TRAC-TR-12-037
20 August 2012

Africa Knowledge, Data Source, and Analytic Effort (KDAE) Exploration

TRADOC Analysis Center
700 Dyer Road
Monterey, CA 93943-0692

This study cost the Department of Defense approximately $77,000 expended by TRAC in Fiscal Years 11-12.
Prepared on 20120801
TRAC Project Code # 614

THIS PAGE INTENTIONALLY LEFT BLANK

TRAC-M-TR-12-037
20 August 2012

Africa Knowledge, Data Source, and Analytic Effort (KDAE) Exploration

Authors

MAJ Thomas Deveans
Ms. Sara Lechtenberg-Kasten
Dr. Samuel Buttrey
Dr. Ronald Fricker
Dr. Jeffrey Appleget
LCDR Walter Kulzy

PREPARED BY:

THOMAS DEVEANS
MAJ, US Army
TRAC-MTRY

APPROVED BY:

JONATHAN K. ALT
LTC, US Army
Director, TRAC-MTRY

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** 31 July 2012	**3. REPORT TYPE AND DATES COVERED** Technical Report, August 2011 to July 2012	
4. TITLE AND SUBTITLE Africa Knowledge, Data Source, and Analytic Effort (KDAE) Exploration			**5. PROJECT NUMBERS** TRAC Project Code 614
6. AUTHOR(S) MAJ Deveans, Ms. Lechtenberg-Kasten, Dr. Buttrey, Dr. Fricker, Dr. Appleget, LCDR Kulzy			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army TRADOC Analysis Center - Monterey 700 Dyer Road Monterey CA, 93943-0692			**8. PERFORMING ORGANIZATION REPORT NUMBER** TRAC-M-TR-12-037
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) Center for Army Analysis			**10. SPONSORING/MONITORING AGENCY REPORT NUMBER** TRAC-M-TR-12-037
11. SUPPLEMENTARY NOTES Findings of this report are not to be construed as an official Department of the Army (DA) position unless so designated by other authorized documents.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			**12b. DISTRIBUTION CODE**
13. ABSTRACT (maximum 200 words) The TRADOC Analysis Center (TRAC), Naval Postgraduate School (NPS), and other Department of Defense (DoD) organizations are currently conducting large data capture and analysis efforts on areas all around the world. As efforts in the US Central Command (CENTCOM) Area of Responsibility (AOR), particularly in both Iraq and Afghanistan draw down, many senior decision makers expect that the US African Command (AFRICOM) AOR will be the focus of future efforts in the coming years. This project will first build an assessment framework focused on the AFRICOM AOR identifying what data we would ideally like to gather and measure in a COIN environment, and then by actually gathering the data points from a multitude of sources we can identify gaps in the available data. Concurrently, this effort will develop the necessary software within the DaViTo (Data Visualization Tool), an open source, government owned exploratory data analysis tool, in order to allow the end user to construct an assessment framework utilizing a customized weighting scheme along with the ability to display results. Finally, this project will develop a scenario methodology and a small Proof of Principle use case in Nigeria by conducting factor analysis of survey data and will use Generalized Linear Models (GLMs) in order to predict future issue stance scores and observed attitudes and behaviors of the population that will directly support TRAC's Irregular Warfare Tactical Wargame (IW TWG).			
14. SUBJECT TERMS Current Operations Analysis, Metric Assessment Framework, Data Visualization, Cultural Geography, generalized linear models			**15. NUMBER OF PAGES** 82
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

NOTICES

DISCLAIMER

Findings of this report are not to be construed as an official Department of the Army (DA) position unless so designated by other authorized documents.

REPRODUCTION

DISTRIBUTION STATEMENT

Approved for public release; distribution is unlimited.

DESTRUCTION NOTICE

When this report is no longer needed, DA organizations will destroy it according to procedures given in AR 380-5, DA Information Security Program. All others will return this report to Director, TRAC, ATTN: ATRC, 255 Sedgwick Avenue, Fort Leavenworth, Kansas 66027-2345.

