Using the Person-Event Data Environment for Military Personnel Research in the Department of Defense

An Evaluation of Capability and Potential Uses

David Knapp, Beth J. Asch, Christine DeMartini, Teague Ruder, Janet M. Hanley

For more information on this publication, visit www.rand.org/t/RR2302

Published by the RAND Corporation, Santa Monica, Calif.

Preface

The objectives of the study described in this report are to determine whether the RAND Corporation's federally funded research and development centers (FFRDCs) for the U.S. Department of Defense (DoD) can effectively and efficiently use the Person-Event Data Environment (PDE) to support DoD-sponsored manpower and personnel research, to assess how using the PDE compares with existing approaches to accessing defense manpower data (notably, via the Defense Manpower Data Center [DMDC]), and to identify what improvements to the PDE would be necessary for it to be used by RAND's FFRDCs for personnel research. The PDE is a computing environment that allows approved users to access and use defense manpower and personnel data. The PDE is currently run by the Army but includes extracts of DoD-wide data. Our approach for assessing the PDE was to (1) identify the data collection and analytical requirements from three in-progress or completed RAND studies typical of manpower and personnel studies conducted in RAND's DoD FFRDCs and (2) replicate the data collection and analysis using the PDE.

At the time this research was undertaken, interest in the possibility of using the PDE as a means of providing data for RAND FFRDC research was growing at DoD. DMDC staff indicated in discussions that future provision of survey data would occur only through the PDE and that DMDC was considering a similar policy for administrative (non-survey) data. The study described in this report ran from October 2015 to September 2017. In February 2016, the Research Facilitation Laboratory released PDE version 2.0, and our analyses were based on our interactions in this newer environment.

RAND Ventures

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND Ventures is a vehicle for investing in policy solutions. Philanthropic contributions support our ability to take the long view, tackle tough and often-controversial topics, and share our findings in innovative and compelling ways. RAND's research findings and recommendations are based on data and evidence and therefore do not necessarily reflect the policy preferences or interests of its clients, donors, or supporters.

Funding for this venture was made possible by the independent research and development provisions of RAND's contracts for the operation of its U.S. Department of Defense federally funded research and development centers.

Contents

Summary

The objectives of this study are to determine whether the RAND Corporation's federally funded research and development centers (FFRDCs) can effectively and efficiently use the Person-Event Data Environment (PDE) to support research studies for the U.S. Department of Defense (DoD), to assess how using the PDE compares with the traditional method of accessing personnel data, and to identify what improvements to the PDE would be necessary for it to be used by RAND's FFRDCs for personnel research. The PDE is a secure, remote-access, analytical computing resource that is made available to researchers and institutions conducting research using Army and other DoD manpower data. Traditionally, RAND's DoD FFRDCs have received data directly from data providers, such as the armed services and the Defense Manpower Data Center, and handled research in RAND's own secure analytical environment. Our approach for assessing the PDE was to (1) identify the data collection and analytical requirements from three in-progress or completed RAND studies typical of manpower and personnel studies conducted in RAND's DoD FFRDCs and (2) replicate the data collection and analysis using the PDE.

We developed a framework to show conceptually how the PDE fits within a DoD-sponsored study. This framework contrasts the costs and benefits of integrating the key elements of a DoD-sponsored study: the sponsor, researcher, analytical environment, and data provider. The PDE reflects specialization in provision of an analytical environment. We contrast the PDE concept's separation of the researcher's and the analytical environment's institutions with the typical RAND FFRDC setup, which integrates researcher and environment.

Our research team was able to acquire data from Defense Manpower Data Center and Army Human Resources Command databases, validate that these providers' data held in the PDE and in the RAND environment were of similar quality, and conduct the analyses for the three studies. Compared with the process when using RAND FFRDCs' analytical environment, the process for acquiring data and conducting analyses in the PDE had many similarities, but there were two recurring themes. First, transaction costs, such as additional required approvals and operational and coordination steps, can be high and lead to delays when the research organization is separate from the organization providing the environment. Second, incentive alignment between the research and analytical environment is important so that transaction costs, though they exist, are resolved more quickly and efficiently. When priorities are not aligned between the research organization and the organization providing the environment, research progress can be limited unilaterally by either organization. Research is a very iterative process, meaning that incentive misalignment and transaction costs between researchers and their analytical environment can lead to substantial delays in providing results to research sponsors, as well as an inability to provide quick support to the FFRDC's sponsor, a contractual requirement of FFRDCs.

We conclude that a nonintegrated analytical environment, such as the PDE, impedes the efficient operation of a RAND FFRDC study relative to existing arrangements by eliminating the researcher and analytical environment's alignment of incentives. Additionally, we make two observations based on this evaluation: (1) A single centralized analytical environment for research processes within DoD has the potential to reduce analytical capacity and discourage researchers from accumulating database-specific knowledge; and (2) the PDE represents a potential opportunity for DoD to engage academic researchers outside of existing research support organizations with recurring DoD sponsor relationships.

Acknowledgments

This project was managed jointly by RAND National Defense Research Institute's Forces and Resources Policy Center and RAND Arroyo Center's Personnel, Training, and Health Program. The authors would like to acknowledge the support during project development and implementation from RAND colleagues John Winkler, Michael Hansen, Ray Conley, Susan Marquis, and Howard Shatz. We also want to thank Michael Mattock for implementing the Dynamic Retention Model in Mattock and Arkes (2007) using the Person-Event Data Environment, Heather Krull for her assistance during project development, and Linda Cottrell for her help comparing analytical outcomes. We deeply appreciate the willingness of the Research Facilitation Laboratory to allow and support this evaluation of the Person-Event Data Environment, including MAJ Paul Lester, Jennifer Murguia, and Brent Ivester. Finally, our report benefited from the excellent comments and input we received from three reviewers: Michael Hansen of RAND, Carra Sims of RAND, and Susan Rose of the Institute for Defense Analyses.

Abbreviations

CAC	Common Access Card
DMDC	Defense Manpower Data Center
DoD	U.S. Department of Defense
DRM	dynamic retention model
EDO	exempt determination official
ETL	extract, transform, and load
FFRDC	federally funded research and development center
GB	gigabyte
HRC	Human Resources Command
HRPP	Human Research Protection Program
IRB	institutional review board
MEPCOM	Military Entrance Processing Command
NDRI	National Defense Research Institute
OCIO	Office of the Chief Information Officer
OSD	Office of the Secretary of Defense
OUSD (P&R)	Office of the Under Secretary of Defense for Personnel and Readiness
PDE	Person-Event Data Environment
PII	personally identifiable information
PHI	protected health information
RAM	random-access memory
RFL	Research Facilitation Laboratory
RMC	regular military compensation
SPA	specific project amendment

Chapter One

Introduction

A substantial amount of analysis and research[1] in support of manpower and personnel policy in the U.S. Department of Defense (DoD) is carried out by federally funded research and development centers (FFRDCs). An FFRDC is a private-sector organization that meets special long-term research or development needs of a government agency that cannot be met as effectively by existing in-house or contractor resources. According to Code of Federal Regulations, Title 48, Section 35.017,

> FFRDC's [sic] enable agencies to . . . accomplish tasks that are integral to the mission and operation of the sponsoring agency. An FFRDC . . . has access, beyond that which is common to the normal contractual relationship, to Government and supplier data, including sensitive and proprietary data, and to employees and installations equipment and real property. The FFRDC is required to . . . operate in the public interest with objectivity and independence. . . . Long-term relationships between the Government and FFRDC's are encouraged in order to provide the continuity that will attract high-quality personnel to the FFRDC. This relationship should be of a type to encourage the FFRDC to maintain currency in its field(s) of expertise, maintain its objectivity and independence, preserve its familiarity with the needs of its sponsor(s), and provide a quick response capability. (Code of Federal Regulations, 2017)

Much of the manpower and personnel research conducted by FFRDCs makes extensive use of administrative records of military personnel, provided by the Defense Manpower Data Center (DMDC). Traditionally, data are provided by DMDC to the FFRDCs either through regular data extracts to support project work for DoD sponsors or through custom-made extracts that are provided as needed. Given the nature of the projects conducted in FFRDCs, custom-made extracts are frequently needed.

In 2005, the Army created the Person-Event Data Environment (PDE) in cooperation with DMDC and the Navy. According to the 2010 memorandum of agreement between DMDC and Army Data Center Fairfield,

> The purpose of the PDE was to establish an environment where research activities could share datasets for study analysis primarily in Manpower and Medical areas. The objective was to bring the analyst to the data, and minimize the practice of sending data to the analyst, exposing the DoD to loss of privacy data. (DMDC, 2010)

[1] By *research*, we mean a systematic investigation to develop or contribute to generalizable knowledge, including investigations for purposes of the federal policy. According to Shelton (1999), research can include a wide variety of activities, including experiments, observational studies, surveys, tests, and recordings designed to contribute to generalizable knowledge.

Under this concept, researchers access and analyze data within the PDE so that the responsibility for maintaining an analytical environment is placed with a third party instead of with the FFRDC or the researcher's institution.

An increased awareness of data security has led to increased vetting, as well as new and unclear requirements, while tight budgets have also limited data providers' staff time. Consequently, receiving custom-made extracts from DMDC and other data providers has become significantly more time-consuming in recent years. DMDC has said that it would like to replace the custom-made approach to the extent possible by requiring researchers to access data primarily through the PDE as a means of improving data security and reducing the staffing burden of creating custom-made data sets for researchers. This would mean that FFRDCs would be required to access data mostly through the PDE.

A critical question is whether this would degrade or enhance the ability of FFRDCs to meet their mission of providing high-quality and objective research in a timely manner to DoD sponsors. Thus, it is important to determine whether the PDE performs as intended and as DMDC expects, whether the PDE is suitable for studies conducted by FFRDCs, and whether the PDE is superior to the traditional approach to accessing requisite manpower and personnel data. The purpose of the study summarized in this report is to conduct analyses to address these questions.

Background on the PDE

The PDE is a secure, remote-access, analytical computing resource that is made available to researchers and institutions conducting research using Army and other DoD manpower data. The PDE's design—a remote access server with individualized data sets generated to fit each user's data request—is becoming a popular method of providing researchers with access to restricted data while permitting the data-owning institution to maintain control of its data. In recent years, the Center for Medicare & Medicaid Services and the Health and Retirement Study have separately considered or are currently considering a similar system.

The PDE is supported by the Army and administered by the Research Facilitation Laboratory (RFL), an organization that is primarily composed of Northrup Grumman contractors. The Army Analytics Group oversees RFL.[2] According to RFL, a purpose of the PDE is to "streamline the data access, data management, and data governance bureaucracy" of projects in support of DoD (Research Facilitation Laboratory, 2015). The PDE is meant to provide a space where authorized researchers for DoD-sponsored studies can access and merge data and conduct analysis in a protected environment.

[2] In 2011, the Army Data Center Fairfield's responsibilities were transferred to the Army Analytics Group.

Study Objectives and Approach

The objectives of the project are to determine whether FFRDCs can effectively and efficiently use the PDE to support research studies for DoD, to assess how using the PDE compares with the traditional method of accessing DMDC data, and to identify what improvements to the PDE would be necessary for it to be used by RAND's FFRDCs for personnel research. An appropriate test of the PDE requires

- determining whether FFRDC personnel studies can be conducted within the PDE
- validating whether required personnel data are available or can be collected for use in the PDE
- comparing, relative to alternative analytic arrangements, the process for adding data, conducting analysis, and reporting results to research sponsors.

Thus, the approach we used to assess the PDE was to (1) identify the data collection and analytical requirements from three in-progress or completed RAND studies typical of manpower and personnel studies conducted in RAND's DoD FFRDCs and (2) replicate the data collection and analysis using the PDE. Each of the three projects represented studies typically conducted in three of RAND's FFRDCs—namely, the U.S. Army's RAND Arroyo Center, the Office of the Secretary of Defense (OSD)'s National Defense Research Institute (NDRI), and the U.S. Air Force's RAND Project AIR FORCE.[3]

RAND1

The first project, denoted RAND1, is a study of Army recruiting that made use of DMDC contract and accession data, Army contract and accession data, and Army enlisted personnel data. Given that almost one-third of Army enlistments do not complete their first enlistment term, and not all enlistment contracts actually access, the project evaluated the factors associated with contract attrition, first-term attrition, and continuation in service, as well as promotion timing during the first term. Specifically, RAND1 linked Army enlistment data, DMDC Military Entrance Processing Command (MEPCOM) enlistment data, and Army Total Army Personnel Database personnel data to assess the relationship between characteristics measured at the time of enlistment and subsequent continuation and promotion outcomes. This involved creating a longitudinal history of Army first-term enlistees who enlisted during or after 2009 and tracking their service record through 2015.

RAND has a long history of conducting studies of Army recruiting in the RAND Arroyo Center and NDRI (e.g., Polich, Dertouzos, and Press, 1986). Thus, the RAND1 project provides a PDE test case for analysis commonly done in these two FFRDCs.

[3] The projects do not map perfectly to in-progress or completed RAND studies, so we use generic names to prevent incorrect association.

