYX needs to maintain a data tracking log, one for importing/receiving/pulling data and one for exporting/sending data. This log needs to be maintained and updated anytime data is sent/received/pulled by AFPC/DSYX.

The following information should be contained in these logs:

- 1) Tracking number
- 2) Date project/pull/submission was Initiated
- 3) Customer (to whom it was sent/received/pulled from)
- 4) Method of reception or exportation
- 5) How the data was sent/received
- 6) Brief description of the data (file name)
- 7) Content of the data (SSANs, names, test scores, etc.)
- 8) DRILS Job Family code for associated DUA pulls (import log only)
- 9) Date the data/file was sent/received
- 10) *Contract number (for ad hoc/external projects)
- 11) Notes – include mail tracking number on exported data if possible

*The contract number is associated with the export log only and refers to the ad hoc/as-needed data pulls that are done for outside contractors. This number should be associated with the

contract/project/analysis for which the data is being pulled. This will help with making later “project files” for AFPC/DSYX if/when the contractor returns their data analyses, reports, etc. for the contract/project.

There are certain rules/differences for pulling specific types of data, for example, Controlled Test Material (CTM) and historical SPARK data that is not actively updated. In addition to logging the data into the import/export log, saving the output in the appropriate fashion and maintaining it in SPARK, it is necessary to append the data to the Project Files that the customer may send back to AFPC/DSYX. This will allow for proper Project files to be created and maintained by DSYX, containing the original data file created, the data output/manipulation done by the customer, the customer’s results/analysis, and any reports or outcomes that may have resulted from the project.

7.3 Process for Pulling Archival Data (the importance of Date Ranges)

The DSYX Restricted Drive has a “reasonable coverage” of data from around 1998-present – although some Job Families go back much further. The ASVAB dataset (Job Family 0843, a primary set of enlisted force predictor variables) goes back further but with national youth population norming changes (PAY 97) and a volatile set of the subtests (loss of NO & CS, adding AO) from 1998-2004, it is recommended to only pull ASVAB data from 2005 to the present. Our primary set of enlisted “training outcomes” is taken from the AETC’s ADSS (Archived Data Support System?) and its coverage spans from FY01-FY20 – that is October 1, 2000 to present. Other emerging data sets have their own coverage spans. The TAPAS scores began experimentally in 2009 for a few AFSCs. In 2014, TAPAS Baseline (*psychometric*) data collection rose to 80% of USAF applicants. After several Predictive Success Models (PSM) were experimentally validated, operational collection of the TAPAS began in January 2016. This is an *operational* TAPAS Job Family and became an official screener in the Air Force Enlisted Classification Directory (AFECD) and is used by Recruiting Service for bringing applicants into selected AFSC.

The process for pulling archival data is as follows:

- 1) Need to determine what type of data needs to be pulled and from what specific time frame (E.g., AFOQT from 1964-1998). This will be based upon the customer’s needs.
- 2) Based on this information, find the job code for the data (the same 4-digit job code used for labeling files) in the DRILS Historical Master Roster.
- 3) After identifying the correct job code, use the DRILS Master File to identify the specific file within that Job Code that captures the specific dates for the data needed.
- 4) Once the appropriate file is identified, look up that series in the historical master archives and determine which spindle the CD will be on. (Note: There are seven spindles of CD-ROMs in the historical SPARK archives.)
 - a. This is done by looking at each file contained in the spindle and identifying the correct .txt file. Once this is found, find the corresponding CD title (should be in a XXXX_XXXXXA or B format) on the spindle that contains the appropriate file that is needed. Items in bold indicate the file name and CD number found in the spindles. There is also a hardcopy of this file in a blue binder.)

- 5) Locate the spindles in the Scanner Room, find the correct spindle, and then locate the correct CD based upon the title identified in the previous step (Note: Please ensure the CD-ROMs stay in the indexed order found in the spindles and return them in the same order to ensure efficient future use of the spindles).
- 6) Pull out the appropriate CD on the spindle and copy the appropriate data file from the CD. At this point, you need to copy the appropriate master layout that goes with the corresponding data file needed.
- 7) After copying the correct file, identify the correct master layout that corresponds to the archived data file. This is done by seeing which layout goes with the master file identified in Step 3 (above).
- 8) Identify the layout series the file falls under, based upon the Job Family code and Master File name.
 - a. The master layout that needs to be copied is found on the Restricted Drive under DPPPWT/DICE/LAYOUTS
 - i. Once you access this folder, identify the appropriate series the layout/data falls under, and then open that directory.
- 9) From that directory, find the appropriate file (e.g., .doc/.docx or .txt) and copy those files.
- 10) Once the correct data file/s and layout/s have been found, copy all those files onto a CD/DVD and provide to the AFPC/DSYX technician.
- 11) Document in the internal data log if this CD is sent to anyone other than AFPC/DSYX.

7.4 Pulling Data from Controlled Test Material Files

In addition to the archived files not actively maintained by AFPC/DSYX, there are numerous files that contain CTM and are stored separately from the other AFPC/DSYX inactive CD-ROMs and files. Although these are stored in a separate area, the DRILS Master File does identify every Master File for a Job Family also indicates whether the file is considered CTM or not.

The CTM files are stored on CD-ROMs within the AFPC/DSYX safes. CTM files contain any dataset that has item level data/responses that could potentially compromise the security of a test. Therefore, this inactive material is stored in the safes, separate from the archived extract files. Within the safes are the historical project files and Master Files that contain CTM, for example, the AFOQT history file. When data from the files needs to be retrieved then the following steps should be followed:

1. Determine which type of data needs to be pulled and from what specific time frame (e.g., AFOQT from 1964-1998).
2. Based upon this information, look up the Job Code for this data (the same four-digit Job Code used for labeling files) in the DRILS Master Roster.
3. After identifying the correct Job Code, use the DRILS Master File to identify the specific file within that Job Code that captures the specific dates for the data needed.
4. Using the index linked above, identify which CD-ROM the file/data is stored on

5. Access the AFPC/DSYX safe to locate the applicable CD-ROM.
6. Open the CD-ROM and pull off necessary files and re-write them to a blank CD
7. Document the CD-ROM that was created in the AFPC/DSYX export tracking log and conduct internal approval before sending to the external agency.

8.0 MOU or DUA

When DSYX establishes a new internal or external MOU or DUA, there are certain requirements that must be included in these agreements in order to maintain appropriate DSYX data standards. (Note: AFPC/DSYX uses MOA and MOU interchangeably; yet, there are two practical differences: 1. If the bilateral agreement involves the transference of equipment and/or Controlled Test Material (CTM), then the custodian is responsible for that equipment/CTM and a MOA is used to establish the agreement between AFPC/DSYX and the other party. The language used to confer the MOA or MOU between parties, however, is comparable. 2. The MOU can always be signed by the Division Chief; yet, the MOA may necessitate signature by the organization's commander in accordance with protocol. On occasions when CTM is involved, a Test User Agreement (TUA) may also be invoked. A TUA has similar provisions as a MOA or MOU; however only pertains to agreements over test materials and resources.) For any new or updated agreements, the following requirements must be adhered to:

1. A MOU/DUA must be established with each agency that will provide the data.
2. A MOU/DUA must set up routine standard pulls/draws of the data elements/fields that will be provided to AFPC/DSYX; the layouts that will be provided; responsibility of the agency providing the data to answer any questions; data element definitions and codes; and the frequency with which the data will be provided.
3. A MOU/DUA must specify how the data will be provided to AFPC/DSYX (CD-ROM, encrypted e-mail, etc.)
4. A MOU/DUA must detail how the data will be incorporated into AFPC/DSYX (i.e. appended to an active file, maintained as an historical inactive file, etc.)
5. MOU/DUA should identify most recent point of contact that pulls/provides the data to AFPC/DSYX. This needs to be updated on an as needed basis to make sure most current POC is identified in the MOU/DUA.

Once a MOU or DUA has been established, the document needs to be reviewed/coordinated through AFPC/DSYX and staffed to appropriate agencies for approval. Once that process is completed, the MOU/DUA needs to be published onto the Community of Practice (CoP) website into the appropriate directory.

Not every AFPC/DSYX DRILS file will have a MOU/DUA. There are certain databases or file types that DSYX has direct access to and can pull/receive at their convenience, without needing a direct agreement with an external agency to access the data.

Typically, DUAs must be tailored to the specific requirements involved and the terms are sometimes negotiated, yet contain the elements in the example below.

8.1 Sample Data Use Agreement

**DATA USE AGREEMENT
BETWEEN
AIR FORCE PERSONNEL CENTER/STRATEGIC RESEARCH AND ANALYSIS
BRANCH (AFPC/DSYX)
JOINT BASE SAN ANTONIO, RANDOLPH AND
AIR EDUCATION AND TRAINING COMMAND, STUDIES AND ANALYSIS
SQUADRON (AETC/SAS)
JOINT BASE SAN ANTONIO, RANDOLPH**

SUBJECT: Collaboration and Data Use Agreement for personnel testing data exchange to support Air Education and Training Command, Studies and Analysis Squadron (AETC/SAS) Maintenance Next (MXN) initiatives by the Air Force Personnel Center, Strategic Research and Assessment Branch (AFPC/DSYX).

1. Situation

AETC/SAS has been tasked to identify predictors of successful training performance for enlisted Maintenance (2AXXX) career fields as part of Maintenance Next (MXN) training initiatives. This includes (but is not limited to) identification of optimal combinations of demographic, cognitive, personality, and vocational interest predictors associated with 2AXXX training success. AFPC/DSYX internally maintains up-to-date databases with relevant personnel testing data including the ASVAB and TAPAS. AETC/SAS has requested AFPC/DSYX to provide said personnel testing data and materials in support of MXN initiatives.

2. Use of Personnel Testing Data and Materials

Personnel testing data and materials received by AETC/SAS from AFPC/DSYX will be used only for the research effort defined in the *Situation*. AETC/SAS may make recommendations extending beyond the research effort defined in the *Situation*; however, personnel testing data and materials provided may not be used to develop potential, operational pre-accessions selection, assessment, or classification models for 2AXXX career fields. Any future efforts to apply provided personnel testing data and materials to 2AXXX career fields must comply with existing Air Force guidelines including *AF-MAN 36-2664*.

3. Agreement Responsibilities

a. AETC/SAS Personnel:

- (1) Will review all documents provided by AFPC/DSYX pertaining to to-be- provided test material and personnel data prior to engaging in data analysis activities.
- (2) Will submit an encrypted file containing relevant individual identifiers (i.e., Social Security numbers) of enlisted Airmen to AFPC/DSYX to match with relevant personnel testing data.
- (3) Will regularly consult with AFPC/DSYX during data analysis to ensure AETC/SAS interpretation of provided personnel testing data are reasonable and accurate.

b. AFPC/DSYX Personnel:

- (1) Will identify available personnel testing data for AETC/SAS using provided individual identifiers (i.e., Social Security number).
- (2) Will provide an encrypted file with paired personnel testing data to AETC/SAS for use in the effort defined in the *Situation*.
- (3) Will work closely with AETC/SAS in a “peer review” and advisory role throughout data analysis steps involving use of provided personnel testing data.

4. Resources

Each party will fund its own expenses related to this agreement.

5. Use by Third-Parties

AETC/SAS is prohibited from sharing personnel testing data and materials to third party entities without written consent of AFPC/DSYX. Release of personnel testing data are restricted to parties listed in this DUA.

6. Term of Test & Data Usage Agreement

AFPC/DSYX has designated that requested personnel testing data and materials be provided to AETC/SAS on a date and time agreed to by both parties. Either AFPC/DSYX or AETC/SAS may, at their discretion, change such dates or terminate the agreement, with notice to the other party, based upon relevant situational factors. The data shall (1) not be used to make individual personnel decisions about the Airmen whose records are included in the dataset; (2) only be reported in aggregate.

7. Amendments to Data Use Agreement (DUA)

Amendments to this DUA may only occur by written agreement signed by authorized representatives of both parties.

8. Signatories

The signatories agree to the stated terms and conditions governing the furnishing of AFPC/DSYX personnel testing data, and are as follows:

AFPC/DSY:

Dr. Peter A. Bosse
Chief Data Officer & Chief, Research,
Analysis & Data Division, HQ AFPC/DSY

AETC/SAS:

Lt Col Steven P. Dillenburger
CDR, AETC Studies & Analysis
Squadron, HQ AETC/SAS

DATE:

AFPC/DSYX Action Officer:

Dr. James F. Johnson
Senior Personnel Research Psychologist
Strategic Research and Analysis Branch
HQ AFPC/DSYX

DATE:

DATE:

AETC/SAS Action Officer:

Capt. Jacob S. John
Operations Research Analyst
AETC Studies & Analysis Squadron
HQ AETC/SAS-CS

DATE:

8.2 Sample Memorandum of Understanding

**MEMORANDUM OF UNDERSTANDING
BETWEEN
AIR FORCE PERSONNEL CENTER, STRATEGIC RESEARCH AND
ANALYSIS BRANCH (AFPC/DSYX)
JOINT BASE SAN ANTONIO AND

AIR FORCE RECRUITING SERVICE MARKET RESEARCH AND
ANALYSIS (AFRS/RSOA)
JOINT BASE SAN ANTONIO**

SUBJECT: Memorandum of Understanding (MOU) for mutual personnel testing data exchange to support operational initiatives for the Air Force Recruiting Service (AFRS/RSOE) and the Air Force Personnel Center, Strategic Research and Assessment Branch (AFPC/DSYX).

1. Situation

AFRS/RSOA and AFPC/DSYX collaborate to determine optimal testing processes to support selection and classification initiatives. AFPC/DSYX has been tasked to identify predictors of successful training/job performance and other positive outcomes for enlisted and officer career fields. This includes (but is not limited to) identification of optimal combinations of demographic, cognitive, personality, and vocational interest predictors. AFPC/DSYX internally maintains up-to-date databases with relevant personnel testing data including ASVAB, AFOQT, TAPAS, and the TBAS among others. AFRS/RSOA also maintains similar databases and the need frequently arises for the two Air Force entities to collaborate and share data to meet mission objectives.

2. Use of Personnel Testing Data and Materials

Personnel testing data and materials received by AFRS/RSOA from AFPC/DSYX (and vice versa) will be used only for the operational efforts defined in the *Situation*. All efforts to apply provided personnel testing data and materials to career fields must comply with existing Air Force guidelines including *AF-MAN 36-2664*. Data sharing for operational, individual model score initiatives will be covered under a separate MOU or other documentation.

3. Agreement Responsibilities

a. AFRS/RSOA Personnel:

- (1) Will identify requested personnel testing data for AFPC/DSYX using existing databases.
- (2) Will provide an encrypted file with personnel testing data to AFPC/DSYX for use in the effort defined in the Situation.

b. AFPC/DSYX Personnel:

- (1) Will identify requested personnel testing data for AFRS/RSOA using existing databases.
- (2) Will provide an encrypted file with personnel testing data to AFRS/RSOA for use in the effort defined in the *Situation*.

4. Resources

Each party will fund its own expenses related to this agreement.

5. Use by Third-Parties

AFRS/RSOA is prohibited from sharing personnel testing data and materials to third party entities for purposes outside the scope of this MOU without written consent of AFPC/DSYX. AFPC/DSYX is prohibited from sharing personnel testing data and materials to third party entities for purposes outside the scope of this MOU without written consent of AFRS/RSOA.

6. Term of Test & Data Usage Agreement

AFPC/DSYX and AFRS/RSOA have designated that requested personnel testing data and materials be provided on a date and time agreed to by both parties. Either AFPC/DSYX or AFRS/RSOA may, at their discretion, change such dates or terminate the agreement, with notice to the other party, based upon relevant situational factors. The data shall (1) not be used to make individual personnel decisions about the Airmen whose records are included in the dataset; (2) only be reported in aggregate. Additionally, the purpose, retention, and populations contained in the data must remain consistent with the respective organization's System of Records.