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The TRADOC Analysis Center (TRAC), Naval Postgraduate School (NPS), and other Department of Defense (DoD) organizations are currently conducting large data capture and analysis efforts on areas all around the world. As efforts in the US Central Command (CENTCOM) Area of Responsibility (AOR), particularly in both Iraq and Afghanistan draw down, many senior decision makers expect that the US African Command (AFRICOM) AOR will be the focus of future efforts in the coming years.

This project will first build an assessment framework focused on the AFRICOM AOR identifying what data we would ideally like to gather and measure in a COIN environment, and then by actually gathering the data points from a multitude of sources we can identify gaps in the available data. Concurrently, this effort will develop the necessary software within the DaViTo (Data Visualization Tool), an open source, government owned exploratory data analysis tool, in order to allow the end user to construct an assessment framework utilizing a customized weighting scheme along with the ability to display results. Finally, this project will develop a scenario methodology and a small Proof of Principle use case in Nigeria by conducting factor analysis of survey data and will use Generalized Linear Models (GLMs) in order to predict future issue stance scores and observed attitudes and behaviors of the population that will directly support TRAC's Irregular Warfare Tactical Wargame (IW TWG).

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

DISCLAIMER ..... **III**
REPRODUCTION ..... **III**
DISTRIBUTION STATEMENT ..... **III**
DESTRUCTION NOTICE ..... **III**
ABSTRACT ..... **V**
TABLE OF CONTENTS ..... **VII**
LIST OF FIGURES ..... **IX**
LIST OF ACRONYMS AND ABBREVIATIONS ..... **XI**
ACKNOWLEDGMENTS ..... **XIII**

SECTION 1. THE AFRICA KNOWLEDGE, DATA SOURCE, AND ANALYTIC EXPLORATION (KDAE) EXPLORATION PROJECT **1**
1.1. BACKGROUND **1**
1.2. OBJECTIVES **2**
1.3. ORGANIZATION OF THIS DOCUMENT **2**

SECTION 2. DEVELOPING THE METRIC FRAMEWORK **5**
2.1. MOTIVATION **5**
2.2. FRAMEWORK **6**
2.3. THE ASSESSMENT FRAMEWORK IN DAVITO **7**

SECTION 3. DEVELOPING THE DATA SOURCES **11**
3.1. FINDING THE DATA **11**
3.2. GAP ANALYSIS **14**

SECTION 4. SCENARIO METHODOLOGY & DEVELOPMENT **21**
4.1 INTRODUCTION **21**
4.2 THE SURVEY DATA **21**
4.3 RECODING AND DATA IMPUTATION **22**
4.4 FACTOR ANALYSIS **25**
4.4.1 The Factor Analysis Model **27**
4.4.2 Determining the Number of Factors **29**
4.4.3 Fitting the Model **30**
4.4.4 Choosing the Preferred Rotation **30**
4.4.5 Factor Analysis of the 2010 Nigeria Survey Data **31**
4.5 PREDICTIVE MODELS **34**
4.5.1 Predicting Issue Stance Scores Using Linear Regression **35**
4.5.2 Predicting Future OABs Using Multinomial Logistic Regression **36**
4.5.3 Proof of Principle Scenario **39**