RAND2

The second project, denoted RAND2, is a study of Air Force officer retention that estimated a dynamic retention model (DRM) using individual-level longitudinal data constructed from DMDC's active-duty master files from 1990 to the present. The DRM is a stochastic dynamic programming model of an individual's decision to stay or leave active duty. It is a life-cycle model in which retention decisions are based on forward-looking behavior that depends on current and future military and civilian compensation and recognizes that decisions are made under uncertainty and that individuals are heterogeneous in their tastes for military service. The DRM estimated in the RAND2 project is a publicly available version for Air Force pilots from Mattock and Arkes (2007), conducted within RAND Project AIR FORCE. In addition to allowing us to test the PDE capability in terms of developing a longitudinal database of Air Force officer retention outcomes, RAND2 allowed us to assess the computational capabilities of the PDE because estimation of the DRM requires significant computational power, such as high memory or central processing unit (CPU) loads. As part of RAND2, we also used DMDC active-duty pay files to construct measures of regular military compensation (RMC) in 2008, 2010, and 2012. RMC is an input to the DRM, and the computation of RMC allowed us to assess the ability to use the pay files in the PDE.

RAND3

The third project, denoted RAND3, is a military health study that examined the relationship between enlisted medical standards and screening and later career outcomes, including early career attrition, promotion, disability separation, and the receipt of disability separation pay or disability retired pay. The study linked individual-level data on service members in all services to create a longitudinal profile of a service member's time since application, including such characteristics as service, gender, enlistment waivers, medical standards, separation reason, disability rating, and disability separation or retirement payments. The study linked data from DMDC's MEPCOM files, active-duty master files, active-duty loss files, active-duty pay files, and retiree pay files covering 1988 to the present.

RAND has conducted several military health studies, including studies of the DoD and U.S. Department of Veterans Affairs disability systems, and RAND3 provides a PDE test case for analysis commonly done in this area.

Analysis Plan

The focus of our replication of these studies was on testing the PDE rather than conducting a full analysis and documenting those analyses per se. Thus, for each study, we requested the necessary data, similar to the data for the original study that had been requested in the traditional way by RAND. We then developed analytic files in the PDE and conducted representative analysis to test the PDE capability. For each project, we evaluated the ease of use of the PDE in

practice and whether the PDE operates as intended, and we examined how conducting analyses in the PDE compared with conducting analyses in the RAND environment. Based on our experience with these projects, we identified unresolved issues and areas where the PDE could be improved, and we identified the types of analyses that seem best suited for the PDE.

Scope

This report evaluates only the RAND research team's experience in the PDE. Comparisons are reported relative to the research team's experience with the RAND FFRDC's analytical environment. The research team includes RAND staff who have worked with military personnel data for more than 30 years, staff who operate on a routine basis in the RAND analytical environment, and staff who have participated in all steps of the research process outlined in Chapter Three.

The research team attempted to acquire data only from DMDC and the U.S. Army Human Resources Command (HRC) for this study. Other typical data providers for RAND FFRDC studies include the Office of People Analytics, Department of Veterans Affairs, Social Security Administration, U.S. Air Force Personnel Center, and Defense Health Agency, among others. The research team also did not conduct its own surveys or focus groups as part of this project or attempt to analyze such data within the PDE. Data collection from other data providers or primary data collection may result in experiences different from those documented in this report.

Additionally, we evaluated only RAND1, RAND2, and RAND3. These projects were chosen in consultation with the directors of RAND's FFRDCs that conduct personnel research. While the requirements of these projects are similar to most RAND manpower and personnel studies, some RAND studies require coordination with more than two data providers or have limitations that we did not explore. For example, some studies require data from providers that do not provide data directly and instead only provide access to query-based data systems where a user from the research staff or analytical environment extracts the data from the data provider's system.

Limitations

Ideally, our study would be able to quantifiably compare time spent in the PDE versus time spent in the RAND FFRDC analytical environment. We are generally unable to make direct comparisons for two reasons:

1. We cannot identify whether delays in the PDE, if they exist, were due to that environment or to a third party, such as the data provider.
2. Because we replicated in the PDE the data collection and analysis from three projects (RAND1, RAND2, RAND3) that reflect broad analytical requirements of in-progress or completed RAND studies, we did not experience the iterative process typical of a research study in which analytical programs are developed and refined over the course of the study.

As noted in Chapter Four, FFRDC personnel studies that require analytical environments like the PDE or RAND's FFRDC analytical environment tend to involve a highly iterative and collaborative process among the research team, often requiring one member of the research team to submit programs in the analytical environment that produce tabulations or estimate models, extract and review the output of those programs for privacy and data safeguarding concerns, discuss that output with the rest of the research team (outside of the analytical environment), and then revise and resubmit the programs in the analytical environment based on the research team's discussions. This process is often repeated many times and cannot be duplicated, even with the same research team, because the lessons learned during the initial research process have been learned.

The few times we do bring up timing or waiting periods during this report are (1) to highlight the benefits or costs of separating the institution providing the researcher staff from the institution providing the analytical environment or (2) to provide the reader a sense of length of time where relative comparisons are possible. Where timing is mentioned, it is important to remember that these are examples and may not be representative of an average user's experience in the RAND FFRDC analytical environment or the PDE.

Finally, we address efficiency or cost only from the point of view of a RAND FFRDC study research team. We do not identify or attempt to quantify the level of risk that DoD is exposed to by having a RAND FFRDC study's data analysis occur in the RAND FFRDC environment relative to the PDE. Regarding cost, as a research institution, RAND would have an institutional review board (IRB), analytical environment, and established privacy and data safeguarding protocols regardless of whether RAND's FFRDCs conduct personnel research in the PDE or the RAND FFRDC's analytical environment. The differences in cost to DoD between using one environment or the other would be primarily driven by the potential cost savings in labor time. Because we do not observe individuals' labor time or cost outside of RAND FFRDCs (i.e., PDE staff, the research sponsors, the FFRDC's sponsors, data providers, or other actors), we cannot quantify the relative costs or impact on efficiency borne by other organizations.

Road Map

The report is organized as follows. Chapter Two provides an overview of the key elements required for conducting a successful study and presents a conceptual framework of the different approaches for organizing those elements. Chapter Three provides an overview of the steps required for conducting a study. Chapter Four shows the results of our analysis. It compares the PDE with the RAND environment in conducting the study steps and assesses whether the PDE works as intended. We offer concluding thoughts and recommendations in Chapter Five.

Chapter Two

Conceptual Framework for a Study

The PDE represents a new approach for accessing defense manpower data and conducting analysis, essentially adding an additional element or organization into the research study process. This new model raises the question of the costs and benefits of separating the elements of a study so that they are performed by separate organizations—for example, specialization by each organization; integration within a single organization; or something in between, in which some elements are conducted within a single organization and some by separate organizations. In this chapter, we discuss this issue conceptually, focusing on research studies that address questions that are sufficiently complex as to require a study. We consider the key elements to a study, the role of specialization across the study elements, and the costs and benefits (i.e., trade-offs) of integrating the elements within a single organization versus specializing elements into separate organizations. These trade-offs highlight that the context surrounding a research question is important for determining when specialization can lead to efficiencies or inefficiencies.

Key Elements in a Study

The four key elements required for answering a research question are as follows:

1. Sponsor: The person or organization that asks the question.
2. Researcher: The person or organization that has the technical capability to answer the question.
3. Data: The raw information required as input to conduct the analysis. Data are held by a data provider, which is the person or organization responsible for holding, protecting, and in some cases collecting the data relevant for the analysis.
4. Environment: The place where the analysis is conducted that meets the sponsor's, researcher's, and data provider's requirements to analyze and protect the data. The data environment facilitates the creation of analytical data sets, including data access, linking, and manipulation.

For many research questions, the four elements are all within one organization. This is a case of complete integration. In the DoD context, an office might be responsible for executing and collecting the results from a program. The office director may ask whether the program was successful. He or she could assign the analysis to one of the office's staff members, who then uses the internal data collected to determine whether the program was successful. The data are held on the office's computer server, and the staff member's computer acts as the analytical environment. The office is responsible for data protection.

Role of Specialization in a Study

As we discuss later, complete integration of all four elements within a single organization has several advantages, such as lower coordination costs across the elements, aligned management oversight, aligned incentives toward achieving the common goals of the study, and integrated (i.e., cross-element) institutional expertise. But specialization can have benefits as well. Specialization can be thought of as incomplete integration, where the sponsor relies on an external person or organization to answer its research question. There are many reasons for a sponsor to want or need specialization. We highlight four:

1. independence
2. lack of technical expertise
3. limited internal resources
4. economies of scale.

Independence means that the study is objective, balanced, and reliable and applies the highest standards of rigor. If the study sponsor has a real or perceived interest in a particular outcome of the study, then the sponsor is not perceived as objective and the value of the study becomes questionable. For example, DoD may request a study to inform the budgeting process through analysis that is not influenced by interested parties, either in fact or by perception. In this case, it is preferable for the study to be conducted independent of the sponsor.

Another instance where complete integration may not be preferred is when the sponsor does not have the technical or analytic capability, data, or environment to answer its research question. For example, personnel managers in DoD have responsibility for the day-to-day management of personnel pay and benefits and the setting of policy but may have limited capabilities to assess the behavioral impacts of alternative pay and benefit policies. These capabilities include analytic methods and specialized knowledge of the relevant research literature, experience and knowledge of the data, and the operational capability to manage and protect the data.

Limited internal resources can also constrain a sponsor's need to quickly respond to an external inquiry. For example, if a government agency receives a request from a congressional committee to answer a technical research question, the agency may not have sufficient staff to answer that question within the committee's time frame. In that case, an external researcher, working in coordination with the sponsor, can augment the staffing of the agency.

Finally, economies of scale can lead to efficiency gains and cost-savings for an organization if the fixed costs of certain elements of conducting the study can be spread across a large number of studies. For example, certain data management tasks, including system hardware, administration, and protection, could be common across many studies, and the cost per study of these tasks is lower when there are more studies receiving data management services from the organization. Furthermore, larger-scale operations can permit gains to specialization (including gains to specialized knowledge about certain analytic methods) and data sets. Economies of scale

are more likely to be achieved when the scale of operations is large enough that an organization can functionally specialize in one or more aspects of the study, such as the data environment or the provision of research. When such economies can be realized, it may be preferred to specialize (incompletely integrate) those elements of a study.

Alternative Cases of Specialization in a Study

The previous section makes clear that there are reasons for specialization when conducting a study to answer a research question. But there may also be potential drawbacks that may lead to inefficiencies or higher costs. In this section, we consider four alternative approaches to specialization or integration and consider the benefits and costs of each approach in terms of independence, required internal technical expertise, required internal resources, required direct cost, data safeguarding, and timeliness.

Case 1: Complete Integration

Complete integration, where the sponsor, researcher, data, and environment are all handled within a single organization, generally leads to lower transactional and coordination costs because policies, procedures, culture, and management structure are common to all elements. Such integration also encourages data safeguarding because data do not have to be transferred to an external environment. Furthermore, incentives across the sponsor, researcher, data, and environment to reach the common goals and objectives of the study are more likely to be aligned than when these elements are specialized in separate organizations. Lower transaction costs and better-aligned incentives can increase efficiency and improve the timeliness and quality of the study. Additionally, complete integration promotes integrated (i.e., cross-element) institutional expertise: the ability of one element of the study (e.g., sponsor, researcher, data, or environment) to understand the issues and limitations facing another element. For example, if researchers know how data are collected and stored, then this knowledge allows the researchers to understand the data's potential analytical limitations. However, complete integration can call into question whether the research is independent and objective, and the performance of the research may be hampered by lack of internal technical expertise, specialized data knowledge, and internal resources.

An example of complete integration is a policy shop within the services, whose objective is to answer questions directed by its leadership.

Case 2: Integration of Researcher, Data, and Environment

Separating the sponsorship from the other key elements can provide independence and objectivity for answering a research question. The sponsor is not responsible for recruiting qualified researchers; collecting, holding, and protecting the data; having the requisite

specialized skills for conducting the analysis; or providing the environment in which the analysis is done.

However, separating the sponsor from the researcher, data, and environment means that the research provider must maintain a suitable environment and a broad research staff, thereby potentially increasing the research provider's cost to the sponsor. In order to sustain the research provider, a consistent stream of research demand would be required. While the integration of the researcher, data, and environment can allow for the efficiencies associated with economies of scale by centralizing common functions, such as interactive systems supporting record-keeping across an enterprise (e.g., a medical history database), the separation of these elements from the sponsor can create a tension between the performance of the research, the provision of the data, and the analytical environment on the one hand and the objectives of the sponsor that is seeking research support on the other. That is, separation of the sponsor from the provision of the research, data, and environment can potentially add transaction costs, leading to delays or research products that do not fully meet the sponsor's needs.

An example of integration of researcher, data, and environment is DMDC providing research to the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD [P&R]).