7. Amendments to Memorandum of Understanding

Amendments to this MOU may only occur by written agreement signed by authorized representatives of both parties.

8. Signatories

The signatories agree to the stated terms and conditions governing the furnishing of AFPC/DSYX and AFRS/RSOA personnel testing data, and are as follows:

AFPC/DSYX Action Officer:

Dr. Katie Gunther
Branch Chief
HQ AFPC/DSYX

DATE:

AFRS/RSOA Action Officer:

Dr. Mark Rose
Section Chief
HQ AFRS/RSOA

DATE:

9.0 SPARK's DATA JOB FAMILIES: OVERVIEW

Job Families are broken up in different pieces, with two purposes for every file: 1. Psychometric analysis; 2. Operational scores. All data (whether used for psychometric analysis or operational purposes) requires a DUA between the sending and receiving agencies if PII is included. The first psychometric phase is experimental in nature, used to derive possible benefits by developing tests and monitoring their validity over time. Typically, there are five or six versions developed before the tests are turned over for operational use. Once the test is implemented operationally, (e.g., by the Recruiting Service), the AFPC monitors the operational use for about one year.

9.1 SPARK: Descriptions by Job Family & Refresh Rates

9.1.1. 00XX Infrastructure and Emerging Job Family Data Systems

0060 Occupational Analysis (OA) Data Archive

Occupational Analysis (OA) Survey Data Archive

Note: Drive #4 for the OA archive; all subsequent databases on the R Drive.

Emerging datasets: Job satisfaction data derived from occupational surveys. Occupational analysis groups routinely do studies; AFPC/DSYX receives this survey data, feeds Rand Corporation data, Rand then extracts job satisfaction data, and then returns the dataset to AFPC/DSYX. The future goals are job satisfaction and occupational structuring of data. AFPC/DSYX has the complete survey dataset yet is currently only using the job satisfaction data. The complete OA dataset is needed to evaluate viable possibilities for career field restructuring.

0099 FIDO-Codes, GDIT-DRILS, Layouts

File Item Data Overview (FIDO) – Codes, General Data Index Table (GDIT) – Data Retrieval and Information Locator System (DRILS), Layouts. FIDO was the historical method of documenting CODES used in various data elements. The CODES column on data layouts replaces the term “FIDO” to identify small text files with the codes used in a variable. Newer layouts also permit explicating listing of code values beneath the entry in the layout for that variable. DRILS is a spreadsheet that matches Job Families into available datasets. The historical GDIT was replaced by DRILS.

9.1.2. 0500 Longitudinal Officer and Enlisted Personnel Data System (PDS)

0501 PDS – Officer – Enlisted

Military Personnel Data System (MilPDS) (Officer – Enlisted)

Data from personnel system that records about 3,000 data elements per person throughout their entire military career.

Refresh: As needed.

9.1.3. 0600 Officer Accession Source Topics

0614 United States Air Force (USAF) Academy

Order of Merit or other data requests on an as-needed basis. Note: The 0614 Job Family is not a

Master File held by DSYX. There are not fixed variables contained in the file and there is not a routine schedule for pulling this file.

Refresh: As needed and negotiated; (no standard list of variables, so not a Master File).

0615 ROTC – OTS – WINGS DB

Reserve Officer Training Corps (ROTC); Officer Training School (OTS); WINGS database (DB). Data available upon request.

Refresh: As needed and negotiated; (no standard list of variables, so not a Master File).

0664 PCSM –TBAS-Pilot Candidate Selection Method (PCSM)

Pilot Candidate Selection Method (PCSM). Test of Basic Aviation Skills (TBAS).

PCSM is the final score submitted to the board comprised of AFOQT (pilot composite score, one of 5 scores), TBAS (pilot psychomotor test, 6 out of ninety scores are used for PCSM equation), and the entries from the Federal Aviation Administration's (FAA's) logbook.

Refresh: As needed.

0665 TIMS Flying Training Outcomes (OBE See ADSS)

Formerly in Training Integration Management System (TIMS) database; yet overcome by events (obe) and currently in the Flying Training Outcomes database. This database contains the outcomes from the flight training program. (Note: General level for initially becoming a pilot and advanced training for fighting pilots). The contents of this database will migrate to the Automated Data Subsystem (ADSS) database in the near future.

0681 AFOQT – Archival AFOQT Form “Q” and Earlier

See AFOQT – Scans – 1

Air Force Officer Qualifying Test (AFOQT)

AFPC/DSYX is the source for this file, i.e., AFPC owns the scanning programs and does not receive them from others. So, AFPC/DSYX has operational (support) involvement with the AFOQT. The files have been collected historically prior to 2005. Note: Scores from older (no longer valid) forms of the AFOQT (Form S) cannot be used for any official AF personnel action. Differentiated from 0987 (Form S), which became 0992 (Form T).

Refresh: Not needed for a static archive.

0686 AFOQT WEB Official Score of Record (SOR) Repository

Air Force Officer Qualifying Test (AFOQT) WEB Score of Record (SOR) in the Military Personnel Data System (MilPDS). AFOQT is a rack-and-stack score, i.e., a percentile that is a norm referenced score. Older versions cannot be normed, since the historical population is too different from the current population of test-takers. This is a snapshot of the website (as-of date). The five composite scores are pilot, combat system operator, academic, quantitative, and verbal. The minimum scores for an officer to enter the AF are at least a 15 on the Verbal and a 10 on the Quantitative, not necessarily obtained in a single test administration. Current data is used to predict what will happen in the future. Scores in 2005 (Form S converted to Form T): Two forms

of test are T1 and T2. S1 is T8; S2 is T9. Snapshot is provided to anyone doing analyses on AFOQT's.

Refresh: As needed.

0687 (Historical) AFPT 987 Common POE (purpose determined by Technical Control Officer (TCO) number)

Air Force Personnel Test (AFPT) 987 Common Point of Entry (POE) (purpose determined by TCO number. Format S AFPT form number of answer sheet (green sheet). Common POE. Everything is scanned and then separated. Some will become 984's, some 985's. All AFOQT answer sheets are scanned. When scanned, the TCO indicates the type of answer sheet, e.g., Job Compatibility Assessment, etc. AFOQT Form S is TCO 4999 on AFPT 987. Form S is no longer in use; however, Form S's answer sheet is still in use. The subset database from 2005-2015 has raw scoring and data for Form S. Form T (blue and charcoal answer sheets) are for AFOQTs administered from 2015 to the present.

Refresh: As needed.

0692 AFOQT Form T

Air Force Officer Qualifying Test Form T

New form started in 2014. Blue and brown (charcoal) sheets, (2 answer sheets per test, one for Part A and another for Part B). Current operational AFOQT form: After the norming, Form S and Form T scores were integrated, so now there are scores from 2005-present.

Refresh: As needed.

9.1.4. 0700 Officer Personnel Topics

0710 BAE-UOR

Officer Personnel File – Uniform Officer Record

When “SPARK” was Human Resources Research Databank (HRRD) and located at the Air Force Human Resources Laboratory (AFHRL), various officer “extract” files—Uniform Officer Record (UOR)—were produced monthly at AFPC and sent to AFHRL. Since “SPARK” is now part of AFPC, the Master MilPDS snapshot (BAE) is directly available when needed.

Refresh: As needed.

9.1.5. 0800 Enlisted Accession Source Topics

0843 DMDC – MAGE – ASVAB

Defense Manpower Data Center (DMDC), located in Monterrey California. Mechanical, Administrative, General, Electronics (MAGE): core aptitude areas measured in the Armed Services Vocational Aptitude Battery (ASVAB). Note: All pre-accession demographics are contained in this database, e.g., Home of Record, zip code, etc.

Refresh: Biannually.

0850 RDB – Recruit Database – “as shipped” to BMT by AFRS

Recruit Database (RDB). Air Force Recruiting Service (AFRS). This database contains “as-shipped” to Basic Military Training (BMT) information. When the recruit arrives at BMT, the information contained in this database is the first record available. Original record of the airman.

Refresh: Annually.

9.1.6. 0900 Enlisted Personnel Topics

0905 ADSS

ADSS - Automated Data Subsystem (ADSS) Technical Training Outcome.

Consolidated Technical Training Outcome (integration of all other training record systems, including the TTMS, Training Integration Management System (TIMS), General Training Integration Management System (GTIMS), and OTA. This database comprises course outcomes for technical training for USAF personnel spanning from Basic Military Training (BMT) through Qualifying courses (AQ, if any) to AFSC Awarding Courses (AB) as well as recurring certification training and advanced Craftsman Level courses (AC) as well as Major Air Command specialized/local training. The actual course outcomes transactional file, including final outcomes, are presented. The full transactional file includes the cost for putting someone through the course, as well as the score obtained. Note: “Transactional” means that every event that happens to the student is contained in the file and a single “SSAN-COURSE-ID-DATE” set of records may have upwards of 20-30 records per unique combination. There are over 400 “status” codes for tracking individuals in the training pipeline. SPARK ADSS Masters attempts to identify and code the “final outcome” for all the records in a “SSAN-COURSE-ID” set. The Salience TTMS CODE is a 2-digit number signifying the probable value of this “final” record for DSYX study analyses. The top codes include:

- 99 Graduated with a Final Course Score (00-100) available
- 98 Graduated with no Final Course Score (Pass/Fail Courses)
- 96 Temporary Delay after Graduation
- 95 Elimination: Academic Deficiency
- 94 Elimination: Self-Imposed Elimination (SIE)
- 93 Elimination: Unsuitability
- 92 Elimination: Performance Deficiency
- 91 Elimination: Medical Reasons
- 90 Elimination: Other Reasons
- 89 Elimination: Compassionate/Hardship/To Mil Academy/Death.

Note: Use of the Technical Training Management System (TTMS) was overcome by events (obe), and so was augmented to include the Oracle Training Assistant (OTA). Currently, ADSS covers all USAF personnel in Air Force provided courses, including Basic Military Training School (BMTS, since 1999). Still used for tracking all USAF personnel while assigned to Air Education and Training Command (AETC) courses. The OTA is the source for training outcomes provided by other DoD services (Navy, Army, etc.), such as EOD and Dive School. ADSS (0905) absorbed both TTMS and OTA datasets.

Refresh: Biannually.

0910 AAE-UAR

Enlisted Personnel File – Uniform Airmen Record

When “SPARK” was Human Resources Research Databank (HRRD) and located at the Air Force Human Resources Laboratory (AFHRL), various enlisted “extract” files—Uniform Airmen Record (UAR)—were produced monthly at AFPC and sent to AFHRL. Since “SPARK” is now part of AFPC, the Master MilPDS snapshot (AAE) is directly available when needed.

Refresh: As needed.

0947 WAPS Testing

Weighted Airmen Promotion System (WAPS). Enlisted promotion testing is a compilation of six factors: two tests (Specialty Knowledge Test (SKT), Promotion Fitness Exam (PFE)), performance reports, decorations, time in service, and time in grade.

Refresh: As needed.

9.1.7. 0950+ Special Projects

0950 TAPAS Psychometric

Tailored Adaptive Personality Assessment System (TAPAS) data. TAPAS is an instrument developed by the Army Research Institute. This database contains the 15 facet scores from the TAPAS’s “Big 5 personality assessment”. The TAPAS Master File allocates 21 facet slots; Version 5 accounts for the first 15, which are the most populated. The TAPAS database contains DSYX psychometric data from 2009 to the present, updated quarterly.

Refresh: Quarterly.

0951 TAPAS Operational

Tailored Adaptive Personality Assessment System (TAPAS) data. This database contains the 15 facet scores from the TAPAS “Big 5 personality assessment”. These data include Scores of Record (SORs). Operational scores are from 2016 to the present, updated daily.

Refresh: As needed.

0952 TAPAS Hybrid

Tailored Adaptive Personality Assessment System (TAPAS) merged (hybrid) data. For DSYX Model Development ONLY. In format of operational scores. If cases before 2016 are required, must use truncated version 0950 record. Armed Services Vocational Aptitude Battery (ASVAB) is included in 0951, not 0950, so 0952 TAPAS hybrid is augmented with the 0843 ASVAB database.

Refresh: As needed.

0960 ICTL Cyber Test

Information Communication Technology Literacy (ICTL). Developed for Manpower Accession Planning Working Group (MAPWG) for use in the ASVAB. Developed cyber test became operational in 2016.

Refresh: As needed.

0983 APAT Computer-Based Datasets

Applied Performance Assessment Testing (APAT) computer data NOT yet controlled by DSYX/SPARK.

Refresh: As needed.

0984 APAT Paper ONLY Data

Applied Performance Assessment and Testing (APAT) facility, located at Lackland AFB. Note: DSYX/SPARK has control over paper-based data, not computer-based data. Two flights of Basic Trainees arrive at Lackland AFB every Saturday and they undergo psychometric testing (as opposed to operational testing that becomes part of their personnel file). This Job Family is a database of all data collected using the AFOQT Form S (green) answer sheets. The APAT facility uses tests with both paper and computer results. The green sheet form goes into a database; every answer sheet has a code (TCO—Test Control Option) that identifies the project. (Note: All 4000 series TCOs are for psychometric data, except for 4999, which is the operational TCO for the Job Compatibility Assessment (JCA).) All data goes into the same database and TCO's are screened to yield information for a project. The particular form identifies which score key to use. Another APAT sample is used for norming. After the test is completed and the norming is applied, then the test can be used operationally. The APAT affords a way to keep psychometric data isolated from personnel files. The test results are used to validate the instrument itself, not to assess the people taking the test. Sexual Assault Prevention Reporting (SAPR) is measured using an anonymous form to find the base rates.

Refresh: Each time answer sheets are scanned.

0985 JCA Operational Job Compatibility Assessment

Operational score developed for AF security forces to screen civilians hired to carry guns on base. This comes in on same Form S (green answer sheets), but not mixed with APAT data since this score is used operationally. Note: The TCO Number that identifies this Job Family is 4999, which is the only operational number in the 4000 series.

Refresh: Biannually.

0989 (Active) Form S (green) answer sheet used for other than AFOQT purposes.

Primarily for APAT data collection from Lackland. Also used for JCA Program for Security Forces (operational).

Refresh: Each time answer sheets are scanned.

0999 Enlisted Master File

This database merges all datasets into one record per person. Goes through each data set for all enlisted items (e.g. Technical Training Outcome, ASVAB, recruit database, TAPAS, ICTL, etc.) to amass the aggregate list of SSANs. Each file with a unique SSAN is taken in order and matched with all ASVAB records to yield 60-70% matching records: matching SSAN and ASVAB and then TAPAS, etc. creates Base 1 file, Base 2, Base 3 file, etc. Every aggregate record matched with a unique SSAN will include everything about that person, e.g., section of cyber, PDS, ADSS (training courses), etc. For each source file, the date of the test will be distributed. Once the files are merged, the date ranges will change because the most recent record will be matched with the unique SSAN: one record per test per SSAN.

Refresh: Every 4 years.

10.0 LAYOUT SYSTEM

The audit process in SPARK for any input data file is driven by a layout file. A layout is a text file that contains data dictionary information about a target data file. The layout is comprised of a list of all the variables within a data file, a description of those variables and the applicable variable codes (i.e., DICE codes), the length of record for each variable in the file, and any applicable notes that explain aspects of the data file and the variables. Layouts are created by importing the data file into specific computer programs that analyze the data and create layouts based upon a pre-designed template. A layout file shows the basic display and variable elements included in a corresponding data file. There is a layout for every data file submission and a master layout for every master file submission. A single layout file may apply to and describe multiple data files that are the same, varying only by the date covered. Therefore, when pulling data, a single layout may work for multiple files. (See Appendix _ for a description of the four standard elements comprising each data field definition in a SPARK layout file, as well as examples of those files.)