SECTION 5. CONCLUSION **46**

APPENDIX A. ASSESSMENT FRAMEWORK USER'S GUIDE **47**

APPENDIX B. NIGERIA FOCUSED DATA COLLECTION EFFORTS **49**

APPENDIX C. R CODE FOR CAPTURING / DOWNLOADING DATA SOURCES.....53

APPENDIX D. SME INPUT & LOOK-UP TABLE.....61

APPENDIX E. R CODE FOR FACTOR ANALYSIS AND REGRESSION MODELS.....62

LIST OF REFERENCES.....81

LIST OF FIGURES

Figure 1. Description of data sources 14
Figure 2. An illustrative example of factor analysis with six observed variables that can be effectively summarized in terms of two latent variables (factors). 26
Figure 3. Parallel analysis where the eigenvalues for 27 factors were greater than those from the simulated data (the blue line is greater than the dashed red line). 33
Figure 4. List of factors and factor names 34
Figure 5. Cumulative issue stance score over time for the 4 key issues. 41
Figure 6. Partial listing of cumulative issue stance changes over time 42
Figure 7. Observed attitude and behavior probabilities over time. 43
Figure 8. Partial listing of observed attitude and behavior probabilities over time 44
Figure 9. Notional SME input values for each factor by event 61
Figure 10. Look-up table for issue stance and OAB by event 61

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AFRICOM	US Africa Command
AO	Area of Operations
CENTCOM	US Central Command
CG	Cultural Geography
CODDA	Complex Operations Data Development Activity
COIN	Counter Insurgency
DaViTo	Data Visualization Tool
EDA	Exploratory Data Analysis
GLM	Generalized Linear Model
HNSF	Host Nation Security Forces
IR	Information Requirement
IW TWG	Irregular Warfare Tactical Wargame
KDAE	Knowledge, Data Sources, and Analytic Efforts
LOE	Lines of Effort
MCO	Major Combat Operations
MPICE	Measuring Progress in Conflict Environments
OAB	Observed Attitude and Behavior
OE	Operational Environment
SME	Subject Matter Expert
TRAC	TRADOC Analysis Center
TRADOC	Training and Doctrine Command
VEO	Violent Extremist Organization

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to especially thank Ms. Sara Lechtenberg-Kasten, a member of TRAC's Complex Operations Data Development Activity, whose tireless efforts in locating sources of data for this project were instrumental in its success. Additionally, I would like to thank LCDR Kulzy, Dr. Fricker, and Dr. Buttrey for all of the advice and guidance regarding the mathematical methods and R code utilized during this effort. Lastly, I would like to thank the sponsor of this effort, the Center for Army Analysis, whose generosity and guidance made this project possible.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 1. THE AFRICA KNOWLEDGE, DATA SOURCE, AND ANALYTIC EXPLORATION (KDAE) EXPLORATION PROJECT

1.1. BACKGROUND

The African continent has not always been at the forefront of American foreign policy. It was not even until the early 1960's that the entire continent was assigned to a responsible military command, long after the other ones. The 1998 terrorist attacks against the two U.S. embassies in the East African capitals of Tanzania and Kenya, and the resulting retaliatory strikes by the Clinton Administration in the Sudan marked a major turning point in U.S. strategic policy and interest in the region (Ploch, 2011). In 2002, then President Bush in his 2002 National Security Strategy laid out future policy and actions in the region necessary to combat terrorism and its catastrophic effects not only upon the continent, but the entire world as well.

> In Africa, promise and opportunity sit side by side with disease, war, and desperate poverty. This threatens both a core value of the United States - preserving human dignity - and our strategic priority - combating global terror. American interests and American principles, therefore, lead in the same direction: we will work with others for an African continent that lives in liberty, peace, and growing prosperity. Together with our European allies, we must help strengthen Africa's fragile states, help build indigenous capability to secure porous borders, and help build up the law enforcement and intelligence infrastructure to deny havens for terrorists (The White House, 2002).

More recently, president Obama has reinforced this view is numerous press releases, policy documents, and public statements. In a speech to the Ghanian Parliament, he said:

> When there is genocide in Darfur or terrorists in Somalia, these are not simply African problems, they are global security challenges, and they demand a global response.... And let me be clear: our Africa Command is focused not on establishing a foothold on the continent, but on confronting these common challenges to advance the security of America, Africa, and the world (Obama, 2009).

Security concerns as well as terrorists and Violent Extremist Organizations (VEOs) are not the only reason for the increase in U.S. interests in Africa, however. America's increasing oil consumption and subsequent increasing reliance on foreign countries for oil has made finding willing trade partners a matter of vital national importance. America has gone from importing 25% of its oil from foreign countries 20 years ago to importing nearly 60% today (Varner, 2007). Add to this the growing influence of China in the region as the Chinese similarly seek to quench

their increasing thirst for oil, and we see a Cold War like “battleground” where the U.S. and China are fighting for both diplomatic and economic influence in Africa.