Case 3: Integration of Researcher and Environment

Further separating the researcher and environment from the sponsor and the provision of raw data allows specialization in research capability separate from research sponsorship and from data collection and holding. It achieves independence and does not require the sponsor or the data provider to maintain internal modeling or data analysis expertise, but it requires internal resources to monitor the research and ensure data provision. Separating the research provider from the data provider can enable specialization in their respective fields to achieve economies of scale. Furthermore, integrating the researcher with the environment enables the researcher to develop and maintain analytic databases in the research environment and tailor the environment in a way that develops long-term capability that supports the sponsor while permitting rapid response to sponsor research needs.

This level of integration requires direct costs to identify and write contracts with qualified contractors for the research and environment provider and data provider separately. Timeliness and other transaction costs can become an issue as coordination across entities introduces the potential for transactional delays. Furthermore, data safeguarding becomes a concern, because the data must be transferred between environment and data provider.

An example of integration of researcher and environment is an FFRDC that provides the research and environment while DMDC acts as the data provider.

Case 4: Complete Specialization

Separating each of the elements into its own specialty achieves independence and does not require the sponsor to maintain internal technical expertise, but it requires internal resources to

monitor the research and ensure data provision. The research provider, environment provider, and data provider can specialize in their respective fields in order to achieve economies of scale. An independent environment provider can promote economies of scale by focusing on streamlining data access, data management, and data governance bureaucracy and can improve data security by minimizing data exchange containing personally identifiable information (PII); such data exchange exposes DoD to loss of privacy-protected data.

The case of complete specialization is likely to lead to the largest costs associated with coordination, affecting the timeliness and possibly the quality of the analysis. That is, coordination across entities introduces the greatest potential for transactional delays.

An example of complete specialization is a university researcher using the PDE with data provided by DMDC.

Summary

In considering the appropriate level of integration across the key elements of study, the sponsor must weigh the costs and benefits associated with different levels of integration. Does the sponsor require independence? Does the sponsor require quick-turnaround work? How often will the research question be asked? How many sponsors have similar questions? How much does the sponsor value integrated institutional knowledge, analytic rigor, and specialized knowledge of the data? The answers to these questions will depend on constraints imposed for data security and privacy protections, as well as how frequently the research question needs to be answered. For example, integration of researchers, data, and environment may be most cost efficient if the research question is asked routinely, the data collection and holding is for a narrow purpose, and the question will be analyzed by a research staff with that specific skill set. However, if the data holdings are broad and the questions diverse, then more specialization is likely to reduce costs through more-efficient resource use. A shortcoming of specialization is that it can lead to higher transactional costs—the potential for one organization to delay addressing the research question—and can reduce integrated (i.e., cross-element) institutional expertise.

In the next chapter, we consider the requirements and steps in a typical DoD personnel study and highlight how the four key elements described in this chapter interact to address a research question.

Chapter Three

General Requirements and Steps of a RAND FFRDC Study

RAND's FFRDCs typically provide both the researcher staff and the analytical environment (as in Case 3 in the previous chapter) for DoD-sponsored research. When OSD, the services, the Joint Staff, the Unified Combatant Commands, and the defense agencies rely on a RAND FFRDC to conduct a study, the general requirements of that study typically include the following:

1. Objectivity: The study should reflect the independent analysis of the researchers. The sponsor can provide institutional background, data, and other information it believes to be important to the research.
2. High quality: The study should use the appropriate methods that are currently used in academic and private industry to ensure that the research reaches conclusions that are robust and defensible.
3. Timeliness: The sponsor typically has important policy issues that it is considering. The results of the study must be presented in a timely fashion so that the sponsor can include the study's insights as part of its consideration of a policy.
4. Low cost: Studies should answer the project's research questions using the most-efficient means possible, both in monetary and manpower terms. In many cases, given large existing administrative databases and routine surveys, analyzing existing data can be sufficient to answer research questions.
5. Security: The sponsor relies on its researchers to establish and maintain secure arrangements that will safeguard records from loss and unauthorized use.[4]

To the degree that a research plan is unable to achieve all of these objectives, it falls to the researcher and the sponsor to determine whether the study is achievable within each organization's institutional and ethical requirements, as well as to determine the appropriate data provider and environment that meet the study's requirements.

[4] The first three elements are derived from the NDRI FFRDC's Sponsoring Agreement, which states that NDRI

> was established to be an independent research institution characterized by objectivity, in-depth understanding of sponsor needs, balanced breadth and depth of technical capability, an interdisciplinary and crosscutting approach, and a mid-to-long range focus together with a quick response capability . . . [and] maintain the capability to perform both fundamental and quick-response policy analysis enabled by the depth of institutional expertise and informed by current understanding of sponsors' needs. (Office of the Secretary of Defense, 2016)

Other RAND FFRDC sponsoring agreements have similar requirements.

Steps of a Study

The common steps of a RAND FFRDC study are as follows:

- study initiation
- IRB approval and privacy review
- data request
- data cleaning, de-identification, and provision
- analytic data set creation
- analysis
- reporting
- data archiving and destruction.

Next, we describe each of these steps in more detail.

Study Initiation

A sponsor initiates a study by identifying the research question and the appropriate research entity to address that question. In consultation with the research entity, the environment and data necessary to answer the question are identified, and a project description is written and approved through the sponsor's organizational hierarchy. The sponsor is responsible for ensuring that the research question either has not been previously answered or should be revisited and that the plan proposed by the researcher is scientifically valid for answering the research question. All studies undertaken by a FFRDC "must be within the purpose, mission, general scope of effort, or special competency of the FFRDC" (Code of Federal Regulations, 2017).

IRB Approval and Privacy Review

All studies that involve human subjects must undergo an IRB review to ensure the protection of those subjects. Studies are first reviewed by a qualified individual to see if they require an IRB review (e.g., studies may be exempt because they do not involve human subjects or are not research). If the study is reviewable but is of minimal risk, it may be assigned to one reviewer or a small subcommittee for a determination ("expedited" review). Otherwise, the study needs to be reviewed by the full IRB committee to ensure that adequate precautions are taken for the protection of human subjects. Typically, the researcher's organization is responsible for the IRB review. A second-level review—that is, a review by a qualified member of the sponsoring organization's Human Research Protection Program (HRPP)—may be required at the request of the sponsor. The study must pass the IRB and second-level review before proceeding. If the environment or data provider is not integrated with the researcher or the sponsor, then an official from these organizations' HRPPs may review the opinion of the researcher's IRB. In the case of the environment's HRPP official, if the project is determined to be exempt from further review,

then the proposed project can move forward and use the environment.[5] If the project is determined to be nonexempt, then the environment's HRPP official verifies that all study members have human subjects research training, a scientific review was conducted, an IRB has reviewed and approved the study, and DoD-specific requirements are met.

If a data exchange agreement[6] does not already exist between the data provider and the environment provider, that agreement must be established before a data request can be made. The data exchange agreement dictates the role of the data provider in reviewing projects that request use of its data. In agreements that require the data provider's review, the data provider's HRPP may conduct a review similar to that of the environment's HRPP official.

Data Request

Once a study is approved, the researcher requests the data necessary from the data provider. If the environment is independent of the researcher, then the researcher requests the data through the environment, because the environment maintains the data exchange agreements. The environment provider and data provider may conduct an additional privacy review when the data are requested to ensure that the request is consistent with the project description approved by the sponsor.

Data Cleaning, De-Identification, and Provision

Once a data request is approved, the requested data are provided to the researcher in the environment for analysis. How this is done depends on whether the requested data have been previously provided to the environment and on the data exchange agreement between the environment provider and the data provider. If the data have not been previously provided to the environment, a data transfer must be established. Depending on the data exchange agreement, de-identification may be done by the data provider prior to the data transfer or by the environment after the data transfer. De-identification typically involves replacing or removing PII. This may take the form of creating a unique identifier to replace a Social Security Number or collapsing identifiable categories into a common category (e.g., collapsing all general officers into a single category). Once the data are de-identified, the information is provisioned to the researcher. Provisioning is the act of making the data available to the researcher for analytical use within the environment.

Data cleaning is the act of editing the data prior to providing to a researcher. Data cleaning may include imputing missing data, checking and correcting for errors, creating variables from broader categorical variables, creating consistent measures across time, eliminating records

---

[5] In the PDE, this individual is referred to as an Exempt Determination Official.

[6] We use the term *data exchange agreement* to broadly refer to agreements to exchange, use, handle, or store data. These agreements may take the form of memorandums of understanding or agreement between the applicable parties.

outside the scope of the project description, or de-identification. Data cleaning may also be done out of operational necessity—for example, to reduce the size of the data file so that it is usable within the environment. Data providers often perform data cleaning as part of eliminating errors or inconsistencies in their administrative records. They may also remove data elements before providing data to the environment in order to protect privacy or reduce the file size. It is important to note that data cleaning can create errors; hence, cleaning procedures should be recorded and reported to the researcher.

Analytic Data Set Creation

Once the data are provided to the researcher, the next step is to import the provisioned data into an analytical program that is capable of editing the data to fit specific research needs. The environment must provide appropriate analytical software to do this editing. The editing may include creating new variables from existing variables; dropping data that, after review, is not necessary; reshaping the data; or merging together multiple provisioned data sets. The result is an analytical data set based on the provisioned data. This derived data set needs to be accessible to the entire research team within the environment. This step requires coordination between the environment and the researcher to ensure that the appropriate access permissions are available to the research team.

Analysis

The research team develops a set of programs to implement its analysis plan using analytical software that exists within the environment. The environment must provide the appropriate analytical software required by the project or must be willing to purchase it if funds are provided. Additionally, the environment may need to add packages to analytical software if required packages are not already part of the standard software program. Finally, the environment needs to be able to add programs designed outside the environment by the researcher. The ability to run, review, edit, and share these programs needs to be possible by all the appropriate members of the research team. The environment also must have the appropriate hardware and settings to allow the research team to conduct the analysis. This may include having enough hard drive space to store temporary analytical files, enough random-access memory (RAM) to run the designed programs, and the ability to execute programs without being continuously connected to the environment.

Reporting

During the analysis, results need to be removed from the environment for the purposes of presenting them to the sponsor, including them in the final report, facilitating peer review, or sharing among members of the research team that do not have access to the environment. For the results to be reported, the researcher must export them from the environment. When the researcher and environment are integrated, this may be done through an internal process (e.g., the

researcher directly downloads results from the environment's server). Data exchange agreements may place restrictions on how the environment allows researchers to receive, report, or share results. When the researcher and environment are not integrated, the researcher must submit a request for the results to be exported. This request is submitted to the environment provider, which reviews the results to ensure that they conform to the environment's data exchange agreements with the data providers.

Data Archiving and Destruction

Once the study is complete, the provisioned data and any derivative products need to be archived or destroyed in accordance with the data exchange agreement. When the analysis may need to be replicated or repeated with new data, archiving is generally required. Archiving means that the environment provider must provide a mechanism to retain the study's analytical programs, analytical data sets (or, alternatively the provisioned data), and any other necessary content. Additionally, most archives have a destruction date. The environment is responsible for holding onto and destroying archived data as required by the researcher, sponsor, and data provider. The environment provider may place restrictions on what it will archive and the duration that it will archive that content.

Summary

This chapter has highlighted the interactions among the key elements of a study: sponsor, researcher, data, and environment. From study initiation to data provision, all four elements routinely interact to ensure that human subject and privacy concerns are addressed prior to starting the research and that the data are provided to the researcher in a way that both conforms to the data exchange agreement between the environment provider and data provider and meets the sponsor's and researcher's requirements. Once the data are provisioned, the interaction is primarily between the researcher and environment provider.

In the next chapter, we compare the current procedure used by RAND's DoD FFRDCs, which integrates the researcher and analytical environment, with the alternative provided by the PDE, in which the PDE specializes in providing the analytical environment while the FFRDC is responsible for the research.

Chapter Four

Comparison of Study Steps Across Environments

This chapter compares how the study steps outlined in the previous chapter work within the RAND FFRDC analytical environment and how these steps work within the PDE. We conduct the comparison across environments for each step. Within each step, we discuss the strengths and weaknesses of each approach within the conceptual framework outlined in Chapter Two. As discussed in Chapter Two, the major distinction between the analytical environments is that, for the RAND FFRDC, the environment and the research organization are integrated (case 3 in Chapter Two) and, for the PDE, the environment and the research organization are independent of one another (case 4 in Chapter Two). In our conceptual framework, the additional specialization of the PDE concept allows for potential increases in economies of scale and data security but can also lead to increased transactional delays affecting the timeliness and possibly the quality of the analysis. Our descriptions are based on our experience using the PDE during our study to conduct the data collection and analysis from the three projects that reflect broad analytical requirements of in-progress or completed RAND studies, and some of the experiences we highlight may change as the PDE evolves.

Ideally, our study would be able to quantifiably compare time spent in the PDE with time spent in the RAND FFRDC environment. Because the RAND FFRDC environment is vertically integrated with the research component, there is typically no direct comparison for two reasons, as noted in Chapter One:

1. We cannot identify whether delays in the PDE, if they exist, were due to that environment or to a third party, such as the data provider.
2. Because we replicated in the PDE the data collection and analysis from three projects (RAND1, RAND2, RAND3) that reflect broad analytical requirements of in-progress or completed RAND studies, we did not experience the iterative process in which analytical programs are developed and refined over the course of a study.