The heading/name of a layout must match the information contained in the corresponding DRILS Master File. This ensures that the proper layout can be identified and matched with the appropriate master data file, which is necessary to use and interpret data and variables utilized by internal/external agencies for statistical analyses and studies. Each row of a layout should contain information on each individual data element/variable listed in the corresponding data file. There are three types of layouts (shown below) that can be found in DSYX (due to historical standards). But EVERY layout needs to start and be stored in the .LAY/.TXT version (Notepad). This allows for ease of transmission and further manipulation of the file, if necessary.

Composition of a “people-friendly” layout:

- a. FLD# = field number for each variable (should be one for each data element established by a MOA/DUA)
- b. NC = Number of characters in the data field in a fixed format ASCII file
- c. SC = Starting column number of the field in the data record
- d. EC = Ending column number of the field in the data record
- e. Name = Column ID provided by the Excel program
- f. Description = Brief description of the variable or listing of the variable ID
- g. DICE = Provides the variable ID and code meanings
- h. Range = Provides expected field content/validation rules

Layouts are created by running a simple “NOTEPAD” text data file through a computer program that examines the data and produces the appropriately expanded layout automatically. The DSYX technician must verify that the variable elements are correct, the length of the record is accurate, and the appropriate description and DICE codes that correspond with each variable are identified appropriately. Furthermore, the layout must be saved in the appropriate format and identified in the DRILS file whenever the layout corresponds with a Master File. When external agencies/contractors engage in projects, the data they request from DSYX might not encompass an entire Master File. Therefore, only the needed part of the Master File data should be pulled, and

only part of the Master layout will apply. When a layout needs to be created from a Master layout, the variables requested from the contractor must be identified, so the program that creates the layout only pulls the specific variables needed.

10.1 Layout Versions

1) .LAY or .TXT version should include the following information:

- a. The original source file (i.e., the title of the data file). (“T” in column 1.)
- b. Optional “Note” lines. (“N” or “*” in column 1.) Examples include:
 - i. Total records within the input file. (“N = 788,235”.)
 - ii. Logical Record Length in characters. (“LRL=482”.)
- c. Optional “Section Headers”. (“H” in column 1 to label related sets of variables to follow.)
- d. One Line for each variable in this dataset, including:
 - i. Size: Length of field displayed in ASCII format, independent of data type.
 - ii. Type: Data type modifier. (Assumes “A” for alphanumeric ASCII text).
 1. “A” alphanumeric – ASCII display/text variable, which is the default and need not be entered.
 2. “S” forces SSAN format to just nine numbers. Remove dashes if present. “9S” would be a target format for a source variable in “11”, imported with dashes in the SSAN.
 3. “L” alphanumeric, but forces LEFT ADJUST. (Remove leading blanks and pad with trailing blanks.)
 4. “R” alphanumeric, but forces RIGHT ADJUST. (Right adjusts non-blanks and pads left with leading blanks.)
 5. “N” numeric, (an integer value). (If negative, length must include negative (“-”) sign.) Will right adjust value with leading zeros.
 6. “Fl.d”, real value. A number with an implied or actual decimal point. l=length of field; d=number of decimal points implied if no actual decimal, which overrides this default.
 7. “E”, emit value, can be used to insert blanks into a target record or ignore unwanted data on an input source record. (Note: use variable ID name, “FILLER”.) To “poke” a value into a target record, set size (such as, 002), use a new variable ID (like “FLAG1”), and in the Notes or Source Field, include the text “EMIT=ST”.
 8. “2C”, insert a two-digit Smart Century Variable. Put either a “19” or “20” prefix on the following date, which only has “YY” data. If the “YY” is 23 or less, the program inserts a “20” to capture years 2000 – 2023. Otherwise, “19” is inserted to capture 1924 to 1999.

9. “8Y” forces an 8-digit year format on target, regardless of date input format on source, such as 7 Date (“03Mar18”), 9 Date (“03 Mar 18”), or Date 11 (“03 Mar 2018”), or 12 Date (“Mar 03, 2018”). “Y” must be used with a target date format specification in the Description, Notes, or Source field (most typically, “YYYYMMDD” or “YYYYMM”). Will control automatic variable distribution produced to only show “month-level” data.
- iii. Separator
 1. Space. (Separates field size/type description from variable ID.)
 - iv. Variable ID: Each variable ID must start with a LETTER (A-Z) and may include additional letters, numbers or underscores, yet NO DASHES and NO SPACES.
 1. Strive for eight characters or less.
 2. Use a UNIQUE variable ID across the SPARK data system, if this layout is to be merged with other SPARK Masters.
 3. Initial import programs (CSV2LAY, LAY2CSV) can handle up to 40 characters and permit the format: “Original_ID/SPARK_ID” to permit automatic conversion of a long original variable ID into a shorter (eight characters or less SPARK system ID). An example might be: “2 Number of Dependents/NUMDEPS”.
 - v. Separator 2: Circumflex (“^”, shift 6). Separator, (to end ID and start description that may include spaces.)
 - vi. Description for variable
 1. Personally identifiable information (PII) note: Set the first character in this description to be an exclamation point to flag a field that has PII (or unique information of no audit value). This will prevent showing values in automatic distributions and provides an audit for data release decisions on this variable. (Figure 5)

```

T ASVAB 2018 Master Layout 181024 TERSE Version - No Special Test Breakout past CC 479

N N= 788,235 - Unique SSAN_Test-Date Combinations and

N LRL=482 w/EOF-LF w/DMDC Coding References

9S SSAN^!F_001 - Social Security Account Number^^F_0001 SOURL (1-9)

008 TDATE^F_078 - ASVAB_TEST_DATE YYYYMMDD^^SOURL (10-17)

008 ENTRY_DT^F_053 - DATE_ENTRY DATE YYYYMMDD^^SOURL (18-25)

002 TESTFORM^ASVAB TEST FORM ID 28 29^Table 11 - DMDC - Test Form ^F_0002 SOURL (26-27)

```

Figure 5. Example of a layout file in the Notepad program

2. Any flat ASCII display text desired (may not include a circumflex (“^”), which would terminate the description as will an end-of-line (ASCII 13/10 = Carriage Return-Linefeed) generated by pressing “RETURN” in Notepad.) *Note: If a double quote is used it must have a closed quote within this description area.*
- vii. SEPARATOR 3: Circumflex (“^”, shift 6). Separator (to end variable description and start Notes Section, which may include spaces.)
- viii. Notes: Optional Field that has any flat ASCII display text desired except may not include a circumflex (“^”) that would terminate the Notes Section, as will an end-of-line (ASCII 13/10 = Carriage Return-Linefeed) generated by pressing “RETURN” in Notepad. *Note: if a double quote is used it must have a closed quote within this description area.*
- ix. Source Note: Optional/Automatic Field. The “Reformat” process, which moves data from a Source Data File (with a Source Layout) to a Target Data File (with a Target Layout) will automatically annotate the new Target Layout with the data columns used in the Source Layout (SOURL). This field may also include any flat ASCII display text desired, sans circumflex (“^”), which would terminate the Notes Section as will an end-of-line (ASCII 13/10 = Carriage Return-Linefeed) generated by pressing “RETURN” in Notepad. *Note: If a double quote is used it must have a closed quote within this description area.* This field is typically closed by encountering an End-of-Text-Line.

2) Test norm on ASVAB Master (Figure 6).

```

"*01","[ASV0218T.DAT]","<-Source File Name / Title: ", "F0120:_002/00410_~_NORM"

"*02","[NORM]","<- Field ID / File Name",[NORM.CSV]

"*03","[ 410]","<- Start Data Column"

"*04","[ 411]","<- End Data Column"

"*05","[ 2]","<- Number of Columns"

"*06","[ 778835]","<-Records Distributed"," Finished?"

"*07","[ 3]","<-Unique Values Found","Yes"

"*08","[ 482]","<-Longest Record Length / Count:",[ 778835]

"*09","[ 482]","<-Shortest Record Length / Count:",[ 778835]

778835,"Total"

"*Freq","[Values]"

130872,[80]

631340,[97]

16623,[?]?

```

Figure 6. Sample Audit of a Variable Field

- i. First nine lines from table above: Information about the field being audited.
- ii. Line 10: Total number of items found in the file and distributed below this entry.
- iii. Header for columns that follow. (Values as they appear in the file.)
- iv. Next 15 lines show the top seven and bottom seven values as found in the file. The center line shows how many unique values were skipped between the top seven and bottom seven values, in lexicographical order.

3) .CSV version is created by running the .LAY/.TXT file through a different program that produces the .CSV layout output. This version should include the following:

- a. Title of the file at the beginning
- b. Date the file was created/saved
- c. Field numbers
- d. Number of cells
- e. Starting Cell Number
- f. Ending Cell Number
- g. Variable ID (eight characters or less)
- h. Description of the variable

Note: There needs to be information about the original source file name, length of the longest data record, number of records in the input file, and the length of the header line.

- 4) .XLSX, .XLS, .DOC or .DOCX (Word document) version must include the following:
 - a. Title/description of the layout followed by titles/names of current and older file versions of similar layouts (if they exist).
 - b. Detailed notes that reference the note title that corresponds to certain variable descriptions within the layout.
 - c. Field #
 - d. Number of cells
 - e. Starting cell number
 - f. Ending Cell Number
 - g. Variable ID/name (eight characters or less)
 - h. Description
 - i. DICE
 - j. Range (signifies the range of responses to expect)

There should also be a Master layout file that is created to go with every Master File DRILS entry. (NOTE: As mentioned previously, a single layout may apply to multiple files, including Master Files). The difference between this layout and other layouts is that the dates covered will differ, record lengths will differ, and the name/label of the file will differ from normal layouts. A master layout will encompass all variables and codes for a particular Job Family, within the specified date range for a DRILS Master File.

In addition to the file types mentioned above, an additional file/CD needs to be maintained by DSYX that is not a direct part of the database. This file/CD should contain the original source file of the data (whether pulled, received, etc.). This source file is the initial step taken by DSYX before any auditing and creation of Master Files and layouts occurs. The source file/CD is maintained as part of a general auditing/backup system that allows administrators to recover the original data that goes into DSYX before any internal changes are made. This procedure is accomplished to prevent data loss due to errors in manipulation, lost files, etc. Rather than go back to the original agency that provided the data to DSYX, the administrators can simply go back to this source file to proceed further.

10.2 Data Cleaning Procedures and Auditing Process

When the AFPC/DSYX receives data from any source, the immediate response is to audit that data to ensure a standard of quality. (Layout files need to be audited to ensure no errors occurred and the proper variables and codes were captured.) Current DSYX functions are to dovetail existing and future data into a usable format that can be appended to the pre-1998 longitudinal data. The

audit is necessary to make sure data is in a usable format, contains appropriate information for the Job Family, and allows for proper file formats to be created that are necessary to DSYX functionality. Auditing is done every time a file/submission is received/pulled in order to ensure data integrity and identify any potential layout/variable changes between current and past submissions. The basic idea is to check the record count of the data file, conduct a review of random sampling of the data elements, and utilize the DICE and DRILS to identify any layout and variable changes that may have occurred.

10.2.1. Auditing Process Steps

The auditing process for every data file received/pulled needs to begin by saving a copy of the original file into flat ASCII text format. Upon doing so, the original source file needs to be saved “as received or pulled” for data integrity purposes. This process allows DSYX personnel to recover the original source data before any internal changes were made and to continue forth with DSYX functionality requirements without causing any gaps in the longitudinal data. This helps prevent any data loss, conversion error, etc. when continuing to manipulate and re-save the file for future DSYX tasks. Furthermore, documentation within the AFPC/DSYX Data Import/Export Log needs to be updated to reflect the data that is received/pulled. The source file should be saved to a CD ROM (entitled “Job Family CD”) along with all work products (e.g., steps taken to extract/manipulate the data into a usable form) and the final file that is created to go forward into the DSYX auditing process. This “Job Family CD” is necessary because the source file may contain extraneous data that is not needed by AFPC/DSYX, yet rather than discard the data, DSYX maintains the file as a record to see what was done to the source file in order to make it usable for DSYX. Also creating the Job Family CD allows for recovery of the original data in case mistakes/errors occur during AFPC/DSYX processing.

Once the appropriate data pull or submission has been saved in the appropriate format and entered in the DSYX tracking logs, the actual auditing of the data by the AFPC/DSYX technician begins. The first step in the actual audit of the data is done by running the audit program LAY2CSV, which generates code for item-level audit based upon the input layout. If minor edits are required then make those changes and run the next (special purpose) audit program FIX2FRQ, which will run the item level audit again and create a frequency file output. (Note: Most minor edits include 1) removing “unique” fields like SSAN from the distribution request; 2) Remove coded fields NOT expected to have the same distributions, like state codes for applicants. 3) Mislabelled DICE codes.) If a previous Master file exists for this Job Family code/submission, then a comparison needs to be done (monthly, quarterly, semi-annually, annually, or other specified/standard scheduled data capture time) by manually evaluating distributions of each data field in the old and new files to ensure that each field contains comparable coding and values, accounting for changes in time. If NO differences are found between the current submission and the old Master file, then the current submission will simply be added to the old Master file and the updating of the number of records and dates covered by the “new” Master file upon addition. Then, for maintenance purposes, the old master and old DRILS entry must be copied to a “retired” folder/directory before continuing. The new master will simply add on the additional records and coverage dates (along with any applicable notes) and then save in the proper format, showing the new ending year of the submission for the Master file. An update to the DRILS entry needs to be done to reflect the new name and record count, yet all other fields of the DRILS Master should remain the same. Then a

copy of the new Master and updated DRILS needs to be added to the DSYX directory on the Restricted Drive.

If there are differences between the current submission and old Master file, then a determination of how to proceed needs to be made by the SPARK PM. If the only differences are in the coding of the variables (but they are exactly the same field length), then only the existing target DICE codes need to be updated. (For example, if gender changed from M/F coding to 1=M and 2=F). Changing the DICE codes is done by changing the “end dates” for the old code from 999999 to whatever the end date is. Then updating the new DICE code to show the starting date as MMDDYY and end date as 999999, which signifies that the code is presently being used. (Note: A preferable alternative for making a simple change in code might be to keep the old DICE code as-is and update the variable of interest in the current submission to reflect the “older” coding that was previously in use.) If there are multiple fields/variables added, missing, or changed in length, etc. then a new Master file chain needs to be created. Start by copying the current Master, DRILS entry, and associated layout into an archived/retired directory or file. (DO NOT DELETE THESE OLDER VERSIONS.) Once the older versions have been “retired” the next step is to create the new Master file, DRILS entry, layout, and appropriate DICE coding, if necessary. These files should be saved in the proper format/name and posted to the appropriate directories upon completion. For every new Master file that is created, two backup copies (CD-ROM version) need to be created: one for the internal AFPC/DSYX storage and another sent to the back-up off-site storage facility, and then properly recorded in the AFPC/DSYX Export Tracking Log.

Note: The DSYX Technician must ensure that ALL data captured in the final layout is identified in the DRILS entry and Master file. Moreover, when making a Master file, be sure to include the latest version of the layout on the CD-ROM version with the DRILS Master file that is to be created for AFPC/DSYX and for the off-site storage facility. Finally, this process is designed to be documented and saved in the appropriate manner to allow tracking and recovery of data if errors occur. The source file and the end AFPC/DSYX product (Master File) should be maintained in their original forms.