The chief national interests of the United States in Africa include an increasing importance on natural and energy resources, mounting threats and the growing concerns over VEOs, that include piracy and illegal trafficking, as well as the many humanitarian crises, brought about by famine and genocide (Ploch, 2011).

1.2. OBJECTIVES

The primary objectives of this effort are to determine the necessary information requirements to be used in order to measure progress in the region should the United States become involved in counterinsurgency operations similar to those currently ongoing in Afghanistan; identify, collect, and consolidate existing data sources in the region; and identify gaps between the required and existing data sources / information requirements. Concurrently, this effort will develop the necessary software within the DaViTo (Data Visualization Tool), an open source, government owned exploratory data analysis tool capable of displaying over 100,000 data points simultaneously from multiple data sets consisting of multiple data types, in order to allow the end user to construct an assessment framework utilizing a customized weighting scheme and to display the results. Finally, this project will develop a scenario methodology and apply it in a small Proof of Principle use case in Nigeria by conducting factor analysis of survey data and will use Generalized Linear Models (GLMs) in order to predict future issue stance scores and observed attitudes and behaviors of the population that will directly support TRAC’s Irregular Warfare Tactical Wargame (IW TWG).

1.3. ORGANIZATION OF THIS DOCUMENT

The remainder of this document will focus on what has been accomplished thus far, including the development of the metric framework and the corresponding data sources (including a quantitative only data source “deep dive” into Nigeria and the surrounding Sahel region), as well as gaps between those information requirements specified in the metric framework and the available data sources. Additionally, the new DaViTo functionality allowing users to build an assessment framework and display the results will be described. Lastly, a “proof of principle” scenario will be developed using survey data and generalized linear models.

Finally, the appendices will go into greater detail about not only the metric framework, including a kind of “User’s Guide”, but also about specific data sources and how and where they can be found and updated. Specific details regarding the analyses with the corresponding R code will also be provided.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 2. DEVELOPING THE METRIC FRAMEWORK

2.1. MOTIVATION

The first task in this project was to determine what information requirements are necessary in order to measure progress, success or failure, in the event of U.S. involvement in some type of conflict on the African continent. For our purposes here, we define and gather our information requirements in a COIN environment, one similar to our current involvement in Iraq and even more so in Afghanistan. The proper development and use of metrics within a carefully nested assessment framework are critical to giving key decision makers the ability to make the correct decisions at the right place and time. This includes tracking or measuring progress against the objectives laid out in a campaign plan, as well as ensuring that subordinate headquarters' metrics and assessment framework are aligned with those of higher. It is critically important that an assessment framework drive the decision making process, otherwise they become irrelevant. Too much information is just as bad as not enough as highlighted in a recent article:

> One persistent criticism is that operational assessments teams have overreached in the pursuit of perfection. Some have tried to measure the universe, attempting to aggregate all the disparate information in the battlespace. Others, at the other end of the spectrum, have thrown up their hands and accepted the constraints of statistical reporting, merely counting events rather than interpreting them. Another criticism is that assessments often proceed from flawed assumptions with little real-world evidence. The varied cast of agencies performing assessments can at once be criticized for being too complex in their methodology and too simplistic in their analysis. This has resulted in understandable disenchantment with the assessments process (Upshur, Roginski, & Kilcullen, 2012).

The U.S. militaries recent experiences in both Iraq and Afghanistan have highlighted the importance of using metrics and their corresponding indicators that are measurable, collectible, and relevant. These two recent conflicts, both occurring in a COIN type environment, where precise knowledge of progress is even more difficult to discern than in an MCO construct, also highlight the difficulty of measuring the correct metrics over time, of maintaining the flexibility to adapt how forward progress is measured as the environment changes and the enemy adapts, and translating the large amounts of raw data into something of use to a decision maker.