The few times we do bring up timing is to highlight the benefits or costs of separating the institutions providing the researchers and the analytical environment.

Step 1: Study Initiation

Studies begin with a project description approved and funded by a sponsor. Once the study is established, the study's researchers can move forward with administrative processes and research planning, such as establishing a study plan, securing an appropriate analytical environment, requesting IRB approval, and establishing the necessary access to the study's required data.

RAND FFRDCs

Once a project description is funded, the administrative processes for the study relating to data request and environment access can begin for the RAND FFRDC. Project staff request access to the FFRDC's analytic servers by contacting the RAND Data Facility. The Data Facility forwards a data use agreement to researchers that will be accessing the environment. The researchers are required to sign the agreement annually in order to retain access to the analytical environment. Among other things, this agreement details rules associated with access, use of identifiable data within the environment, linking of data, and removal of data and derivative results from the environment. Additionally, all RAND researchers are required to complete an information security course every year.[7] Access to the environment does not result in data access. Derived data stored in the environment is in restricted-access folders (discussed in greater detail in step four).

PDE

Under the PDE concept, a researcher with a potential project requests initial access to the PDE by visiting the PDE website and completing a request form. A Common Access Card (CAC) is required for access to the PDE. Once initial access is granted, the researcher needs to electronically sign a PDE Non-Disclosure Agreement, a Health Insurance Portability and Accountability Act Brief, a Privacy Act Brief, and a PDE Acceptable Use Policy.

The PDE is organized into groups. To initiate a study, a researcher must first join at least one group. The groups that RAND is affiliated with include

1. RFL group, which is operated by RFL staff, for non–OSD-sponsored projects (e.g., Department of the Army)
2. DMDC group, which is operated by DMDC staff, for OSD-sponsored projects.

The group owner is responsible for approving the list of affiliated users and the studies within the group. The group owner is also responsible for making regulatory determinations and coordinating and authorizing data access. Group management is split between a primary group point of contact and an exempt determination official (EDO). The point of contact manages interactions with the study team. The EDO makes exempt or nonexempt determinations for research studies.

After joining a group, a project leader initiates a study by completing a form that requests information on the project, including

- project name
- description

[7] RAND, as a federally accredited contractor, must adhere to federal certification and accreditation standards for information security awareness and training. Therefore, RAND and its environment must be compliant with the National Institute of Standards and Technology's Special Publication 800-171, Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations.

- the PDE group to which it belongs
- sponsoring organization
- expected start and end dates
- expected time frame of data requested.[8]

The project is then forwarded to the group owner for approval. As part of the approval process for research, the group owner's EDO reviews the study plan for human subjects protection issues.

From a researcher's perspective, the PDE has two parts:

1. PDE portal, where the researcher can set up a study, join existing studies, request data, and review what permissions have been granted
2. Citrix environment, where the researcher can access analytic programs, use data that have been provisioned to the project, conduct analysis, and save output.

The Citrix environment is a remote connection to the PDE server. This means that the data are never stored on the researcher's computer.

Discussion

In both cases, requesting initial access to the environment requires researchers to sign an agreement specifying how they are permitted to use data within the environment. At this step, neither environment grants access to data, just initial access to the environment. The PDE requires a CAC because the environment can be accessed from any computer. Having a CAC verifies that the user satisfies the level of trust required to hold a CAC, and it provides the PDE with a dual authentication for accessing the analytical environment (i.e., the user must both have the CAC and know the appropriate passcode). In contrast, RAND FFRDCs do not require a CAC because data access and analysis are restricted to individuals working on the RAND network, which requires either a physical presence in a RAND facility (i.e., dual authentication by being able to access the building and knowing a password to access the analytical environment) or a secure virtual private network (i.e., dual authentication by requiring a physical token and passcode). The PDE requirement did result in delays for our study members that were not near a CAC-issuing facility.

Unlike the RAND FFRDC environment, the PDE's process ties administrative processes for the study relating to data request and environmental access to IRB approval (the next step). Consequently, the IRB approval process became a limiting factor for study initiation in step 1.[9] Because IRB reviews can take time, it would be helpful if administrative processes, such as

[8] Similar information is required by the RAND FFRDCs during the creation of the project description or when the RAND IRB reviews the study.

[9] As noted in the next step, Step 2, RAND2 and RAND3 took five months and two months, respectively, to complete the IRB approval process.

getting access to the analytical environment (i.e., Citrix), developing data requests, and adding research team members, could be conducted in parallel with these reviews.

Step 2: IRB Approval

All studies conducted by an FFRDC undergo review by that organization's IRB. Depending on the nature of the research and the sponsor's requirement, the study may require a second-level review by a DoD entity.

RAND FFRDCs

Once a project description is approved, data acquisition cannot begin until the study has received human subjects approval and any conditions or requirements of approval (e.g., informed consent protocols and data safeguarding plans) relevant to the specific data collection activity have been met. Consequently, after the project description is signed, the researcher files a new study application with the RAND IRB. RAND's Federalwide Assurance for the Protection of Human Subjects serves as RAND's assurance of compliance with federal regulations. According to this assurance, the RAND IRB is responsible for reviewing all research, regardless of the source of funding. RAND has signed a DoD Addendum to its Federalwide Assurance covering all DoD-sponsored research performed at RAND. In the addendum, RAND agrees to abide by the human subjects research requirements set by DoD, as well as those of the following components:

- Department of the Army
- Department of the Navy
- Department of the Air Force
- OUSD (P&R).

This study application specifies the research to be conducted, specifies which personnel will be involved, and screens for the involvement of human subjects and eligibility for exemption. The application reviews the study's populations and procedures (including questions about merging data and potential risks) and informed consent. All study staff in contact with human subjects or working with human subjects data are required to complete training in conducting research with human subjects. The RAND IRB makes the following three determinations based on the new study application:

- Is this research?
- Are human subjects involved?
- Does this fall into one of the categories for which exemption from review is allowed?

Once the study receives approval from RAND's IRB, the study is reviewed annually until it is no longer eligible for review due to completion and destruction of PII.

Although DoD adheres to the same principles as RAND's IRB, each agency and component may have its own policies for implementing those principles. Depending on the study, the sponsor, the FFRDC, or the data provider may require an additional review by the applicable DoD or service IRB, also known as an HRPP. Second-level review generally requires similar materials as RAND's IRB. This IRB may choose to concur with the RAND IRB's determination or conduct its own review. Starting in 2017, OUSD (P&R) began requiring all of its studies, even those deemed not research, to undergo a second-level review.

These steps are required of RAND research, regardless of the analytical environment. Additionally, all RAND researchers who will come into contact with human subjects or human subjects data are required to take training in the protection of human subjects. Training is valid for three years, after which taking the training again is required. When RAND's IRB identifies a study as likely to be human subjects research, all study staff listed are checked for this training.

PDE

Once the required IRBs have approved the study from the sponsor's and researcher's perspectives, the study must be reviewed by the analytical environment's IRB. A study intending to use the PDE as its analytical environment may begin this process by "proposing a study" within the PDE.[10] A study must be proposed within a specific PDE group, and the choice of group will determine who acts as the EDO. The proposed study is forwarded to the group's EDO, who reviews the study's documentation to determine whether the study is research, whether human subjects are involved, and whether the study is exempt from further review. If the study is determined to be exempt, then the proposed project is approved. If the study is determined to be nonexempt, then the EDO verifies that all study members have human subjects research training, a scientific review was conducted (this should be done by the study's sponsor), and an IRB has reviewed and approved the study. If the EDO concurs with the FFRDC's IRB, then the study can be approved and access to the PDE portal and Citrix environment can be granted. Alternatively, the EDO can choose to forward the study to the appropriate HRPP for review. In that case, the data cannot be requested and additional team members cannot be added until the study is approved by the HRPP.

Discussion

Allowing a separate organization to specialize in providing the environment requires an additional process for satisfying the privacy and human subjects requirements of the environment provider because its reporting authority may differ from the other key elements of the research process (i.e., sponsor, researcher, data provider). Our experience varied based on the PDE group. For each project, we provided the group's EDO with project background, summary, tasks, data

[10] For the present study, the sponsor did not require second-level review because the RAND IRB deemed that it was not research. However, the authors have experience with the second-level review process from other studies.

requested (including data set and variable names), DoD sponsor, RAND project contact, RAND IRB determination, and project staff list. The time between PDE study initiation and PDE group EDO approval for each project was as follows:

- RAND1 (RFL group owner): 2 weeks
- RAND2 (DMDC group owner): 5 months
- RAND3 (DMDC group owner): 2.1 months.

For RAND1, the RFL group's EDO concurred with the RAND IRB's determination that the study was not research. For RAND2 and RAND3, the DMDC group's EDO review was delayed because an EDO was not assigned until two months after the studies were initiated in PDE. Furthermore, DMDC requested information in a specific format in order to reach a determination, and the requests were forwarded to the HRPP. Specifically, the DMDC group EDO requested

- the study protocol in DMDC's HRPP-specified format
- researcher curriculum vitaes
- a signed project description.

The projects overseen by the DMDC group took longer to initiate because the additional information required to complete the EDO and HRPP determinations were not known up front. We believe that many of the delays were due to errors that are easily fixed and that, as the process evolves, are unlikely to be repeated.

Conceptually, the delays experienced by RAND2 and RAND3 represent transaction costs from specialization. By having separate institutions provide the research and the analytical environment, each institution will require, at least, an IRB concurrence. Additional EDO or HRPP reviews led to further requirements for documentation or rewriting existing documentation into the HRPP's preferred format. Additional HRPP reviews lead to delays, because each HRPP must become familiar with a project, and each HRPP may have its own idiosyncratic rules. For RAND NDRI, this means that a study using the PDE for its analytical environment will be reviewed at least three times: by RAND's IRB, by the FFRDC's responsible HRPP (i.e., an OSD HRPP), and by the PDE group's EDO or HRPP (i.e., an Army HRPP). While it is possible that the additional reviews may also provide additional oversight, the reviews are not costless in terms of the time and resources involved in doing them, and there is a question of diminishing returns of doing three reviews, especially for studies that are deemed not to be research or not to involve human subjects by the first review.

Step 3: Data Request

Data come from a variety of sources. In order to receive data for analysis, the organization providing the analytical environment must sign data exchange agreements. These agreements determine the circumstances under which the data can be used, what analysis can be done using

the data, who determines whether or not results from that analysis can be released, and possibly other restrictions. Once the exchange agreements are in place, the researcher must request specific data elements to be used in the analytical environment.

The federal policy that defines the regulation surrounding the creation, use, review, and termination of an FFRDC, 48 CFR 35.017, states that "an FFRDC, in order to discharge its responsibilities to the sponsoring agency, has access, beyond that which is common to the normal contractual relationship, to Government and supplier data, including sensitive and proprietary data, and to employees and installations equipment and real property" (Code of Federal Regulations, 2017). Each FFRDC must establish its own data use agreements or establish a consistent means of being able to get access to data that it requires to fulfill the contractual obligations of its sponsoring agreement. Consequently, a FFRDC's management establishes data exchange agreements required to fulfill its contract or secures access to environments that can provide that access.

RAND FFRDCs

If a study intends to use the RAND analytical environment, once the study is approved by the appropriate IRBs, the study can request data from a data provider through a pre-established data exchange agreement. If a data exchange agreement does not exist, then the FFRDC arranges an agreement with the appropriate data provider as required. Typical data exchange agreements establish the conditions for the request, transfer, storage, use, reporting, and termination of the data provided. We use the RAND NDRI and DMDC data exchange agreement as a representative example of how a data request is handled (Defense Manpower Data Center, 2016). In the context of a DMDC data request, a RAND NDRI researcher submits a specific project amendment (SPA) that details the project, sponsor, researcher, contract, a summary of the project, and the frequency and time period of the data sets and elements required.[11] The SPA specifies which data sets and elements are being reused under the data exchange agreement from another project and which variables are new requests that are not currently held by RAND NDRI under the data exchange agreement. A representative from the DMDC Office of the Chief Information Officer (OCIO) reviews the SPA to ensure that the request satisfies the data exchange agreement's requirements for requesting data. Once the SPA is approved, the OCIO representative authorizes the provision of the data. As part of this process, the applicable DMDC file managers are consulted. A similar process is established for most data exchanges, although the requirements imposed have varied over time, with a recent trend toward more-formal data exchange processes in order to delineate responsibilities of each party to the agreement.

[11] A SPA is so named because it becomes an amendment to the appropriate data exchange agreement.

PDE

Once a study is approved, the researcher can select the variables necessary for the study from the list of data sets in the data catalog. The data catalogue documents data for which existing data exchange agreements exist. If a data exchange agreement does not exist, then the researcher must contact RFL, and RFL will arrange an agreement with the appropriate data provider. Existing data exchange agreements specify whether data sets are marked as open, restricted, or private in the PDE data catalog. Open and restricted data sets can be viewed in the data catalog and can be requested by PDE users. Private files cannot be seen in the data catalog, so they cannot be requested by other PDE users.