10.2.2. Data Cleaning

Since each data family comes from a different source, there are specific data cleaning procedures to “normalize” the data for each particular Job Family. First, the input (source) file is received as either a flat file that may arrive with its own layout file or the source file is in Comma-Separated Values (CSV) format and must undergo a process to derive the layout file. If a flat file is received, that file may still require modification to reconcile the data format with the Master File system. The layout is updated to record any discrepancies, so that subsequent submissions from the same source can be accommodated by a previous normalization procedure. The output is the layout for the Master File and in between the input and output layouts are the data cleaning procedures.

The layout system is designed to accomplish most of the data cleaning; a relative minor amount of data cleaning is accomplished via hard-coded, unique programs designed to perform specific functions. Most of AFPC/DSYX’s files are saved from a software suite called “SAS” (Statistical Analysis System) and then are saved with a CSV extension, since CSV files can be easily imported into Microsoft Excel. A CSV file is a “flat file”—(i.e., a file that has data formatted in plain-text

(usually with some kind of delimiter) and the only characters in any record are the displayed characters)—and is the preferred import file to minimize data integrity issues. Consequently, if the source file is not in CSV format, AFPC/DSYX converts the file to CSV format. The next step is to change the lengthy labels for columns—the source file might have two or three lines of labels—to relatively terse labels. Once the CSV file has terse labels, the field program is next: 100 labels corresponds to 100 data columns under every record. Sometimes, commas are used as separators (delimiters), sometimes exclamation points, etc. (The delimiter is annotated in the program.) In particular, the CSV to field (CSV2FLDS) program ensures that every data record matches every number of fields in the record: The program searches every record to ensure the designated number of fields is presented in the data. The upshot: CSV2FLDS eliminates the bad records. Null fields are included in the count, as well. The labels are established with the data fields below them.

There are common issues that can present with several Job Families. For example, the name field **John “Doc” Smith** is one field, so must be inspected carefully. Another issue that presents when a SAS file is converted to CSV relates to different meanings assigned to characters; for example, in the Personnel Data System (PDS) Job Family, **HQ CR LF** (Headquarters, carriage return, line feed) is a four character field in SAS; yet, when saved to CSV, the CR LF designates the end of a CSV line. Therefore, if left unnoticed, the record becomes truncated when the variables HQ CR LF present. Nonetheless, CSV is still the preferred import file because the file causes the least amount of data integrity issues. The other option (aside from the CSV format) for converting a flat ASCII file is the fixed format, meaning that every field is a designated size. (The originator of a fixed format source file must also send a layout file in the transmission.) Yet, a fixed format file is not as easy to create as a CSV file and when used historically, has often been accompanied by errors that require DSYX to create utility programs to resolve the problem and maintain data integrity.

As mentioned previously, **CSV2FLDS is a program in the layout system that eliminates bad records**. Once the file is corrected, the program CSV2LAY generates a SPARK layout. (See **Appendix E for SPARK layout programs**.) For every field, the SPARK layout has a length (10 characters), a space, a unique variable ID, a circumflex (annotating a delimiter), and a description, for example: **10_Name ^ Description**. (Note: The import file can have a length up to 40 characters; however, the target SPARK layout only has a length of 10 for operational use. Moreover, the 10 characters must be unique throughout the layout system.) AFPC/DSYX’s goal is to have all imports to SPARK be in a flat ASCII fixed-format file accompanied by a standardized SPARK layout. The **program CSV2FIX generates a fixed format record** and is often used after AFPC/DSYX performs a manual edit to a layout. (Note: A record is one line within a file, representing one case or observation.) For example, when AFPC/DSYX receives an update to an ASVAB file, the following layout may present: **AV1908 – HE.LAY**, and the “HE” represents “hand-edited”. Then the hand-edited layout goes into the fix-format program, i.e., CSV2FIX. Hand-editing becomes important with recent updates to the master files. The goal is for the last hand-edit to work the next time. To check if the most recent hand-edit still resolves the issues, DSYX runs an audit with CSV2FIX on about 2,000 records. Auto-generated layouts will show the problems that still exist. For example, AFPC/DSYX master files have only 25 characters; yet, if a file has 27 characters for every record, AFPC/DSYX technicians must change the master file to update the new requirement. So, if the size of the field changes, the old master must be changed to

accommodate future updates.

Once the hand-edited layout AV1908 – HE.LAY (example above) goes to the LAY2CSV program, the output is a readable format in CSV that can be loaded into the Excel program. When the LAY2CSV program runs, a batch file is also generated: AV1908.BAT. The **FIX2FRQ program is then used to generate an audit** by reading the batch file and creating the frequency distribution. The naming of fields is important because each layout name becomes a file name, showing the distribution of records for each field. So, if “name” should be in the field, yet the distribution is SSAN’s, for example, the AFPC/DSYX technician knows a mistake was made. Once the file has gone through the auditing process, i.e., the batch file has run, the program generates an output with a .DAT extension, to use our example: AV1908.DAT. To summarize, the AV1908L.LAY is the first file and the SPARK layout; AV1908L.CSV is the file in a readable form that can be loaded to Excel, AV1908.BAT is the form that generates an audit, and AV1908.DAT is the audited form of the file.

There are also two SPARK programs used for reformatting: PAIR2MAT and MAT2FIX. The function of the PAIR2MAT program is to audit by creating pseudo-code to do the reformatting. The MAT2FIX program takes the match commands (identified by PAIR2MAT) and generates a new fixed format file.

11.0 RESULTS AND DISCUSSIONS

This report has documented the design philosophies and detailed implementation of the software used in the Strategic Policy Analysis Resource and Knowledgebase system, with the intent of describing the elements and goals that must be preserved in any replacement configuration. Since many of these technical approaches go back to Turbo Pascal 2.0 (16-bit software) and use MS-DOS as the bridge between batch files, these approaches are not supported or permitted on the current USAF Network. This means that all SPARK data processing must take place on Windows XP PCs, which are now dated and no longer supported by Microsoft.

The first effort in this project was to find an XP emulator that would run on a modern I7 grade laptop. It has been demonstrated that this stop-gap measure can avoid mission failure in the event of the actual XP “SPARK” Machine’s no longer working and/or being available as an option. After numerous tests with a variety of Virtual Machine (VM) applications, the VMware Workstation 15 Player was chosen, since scripts could be run at operational speed with this particular Virtual Win XP Pro Machine on a Win 10 host operating system. Instructions on running this VM are included in Appendix A.

This report lays down the foundation for goals and products in the SPARK approach. The way ahead, however, requires some time in order to assess the future platform for the SPARK programs and other resources. The AFPC is considering updating their enterprise software from the current SAS 9.4 analytics, reporting, administration, and coding capabilities to a more modern, scalable software platform. The new capability must include the current SAS 9.4 statistical analysis capability as well as enhanced features, such as predictive modeling, text analytics, and machine learning. Furthermore, AFPC would like to leverage the open-source capability offered by Python, R, Java, etc., in the near future, so an open-architecture is required for data manipulation and analytics development to be callable via the open-source programming language that is ultimately chosen. The open-source programming language capability would allow AFPC’s Industrial/Organizational (I/O) psychologists to share code with other researchers who might suggest improvements to the tools they are currently using.

12.0 CONCLUSION

The original SPARK (HRRD) was developed by the AFHRL in the 1960s and has evolved over the years. Currently, a system-wide update is needed. Industrial/Organizational (I/O) psychology is an emergent field and the current SAS and SPSS applications will not necessarily respond to the needs of the AFPC/DSYX's I/O psychologists, which are often outside the mainstream requirements met by proprietary statistical packages. Moreover, if AFPC/DSYX obtains case- and task- clustering in their toolset, the I/O psychologists would become more innovative contributors to their user group, which would in turn attract the best and most qualified students to enhance the work of the AFPC's Strategic Research & Assessment Branch. Finally, if AFPC/DSYX transitions to an open-source programming language, they will attain User Group software specific to I/O psychology that will markedly enhance the overall functioning of their significant, strategic work in improving the person-job-match in career fields for optimal Air Force mission capability.

13.0 REFERENCES

- Christal, R. E., & Weissmuller, J. J. (1977). *New CODAP programs for analyzing task factor information*. Washington, DC: American Psychological Association Journal Supplement. Abstract Service.
- DoD Scientific and Technical Information Program* [22 Aug 2013]. DoDI 3200.12. Retrieved from <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/320012p.pdf?ver=2019-04-30-073122-220>.
- Goody, K. (1976). *Comprehensive occupational data analysis programs (CODAP): Use of REXALL to identify divergent raters*. Brooks AFB, TX: Department of the Air Force, Air Force Systems Command, Air Force Human Resources Laboratory.
- Personnel Assessment Program* [16 May 2019]. AFMAN 36-2664. Retrieved from https://static.e-publishing.af.mil/production/1/af_a1/publication/afman36-2664/afman36-2664.pdf.
- Staley, M. R., & Weissmuller, J. J. (1981). *Interrater reliability: The development of an automated analysis tool*. Brooks AFB, TX: Air Force Human Resources Laboratory, Air Force Systems Command.
- Turbo Pascal: Version 3.0: Reference manual*. (1985). Scotts Valley, CA: Borland International.
- Weissmuller, J. J., Barton, B. B., & Rogers, C. R. (1974). *CODAP: Programmer Notes for the Subroutine Library on the Univac 1108*. Ft. Belvoir, VA: Defense Technical Information Center.
- Weissmuller, J. J., Barton, B. B., & Rogers, C. R. (1974). *CODAP: Source Program Listings for the Univac 1108*. Ft. Belvoir, VA: Defense Technical Information Center.

Appendix A: Instructions for Virtual (Windows XP Pro) Machine VMware Workstation 15 Player with Veracrypt for Encrypted Files on GUEST Machine

Currently, AFPC/DSYX is using stand-alone Win XP desktop computers because the software to run the SPARK programs is not approved in the Network. (MS DOS version of Turbo PASCAL only runs on 16-bit machines, so must use an XP machine. Windows 7 has an XP-emulator that can run programs; however, cannot complete a script. Currently, DSYX configured a Virtual Win XP Pro Machine on a Win 10 host operating system using the VMware Workstation 15 Player, which can run scripts at operational speed.) The SPARK program is run in Turbo Pascal 3.0, 16-bit. The Open Source program called Open Office is used, since Open Office is written in Python and so can use Excel, etc.

Below is the algorithm for this work-around solution, running a Win XP virtual machine on a Windows 10 host machine.

Preliminary Instructions to launching Veracrypt

- A. Open native USB drive to find the password for encrypted files.
- B. Open folder entitled "Inventory".
- C. Click on "Date Modified" to order the dates with most current on top.
- D. Open the most current Inventory text (txt) file.
- E. Click on Edit and select Find.
- F. Type name of program listed on connected native USB drive (e.g., OA).
- G. Click "Find Next" until password is located.
- H. Keep the Notepad window open (but moved to the side) to later cut & paste the password when launching the Veracrypt program.
- I. Close Windows Explorer.
- J. Disconnect the USB drive, so the window presents in the virtual machine to connect the drive when you connect the cable to the laptop.

Run the Virtual XP Pro Machine: VMware Workstation 15 Player

1. Double-click on VM Workstation on the host machine.
2. Double-click on the Operating System: Windows XP Professional.
3. Connect the USB drive with the 2.0 USB extension cable.
4. Click on "Connect to a virtual machine" when the "New USB Device Detected" window appears, then click OK.
5. Click OK on the window "Removable Devices" (showing devices that can be connected to the VM).

Launch Veracrypt on Guest Machine

6. Open Veracrypt program on the guest machine.
7. Choose a drive letter (e.g., Drive R).
8. Click Select File button.
9. Browse the USB drive that has the encrypted file.
10. Select and open the file (e.g., S3_x.tc).
11. Click the Mount button.
12. Place a checkmark in the True Crypt Mode box.
13. Select Mount Options.
14. Place a checkmark in the Mount Volume as Removable Medium box.
15. Click OK.
16. Input the password for the encrypted file by cutting from the open Notepad window on your screen and pasting to the Veracrypt box provided.
17. Click Display Password to ensure entered correctly.
18. Click OK.
19. Click Exit.

After the mounting completes, the contents of the encrypted file will show as an unencrypted drive letter on the machine.

20. Click on My Computer.
21. Under Devices with Removable Storage, double-click on the SPARK (USB drive) designated by the letter (e.g., R) mounted in Veracrypt.
22. Run programs.

Power-down VM

23. Double-click on Veracrypt.
24. Click Dismount All.
25. Once drive(s) is (are) dismounted, click Exit.
26. Disconnect SPARK USB drive, if desired.
27. Click Player on the VM and select Exit from dropdown menu.
28. Click Power Off.

NOTE: To access a 300,000 character file that is encrypted, e.g., S3.tc (extension “tc” is for true-crypt). Go to Veracrypt (true-crypt). Go to restricted drive and call it Drive R. Load true crypt as a removable media.

Appendix B: Access to AFPC/DSYX's Restricted Drive

Location of new Restricted Drive Master File System

\\tymx-fs-002v\restricted\DS\DS\DSY\DSYX\9 - CAPS_____ SPARK-Data-Masters\
APAT\APAT-18M-Master-180807

Appendix C: Point of Contacts (PoCs)
External to DSYX and Internal Job Families

ADSS

Lawrence “Larry” Dohm
Principal Software Engineer
Contractor – G2S Corp
HQ AETC/A5TE
DSN 487-3602 | Comm. 210-652-3602

ASVAB

DMDC Reporting System (DMDCRS) Helpdesk
Defense Manpower Data Center (DMDC)
DoD Center Monterey Bay - Seaside, CA
COM +1.831.583.2400 | DSN 312.878.2951
Support Hours: 1000 ET - 1900 ET
<https://dmdcrs.dmdc.osd.mil/dmdcrs>

(AND)

Tim Powers
Branch Chief
Decision Support/Business Intelligence
Defense Manpower Data Center
(831) 583-2400 x4490

Appendix D: Elements of Data Field Definition Lines in SPARK Layout Files

A SPARK layout file is a text file which describes the contents of a specific data file to be imported or processed within SPARK. The data file must be a flat-ASCII, fixed length record file. ASCII means American Standard Code for Information Interchange. The term "flat-ASCII" means using only those codes (typically codes 032-127) which represent letters "A" to "Z" (or "a" to "z"), numbers, a few special symbols which may appear on a computer keyboard. The only "non-visible" characters are the Tab character (ASCII 09), a Carriage Return (ASCII 13), and Linefeed (ASCII 10). Many external data files are "saved" into a CSV data format from its source system and SPARK has a sequence of programs to generate the required flat-ASCII fixed-format data file for integration into the SPARK Master File System.

A SPARK layout file is also a flat ASCII text file, which can be created or edited using NotePad or another Text Editor. (Note: A SPARK program can also be used to generate, edit, or modify layouts.) SPARK layouts may not be edited using standard word processing programs like Microsoft Word, unless they are exported to Save As - Plain Txt (.TXT).

A layout combines several record types (defined by the first letter of the line in the layout file.) The first line must be a Title Line ("T ", followed by the file's Title/Description.) Additional lines of text that may be included are Headings ("H") for introducing various sections to follow. Notes ("N ") line may be used to provide information about the full file or about individual data fields to follow.

Each data field definition line has four standard elements. Each data field definition line begins with an integer number on the left margin. This first element gives the length of data field (visible characters as displayed in ASCII Text in the file). Optionally attached to that number is a type indicator:

- The default Type is "A" for alphanumeric, which means the field is treated as TEXT only.
- Numeric Type codes (used later by SAS or SPSS to translate variables) include "N" for numeric (integer) or "F" for floating point (values with decimal points, either explicit or implicit).
- Type codes "R" or "L" are for "Right justify a number, adding leading zeros" or "Left justify text, padding with blanks."
- Type code "S" (on a "9S") is used to force the field to a nine-digit SSAN (adding leading zeros if necessary or even if input has 11 characters with dashes).
- Type code "Y" ("8Y") is used to force the field to an 8-digit YEAR-MONTH-DAY format of YYYYMMDD, regardless of the input field format that may be YYMMDD, DD MMM YY, MM-DD-YY, etc.
- Type code "E" is for an EMITTED Field. This will generate a data field with the number of blanks specified (unless an "EMIT=" command is specified.)
- A blank space is used to terminate the field length/Type code element.

The second element in a data field definition line is the unique Variable ID that will be used

throughout the SPARK system. Ideally, this is eight characters or less and begins with a letter and should be unique among the files to be merged. Note that "letters" include "A" to "Z" and Underscore ("_"), but not a dash ("-"). Layouts, especially auto-generated layouts from pre-existing data dictionaries or older layouts, may have Variable IDs with up to 40 characters. There are SPARK programs that allow the inclusion of a slash following the original, long-name with a manually generated "short" Variable ID: The output is a replacement layout with only the short IDs for the follow-on programs to use. The "slash" file documents the before and after names.

There are three "reserved word" Variable ID Identifiers: "FILLER", "CR", and "LF". "FILLER" identifies data columns from the INPUT file to be ignored or on the OUTPUT file to be created and filled with blanks. The "CR" and "LF" Variable IDs must be the last two fields specified in the layout and represent the Carriage Return and Linefeed used in MS-DOS ASCII to denote the end of a data record. Data records from some IBM mainframe systems leave off the Linefeed. (See CRS2LF.) Variable IDs may also include numbers. Numbers in a Variable ID typically represent multiple occurrences of a given variable, like SSAN, which is used from every Master File to merge records. (So, SSAN, SSAN2, SSAN3, etc., is a common occurrence.) When a layout is generated without a pre-existing data dictionary, SPARK generates Variable IDs that identify the Field number within the record, such as, F0001, F0002, F0003, etc., which also occurs when bringing in repeating records from a given source, like the ADSS Training Course Outcomes that have the same format yet pull out separate records for courses like Basic Military Training (BMT), Airman Qualifying (AQ) Courses, and AFSC Awarding Courses. When time permits, the Variable IDs are modified to use a one-letter suffix like "B" for BMT, "Q" for AQ, and "W" for AFSC Awarding Course. When prep time does NOT permit, the Variable IDs are automatically modified to use a 3- or 4-digit number SUFFIX to document its location in the combined file. For example, what might be "SSANB" could be "SSAN2048" for BMT. "SSANQ" could be "SSAN2855" for the AQ section. Finally, "SSANW" could be "SSAN3118" for the AFSC Awarding Course. The benefit is that even if the Variable IDs are not unique within the first four characters, the 4-digit starting column for that field is guaranteed to generate a unique identifier within the merged file. The Variable ID element is terminated with a circumflex (aka, carrot) character ("^").

The third element in a data field definition line is the description of the Data Element. This is a flat ASCII text description field that should NOT contain any of the following symbols: a double quote ("), a circumflex (aka carrot) ("^"), a tab (ASCII 09), a carriage return (ASCII 13), or a linefeed (ASCII 10). If the data field contains Personally Identifiable Information (PII), IAW with the Privacy Act, set the first character of this Description field to an exclamation point ("!"). Also, if this is a relatively unique field of little information to routine study needs (like Assignment ID), also set the PII flag with an exclamation point in the first character. The PII flag will prevent a long, useless distribution report.

If the field is for a DATE, the field should include the format of the data stored. For the standard 8Y field, the format is YYYYMMDD. The Audit program will only report the YYYYMM portion. Other typical date formats include YYMMDD and YYMM. If the field is named "FILLER" or uses an "E" Type code (emit), then the default of all blanks for this field can be overridden. Example: Enter text of "EMIT=xxxx". Every character after the Equal Sign = will be used to the length specified by the field length at the front of the line. Example: "EMIT=YES" would output "YES" if the field length were 3, but only "YE" if the field length was 2.

Also, assume "b" means blank, so if "EMIT=bYES", (because the space after the equal sign is taken as part of the field to be emitted) only "bYE" would be output for length 3 or "bY" for length 2.

The fourth element in a data field definition line is the optional feature for up to two additional columns. If a fourth element will be used, the description element is terminated with a circumflex (aka, carrot) character ("^"). When performing a file REFORMAT process, the REFORMAT process automatically documents where each TARGET data field came from in the SOURCE data file by appending the circumflex and specifying "SOURL (s-e)" showing the starting and ending data columns in the source file. A second optional comment column may also be used by inserting another circumflex and adding another description/note/comment. In the example below, the first comment is the Field ID (e.g., F0001) from automatically imported data file and the second comment reports the data columns from that source file [e.g., SOURL (1-9)].

=====
The Raw layout file from NotePad. (Layout.lay) is input to the LAY2CSV program to generate audit materials.

Example: **Lay2csv 999_860L.LAY 999_860L.CSV 0999-860.DAT**

```
T 999_860L.LAY - Layout for Master Enlisted Applicant to AFSC Award
Master File: 0999-860.DAT
  N This is the original EMAF Master with Enlisted Separation Data
  Appended to the end (0860)
  N This is a Combination of Master Files from the SPARK System - Job
  Families like 0843=ASVAB
  N SPARK: Strategic Policy Assessment Resource and Knowledge Base
  N Per Air Force Instruction - AFI 36-2661 13 Sept 2018 Section 1.3 page
  3
  .....
  N To locate the head of any particular Master File section, locate the
  Variable ID shown below
  N ...
  N .
  N A: PLUS02 A ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-
  MEPS_0843)
  N T: PLUS04 T TAP_Tailored_Adaptive_Personality_Assessment_System
  (TAPAS_0952)
  N I: PLUS06 I ICT_Information-Communication_Literarcy_Test (ICTL_0950)
  N R: PLUS08 R RDB_Recruiting_Service_As_Shipped_to_BMT_Database
  (RDB_0850)
  N D: PLUS10 D PDS_Personnel_Data_System (PDS_0501)
  N B: PLUS12 B BMT_Basic_Military_Training (ADSS_0801)
  N Q: PLUS14 Q AQR_AFFC_Qualifying_Course (ADSS_0801)
  N W: PLUS16 W ABR_AFSC_aWarding_Course (ADSS_801)
  N ...
  N Part 1: Backbone.lay Captured all Unique SSANS from ASVAB TAPAS ICTL
  RDS and ADSS - BMT AQ ABR
  9S SSAN^!F_001 - Social Security Account Number^F_0001^SOURL (1-9)
  1 PLUS01^PLUS01 Above - SSAN Backbone File
  1 PLUS02^PLUS02 Below - A
  ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-MEPS_0843)
```

```

      N Part 2: A ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-
MEPS_0843)
      N LRL=481 w/EOF-LF
      9 SSAN_2A^!F_001 - Social Security Account Number^F_0001^SOURL (1-9)
      008 TDATEA^F_078 - ASVAB_TEST_DATE YYYYMMDD^^SOURL (10-17)
      008 ENTRYDTA^F_053 - DATE_ENTRY DATE YYYYMMDD^^SOURL (18-25)
      002 TESTFOAM^ASVAB TEST FORM ID 28 29^F_0002^SOURL (26-27)
      001 ASVABTEA^F_061 - ASVAB_TEST_VER^F_0005^SOURL (28-28)
      002n AFQTPCTA^AFQT Percentile Score 30 31^F_0002^SOURL (29-30)
      001 MENTCATA^AFOQT Mental Category 32 32^F_0002^SOURL (31-31)
      001 ENTRYSTA^F_042 - ENTRY_STATUS^F_0004^SOURL (32-32)
      002 EXAM_STA^F_052 - EXAM_STATUS^F_0004^SOURL (33-34)
      002n MAGE_MA^F_125 - USAF_COMP1_M2^F_0014^SOURL (35-36)
      002n MAGE_AA^F_126 - USAF_COMP2_A2^F_0014^SOURL (37-38)
      002n MAGE_GA^F_119 - USAF_COMP3_G^F_0013^SOURL (39-40)
      002n MAGE_EA^F_120 - USAF_COMP4_E^F_0013^SOURL (41-42)

```

Running the LAY2CSV on the Main Layout File will produce the Layout.csv for a human-readable file (after loading into EXCEL).

```

      ,,,,,,"To locate the head of any particular Master File section, locate
the Variable ID shown below"
      ,,,,,,"..."
      ,,,,,,"..."
      ,,,,,,"A: PLUS02 A ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-
MEPS_0843)"
      ,,,,,,"T: PLUS04 T TAP_Tailored_Adaptive_Personality_Assessment_System
(TAPAS_0952)"
      ,,,,,,"I: PLUS06 I ICT_Information-Communication_Literarcy_Test
(ICTL_0950)"
      ,,,,,,"R: PLUS08 R RDB_Recruiting_Service_As_Shipped_to_BMT_Database
(RDB_0850)"
      ,,,,,,"D: PLUS10 D PDS_Personnel_Data_System (PDS_0501)"
      ,,,,,,"B: PLUS12 B BMT_Basic_Military_Training (ADSS_0801)"
      ,,,,,,"Q: PLUS14 Q AQR_AFFC_Qualifying_Course (ADSS_0801)"
      ,,,,,,"W: PLUS16 W ABR_AFSC_aWarding_Course (ADSS_801)"
      ,,,,,,"..."
      ,,,,,,"Part 1: Backbone.lay Captured all Unique SSANS from ASVAB TAPAS
ICTL RDS and ADSS - BMT AQ ABR"
      1., "9", "    1", "    9", "SSAN", "!F_001 - Social Security Account
Number", "F_0001", "SOURL (1-9)"
      2., "1", "    10", "    10", "PLUS01", "PLUS01 Above - SSAN Backbone File"
      3., "1", "    11", "    11", "PLUS02", "PLUS02 Below - A
ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-MEPS_0843)"
      ,,,,,,"Part 2: A ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-
MEPS_0843)"
      ,,,,,,"LRL=481 w/EOF-LF"
      4., "9", "    12", "    20", "SSAN_2A", "!F_001 - Social Security Account
Number", "F_0001", "SOURL (1-9)"
      5., "8", "    21", "    28", "TDATEA", "F_078 - ASVAB_TEST_DATE
YYYYMMDD", "", "SOURL (10-17)"
      6., "8", "    29", "    36", "ENTRYDTA", "F_053 - DATE_ENTRY DATE
YYYYMMDD", "", "SOURL (18-25)"
      7., "2", "    37", "    38", "TESTFOAM", "ASVAB TEST FORM ID 28
29", "F_0002", "SOURL (26-27)"

```

```

      8., "1", "    39", "    39", "ASVABTEA", "F_061 -
ASVAB_TEST_VER", "F_0005", "SOURL (28-28)"
      9., "2", "    40", "    41", "AFQTPCTA", "AFQT Percentile Score 30
31", "F_0002", "SOURL (29-30)"
      10., "1", "    42", "    42", "MENTCATA", "AFOQT Mental Category 32
32", "F_0002", "SOURL (31-31)"

```

Lay2CSV also generates the Audit Distributions for all non-flagged (PII) variables.

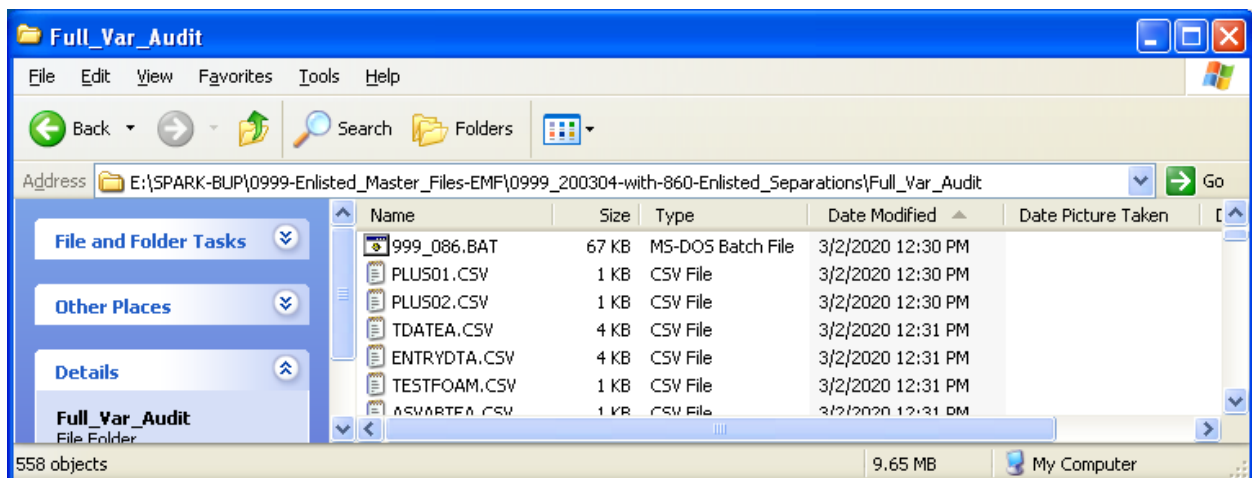
This file is 999_086L.BAT

```

Echo N ...
Echo N Part 1: Backbone.lay Captured all Unique SSANS from ASVAB TAPAS
ICTL RDS and ADSS - BMT AQ ABR
FIX2Frq 0999-860.DAT PLUS01 10 10 F0002:_001/00010_~_PLUS01-
Above---SSAN-Backbone-File
FIX2Frq 0999-860.DAT PLUS02 11 11 F0003:_001/00011_~_PLUS02-
Below---A-ASV_Armed_Services_Vocational_Aptitude_Batt
Echo N Part 2: A ASV_Armed_Services_Vocational_Aptitude_Battery (DMDC-
MEPS_0843)
Echo N LRL=481 w/EOF-LF
Echo Date Field - Reporting only YYYYMM
FIX2Frq 0999-860.DAT TDATEA 21 26 F0005:_008/00021_~_TDATEA
Echo Date Field - Reporting only YYYYMM
FIX2Frq 0999-860.DAT ENTRYDTA 29 34 F0006:_008/00029_~_ENTRYDTA
FIX2Frq 0999-860.DAT TESTFOAM 37 38 F0007:_002/00037_~_TESTFOAM
FIX2Frq 0999-860.DAT ASVABTEA 39 39 F0008:_001/00039_~_ASVABTEA
FIX2Frq 0999-860.DAT AFQTPCTA 40 41 F0009:_002/00040_~_AFQTPCTA

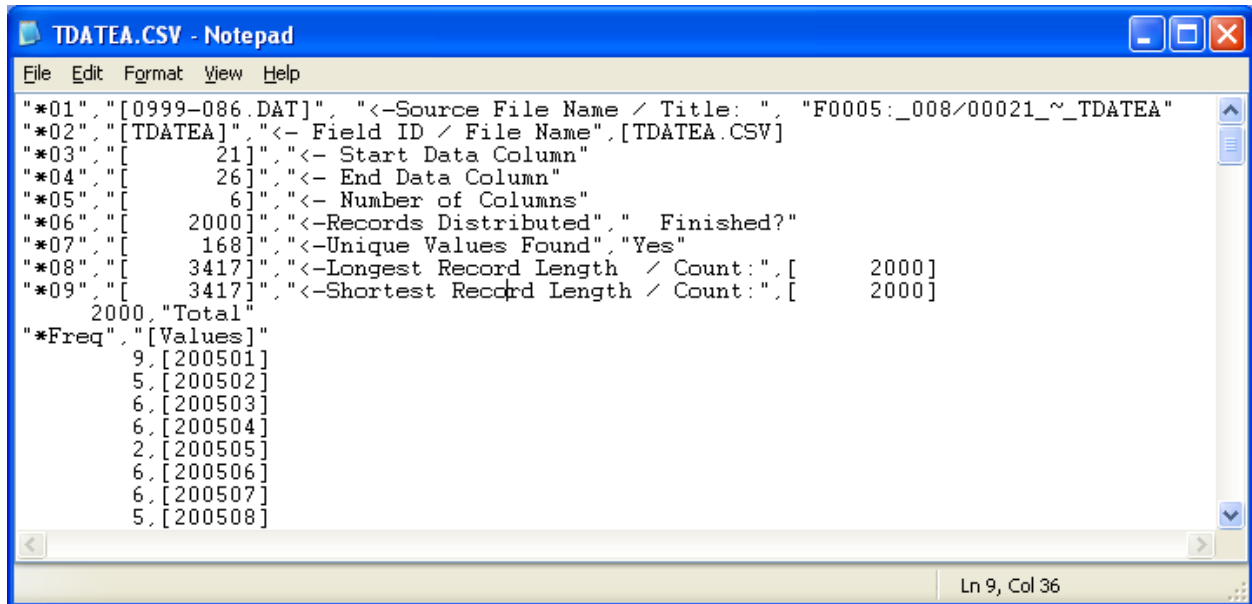
```

Running the Audit Batch file creates an Audit File for each data field.