Once the study's required data are selected, the PDE portal requires the researcher to provide a brief rationale for the required data and the applicable military services and time period covered by the request. Use of open data sets and specific variables is approved directly by the group owner. The data owner is not required to approve the request. The PDE aims to have as many data sets as possible classified as open to promote efficiency of use. Use of restricted data sets and specific variables require the approval of the data owner in addition to the group owner. In the case of restricted DMDC data requests, this process is first reviewed and approved by the applicable group owner, then by the applicable DMDC file manager and OCIO representative.

Discussion

The data request process is similar for each environment, but requests in the PDE must receive additional intermediate approvals before the data request is approved. We consider three types of data request experiences: (1) reuse of data already in the analytical environment, (2) new data requests under an existing data exchange agreement, and (3) new data requests requiring a new data exchange agreement.

In the case of reusing data already in the analytical environment, the data sets required by RAND FFRDC studies fall under both open and restricted data in the PDE.[12] Within the PDE, the *restricted* designation means that the data request requires the review and approval of the data owner in addition to the PDE group owner, while the *open* designation requires the approval of only the group owner. As a comparison, in the RAND Project AIR FORCE and RAND Arroyo Center FFRDCs, the review and approval process for reuse of data from the Air Force and Army, respectively, is handled internally, requiring the approval of the research center director,[13] while the process for DMDC data for all three FFRDCs (RAND Project AIR FORCE, RAND Arroyo Center, and NDRI) requires the review and approval of DMDC's OCIO representative (as the data provider). Consequently, in the RAND FFRDC environment, a data

[12] For example, DMDC's Active Duty Pay and Defense Enrollment Eligibility Reporting System files are restricted, while HRC's Regular Army Analyst file and DMDC's Active Duty Master file are open.

[13] This process continues to evolve. As of the writing of this document, the RAND Arroyo Center was negotiating a new data agreement that would be closer in spirit to the DMDC data exchange agreement.

reuse request must satisfy only one reviewer; in the PDE, open data sets require only one approval and restricted data sets require two approvals. Additional reviews add coordination and transaction costs. Our experience in the PDE was complicated by the requirement to submit a data request digitally, which was susceptible to system errors and connection failures.[14] In the RAND FFRDC environment, reuse requests without any new data requests are approved by DMDC's OCIO representative usually in one or two days, and most are within one week. Provision can be done immediately by the RAND Data Facility once permission is granted. The requests within the PDE were generally reviewed in a similarly short time frame by the group owner based on the data we were able to collect. Based on what could be observed in the PDE portal, delays primarily arose between receiving authorization for reuse from the data owner on restricted data sets and the provision of the data. The time from data request to provision ranged from five days to 206 days, with an average of 58 days and a median of 46 days for RAND1, RAND2, and RAND3.[15]

Receiving review and approval of data requests for new (as opposed to reused) data under existing an data exchange agreement can be time-consuming in both analytical environments. For example, in both environments, we made a request for DMDC's military retiree pay file, which had not previously been received in either environment but for which there existed a previously approved agreement with DMDC. In both cases, it took a long time for the environment to receive approval and for the data to be provided. For the RAND FFRDC environment, it took 144 days from data request to initial provision of the data file, and in the PDE, it took 206 days from data request to initial provision.[16] This is just one example and is not likely representative of all new data requests in either environment. However, it does point to a

[14] For example, the PDE digital request form encountered an error twice during the request and approval process for Military Entrance Processing Command data. As a result, the study team had to complete the request process again. Because some requests can involve dozens of elements out of hundreds of possible elements, completing and recompleting a request can be a cumbersome process. In the RAND FFRDC environment, this process is handled by copying variable names and submitting a digital document, which is not dependent on a network connection for variable selection.

[15] These numbers are based on the time difference between when the file was requested by the research team and PDE's notation in the PDE portal for when the data set was provisioned. We omitted requests that were denied, removed, or resubmitted due to a system error. It appears that PDE staff made a concerted effort to document provision time in August 2017. It is possible that the provision times that they report reflected the most recent provision of the data as of August 2017, potentially misreporting the first provisioning of the data and causing the reported average, median, and maximum to be artificially high.

[16] Measuring the time for the PDE is difficult because we do not have vantage on when PDE staff actually made the request to DMDC for the file. The project requesting the retiree pay file (RAND3) was approved on May 24, 2016, but the initial request for the data set was provided to the PDE on December 17, 2015. A PDE-DMDC data use agreement draft from March 2016 includes the retiree pay file. The ability to request the file appears to have occurred in November 2016, and the initial provision of the retiree pay data occurred on December 16, 2016. Our computation of time for the data request is based on the time that the project received approval in the PDE system to when the data were initially provisioned—May 24, 2016, to December 16, 2016.

joint difficulty in environments receiving data, even conditional on having data exchange agreements in place.

The RAND FFRDC's analytical environment and the PDE experience extended delays when requesting new data exchanges. For example, for RAND1, it took PDE staff six months to receive enlistment data from the U.S. Army's HRC because the data RAND requested required a new data exchange agreement; of those six months, at least two to three months involved back and forth with the appropriate HRC office.[17] RAND FFRDCs have experienced similarly long, and in some cases longer, waiting periods for getting data exchange agreements approved. For example, a new data exchange agreement with HRC required nine months, and, as of June 2018, a data exchange agreement with Social Security Administration and DMDC had been in process for 22 months. Additionally, we did not evaluate the process for acquiring data from other data providers. The RAND3 project required data from the Defense Health Agency, a process that requires a new data exchange agreement for each project. At the request of PDE staff, the research team agreed that this evaluation should not include Defense Health Agency data because of the RAND FFRDC's experience in the additional burden required for completing this process.[18] Given the multitude of complexities associated with forming data exchange agreements, it is unclear whether any environment is necessarily better suited to securing these agreements in a timely manner. The PDE, because it is internal to DoD, may avoid contractor-related technical hurdles that are typically experienced by the RAND FFRDCs, despite the language of 48 CFR 35.017, which requires FFRDCs to have access to data beyond what is common in normal contractual relationships. Additionally, economies of scale are associated with the PDE securing these data exchange agreements that are written to cover all PDE users. This advantage does not currently apply to RAND FFRDCs because these agreements are largely in place. On the other hand, because the PDE is external to the RAND FFRDC, it is more difficult to ensure that incentives for timely research are aligned. Thus, there could be less sense of urgency in getting agreements in place when the PDE oversees the formation of data exchanges.

A data request when the data are not currently available in an analytical environment is a process with an uncertain time frame for completion that is likely measured in months, regardless of whether it is done for the RAND FFRDC's analytical environment or the PDE. Although the research team cannot observe the experience of PDE staff in forming data

[17] The initial request for the data set was provided to the PDE on December 17, 2015. Specific elements from the data sets were provided on January 29, 2016. PDE staff attempted to establish contact with HRC starting in February 2016 and were successful by April. The HRC SPA was certified on July 18, 2016. RAND staff submitted the request in the PDE portal on July 21, 2016, and the data were provisioned on August 1. Our time estimate is measured from January 29, to August 1, 2016.

[18] From discussions with a RAND FFRDC researcher involved in this process, the process of getting a Defense Health Agency data exchange agreement approved requires five months, on average, although the agreement for the project that RAND3 is designed to mirror took 16 months.

exchange agreements, the RAND FFRDCs' experience has been that the uncertain time frame partially results from continuously evolving requirements from data providers in light of a broader demand for more-rigorous data security, including a common desire of data providers to want oversight in how the data they are responsible for are transferred, handled, and used by the analytical environment's owners and users (e.g., researchers). Additionally, delays are often incurred because of the multitude of offices included on each side of the agreement (i.e., legal, privacy, functional, and technical staff), leading to transactional delays from iterations among multiple parties within and across the organizations. Although a nonalignment of incentives between the researcher and the analytical environment is a concern, the experiences documented here suggest that reaching the agreement for the use of new data remains the dominant source of delays in data requests.

Step 4: Data Cleaning, De-Identification, and Provision

Once a request for data is approved, data are provided to the researcher for analysis. Before or after the data files are provisioned to the environment, they may need to be cleaned or de-identified. Data cleaning involves ensuring that the data are consistent across time and are validated (e.g., reflect aggregate statistics, are not missing key elements or records), as well as noting any irregularities. De-identification requires ensuring that PII is not available unless required as part of the research and approved by the data provider and IRB.

RAND FFRDCs

Data are provisioned to the RAND Data Facility. Depending on the sensitivity of the data request, the Data Facility decrypts the data following transmission and transfers the information to the RAND Data Facility's cold room—a computing environment without a network connection—for review. A cold room review includes a check of the data quality (e.g., data may be in an inconsistent format that prevents the information from being read in), a check for PII, and de-identification of the data as necessary. Consistent with NIST Special Publication 800-171, the Data Facility and cold room reviewers are not study research staff, to reduce the risk of unauthorized activity without collusion. The data are transferred to the RAND FFRDC analytical environment, where the data are stored in restricted-access folders (access to the environment does not equate to data access). RAND programmers who are experienced in working with the appropriate data files then review the data for consistency with previously received data (if applicable) and conduct any further data cleaning typically conducted for that data set (e.g., creating a consistent set of data elements across time). Typically one authorized research staff member per study is granted access to the folders for which data access has been approved. That researcher is permitted to extract the required elements from the data set. The researcher responsible for extracting the required elements creates a derived data set that contains the necessary and approved elements. The derived data set is made available to the research team in

the environment. Depending on the sensitivity of the derived data set, the RAND IRB or Data Facility may require that it be placed in a restricted-access folder in the environment.

Because many RAND FFRDC studies require access to bulk panel data (i.e., many records with repeated observations over time), the data could potentially be identifiable by inference with sufficient external information. Per the RAND FFRDC environment's data use agreement, this data must be retained in the environment for analysis until the data are aggregated. Aggregate data exported outside of the environment for publication must employ best practices for masking small cell sizes.[19] All aggregated data removed from the environment can be moved only to another RAND-owned resource. No identifiable data may be stored outside of the environment without special permission.[20]

PDE

RFL extract, transform, and load (ETL) specialists are responsible for handling the intake of new data into the PDE. The data are moved to the PDE-S, a staging environment, where the data files are imported into a common software system, Social Security Numbers are replaced with a PDE-specific identifier, and data are further transformed and loaded as a conformed data set into the master tables in PDE-S (Research Facilitation Laboratory, 2015). These master tables are then moved from the PDE-S to the PDE-A, the analytical environment. After the use of a data set for a study is approved by the PDE group's EDO and the data set has been loaded into the PDE-A, an ETL specialist is responsible for extracting the appropriate data elements, creating a study-specific identifier from the PDE identifier, and conducting required transformation to data elements. The ETL specialist applies a set of business rules to the data, including standardizing unique identification codes based on Army rules, eliminating day of month from birth date, consolidating high-rank and pay-grade groups into common categories (i.e., officers in pay grades above O-5, warrant officers above W-2, enlisted above E-7). After this is complete, the ETL specialist moves the derived data from the PDE-A into the PDE-A Oracle study schema. It is from the study's Oracle schema that the researcher is able to access the derived data set for analysis.

---

[19] A RAND FFRDC's work must be cleared for public release before it is published, presented, or released. The requirements for clearance differ across RAND's FFRDCs but generally include the sponsor approving the document for release. Additional clearance is required for publications that are unclassified and have no dissemination restrictions, a process that is typically conducted by an applicable DoD public affairs or security office. The purpose of the clearance is to ensure that the work contains no classified or sensitive material or substantive errors, not to censor the results of the FFRDC's analysis or suppress findings that are critical of DoD.

[20] When required by a business need, studies must notify the RAND Data Facility of the location where files will be stored. Files may be moved only to a RAND-owned resource. Person-level files stored outside of the environment must have the unique identifiers removed, must be kept encrypted when not in use, and must be removed from the resource as soon as the business need is met. Any projects granted an exception must also file and adhere to a project-specific data safeguarding plan with RAND's IRB. It is the responsibility of the RAND Data Facility and FFRDC management to ensure that any exception satisfies existing data sharing agreements with the data provider and institutional and federal requirements for the safeguarding of data.

As part of the provisioning process, a project-specific unique identifier is created for individuals in the data. This identifier is meant to facilitate the merging of data sets. Because it is project-specific, it prevents the merging of data sets across studies.

Discussion

Data that the research team received access to through the PDE were of similar quality to what RAND FFRDCs have previously received from DMDC and HRC. The RAND team compared tabulations of key records (e.g., total number of accessions, disability separations) with aggregate reports or intermediate tabulations done in the RAND FFRDC environment. RAND FFRDC and PDE processes for data provision follow similar approaches for safeguarding data. Both entities review the data for PII, transform unique identifiers before data are provided to a researcher, and have an individual not associated with the analysis handle the review and transformation of data received by the environment.

There are also important differences. The first key difference is in how data are cleaned. A researcher must be able to fully understand the limitations of the data provided in order to produce high-quality analysis, one of the requirements of FFRDC studies. Decisions made by the analytical environment before provision to the researcher should be clearly documented and communicated, and someone familiar with the data cleaning process should be available to answer questions, quickly if need be, regarding those decisions. Integration between the institution providing the research and the institution providing the environment can help resolve this coordination problem.