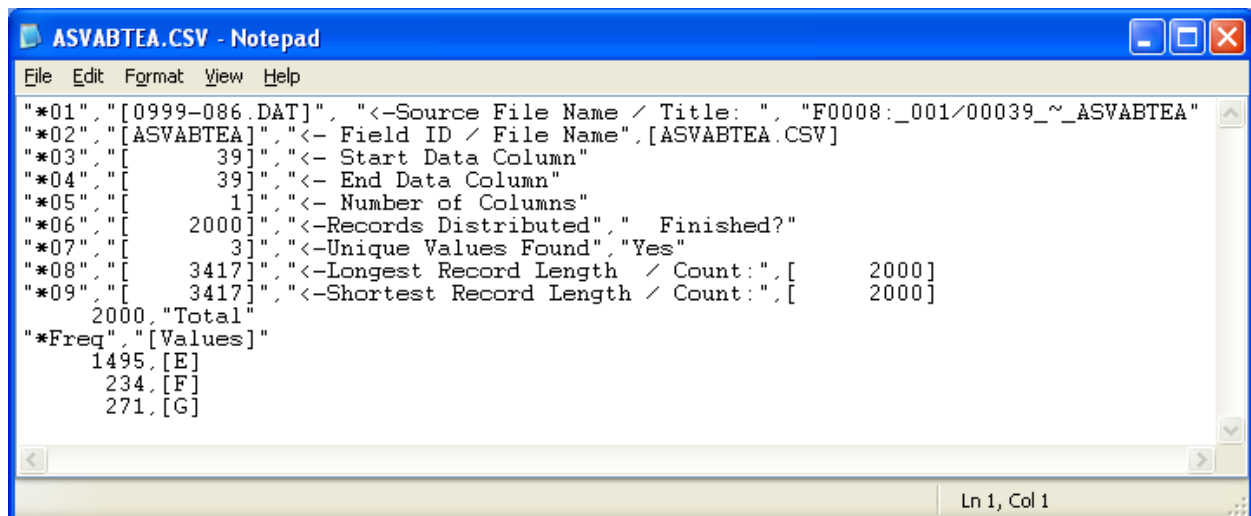


Each data field has a file that shows all the "Unique" values within the field on the file. Below are three examples. (Note: Date fields are automatically reduced to show only year and month.) The *nn Records show

header information for that field (citing only columns actually reported).



```
*01","[0999-086.DAT]", "<-Source File Name / Title: ", "F0005:_008/00021_~_TDATEA"
*02","[TDATEA]", "<- Field ID / File Name",[TDATEA.CSV]
*03","[      21]", "<- Start Data Column"
*04","[      26]", "<- End Data Column"
*05","[      6]", "<- Number of Columns"
*06","[      2000]", "<-Records Distributed"," Finished?"
*07","[      168]", "<-Unique Values Found","Yes"
*08","[      3417]", "<-Longest Record Length / Count:",[      2000]
*09","[      3417]", "<-Shortest Record Length / Count:",[      2000]
2000,"Total"
*Freq","[Values]"
9,[200501]
5,[200502]
6,[200503]
6,[200504]
2,[200505]
6,[200506]
6,[200507]
5,[200508]
```



```
*01","[0999-086.DAT]", "<-Source File Name / Title: ", "F0008:_001/00039_~_ASVABTEA"
*02","[ASVABTEA]", "<- Field ID / File Name",[ASVABTEA.CSV]
*03","[      39]", "<- Start Data Column"
*04","[      39]", "<- End Data Column"
*05","[      1]", "<- Number of Columns"
*06","[      2000]", "<-Records Distributed"," Finished?"
*07","[      3]", "<-Unique Values Found","Yes"
*08","[      3417]", "<-Longest Record Length / Count:",[      2000]
*09","[      3417]", "<-Shortest Record Length / Count:",[      2000]
2000,"Total"
*Freq","[Values]"
1495,[E]
234,[F]
271,[G]
```

```
TESTFOAM.CSV - Notepad
File Edit Format View Help
"*01","[0999-086.DAT)","<-Source File Name / Title: ", "F0007:_002/00037_~_TESTFOAM"
"*02","[TESTFOAM)","<- Field ID / File Name",[TESTFOAM.CSV]
"*03","[      37)","<- Start Data Column"
"*04","[      38)","<- End Data Column"
"*05","[      2)","<- Number of Columns"
"*06","[      2000)","<-Records Distributed", "  Finished?"
"*07","[      28)","<-Unique Values Found","Yes"
"*08","[      3417)","<-Longest Record Length / Count:",[      2000]
"*09","[      3417)","<-Shortest Record Length / Count:",[      2000]
      2000,"Total"
"*Freq","[Values]"
      64,[01]
      82,[02]
      62,[03]
       2,[04]
      27,[05]
      26,[06]
      27,[07]
      30,[08]
       4,[09]
      17,[23]
      20,[24]
      76,[25]
      63,[26]
       1,[27]
       1,[31]
       7,[33]
       8,[34]
     269,[35]
     283,[36]
      96,[37]
     284,[38]
     215,[39]
       8,[40]
      90,[53]
      91,[54]
      66,[55]
      49,[56]
      32,[57]
```

Appendix E: Portal to SPARK: Layout Programs

Note: Lay CSV program: Source code from layout to CSV. CSV file describes the input (layout file). CSV LAY converts CSV files into flat (ASCII) files.

Program CSV2LAY;

{10 Aug 2017 - Removed audit and non-fatal PAUSE for Acknowledgement of intended actions/problems}
{16 Nov 2016 - Added a Separator of Vergule | as well as TAB Use V or v as separator on command line}
{31 Dec 2015 - Removed comment that & is a Tab, now only t or T is a TAB (09) like CSV2FIX }
{14 Aug 2015 - Made BIGOUT unconditional - Removed "Write" of EOF to ERROR file}
{03 May 2012 - Realized "&" means something to DOS on a command line and will NOT feed into the program...}
{03 May 2012 - will use a "t" or "T" to request a TAB separator}
{03 May 2012 - Corrected Logic to use & as "TAB Separated Values" input }
{24 Oct 2011 - Inserted and set MaxFldId to 40}
{29 Sep 2011 - Added a screener for TTMS to remove blank lines or ones that start with '/' or '-'}
{15 Sep 2011 - Allowed to Specify an Alternate SEPARATOR to comma or tab }
{ The TTMS file uses a tilde separator so it is not a .CSV but a .TSV}
{ Also - now, if a non-comma is the separator - changes incoming commas to spaces}
{15 Aug 2011 - Allowed a record to be short ONE Column which is assumed to be a NULL}
{9 Aug 2011 - Added Code to change "NULL" to single space or "#NULL" to a "0"}
{Inspect CSV Format File - Create Fixed-Length Flat File Format Layout 03 Nov 2008}

```
writeln;  
writeln('Proper Usage: ');  
writeln('CSV2LAY [In-CSV File] [Out-Layout File] [IStart] [Separator] [Cutoff]');  
writeln(' Creating Layout for Fixed-length Record version of the CSV File');  
writeln(' See LAYC2TXT for CSV File Conversion to Fixed Text using this layout.');
```

```
writeln;  
writeln(' If you wish to start at a case count OTHER THAN "1",);  
writeln(' specify your desired starting record number in "IStart"');  
writeln(' Example: For *.CSV Files with a LABEL Line, set IStart=2');  
writeln(' Note: The program assumes the SEPARATOR is a COMMA (ASCII 44)');  
writeln(' If you have a DIFFERENT separator,);  
writeln(' say, a tilde ~ (as in TAPAS), enter that SEPARATOR CHARACTER');  
writeln(' as the SEPARATOR following the ISTART designation.');
```

```
writeln(' Separator is a single character to replace COMMA in CSV Format.');
```

```
writeln(' Use Separator = "T" for a TAB. Note - TAPAS uses a tilde ~ PDS uses a $');  
writeln(' Use Separator = "V" for a Vergule or Pipe | (ASCII 124).');
```

```
writeln(' CUTOFF is a command to cut-off file processing at the target CASE NUMBER');  
writeln(' This is helpful in diagnosing file problems.');
```

```
writeln;
```

Program CSV2FLDS;

{11 Aug 2017 - Added more diagnostics to SUMFile}
 {10 Aug 2017 - Removed any PAUSES that are non-fatal}
 {04 Apr 2017 - Tailored BadOUT.DAT to be <Output-File-name>.BAD }
 {04 Apr 2017 - Cleaned up the Intro Help screen and Error Messages}
 {16 Nov 2016 - Augmented Summary File outputs}
 {16 Nov 2016 - Added a Separator of Vergule | as well as TAB Use V or v as separator on command line}
 {30 Dec 2015 - Same Mods as CopyFile to output final record even if it does not end with a CR/LF GETLINE}
 {13 Feb 2015 - Friday the 13th - Added the BADOUT.DAT file for bad records}
 {16 Dec 2013 - Added Bad Record Info into Summary File}
 {16 Dec 2013 - Added Alternate ASCII Separater Character as field #3. Also added Summary File}
 {16 Oct 2013 - changed output error file name to CSV2ESum.txt}
 {Cloned from CSV2LAY on 15 Oct 2013 - rewrite input file purging records with wrong number of fields}
 {03 May 2012 - Realized "&" means something to DOS on a command line and will NOT feed into the program...}
 {03 May 2012 - will use a "t" or "T" to request a TAB separator}
 {03 May 2012 - Corrected Logic to use & as "TAB Separated Values" input }
 {24 Oct 2011 - Inserted and set MaxFldId to 40}
 {29 Sep 2011 - Added a screener for TTMS to remove blank lines or ones that start with '/' or '-'}
 {15 Sep 2011 - Allowed to Specify an Alternate SEPARATOR to comma or tab }
 { The TTMS file uses a tilde separator so it is not a .CSV but a .TSV}
 { Also - now, if a non-comma is the separator - changes incoming commas to spaces}
 {15 Aug 2011 - Allowed a record to be short ONE Column which is assumed to be a NULL}
 {9 Aug 2011 - Added Code to change "NULL" to single space or "#NULL" to a "0"}
 {Inspect CSV Format File - Create Fixed-Length Flat File Format Layout
 03 Nov 2008}

```
writeln;
writeln('Proper Usage: ');
writeln('CSV2Flds [In-CSV File] [Out-CSV File] [Character Separator]');
writeln(' Creating clean copy of the CSV File');
writeln(' Default Character Separator is a Comma');
writeln(' Use Separator = "T" for a TAB. ');
writeln(' Use Separator = "V" for a Vergule or Pipe | (ASCII 124).');
writeln(' Note - TAPAS and ADSS use a tilde ~ Character Separator ');
writeln('      PDS uses a $');
writeln;
writeln(' See CSV2FIX for CSV File Conversion to Fixed Format using this');
writeln('      Clean Output File: ',Output_File);
writeln;
writeln;
```

Program CutTilde;

{For quick edits on over-verbous Layout file titles}
 {07 Feb 2018 - Cloned from CutSpace -- will drop line segment from ^ (leaving ^) through
 closing Tilde ~ }

```
Var IFile, OFile1, OFile2 : Text;
    ILine, IName, OName1, OName2, OLine, JLine : String[255];
    I, J, K: Integer;
    IsAt, ICut, NTab, NChg, NVals, NOut, NDrop, AtRoof, AtTilde, N2Cut : Integer;
    IChar, ARoof, ATilde : Char;

    writeln;
    writeln('Proper Usage: CutTilde [Input-Layout-File] [Output Layout File]');
    writeln;
    halt;
```

Program CSV2FIX;

{02 Feb 2018 - Cloned from CSV2FIX -- now all INPUT files are from C:\C\name }
 {10 Aug 2017 - Removed all non-fatal error PAUSES}
 {16 Nov 2016 - Added a Separator of Vergule | as well as TAB Use V or v as separator on
 command line}
 {30 Dec 2015 - Made Changes same as CopyFile -- to output LAST RECORD if it has no
 CR/LF }
 {30 Jan 2015 - Trying 1600 variables with 10 char IDs to Match CSV2LAYX: Experimental was
 CSV2FIXX }
 {16 Dec 2013 - Removed pause after reporting special character separator. Added Summary
 File}
 {03 May 2012 - Incorporated the "special characters" as separators logic. Using t or T to
 represent TAB delimiters}
 {==== the following note is not true =====}
 {7 Nov 2011 - This program does NOT support Edits-on-the-fly. Take this output to Pair2MAT
 and MAT2Fix}
 {This may have been the start of a new program logic that never went anywhere like
 FIX2CSV...}
 { This program does dynamic reformats from a CSV format file where every RECORD is a
 different layout}
 {==== continue OK notes... =====}
 {24 Oct 2011 - Set MaxFldID Parameter to 40 up from 10}
 {29 Sep 2011 - Added a screener for TTMS to remove blank lines or ones that start with '/' or '-'}
 {16 September 2011 - Allowed an alternate SEPARATOR in place of Comma or TAB}
 { Used in TTMS where the tilde ~ is the separator and COMMAS will become BLANKS}
 {15 August 2011 - Allowed input record to be short one NULL Field without erroring off}
 {24 Jan 2011 - 0927 - Fixed program to NOT output a Lab Head Record}
 {Recompiled without change 24 Jan 2011 - second round expanded AUDIT prints}
 {CSV Format File to Fixed-Length Flat File Format Layout loaded via LayTemp}
 {LayTemp input was from CSV8LAY on 0323-1103L.LAY with possible hand-edits after
 creation}

{20 Dec 2008}

```
writeln;
writeln('Proper Usage:');
writeln(' CSV2FIX [In-CSV File] [Out-Fixed File] [Temp-Layout] [IStart] [Separator]');
writeln(' Creating Fixed Length File from Layout from a CSV File');
writeln(' See LAYTEMP for pre-loading Temporary Layout to be used in this process. ');
writeln;
writeln(' If you wish to start at a case count OTHER THAN "1", ');
writeln(' specify your desired starting record number in "IStart" ');
writeln(' Example: For *.CSV Files with a LABEL Line, set IStart=2 ');
writeln;
writeln(' Separator is a single character to replace COMMA in CSV Format. ');
writeln(' Use Separator = "T" for a TAB. Note - TAPAS uses a tilde ~ PDS uses a $ ');
writeln(' Use Separator = "V" for a Vergule or Pipe | (ASCII 124). ');
```

Program LayAudit;

```
{10 Aug 2017 - Removed all non-fatal PAUSE Messages}
{28 Oct 2011 - Accommodates "<short-name>" syntax - outputs REVISED Layout}
{New Logic - load current layout to Temp files first, THEN Check for Variable ID Conflicts}
{21 Oct 2011 - Began upgrade to 40 Character Variable IDs replaced FldRec.REC with
LayFlds.INC}
{15 Aug 2011 - Ensured the "N" and "Z" options give same results}
{10 Aug 2011 - Added "Y" code to fix date formats & tightened logic on EMIT detection}
{Re-compiled without changes - 24 Jan 2011}
{Revised 28 Dec 2008 to incorporate FldSource and FldEmit fields in Entry Structure}
{Revised at 9:19 PM to include FldPreHat in Entry Structure}
{14 July 2009 - attempted to copy with very long (over 254 chars)}
```

```
writeln('Proper Usage:');
writeln('LayAudit 82910.lay /Audit');
writeln(' where 82910.lay = a RAW layout format (created with NotePad)');
writeln(' /audit causes the program to print diagnostics as it runs. ');
writeln(' Specifying the input file is required. ');
writeln(' REVISED output layout is named <input-name>,REV ');
writeln;
write('Press [ENTER] to halt this program. --> ');
readln(IName);
```

Note: To print the program, take CSV file into Excel.