The RAND FFRDC process requires a file manager to review the data for consistency and create consistent elements over time. These file managers are part of the research programming staff and often work with researchers on studies. They are familiar with the intended use of the data and serve as a repository of institutional knowledge.

In the PDE, ETL specialists can only indirectly learn this information from interactions with researchers. Our study required iterative discussions with ETL specialists to ensure the data were provided in a usable way. For example, our study requested monthly versions of one file type, only to discover that it was provisioned in a quarterly format; files were occasionally reduced in size to accommodate space constraints, and the consequences of these file reductions were not known to the researchers; in some instances, multiple versions of the same derived data set existed because a revision to the data had been made but the original version had not been removed. These issues were generally resolved through iterative email exchanges. In some instances, issues or questions regarding the data provided were answered quickly; at other times, the iterative process played out over days. Many of these issues arise within the RAND FFRDC's analytical environment, but they are generally resolved more quickly, or additional technical support can be brought in as required because of an institutional alignment of incentives between the analytical environment and the research team.

One of the examples just offered is what we call passive data cleaning—that is, data cleaning intended to accomplish a technical issue, such as accommodating space constraints in the analytical environment, but that can have consequences for analysis. As part of RAND3, we required the retiree pay file, which tracks the monthly payments of all military retirees. The file contains mostly repetitive information because a retiree's pay is fixed. Consequently, in order to save space, we agreed to a snapshot at the beginning of the year. However, if individuals make an election or other adjustment, taking a snapshot at the beginning or end of the year may miss within-year changes. This type of data cleaning occurs in the RAND FFRDC environment as well. Because we had the opportunity to play a part in the discussion of how the data would be stored, we understood the data cleaning choices that had been made. Passive data cleaning, however, can affect subsequent data users that were not part of the data cleaning decision. It is important that these types of data choices are recorded and that mechanisms exist so that the more detailed data can be recovered for projects that require that information.

A second key difference between the RAND FFRDC analytical environment and the PDE is how the data are provisioned to the research team. The PDE process for providing the derived data through the Oracle schema is a systematic approach for ensuring that requested data elements do not exceed or differ from those requested and approved. In the RAND FFRDC process, the responsibility for ensuring that unapproved elements are not in the study's derived data set is delegated to the researcher authorized to extract the data from the restricted-access folders. While the use of unapproved elements would be a violation of the RAND FFRDC data use agreement, the PDE process is a more direct way of ensuring compliance.[21]

A third key difference between the RAND FFRDC analytical environment and the PDE is that the PDE generates a unique study record identifier for each project. A study-specific identifier is intended to prevent a mal-intent researcher from merging data that he or she is not permitted to access for a study or from merging data from one study to a different study. Within the RAND FFRDC environment, the record identifier is not unique to the study, but it is understood by the research staff that data cannot be merged for research purposes outside the scope of the IRB ruling, the project amendment to the DMDC–RAND FFRDC data exchange agreement, and the RAND FFRDC's data use agreement. This is consistent with the trust environment of the RAND FFRDC.

A cost of the study-specific identifier is that it limits a researcher's ability to engage in activities that can improve consistency and efficiency of the research. First, to ensure consistency across researchers, merging data sets to reconcile analytical differences can be the most straightforward method for identifying empirical discrepancies. This requirement can be particularly important in the data cleaning phase or when a researcher is attempting to replicate

[21] As of the writing of this document, RAND's information technology infrastructure has the ability to support the automatic issuance of derived data sets. The RAND Data Facility is developing processes to institutionalize this capability.

past work. Additionally, RAND FFRDC research commonly builds on past work, often over several decades; indeed, one purpose of an FFRDC is to more easily allow the development of subject-matter expertise, including expertise about data. Study-specific identifiers require the regeneration of the data from the original file. While there may be a benefit to having study-specific identifiers, it is unclear that the benefits outweigh the costs, especially because the researchers are required to abide by user agreements that already specify that the researchers cannot use data for research purposes for which they have not been granted permission.

A final distinction between the RAND FFRDC environment and the PDE is that, in the RAND FFRDC environment, the extracting of specific elements to create the derived data set is left to the project programmer. In the PDE, an ETL specialist is required to produce the derived data set. Therefore, the response time depends on the PDE ETL specialist's time and how the PDE prioritizes the work. This has the potential to result in a misalignment of prioritization between the research team and the analytical environment.[22]

Step 5: Analytic Data Set Creation

Researchers may require a subset of the data provisioned for a study or wish to reorganize the data in an alternative format prior to conducting analysis (e.g., creating indicator variables for a specific category from data elements with multiple categories consolidating in a single element). The final analytical data set is designed and created in the environment to reflect the data format required by the study. It should be accessible to the study's researchers in the environment. The key benchmarks we examined include the ability to write and execute code, save and use intermediate output files, merge data, and evaluate the analytical data sets' external validity (e.g., do observed aggregate statistics track statistics reported by other publicly available, trusted sources?).

RAND FFRDCs

The researchers involved in this study typically develop code outside of the environment on RAND-owned devices. After developing the code, a researcher uploads code to the environment using a secure file transfer program and then executes code in batch or by command line.[23] File

[22] For example, during RAND3, we had to request additional elements from certain data sets after we discovered that the initial elements requested were insufficient. Consequently, the research team had to wait until the data request was reviewed by PDE staff and the data were provisioned by an ETL specialist. A similar issue in the RAND FFRDC environment would require an amendment and approval of the original request by the appropriate review authority (see step 3) before the project programmer could add the required element.

[23] To execute code in batch program means that a program is written that will execute the underlying code and produce the output on the server without interaction with the researcher. A batch program is submitted by the user from the command line. The output is not displayed on the screen but is available only in the output files that are specified in the batch program or the program that the batch program calls. To execute code by command line means to open up the analytical program first and then execute code from within the program. In this case, the output is

transfers can occur only when a RAND-owned device is connected to the RAND computing network and the researcher logs into the RAND FFRDC computing environment. Some researchers activate a remote desktop and execute code directly, although this method is less common due to latency issues. (Latency is the delay between request and execution, such as the delay between a typing and the appearance of words on screen.)

Researchers are allowed to reuse code developed previously, and the Research Programming Group at RAND has established research staff that are knowledgeable about different data sets. As mentioned in step 4, the research staff are authorized to extract the required elements from the approved data sets. Typically, a researcher is also responsible for creating the analytical data set, which may require merging approved data sets. This analytical file is then stored in folders specified by the researcher. The researchers have the ability to create folders in their own personal folders and assign controlled access to project folders in their personal folders. This allows the researchers to grant and revoke access to other members of the environment.

In creating an analytical data set and reviewing it for external validation, the researchers can reshape the data, create new variables, run descriptive statistics, and review lines of data for cross-record consistency. This can be done in a sequential manner from the command prompt of the environment or in bulk by executing the code and downloading a file reporting the output's results. If the latter method is chosen, that reporting output can be downloaded for review (this will be described in greater detail in step 7).

PDE

Code can be written outside the PDE and imported into it, or code can be written within the PDE. Any code written outside the PDE or external data files must be submitted to the PDE help desk to be uploaded into the environment.[24] An RFL PII review team member will review the file and then place the file in an import folder, at which point the researcher must move the file to an appropriate folder. If the code is written within the environment, no interaction with PDE staff is required.

The PDE operates only with a remote desktop, and provisioned data are accessible by an Oracle schema, which requires a user-specified password. Once the user-specified password is provided, the data are accessible within remote desktop by analytical programs that can read in Oracle databases. Within an Oracle schema, the provisioned data are organized into tables that correspond to each provisioned data set. Accessing the remote desktop requires the user to have access to a CAC reader, up-to-date Citrix software (this is required to activate the remote

---

generally displayed on screen. In both cases, the code and the data are executed within the environment, and the data do not leave the environment.

[24] Files larger than 10 megabytes need to be sent via the Aviation and Missile Research, Development, and Engineering Center's Safe Access File Exchange (SAFE).

desktop, and appropriate software exists for Windows and Macintosh computers), a current CAC, and appropriate CAC certificates.

Project folders and permissions have to be initially established by PDE staff. Once the folders are established, authorized users can read, write, and execute programs from these folders. Additionally, users can establish new folders, and analytic data sets can be saved in these folders.

Data sets can be merged within a PDE study's provisioned data using the study's unique identifier. Analytic data sets cannot be merged across PDE studies using the unique study identifiers.

In creating an analytical data set and reviewing it for external validation, the researcher can reshape the data, create new variables, run descriptive statistics, and review lines of data for cross-record consistency. This can be done in a sequential manner by running the code within the program or in bulk by executing the code and requesting that the file reporting the output's results be exported from the PDE. If the latter method is chosen, that reporting output can be downloaded for review once the PDE has approved its release (this will be described in greater detail in step 7).

Discussion

The study team's experience when writing and executing code, saving and using intermediate output files, and merging data was that performing these tasks was more difficult in the PDE than in the RAND FFRDC environment. The main differences involved (1) consistency of PDE access, (2) redundant password requirements, (3) difficulties with project folder creation and permissions, (4) difficulties with connections to project folders from analytic programs, (5) remote desktop latency, and (6) CAC requirements. We note that the first four points—while they were persistent issues during our evaluation period and across all projects (RAND1, RAND2, RAND3)—are potentially resolvable with appropriate time and resources.[25] The last

[25] Regarding difference (1), there were persistent problems with study staff being able to access the Citrix environment, which is a precursor to doing work. Often, the inability to access the environment would be resolved on the second or third attempt, but the Citrix environment occasionally remained inaccessible for days. Related to (1) and (2), the password requirements for the Oracle database led to several follow-on problems because the password was required to be routinely reset—at least once every two months. For users of SAS (a common analytical program), the password for the Oracle database needs to be included in the SAS library name statement to make the connection to the Oracle databases. The user must update the password in all SAS programs where SAS library name statements appear. If the user does not update all SAS library name statements and the user attempts to execute the SAS program, then the user will be locked out and must contact PDE support staff to reset his or her Oracle password. This issue arose on March 16, 2017, resulting in several iterations between one of our SAS users and PDE support that did resolve the issue and prevented further work. While the delay was due, in part, to the user repeating his or her error a second time, the issue was not resolved until June 20, 2017. On that day, the PDE released a new feature automating this password reset process; however, the requirement remains to update all SAS library name statements each time the Oracle password is reset. Given the requirement for CAC with passcode access, it was unclear to the study team why this additional password requirement was needed, given the repeated delays induced by it. Regarding (3), it took a long time establishing folders and securing the appropriate sharing permissions. The original request was submitted on July 21, 2016, and the file permissions issue was not resolved

two points represent persistent differences between the RAND FFRDC's analytical environment and the PDE.

A key distinction between the PDE and an environment like RAND's is the ability to access the PDE from anywhere, including on a nonsecure network. This is achieved because the user receives only a visual representation of what is occurring remotely on the PDE's computers. There is no ability for the user to extract the data because the user cannot directly transfer data outside the environment. Data transfer must go through PDE staff. A key challenge with a remote desktop environment is end-to-end latency—that is, the time from the user typing or clicking and the realization of that action in the environment. Location, internet speed, and the number of other users can influence the PDE experience from the researcher's perspective. This issue can be amplified for researchers who work in a secure, non–military network environment, because the transmission of data has to pass through multiple routers, including the researcher's network and the PDE's network. Latency issues are difficult to determine because a multiplicity of factors may cause them, and not all are within the PDE's control.

The implication of latency issues is a reduction in the ability of the user to effectively go about his or her work in the analytical environment. In a computing environment, a user is continuously interacting with the environment, so latency issues can be unobservable, mildly disruptive (e.g., a short delay between typing and observed action of less than one second), or prohibitive. During our study, we experienced the full range of these situations while using the PDE, although prohibitive delays were not common. Generally speaking, latency is not observable in the RAND FFRDC environment, likely because the communication between the analytical environment's server and the end user's computer is occurring within the same network.

During our study, the CAC requirement caused several of our team members to have repeated issues accessing the PDE. As noted in step 1, CACs are not standard issue for RAND FFRDC researchers, although they can be acquired with a stated need. One issue is that some RAND FFRDC researchers are not located near CAC-issuing facilities. A second issue is that the required certificate authorities for accessing the PDE were not well understood, which resulted in inconsistent advice and difficulty accessing the environment. While DoD-issued computers may have the required features already installed, this step required iterative interactions between the PDE help desk and affected researchers. Despite the PDE being mostly a remote environment, the end user's machine introduces additional complications in establishing a consistent connection.

---

from RAND1 until September 2, 2016, or for the other projects until September 13, 2016. In the RAND FFRDC analytical environment, initial project folders and sharing permissions can be created within a day by contacting the Data Facility or coordinating with RAND Information Services staff. Regarding (4), as part of the analysis stage, a RAND3 researcher after June 2017 was unable to output results from SAS due to an inability to read from or write to the project folders using the Output Delivery System (ODS), a common SAS data outputting function. At approximately the same time, the RAND2 researcher lost the ability to save or reuse code. Additional notes are available from the research team upon request.