Appendix F: The History of SPARK Software

There are very few published references relevant to this project, since the “SPARK” (i.e., Master File importing, cleaning, and aggregating) mission was accomplished by specialized software that evolved over multiple decades and was used only to process incoming databases. This process was invisible to those who relied upon its products, and assumed to be done correctly by the Air Force Human Resources Laboratory (AFHRL) Task Scientists, who drew data from the Human Resources Research Databank (HRRD). Documentation for this suite of programs was mainframe computer-based and easily printed, until the AFHRL computer system was “unplugged” and the documentation did not survive the data reduction of 1998. However, in the mid-to-late 1980s, many of these mainframe (Univac 1110) programs were rewritten from COBOL into the PRISM language, developed by Steven O’Hara in the AFPC/SMASZ Programming Unit. Two of the few items that were ever published in the AFHRL Technical Report series were the announcement and description of the PRISM language itself, as well as a description of the Generalized Layout System, both by Steve O’Hara:

O’Hara, S.A., General Purpose File Layout System., AFHRL-SR-82-01, Brooks AFB, TX: Technical Sciences Division, Aug 1982

Rogers, C.R. & O’Hara, S.A., PRISM: A general purpose programming system. AFHRL-TP-82-44, Brooks AFB, TX: Technical Sciences Division, March 1983

Programmers arriving for duty in the Computer Programming Branch of the AFHRL Technical Services Division took courses on how to use these specific programs. Later, as computer terminals to the AFHRL mainframe were replaced by PCs, the opportunity to write one’s own software emerged. When inexpensive computer language compilers like Turbo Pascal became available, these mainframe programs were recoded into Turbo Pascal, which also generated a flurry of special purpose programs. Since the last vestige of the AFHRL mission shut down in 1998, not much of that source code (main program or special purpose) remains. In fact, the only remnant from the former AFHRL mission is a notebook of hard copy program write-ups, explaining the control card options for the primary AFHRL programs.

Transition from HRRD to “SPARK” Software

From this inherited legacy, Johnny Weissmuller of the Air Force Personnel Center (AFPC) undertook the task in 2004 to model new Turbo Pascal programs in the spirit of the AFHRL HRRD approach. His goal was to reclaim the computing power needed to efficiently import data files from external agencies, by converting, reformatting, and recoding them to dovetail into the legacy AFHRL database, which was salvaged. Weissmuller was “one-deep” in this programming effort. He was also the one who both wrote the programs and used the programs to pull data for requesting scientists. Yet, there was never time enough to put together formal documentation. He did adopt, however, “just in time” program information using the practice of having each major program (when executed without parameters) automatically print out the needed guidance for the parameters and input/output file specifications. So, the “documentation” generated for SPARK in this contract is actually a printout of the Turbo Pascal source code, which displays program usage guidance in the absence of parameters.

Since Mr. Weissmuller was one-deep in his support capacity—(not to mention he was hired as a Psychometrician, not a Programmer)—he built software routines that could be easily “parameterized” to service a family of functions for multiple Job Families. Currently, these utility programs form the backbone of the SPARK software package. There are, unfortunately, a number of special purpose programs to do intricate reformatting or recoding that is too complex for a utility program.

The primary reference material for this project was the Turbo Pascal Source Code, which is stored on the AFPC/DSYX SPARK Machine. The main software directory was used for programming the SPARK System and for also the special purpose programming that supports analysis projects, psychometric programming, developing scoring algorithms for new instruments, as well as the historical atCODAP software used in Occupational Analysis. A first review of this software directory showed more than 742 Turbo Pascal programs. Removing the 214, which were last edited prior to 2004, left 526 Turbo Pascal programs as possible SPARK candidates. To identify the most salient possibilities, the 46,313 batch files (in the approximately 30 main SPARK Job Family directory) were searched and all the program invocations in those batch files were alphabetically sorted, with the unique entries identified. This SPARK superset was then inspected and programs were classified as “utility”, “unique-special purpose”, or “logistical support”. The focus of this project was to identify the functionalities needed to replace the SPARK System 1.0, and the following section (Appendix G) shows the final categories for “SPARK” software.

Appendix G: SPARK Software Categories & Commands

Utility (used for file processing, usually INPUT/OUTPUT files)

COPYCRLF.COM
COPYEQUX.COM
COPYFORC.COM
COPYFORK.COM
COPYFOSM.COM
COPYGE.COM
COPYMERG.COM
COPYT2TL.COM
CSV2COMP.COM
CSV2FIX.COM
CSV2FIXC.COM
CSV2FLDS.COM
CUTTILDE.COM
FORKSAME.COM
MAT2FIX.COM
MAT2FIXC.COM
MERMATVC.COM
MIX2PAD.COM
POKEONLY.COM

Psychometric or Special Purpose

DMDC123B.COM
FORK4SQL.COM

Layout Generator (.LAY) or Layout-driven Output (.SAS, .SPS, .LAY)

CSV2LAY.COM
LAY2CSV.COM
LAY2SUFC.COM
MASH2LAY.COM
MASHLEN2.COM

Layout & Script/Code Generation Macros (usually Directory-driven Outputs (.BAT, .PAS))

LAY2PAS2.COM
LAY2SAS.COM
LAY2SASD.COM
LAY2SASW.COM
LAY2SASX.COM

Layout and Utility to Audit/Report File Contents (View/Inspect File; Distribute Contents)

LAY2FREQ.COM
LAYAUDIT.COM

Script/Code Generation Macros

DIR3HXP.COM
DIRLIST.COM
EDITBAT.COM
GEN_EDIT.COM
GRAB_IT.COM

Utilities to Audit/Report File Contents (Output = Audit Display)

CHEK1000.COM
CHEKLENG.COM
FOUR2FRQ.COM

Appendix H: SPARK Data Pull Request Form

1. POINT OF CONTACT (POC) of REQUESTOR:

NAME: _____ TITLE: _____

PHONE: _____ E-MAIL: _____

2. Project ID: _____

3. BRIEF DESCRIPTION OF YOUR PROJECT/OBJECTIVE AND DESCRIPTION OF DATASET REQUIRED. (ATTACH AN EXPLANATION AND/OR PROJECT CHARTER.)

4. TYPE OF RECORDS:

☐ () APPLICANTS

☐ () ACCESSIONS

5. TIME-FRAME: FROM _____ TO _____

6A. SOURCE OF DATA BY JOB FAMILY NUMBER(S):

6B. LAYOUT NAME TO BE ANNOTATED WITH VARIABLE ID REQUEST:

(NOTE: FOR 6A AND 6B ABOVE, SEE APPENDICES A AND B FOR JOB FAMILY DESCRIPTIONS AND LAYOUT NAMES, RESPECTIVELY.)

7. FORMAT OF OUTPUT:

- ☐ () (.CSV) TEXT DELIMITED FILE
- ☐ () (.TXT) TEXT FIXED WIDTH FILE
- ☐ () (.XLS) EXCEL SPREADSHEET
- ☐ () (.SAS) SAS DATASET

8. DIGITAL SIGNATURE & DATE ACCEPTED BY DSYX

KATIE GUNTHER, PhD
Chief, Strategic Research & Assessment Branch (DSYX)
Air Force Personnel Center (AFPC)

Appendix I: SPARK Quick Reference Guide

Overview of Data File Nomenclature

Quick Access Links:

[Locating SPARK Data](#)

[List of SPARK Families](#)

[FISCAL YEAR](#)

[Fiscal Month Coding](#)

[File Naming Standards](#)

[MORE on Job Families](#)

[OPERATIONAL VS PSYCHOMETRIC Files](#)

IMPORTING INTO SPARK

External data files are imported into SPARK from many sources. The raw data file from the source agency is copied onto the DSYX Restricted Drive by the SPARK Analyst with external system access privileges. The SPARK technician then copies the raw master to a CD-ROM for off-line archiving, cleaning, recoding, and standardization on an external USB (labeled “R” in the SPARK clean-room.) This process converts the raw input file using various protocols to create a “normalized standard SPARK Master File.”

LOCATING SPARK DATA (Job Families)

After an imported data update file has been “normalized” for use in SPARK, the standardized format file is merged with the prior Master file for that Job Family. The new, comprehensive Master file is copied onto the DYSX Restricted Drive under the “**9 CAPS/SPARK Masters**” directory. Under this directory, each major type of data is filed in its “Job Family” directory.

DIFFERENT FILE TYPES IN SPARK

1. Import update files (formats determined by the supplying organization).
2. Standardized SPARK update files (after in-processing).
3. SPARK Job Family Master file (flat ASCII file with all prior records merged in).
4. SPARK Layout File (a data dictionary giving field sizes, notes, coding logic and sometimes actual frequency distributions).
5. SPARK Master Converted into SAS automatically by using the SAS Layout and the SAS Job Family Master Data File.
6. SPARK Master Converted into CSV as Exported from SAS.

JOB FAMILIES

Note that each Job Family represents both a “type of data” and a specific format or layout. Sometimes a given test, like the Tailored Adaptive Personality Assessment System (TAPAS, Job Family 0950) undergoes minor modifications during its psychometric development. Minor modifications are typically handled by creating a new Master Layout, which can accommodate both the old and the new data fields in one Master File. TAPAS in Job Family 0950 had about six such modifications and layout adjustments during its development (2006-2016). In 2016, TAPAS went from a PSYCHOMETRIC DEVELOPMENT STAGE into the OPERATIONAL STAGE. The Operational Stage, managed by the Air Force Recruiting Service became a new TAPAS Job Family, 0951. Examples of this “fuzzy” definition can be seen in the AFOQT Job Families (0641, 0986, 0987, 0989, 0991, 0992, 0993, and 0988). Although these are all for the Air Force Officer Qualifying Test (AFOQT), that test battery has undergone major revisions over the years. Each major revision is assigned a new Form identifier such as “O”, “P”, “Q”, “R”, “S” and (as of 2015), “T”. Each Form of the AFOQT has two parallel Versions (e.g., Version, T1, T2). Each version has its unique SCORE KEY, since the versions contain different items or the same items in a different order. To minimize confusion, each Form of the AFOQT is assigned its own Job Family.

NAMING SPARK DATA FILES

In the **SPARK** Job Family directory, data files are labeled by the Job Family ID (000-999) and the **FISCAL YEAR (FY)** (not Calendar Year) and **FISCAL MONTH OF THE MOST RECENT DATA FOUND IN THE FILE**.

FISCAL YEARS and FISCAL MONTHS

Because **SPARK** receives routine data snapshots from federal agencies that base their operational data on **fiscal year boundaries (1 October through 30 September)**, DSYX labels **SPARK** import files with an “as-of” date in terms of Fiscal Year and Fiscal Month (FY FM). Note: In assigning the 12 alpha codes “A”, “B”, “C”...“N” to fiscal months, the letters “I” and “L” are omitted, since *in the past, they have caused confusion and errors for human readers*. The example FY FM designator, “**20H**”, would contain data up into (yet not necessarily completing) May 2020.

FILE NAMING FOR DIFFERENT FILE TYPES

Use the AFOQT Form T for May 2020 as an example. The main AFOQT Form T Master Data file (in flat ASCII) will be named 0992-20H.DAT. Note that this format uses a DASH between the Job Family and the Fiscal Year-Fiscal-Month designation. The extension is “.DAT”, which indicates a text or flat ASCII file in the SPARK system.

The Master Layout File in Excel Format will be named 992L_20H.XLS (perhaps with extra words describing the file as the LAYOUT. Important to note here: When SPARK names files that DESCRIBE the layout of the data file, these files DROP the leading zero on the Job Family ID and use an UNDERSCORE (not a dash) between the Job Family and Fiscal Year. This master Excel Layout is automatically generated from a NOTEPAD “Layout” (992L_20H.LAY) file that has one line for each variable. Each line includes:

- 1) The length of the variable (the fixed number of character positions allocated in the data file for this variable);
- 2) A variable ID to be used in all subsequent processing (usually eight characters or less, using only letters, numbers, or an underscore (again, no dash) and terminating with a circumflex “^”;
- 3) The description of the variable, consisting of text as desired (using no double quote marks);
- 4) Optional information on data format (such as, YYYYMMDD for dates) or codes like M=Male, F=Female. For example: “8 DOB^Date of Birth YYYYMMDD”. Using the LAY2CSV SPARK program generates the 992L_20H.CSV file, which can be clicked on and loaded into Excel and saved in the .XLS format for easy printing.

There is an AUDIT subdirectory within the Job Family Directory. This contains a small text file for most variables in the layout (on the Master file). Variables flagged as PII (or uniquely valued) are skipped. These small text files are named with the VARIABLE-ID dot “FRQ” or “xxx” and show the distribution of values found on the Master file for this variable.

OPERATIONAL (PII) VS PSYCHOMETRIC (CTM) TEST SCORES

In general, the most important distinction is between “operational” and “psychometric” data within the Job Families. The “**operational**” datasets are the official score of record, documenting the final scores from the administration of a given test. These score file records are considered Personally Identifiable Information (PII) and are exported for many operational personnel career-affecting decisions in accordance with published Air Force Policy. Approximately 97% of people who use these scores consider them the **ONLY** scores for that test.

DSYX is responsible for creating, monitoring, and/or upgrading these tests from time to time. When DSYX undertakes a project to create/change a test, there is a time period allotted for trying out new test items, scoring algorithms, benchmarking on a new norming sample, and equating or launching a new percentile conversion process. During the entire time that DSYX is accomplishing these tasks, **NONE** of the resulting (in development) test scores will ever go into a subject’s personnel record, because these scores are not comparable to (i.e., on the same metric as) the current operational test. DSYX labels this whole process as the **psychometric** development of a new/replacement instrument. These datasets include individual item responses so that the effectiveness of proposed distractors can be evaluated and adjusted or replaced, based on the item statistics. In general, any test data file that contains item-level responses is declared to be Controlled Test Material (CTM), iaw AFMAN 36-2664, 16 May 2019). The AFOQT History File is the (CTM) item-level data file used to create the AFOQT Composites. The composite scores are exported (in an 89-character record) to the network and uploaded to the AFOQT Official Score website each morning. The AFOQT History file (with item responses and SDI facet scores) resides **ONLY** on the air gap AFOQT scanner system in the Testing Vault. To use SDI facet scores, these daily History files must be merged, copied to a USB drive, and imported to the air gap SPARK machine in DSYX. The goal is to do this about once a quarter.

Typically, this means that there is **NOT** complete data for that last month. For example, the Enlisted Master Applicant File (EMAF, Job Family 0999) combines data from many other Job

Family sources: The ASVAB data (Job Family 0843) from DMDC and the TAPAS data from either the Army (Job Family 0950-Psychometric TAPAS) or the Recruiting Service (Job Family 0951-Operational TAPAS or Job Family 0952- TAPAS Hybrid). The date on the merged file is set for the most recent data of any of the sources.