Step 6: Analysis

The study's researchers conduct analysis within the environment using approved data. To conduct the analysis, the researchers require access to current analytical software and hardware. This may include software licenses, additional software modules that must be downloaded from approved internet sites, computer memory, or computer processing power. The researchers must also be able to access the environment to conduct the analysis on a regular basis and have a mechanism by which technical issues can be addressed in a timely manner. Step 5 focused on the manipulation and storage of data. This step emphasizes key analytic benchmarks: software and hardware capabilities, the ability to share within study teams, and the consistency and usability of the user's experience.

RAND FFRDCs

RAND's network consists of dozens of servers, of which the FFRDC's servers represent a small fraction. Each RAND FFRDC operates its own server. The servers vary in capability but typically include 32–64 cores and at least 256 gigabytes (GB) of RAM. Additional analytical servers can be used if required (conditional on data security restrictions). Approved researchers are able to use all or a fraction of the server's resources (no artificial limitations exist), and busy times are handled through internal communication and reallocation. Analytical jobs can continue without the researcher's computer remaining connected to the environment. If additional software or hardware is required, then the FFRDC arranges to get the required licenses and bears the cost.[26] Some analytical programs require additional plugins, or software downloaded from the internet. In some cases, certain downloads are permissible, which allows the researcher to directly install the required support package, but usually only in a user file (the downloads are not made available systemwide). Alternatively, RAND's Information Services staff can arrange to have these installed during normal business operations. In most instances, the response is same day and can occur outside of business hours if necessary.

The RAND FFRDC environment is not user-friendly. It consists of a command prompt interface, requiring either batch programs or command line interaction. Outages are generally scheduled in advance and done during nonbusiness hours.

PDE

Prior to logging into the Citrix environment, a PDE user must choose between using a SAS desktop or other analytical program (e.g., R, Stata, SPSS). Specifically, SAS has its own separate environment. If a user intends to use the SAS environment, RFL must add the user to the SAS

[26] The cost of hardware and software purchases is not generally placed on specific RAND FFRDC projects, because the resource can be reused by the RAND FFRDC or the RAND Corporation. Reusable resources are purchased using corporate or RAND FFRDC overhead charges placed on all corporate and RAND FFRDC projects' labor costs. Consequently, they are borne indirectly and partially by the project.

user list. A PDE user is restricted to one computing core and 4 GB of RAM during a typical session. As needed, the PDE works with researchers to secure appropriate licenses.

The analytical method and program are left up to the PDE user.[27] Output should be stored in project folders. Additional productivity programs, such as the Microsoft Office suite, are available. Study teams can share analytical code, Microsoft Office documents, and other files within the team's folders. Users must remain connected to the PDE for it to continue to run; that is, a process cannot be run in the background.

RFL has downloaded a large list of toolkits and packages from public sites for use by supported analytical programs (e.g., RFL downloaded the full set of R packages from the CRAN library). Other analytical program add-ons have to be requested through the PDE help desk.

Discussion

The major differences between conducting analysis in the PDE versus the RAND environment included user-friendliness, hardware limitations, and PDE help desk response times. Some of the comparisons highlighted in the discussion of step 5 are still relevant for the analysis part of a study; these comparisons include consistency of PDE access, connections to project folders from analytic programs, and remote desktop latency.

The PDE is designed to provide the user with a standard Microsoft Windows desktop experience. This means that researchers can compose code, conduct analysis, and record results within an environment that is familiar to Windows users. In practice, due to latency issues described in step 5, composing code or drafting results can be difficult within the PDE.

The PDE also imposes more-stringent hardware limitations. During our study, the limitations became apparent when attempting to do memory-intensive activities, such as generating graphs. This was particularly notable for estimating complex, dynamic programs like the DRM in RAND2. The model estimated as expected, but it was substantially slower given the single computing core restriction.[28] These hardware limitations could be relieved as additional computing resources are made available. Some specific examples are provided in the appendix.

Finally, the time it took for the PDE help desk to respond to and follow up on issues or requests varied. In some cases, an issue or request was resolved the same day. In other cases, issues remained unresolved for an extended period of time, sometimes weeks, which led to significant operational delays. For one member of the research team, it took in excess of one month to resolve an Oracle password reset issue that prevented the researcher from doing any

[27] It was unclear if the PDE would support low-level programming languages, such as FORTRAN or C++.

[28] A computer's processor is made up of cores. A computing core can execute a single computational process. Computers can process the data serially (i.e., one process after another) or in parallel (assumes the two processes are independent). The execution of an analytical program can be sped up by using a faster processor that takes less time to complete each serial computation or by greater parallelization. Modern analytic programs are making use of multi-core processers to handle independent processes.

follow-up work during this period.[29] Extended delays can arise in the RAND FFRDC's analytical environment, but they are generally rarer. This is because additional technical support can be brought in quickly as required due to an institutional alignment of incentives between the analytical environment and the research team.

Step 7: Reporting

Once analysis is completed, the results must be extracted from the analytical environment so that they can be included in reports, presentations, or other manners of sharing what has been learned from the analysis. Release of results from the environment are governed by the operator of the analytical environment based on the data exchange agreements and institutional requirements.

RAND FFRDCs

Researchers are permitted to remove aggregated output from the RAND FFRDC environment for study purposes. They are required by their data use agreement to export only de-identified aggregate data and employ best practices for masking small cell sizes. There are no restrictions on importing or exporting analytical programs or on using data from external sources that do not require data exchange agreements (e.g., publicly available data sets, such as U.S. Census data). The output cannot be shared publicly until it has been cleared for public release.

PDE

This process was formalized during the study's evaluation period. PII and protected health information (PHI) cannot be removed from the PDE, and neither can any output that includes ranks, unique identification codes, or PDE and study-specific identifiers. Code, analytical output, and other descriptive statistics are exported from the PDE using the following method (Research Facilitation Laboratory, 2017):

1. A researcher creates a PDE support ticket via the PDE portal.
2. The study team member places the file or folders to be exported in a specific folder.
3. An RFL PII and PHI review team member reviews the files for PII and PHI.
4. The approved files are moved out of Citrix and emailed to the PDE help desk.
5. The PDE help desk sends the output files to the requesting researcher.
6. The PDE help desk removes the exported file or folders from the review folder.

(The analogous process for importing is defined in step 5.)

[29] From March 16 until April 24, 2017, the RAND3 researcher communicated and followed up with the help desk regularly. After an extended conversation on April 24 when the issue was unresolved, the researcher stopped engaging with the PDE for an extended period. On June 20, 2017, a new feature was announced that automated password resets, which allowed the researcher to reset his password on June 21. At that time, the reset was successful.

Since January 30, 2017, the response time was generally within one business day. For RAND FFRDC projects, output from the PDE cannot be shared publicly until it has been cleared for public release.

Discussion

The reporting processes for the RAND FFRDC environment and the PDE differ significantly. The most significant difference is the level of trust placed on researchers using the environment. Personnel studies that require analytical environments like the PDE or RAND's FFRDC analytical environment tend to involve a highly iterative and collaborative process among the research team, often requiring one member of the team to submit programs in the analytical environment that produce tabulations or estimate models, extract and review the output of those programs for privacy and data safeguarding concerns, discuss that output with the rest of the research team (outside of the analytical environment), and then revise and resubmit the programs in the analytical environment based on the research team's discussions. In the RAND FFRDC environment, review of output is entrusted to research staff with approved analytical environment access, meaning that results can be downloaded and reviewed by the qualified researcher in minutes and then immediately shared with the research team. Alternatively, in the PDE, reporting of results for discussion with the research team outside of the PDE requires filing a request and having it fulfilled by PDE staff. Despite the one-day response time by the PDE staff, this transaction cost has the potential to significantly slow the research process. Because the PDE caters to a broad research audience, its policies and procedures are necessarily more protective and manpower- and time-intensive.

Step 8: Data Archiving and Destruction

Once a study has been reported on, the data and analytical programs should be archived or destroyed in a timely manner. The timing and the consideration of whether or not they are archived or destroyed depends on the need to reproduce the results, the ability to recreate the analytical data sets in the future from existing programs, and the types of data that were used in the analysis. Because the analytical environment is responsible for holding the data, it handles data archiving and destruction. The study sponsor or researcher may also have data archiving and destruction requirements based on the IRB process and determination status.

RAND FFRDCs

Data safeguarding plans, which are drafted during the IRB application phase, typically specify a study's method of storage and data destruction plan. In accordance with the plan, as well as the data destruction plan required as part any data exchange agreement, the original data and any derivative files may need to be destroyed after a specified period. For example, for the data exchange agreement between NDRI and DMDC (DMDC, 2016), the data destruction period

is three years after the study (or successor studies) concludes. During this period, the researcher can elect to move the analytical files (not the original or restricted-access de-identified files discussed in step 4) from the RAND FFRDC's environment to the RAND Digital Archive, which stores unclassified RAND research or administrative materials in a secure repository. This repository can include archiving project data, related code, analytic files, key interim reports, and email. The RAND Data Facility reviews the available data within the restricted-access folders on a routine basis and notifies relevant researchers of the deletion deadline. Unless the appropriate authority authorizes an extension, the data are deleted. Individual project folders are the responsibility of the researchers managing those folders and are to be deleted as required by the applicable data exchange agreements.

PDE

At the end of a project, the researcher initiates the project closeout process with the PDE. If the project was deemed to be research, then it must complete both the group's HRPP closeout (in addition to the research organization's IRB and any applicable second-level review) and an RFL closeout. The RFL closeout documentation requires the researcher to report

- project status: project title, sponsor, HRPP determination
- study personnel information: research team members' names, contact information, and affiliation
- closeout status: reason for closing the project, final sample size and data range used and analyses conducted on the sample, list of reports submitted to DoD, list of reports published, whether the data will be placed in a repository, location and management plan for the data, list of data exchange agreements with a closure process, and requirements for data reuse.

Once the closeout documentation is processed, the research team is locked out of the project folder. The PDE retains project folders for up to three years, including code, analytical files, and study-specific identifier crosswalks, unless otherwise stated by IRB documents or relevant data exchange agreements. If a researcher requires access to the folder, he or she has to provide acceptable justification. The PDE destroys original and derived data within the PDE-S, PDE-A, and Oracle database based on the requirements of the data exchange agreement with the data provider.

Discussion

The study team did not have a chance to evaluate this step. In both environments, data destruction requirements are determined by data exchange agreements and the study's data safeguarding plan. In the case of the data exchange agreements, it is the environment's responsibility to ensure that the original and derived files are deleted, while responsibility for ensuring that derived analytical files are destroyed on schedule is delegated to the researcher.

One potential concern is the ease with which previous folders can be accessed and data can be reused.

Summary

This chapter examined the steps of the research process for three RAND studies conducted in the RAND FFRDC analytical environment and contrasted it with the steps for the same studies conducted in the PDE. The research team was able to acquire DMDC and Army HRC data, validate that these providers' data held in the PDE and in the RAND environment were of similar quality, and conduct the analyses for the three studies. Overall, the processes for the two environments had many similarities, but there were two recurring themes. First, when the research organization is separate from the organization providing the analytical environment, transaction costs can be high, involving additional operational steps leading to delays. This is significant because steps 5–7 of the research process, outlined in Chapter Three, can be an iterative process; in some cases, the procedures were evolving or not specified in advance, adding to transaction costs. Introducing another organization into the research process results in additional review burden with a third IRB or HRPP review, even for studies that the first review determined were not research or did not involve human subjects.

Second, incentive alignment between the researcher and analytical environment is important so that transaction costs, though they exist, are resolved more quickly and efficiently. When priorities are not aligned between the research organization and the organization providing the environment, then research progress can be limited unilaterally by either organization. Research is a very iterative process, meaning that incentive misalignment and transaction costs between researchers and their analytical environment can lead to substantial delays in providing results to research sponsors, as well as an inability to provide quick support to the FFRDC's sponsor, a contractual requirement of FFRDCs.

Chapter Five

Summary and Conclusions

The objective of this study was to determine whether RAND FFRDCs can effectively and efficiently use the PDE to support research studies for DoD, assess how using the PDE compares with the traditional method of accessing DMDC data, and identify what improvements to the PDE would be necessary for FFRDCs to effectively use it for personnel studies. Our approach for assessing the PDE was to (1) identify the data collection and analytical requirements from three in-progress or completed RAND studies typical of manpower and personnel studies conducted in RAND's DoD FFRDCs and (2) replicate the data collection and analysis using the PDE. The focus of this report was on the process of conducting a study (outlined in Chapter Three) and not on the specific findings from those analyses. Our study ran from October 2015 to September 2017. In February 2016, the PDE released PDE version 2.0, and our analyses were based on our interactions in this newer environment. This chapter summarizes our approach and details our broad conclusions regarding whether RAND's FFRDCs can effectively and efficiently use the PDE for DoD-sponsored research. The appendix details specific recommendations for improvements to the PDE.

We developed a conceptual framework in Chapter Two as a way of framing how the PDE fits within a study. This framework contrasts the costs and benefits of integrating the four key elements of a study: sponsor, researcher, environment, and data. The PDE reflects specialization in provision of an analytical environment. Several rationales for the PDE exist, including (1) providing economies of scale by streamlining data access, data management, and data governance bureaucracy and (2) improving data security by minimizing the practice of sending data to the analyst, which exposes DoD to loss of privacy-protected data. In Chapter Four, we contrasted the PDE concept's separation of the researcher's and the environment's institutions with the typical RAND FFRDC setup, which integrates researcher and environment.