Major Data—Type Source is assigned a “Job Family” designation under which all standard processing protocols, true copies of their imports, and the final, exportable SPARK Master File is a flat-ASCII Text file. (“ASCII” stands for American Standard Code for Information Interchange, which has 127 unique character codes.) This standard exists to ensure that whatever changes may occur in data processing standards or application packages (e.g., SPSS or SAS or EXCEL), data can be exported into format, and then imported back into application.

- 1) Regardless of the original file format when received at AFPC, the RAW ORIGINAL file is saved into Comma Separated Value (CSV) format, with Row 1 reserved for Variable IDs.
 - a. The CSV file is an ASCII data file.
 - b. Each record in the CSV file contains only displayable characters. Other than letters, numbers, and displayable special characters, this FLAT ASCII file may contain the following characteristics:
 - i. The TAB character (ASCII 09), which looks like spaces in the text;
 - ii. A Carriage Return (ASCII 12) and a Line Feed (ASCII 10) terminating each record;
 - iii. All other ASCII codes in the range 00-31 and 128-256 are purged from the file during standardization processing.
 - c. A CSV may sometimes use a comma as the separator; however, a TAB (ASCII 09), and a tilde (“~”) are sometimes used, as well.
- 2) During initial importing:
 - a. The first row in the file is parsed to create data variable IDS to be used in the Master Layout. (Note: The user may change these later, but this documents its roots.)
 - b. All data rows in the CSV input file are scanned and the maximum length (i.e., number of characters) needed for each data column is determined.
 - c. A proposed layout (TARGET-1.LAY) is generated in order to define specifications for the FIXED FORMAT, FLAT ASCII Master File.
 - d. The CSV2FIX program is run, inputting the CSV File and the TARGET-1.LAY layout. The SPARK Master File (fixed format, flat ASCII file) is the output.
- 3) Variable-IDs from the SOURCE IMPORT File can be changed to standard SPARK identifiers during import (e.g., ASC or SSN into SSAN), which is accomplished by editing the TARGET-1 layout after import.
- 4) Standardize SSAN variables (numeric, 9 or 11 characters) to-9 ASCII digits with leading zeros (if required) and no dashes, which will be cleaned and removed in the next two steps, (accomplished with “9S” action code in TARGET-1 Layout).
- 5) Assignment of a unique 9-character SPARK-ID for each unique individual in the dataset (separating ID cross-walk to original file into a “clean-room only” .IDS file), accomplished

by the SPARK_ID program after input ID is specified. (Input could be SSAN, DoD ID, Air Force Service Number, etc.).

- 6) Removal of standard or statistically determined Personally Identifiable Information (PII), accomplished by changing the field size to “N” for “Note” in the TARGET-2 Reformat step.
 - a. Person Names, SSAN, DoD ID, Air Force Service Number.
 - b. Any personal street address or phone number (Home of Record, Current).
 - c. Other data fields when the all-variable audit distribution shows a high percentage of unique values, when ignoring invalid codes.
- 7) Standard data field types to SPARK standard format:
 - a. Dates (from various input formats) to a YYYYMMDD format, accomplished by the 8Y action code in TARGET-2 Reformat.
 - b. Whole number numeric fields (age, years of education, ASVAB scores) to standard maximum field length, right adjusted with leading zero, accomplished by nN action code added to field length n on TARGET-2 Reformat.
 - c. Recoded variables due to a code (DICE/FIDO) change by the source agency, accomplished by the POKEONLY or RECODE.
 - d. New coding of input variables, like ASVAB Form and Version IDs, into newly inserted SPARK salient codes (e.g., Norm Scoring Rubric: PAY 80 or PAY 97; Administration method: Paper, Computer; and Testing Protocol: Enlistment Testing Program or Career Exploratory Program.) The new coding is accomplished by special purpose programs, such as CODENORA, CODENORM, CODENOR1-9... or for RACE/ETHNICITY codes, RACEETH4, RACETH9, depending on the input format.
 - e. Text (literal value – like Course Name) variable forced to all uppercase to ensure maximum probability of future matching, accomplished by the nnU action code added to field length nn on the TARGET-2 Reformat.

List of SPARK Job Families

Quick Links Menu for Job Family Descriptions

- [Files Containing both Airmen and Officers](#)
- [Officer-Only Files](#)
- [Airmen-Only Files](#)

Data for Both Enlisted and Officer (including Aircrew Testing)

Data / System ID	Data Description	Job Family	SPARK Master Data Format	Controlled Test Material (Item Level Data)	Personally Identifiable Information	Notes
DRILS	Data Retrieval and Information Locator System	0099	Several	No	No	Overview of all SPARK data files and Processing Procedures
PDS	Personnel Data System Extract	0501	0510L.LAY	No	PII	Comprehensive SSAN Master (Enlisted /Officer)
ADSS	Consolidated Training Outcome File	0802	0802L.LAY	No	PII	
PCSM-TBAS	Pilot Applicant Data from the TBAS Automated Test Station	0664	0664L.LAY	No	PII	Subject to version updates
TIMS GTIMS Going into ADSS	Aircrew, Flying Training Information Management System (obe)	0665	0665L.LAY	No	PII	Historical only. See ADSS or George Michaels/Training Manager for current data
Answer Sheet Scanner Imports	Common Point of Entry for Data on AFOQT Form S Answer Sheets –	0987	0987L.LAY Generic Layout	CTM	PII	Including operational AFOQT Form S until Feb 2015 and JCA and psychometric

APAT,
Experimental,
etc.

Data for
Enlisted
Personnel

AAE / UAR	MILPDS Personnel Data Personnel	0610	Varies	No	PII	Pulled from MILPDS SAS extracts when needed
ASVAB	Accession Testing at MEPS with ASVAB scores and demographics	0843	843L.LAY	No	PII	DMDC just changed format of ASVAB data file & we need to re-tool.
ICTL		0960	960L.LAY	CTM	PII	Psychometric Development of the Information & Communicatio n Technology Literacy Tests
Cyber		0960	960XL.LAY (Extract Format)		PII	Operational Cyber Test Scores
TAPAS	Tailored Adaptive Personality Assessment System	0950	950L.LAY	CTM	PII	Psychometric
	2009-Present					
	2016-Present	0951	951L.LAY		PII	Operational
	2009-Present	0952	952zl.LSY		PII	Operational Fields only
AFRS-RDB	Recruiting Service records for	0850	850L.LAY	No	PII	1 pull annually

“as- shipped”
to BMT

Data for
Officers
(including
USAFA,
ROTC,
OTS, etc)
(Go Back)

BAE / UOR	MILPDS Personnel Data Master	0710	Varies	No	PII	Pull as needed from MILPDS SAS Master BAE
AFOQT (S & T)	Operational SCORE File- On demand data pull from AFOQT Official Score Web Page as needed	0686	AFOQTWEB - Export	No	PII	PII Operational Administration of AFOQT Forms S & T Composites and SubTest #Correct Only
AFOQT (O & R)	Operational SCORE File- On demand data pull from AFOQT Old Score Archive as needed	0686	AFOQTWEB - Export	No	PII	PII Operational Administration s of AFOQT Forms S & T Composites and SubTest #Correct Only
AFOQT Form T	Psychometric AFOQT History Masters	0992	0991L.LAY, 0993L.LAY, and 0992L.LAY	CTM	PII	(0991- Officerhip; 0993- Classification, and 0992- Combined- Form_T_CTM- History File.

AFOQT (O & S)	Psychometric (CTM) AFOQT History Masters – Forms O through S	0641	Multiple	CTM	PII	CTM See Job Family 0992 for current AFOQT Form T data
USAFA	Data on USAF Academy Cadets	0614	0614L.LAY	No	PII	Very rarely obtained
AFROTC	Data from AFROTC Wing Db	0615	0615L.LAY	No	PII	Very rarely obtained

Fiscal Year - Month Coding

Calendar Month	Calendar Year (CY)	Fiscal Year (FY)	Fiscal Month Code	Fiscal Quarter	Notes
October	2019	FY20	A	Q1	(First month of the fiscal year)
November	2019	FY20	B	Q1	
December	2019	FY20	C	Q1	
January	2020	FY20	D	Q2	Calendar Year and Fiscal Year now Match)
February	2020	FY20	E	Q2	
March	2020	FY20	F	Q2	
April	2020	FY20	G	Q3	
May	2020	FY20	H	Q3	(This is the example used in the text)
---	---	---	I	--	"I" - do not use – may confuse humans
June	2020	FY20	J	Q3	
July	2020	FY20	K	Q4	
---	---	---	L	--	"L" - do not use – may confuse humans
August	2020	FY20	M	Q4	
September	2020	FY20	N	Q4	(Last month of the fiscal year)
October	2020	FY21	A	Q1	(First month of the fiscal year)

Fiscal Year (Oct-Sep) is abbreviated as "FY"; Calendar Year (Jan-Dec) is abbreviated "CY".

SPARK File Naming Standards

Each Job Family directory should contain:

- 1) The SPARK Master Data File in three formats (as of FY-FM)**
 - a. Flat ASCII format (to ensure maximum portability)
 - b. Labeled as 0<JOBFAMILY#>-FYM.DAT:
 - i. Example 0843-20J.DAT – ASVAB Job Family with data into fiscal month “J” = June 2020
 - ii. Zero+3-digit-Job-Family ID+
 - iii. DASH +
 - iv. Fiscal Year (FY) and
 - v. Fiscal Month (A-N)
 - vi. Dot “DAT”
 - c. SAS Format (for immediate use by AFPC/DSYX researchers)
 - d. Excel CSV format (for export to external researchers)
- 2) The Master Layout File in Excel Format**
 - a. A simple text file created with Notepad or automatically by SPARK
 - b. Data dictionary with data field size, type, variable ID, and description
 - c. Optionally, comments, “codes” or links to coding files
 - d. Annotations or audit frequency distributions for key variables
 - e. Standard Layout Labeled:
 - i. Example 843L_20J.LAY – ASVAB Job Family Layout for fiscal month “J” = June 2020
 - ii. 3-digit Job Family ID (no leading zero) +
 - iii. The Letter “L” for Layout
 - iv. UNDERSCORE
 - v. Fiscal Year (FY)
 - vi. Fiscal Month (A-N)
 - vii. Dot “LAY”
 - viii. Note: The “generic” layout for this Master file would be 843L.LAY without the FY or FM.
 - f. A “Frequency-Annotated” Layout Labeled:
 - i. Same as Standard, except the “L” following the Job Family is changed to an “F”.
 - g. The original text file is processed by SPARK programs to create a CSV version that is loaded into Excel for final form.
- 3) The Audit Subdirectory** (complete or stratified random subset file with distributions showing actual values for each variable.
- 4) The “Location of” Word doc**, with summary information showing this location and the materials found here.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

AAE	Airman Active Extract (Enlisted Personnel File)
ADSS	Automated Data Subsystem
AETC	Air Education and Training Command
AF/AIPT	Air Staff Directorate Force Management Policy for Accession & Training
AFCD	Air Force Classification Directory
AAE	Active Airmen Extract (Enlisted Personnel File)
AFCAPS	Air Force Center for Applied Personnel Studies
AFECD	Air Force Enlisted Classification Directory
AFHRL	Air Force Human Resources Lab
AFI	Air Force Instruction
AFIT	Air Force Institute of Technology
AFMAN	Air Force Manual
AFOCD	Air Force Officer Classification Directory
AFOMS	Air Force Occupational Measurement Squadron
AFOQT	Air Force Officer Qualifying Test
AFPC	Air Force Personnel Center
AFPT	Air Force Personnel Test
AFRL	Air Force Research Laboratory
AFRS	Air Force Recruiting Service
AFSC	Air Force Specialty Code
AF-WIN	Air Force Work Interest Navigator
APAT	Applied Performance Assessment and Testing (Facility at Lackland AFB)
ASCII	American Standard Code for Information Interchange
ASVAB	Armed Services Vocational Aptitude Battery
BAE	Officer Active Extract (Officer Personnel File)
BMTS	Basic Military Training School

CFETP	Career Field Education and Training Plan (Available online from Air Force e-pubs)
CFM	Career Field Manager
CoP	Community of Practice
CSV	Comma Separated Value (May substitute a TAB, \$, or ^ as the delimiter vice the comma)
CT	Cyber Test (formerly developed as ICTL)
CTM	Controlled Test Material
DICE	Data Item Code Explanation
DMDC	Defense Manpower Data Center
DUA	Data Use Agreement
DoD	Department of Defense
DoDI	Department of Defense Instruction
DoD ID	Department of Defense Identifier (cited on back of Common Access Card (CAC))
DRILS	Data Retrieval and Information Locator System
DSYX	Strategic Research and Assessment Branch (HQ AFPC/DSYX)
DTIC	Defense Technical Information Center
EAP	Examining Activities Program (USAF EAP is currently located in HQ AF/AIPT)
EDIPI	Electronic Data Interchange Personal Identifier (now called DoD ID)
EMAF	Enlisted Master Applicant File
FAA	Federal Aviation Administration
FAIR	Findable, Accessible, Interoperable, and Re-usable
FIDO	File Item Data Overview
FOUO	For Official Use Only
GDIT	Generalized Data Index Table
GTIMS	Graduate Training Integration Management System
HRRD	Human Resources Research Databank

ICTL	Information-Communication Technology Literacy
IJOA	Institute for Job and Occupational Analysis
IMTA	International Military Testing Association
IRB	Institutional Review Board
I/O	Industrial/Organizational (I/O Psychologists)
JCA	Job Compatibility Assessment
MAGE	Mechanical, Administrative, General, and Electronics (US Air Force ASVAB classification composites)
MAPWG	Manpower Accession Policy Working Group
MOA	Memorandum of Agreement
MOU	Memorandum of Understanding
MEPS	Military Entrance Processing Station
METS	Military Entrance Test Sites
MilPDS	Military Personnel Data System
MOU	Memorandum of Understanding
MXN	Maintenance Next
OA	Occupational Analysis
OBE	Overcome by Events
ORDB	Occupational Research Databank
OTA	Oracle Training Application
OTS	Officer Training School
PAF	Project Air Force
PCSM	Pilot Candidate Selection Method
PC	Personal Computer
PDS	Personnel Data System
PIA	Privacy Impact Assessments
PII	Personally Identifiable Information
PM	Program Manager

POE	Point of Entry
PSM	Predictive Success Model
R&D	Research and Development
RDB	Recruit Database
ROTC	Reserve Officer Trainer Corps
RPR	Request for Personnel Research
RUID	Random User Identification
SAPR	Sexual Assault Prevention Reporting
SAS	Statistical Analysis System
SIMF	SPARK Identify Master File
SIOP	Society for Industrial and Organizational Psychology
SJT	Situational Judgment Test
SKT	Specialty Knowledge Test
SOR	System of Record
SORN	System of Record Notice
SOURL	Source Layout
SPAP	Strategic Personnel Assessment Program
SPARK	Strategic Policy Analysis Resource and Knowledgebase
SPARK-ID	Strategic Policy Analysis Resource and Knowledgebase Identifier
SPSS	Statistical Package for the Social Sciences
SSAN	Social Security Account Number
STIP	Scientific and Technical Information Program
TAPAS	Tailored Adaptive Personality Assessment System
TBAS	Test of Basic Aviation Skills
TCO	Technical Control Officer
TUA	Test User Agreement
TIMS	Training Integration Management System
TTMS	Technical Training Management System

TXT	Text (File Extension)
UAR	Uniform Airman Record (Now use AAE)
UMD	Unit Manpower Document
UOR	Uniform Officer Record (Now use BAE)
USAFA	United States Air Force Academy
USMEPCOM	United States Military Entrance Processing Command
VM	Virtual Machine
WAPS	Weighted Airman Promotion System
XLS or XLSX	Excel Spreadsheet