We conclude that a nonintegrated analytical environment, such as the PDE, impedes the efficient operation of a RAND FFRDC study relative to existing arrangements by eliminating the alignment of incentives between researcher and analytical environment. Additionally, we make two observations based on this evaluation: (1) A single centralized analytical environment for research processes within DoD has the potential to reduce analytical capacity and discourage researchers from accumulating database-specific knowledge, and (2) the PDE represents a potential opportunity for DoD to engage academic researchers outside of existing research support organizations with recurring DoD sponsor relationships (e.g., FFRDCs, DoD policy shops).

In Chapter Three, we stated the five general requirements of an FFRDC study: objectivity, high quality, timeliness, low cost, and security. Objective research requires that a study reflects

independent analysis of the researchers. High-quality research requires that a study use current methods from academia and private industry. Timely research requires that a study's results are presented in time for the sponsor to include the study's insights as part of its consideration of policy. Low-cost research requires that a study answer the project's research questions using the most-efficient means possible. Secure research requires that the sponsor can rely on its researchers to safeguard records from loss and unauthorized use.

Existing RAND FFRDC arrangements combine researchers and analytical environments within a common institution, which has the effect of aligning incentives so that transaction costs, though they exist, have less of a detrimental effect on efficiency because issues that give rise to these costs are resolved more quickly and efficiently. A nonintegrated analytical environment, such as the PDE, impedes the efficient operation of a RAND FFRDC study relative to existing arrangements by eliminating the alignment of incentives between researcher and analytical environment. Nonintegration places the onus for securing the required data and providing analytical tools on an entity with no direct commitment to the DoD research sponsor. Federal regulation pertaining to DoD's use of an FFRDC states that an FFRDC, "in order to discharge its responsibilities to the sponsoring agency, has access, beyond that which is common to the normal contractual relationship, to Government and supplier data," which it uses to "maintain currency in its field(s) of expertise, maintain its objectivity and independence, preserve its familiarity with the needs of its sponsor(s), and provide a quick response capability" (Code of Federal Regulations, 2017). The PDE, or any similar environment, requires satisfying the agency operating the PDE's requirements in order to be able to carry out the FFRDC's mission. When priorities are not aligned between the research organization and the organization providing the environment, research progress can be limited unilaterally by either organization.

In addition, a nonintegrated analytical environment generates transaction costs through additional bureaucratic requirements associated with additional approvals, operational steps, and coordination. These requirements are limited in some cases, but in all cases, they represent equal or additional burden relative to existing operations. Given the established trust relationship that is intended to exist between the FFRDCs and the departments that sponsor them, the additional burdens imposed by the PDE process—a process generally geared toward researchers and organizations without the established trust relationship—require additional delays and regulatory burden that do not improve the final research product. These additional costs could be outweighed by benefits from efficiencies or data safeguarding; however, we find limited evidence of realized efficiencies in the research process or improvements in data safeguarding relative to existing, required RAND FFRDC processes. The experiences documented in Chapter Four highlighted additional constraints imposed by the PDE, including having a CAC, redundant reviews, transactional costs, environment-imposed limitations on data, restrictions on parallel processing, end-to-end latency issues, and consistency of access. A well-known finding in the economics literature is that when transaction costs are high between organizations, it can be more efficient for the organizations to combine through vertical integration (Williamson, 1985),

thereby internalizing incentives as a means of reducing transaction costs. RAND's FFRDCs, by combining the analytical environment and the researcher, are an example where efficiency is gained through vertical integration and the better alignment of incentives by internalizing transactions within the same organization.

Our first observation based on this evaluation is that a single centralized analytical environment for research processes within DoD has the potential to reduce analytical capacity and discourage the accumulation of DoD institutional expertise among subject-matter experts. It is based on the theory developed in Chapter Two and the documented transaction costs resulting from working in another institution's analytical environment during this research. There have been several recent efforts that move toward a single centralized analytical environment for DoD's research community.[30] If DoD moves toward a single centralized environment, then environmental limitations (e.g., incidental server downtime) broadly affect access and workloads. Additionally, future funding limitations for that environment might constrain its ability to respond to researcher needs. Such limitations reduce the ability of a research organization to respond to sponsor needs. Finally, as the number of users increases, the competition for finite resources at peak time periods will lead to environmental latency, and there exists no clear prioritization among projects. These factors reduce DoD's analytical capacity.

Additionally, the move toward one centralized analytical environment creates an intermediary between the data provider and the researcher, eliminating contact between these entities. Interacting with data providers is important for research staff who want to understand changes in administrative record-keeping and survey methodology that may not be directly discernible from data or information provided by the analytical environment. Retaining this connection can be a valuable way for providing feedback that informs the research process (e.g., data providers may alert the research team to better data sets or identify elements that are not likely to be well recorded for the purposes of the study). Additionally, because researchers often use data in novel ways, they may identify issues or concerns regarding data collection that were not previously well known. Retaining a connection between data providers and researchers can be a valuable way of improving data quality and promoting the accumulation of DoD institutional expertise among researchers who are to be DoD's subject-matter experts.

On the other hand, the addition of the PDE as an alternative environment (i.e., a complement) instead of as a substitute strengthens DoD's analytical capacity. For example, if an FFRDC were to exhibit a data breach and if data transfer and use within its environment were to be restricted

[30] Examples during our study period include the following: An ongoing review of the Academy of Sciences (National Academies of Sciences, Engineering, and Medicine, 2017) recommended steps for DMDC to take to enhance the usability and value of the PDE, implying that DMDC should apply greater effort to improving the PDE; additional meetings within DoD considered the potential of serving the data requirements of the RAND FFRDCs through the PDE; and DMDC file managers, in response to RAND FFRDC data requests, redirected researchers to the PDE rather than providing the data to the FFRDC.

for an indeterminable period, the PDE could represent a means for the FFRDC to continue analytical work that was unrelated to the breach.

Our second observation based on this evaluation is that the PDE represents a potential opportunity for DoD to engage academic researchers outside of existing research support organizations with recurring DoD sponsor relationships. Existing research support organizations with recurring DoD sponsor relationships typically establish a secure environment in order to provide recurring research support. A major barrier to a non-DoD researcher conducting research in support of DoD is access to data in an environment that can safeguard the data and protect personal information. The PDE represents a means of providing that access while the DoD sponsor and data provider can be assured that the PDE's administrative steps will ensure an appropriate HRPP review, limit access to appropriate individuals, and ensure review of any output removed from the environment. To our knowledge, no such environment exists for the support of DoD-sponsored research separate from existing research organizations. In lieu of the PDE, non-DoD researchers would have to pay a substantial fixed cost to establish an environment, or the research sponsor would have to provide the analytical environment.

In the past decade, there has been a greater awareness within DoD of the need for strong data safeguarding protocols that protect PII. This awareness has coincided with tightening budgets and increased demands to quantify the efficacy of DoD-sponsored programs. Providing a common analytical environment for the DoD research community as an approach to establishing data safeguarding protocols attempts to bypass the complexity of this problem with a seemingly simple solution. But complexities remain even with a common environment. Providing consistent requirements for data provision to an analytical environment still remains an issue, as does determining how those requirements should vary based on the level of trust afforded within existing contractual relationships. The repeated need for data exchange agreements has led to an evolving set of requirements. Ideally, future DoD efforts would focus on establishing clear requirements and a set of environments that (1) satisfy the diverse range of analytical needs of DoD sponsors (e.g., broaden access to new researchers through the PDE), (2) provide quick-turn capabilities and the accumulation of DoD institutional expertise among subject-matter experts (e.g., established long-term relationships, such as FFRDCs and University Affiliated Research Centers), and (3) can be consistently and routinely monitored to ensure compliance with DoD data safeguarding and privacy policies. Once requirements are clearly established and the set of analytical environments are known, sponsors of DoD research can choose the combination of research expertise and analytical environment that produces the quality research they require in an efficient, timely, and secure manner.

Appendix

Recommended Changes to the PDE

In this appendix, we list recommended additions or improvements to the PDE from the perspective of RAND's FFRDC personnel researchers based on their research environment needs. A version of this list was originally provided to PDE staff in response to a request from OUSD (P&R) in November 2016.

Recommended Software

The following list represents regularly used analytical software by the RAND FFRDCs:

1. SAS (Base SAS 9.4)
    a. Analytical products
        i. SAS/STAT 14.1
        ii. SAS/ETS 14.1
        iii. SAS/OR 14.1
2. Stata MP
3. R
4. Rstudio
5. GAMS
6. Mathematica
7. MS Office products
8. UltraEdit Studio
9. Git.

Additionally, there should be an ability to add toolkits and packages easily.

Recommended Hardware

There should be levels of virtual desktops that analysts can choose based on their expected need. Standard virtual desktops should be assigned with at least 16 GB of RAM and one computing core. More-advanced desktops should have access to 32 GB of RAM and four cores. Finally, there should be a special virtual desktop available that has access to more memory and computing power (e.g., 256 GB of RAM and 32 cores). The larger hardware demands would be for computationally intensive projects. Access to multiple cores is very important for utilizing the full potential of Stata MP and SAS.

Hard-disk space should be sufficient to handle monthly files of major data sets (e.g., the Active Duty Master File, Active Duty Pay File) over a long period, such as from 2000 to the present. In some cases, a large number of variables is required from each file (e.g., all of the pay

elements in the pay file), so disk space must be sufficient to permit entire files or nearly entire files over a long period. (This is a problem with the existing PDE.) SAS uses hard-disk storage for temporary working files. The system should be configured with adequate free space to hold data sets at least three times the size of the active files. This is particularly important when sorting data sets.

Recommended Process Improvements

1. Significantly reduce response times. This includes
    a. three days for project approval (by project approval here, we mean the creation of a project within the PDE, not the IRB process or request for data access)
    b. three days for data extracts following data access approval in the PDE
    c. one day for privacy review when removing tables and regressions from the PDE
    d. two hours for loading data into the PDE and other simple tasks, such as creating folders
    e. immediate help from a help desk.
2. Provide consistent and easy access to the PDE for qualified users. Currently, access to the PDE is intermittent, with blackout periods sometimes lasting weeks. (We believe that this is more of a problem for people outside the Secret Internet Protocol Router Network, or SIPRNet.)
3. Send proactive communication about planned or unplanned outages. Communication of outages or system problems helps end users to (1) identify if interrupted access to the environment is due to an issue on their end or on the environment's end and (2) plan their daily work flow and update the rest of their research team when an analysis may be delayed.
4. Provide more user-controlled processes. For example, the ability to initiate an Oracle password reset was helpful.

Recommended Improvements for User-Friendliness

1. Build the ability to execute programs from a command prompt so that projects can be run in the background without having to retain a connection to the PDE.
2. Provide tutorials and example code for accessing the PDE from Windows or Macintosh computer and for accessing data through Oracle (Toad) for use in one of the programs (e.g., SAS, Stata, R).

References

Code of Federal Regulations, Title 48, Chapter 1, Subchapter F, Part 35, Section 017, Federally Funded Research and Development Centers, 2017. As of May 31, 2018: https://www.law.cornell.edu/cfr/text/48/35.017

Defense Manpower Data Center, "Memorandum of Agreement Between Army Data Center Fairfield and Defense Manpower Data Center," July 2010.

———, "Memorandum of Agreement Between Defense Manpower Data Center and Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics for Data Sharing with RAND National Defense Research Institute Federally Funded Research and Development Center, Agreement Number A1604," January 2016.

DMDC—*See* Defense Manpower Data Center.

Mattock, Michael G., and Jeremy Arkes, *The Dynamic Retention Model for Air Force Officers: New Estimates and Policy Simulations of the Aviator Continuation Pay Program*, Santa Monica, Calif.: RAND Corporation, TR-470-AF, 2007. As of May 31, 2018: http://www.rand.org/pubs/technical_reports/TR470/

National Academies of Sciences, Engineering, and Medicine, *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions*, Washington, D.C.: National Academies Press, 2017.

Office of the Secretary of Defense, "Sponsoring Agreement Between the Office of the Secretary of Defense and the RAND Corporation for the Operation of the National Defense Research Institute Federally Funded Research and Development Center," September 2016.

Polich, J. Michael, James N. Dertouzos, and S. James Press, *The Enlistment Bonus Experiment*, Santa Monica, Calif.: RAND Corporation, R-3353-FMP, 1986. As of May 31, 2018: https://www.rand.org/pubs/reports/R3353.html

Research Facilitation Laboratory, "The RFL, PDE, & RAND: Opportunities for Collaboration," presentation, Arlington, Va., April 2015.

———, "Import/Export Process for the Person-Event Data Environment," user guide, January 2017.

Shelton, James D., "How to Interpret the Federal Policy for the Protection of Human Subjects or 'Common Rule' (Part A)," *IRB: Ethics and Human Research*, Vol. 21, No. 6, 1999, pp. 6–9.

Williamson, Oliver, *The Economic Institutions of Capitalism*, New York: The Free Press, 1985.