2.2. FRAMEWORK

In light of this necessity, the members of this project team have built an assessment framework, with the goal of determining the information requirements that are needed to measure progress. The framework was built with input from many sources (described below and listed in References), and throughout the process every attempt was made to remain general enough and applicable to the entire continent of Africa. Consequently, this framework is just a tool or starting point, with the option, or necessity really, of "down selecting" certain measures that may be irrelevant or inane to a specific area, or impossible to find. A word of caution, however, using all 158 measures in the framework developed by this project as an assessment tool is ill-advised. There is way too much information here to synthesize down to a level whereby a staff or commander can make any sense of the whole.

The first section in the framework is a section with information requirements that describe the overall operational environment and is intended to provide the user with a better understanding of the environment and a context with which they can more effectively measure progress. This section includes such information requirements as general information on the terrain in the area of operations (AO), population demographics, ongoing developmental projects in the AO, insurgents and violent factions operating in the AO, as well as organizations and government structure in the AO. Every attempt was made to tie each information requirement to specific data sources from the widest possible locations including U.S. government, various international organizations, academia, and Subject Matter Expert (SME) input. In Section 3 we will provide a more detailed description of the data collection efforts.

The second section in the framework describes Lines of Effort (LOE), broken down into tasks, indicators, and measures. As a starting point, FM 3-24 Counterinsurgency was used to develop the overall Lines of Effort (LOE) for the assessment framework (United States, 2006). These include establishing civil security and control, support to Host Nation Security Forces (HNSF), support to governance, restoration/establishment of essential services, and support to economic and infrastructure development. We then broke down each LOE into specific tasks, each with its own indicator(s) and associated measure(s). The tasks, indicators, and measures were gathered from a variety of works and sources including FM 3-24, Counterinsurgency and FM 3-24.2, Tactics in Counterinsurgency (United States, 2009), a U.S. joint and interagency

effort called Measuring Progress in Conflict Environments (MPICE) (Dziedzic, Sotirin, & Agoglia, 2008), various articles from COIN SMEs including David Kilcullen (Kilcullen, 2009), input from several former Special Forces officers (D. McCracken & S. Whitmarsh, personal communication, September 15, 2011) and U.S. government employees with extensive experience on the continent of Africa (S. Kasten, personal communication, September 30, 2011), and recently deployed operational research analysts (G. Kramlich, personal communication, October 15, 2011) to include my own experiences in Afghanistan as a member of an assessments cell. All of this input was pulled together and used to create the assessment framework itself, and subsequent efforts, as described in the paragraph above and in Section 3, attempted to tie each measure to a specific data source from a multitude of locations.

The document containing the actual assessment framework is too large to be incorporated in this technical report, and so it can be viewed or downloaded at this link.

2.3. THE ASSESSMENT FRAMEWORK IN DAVITO

In direct support of the efforts of this project, TRAC's Data Visualization Tool (DaViTo) was modified to include the capability to not only build an assessment framework, but to visualize chosen measures and subsequent "progress" as well. DaViTo is an exploratory data analysis tool employed across DoD that enables analysts to conduct first level analysis into data sets as well as to visualize and refine second order analyses.

TRAC modified DaViTo in order to facilitate a value hierarchy deconstruction of a data set. This value hierarchy methodology is based upon multi attribute utility theory and allows users to define a structure which allows attributes to be weighted according to their relative importance. We describe the method employed in more detail below.

Given a vast data set with attributes for various regions within the AFRICOM AOR, it is extremely difficult if not impossible to compare all attributes to one another in a meaningful and methodical manner. For this reason, we chose to utilize concepts from multi attribute utility theory in which comparisons are continually broken down into smaller and smaller chunks until the analyst must only compare a small subset of the overall data set wherein all elements are somehow related (Luce & Tukey, 1964).

For example, given the measures of population growth rate, HIV deaths and education expenditures, among others it is often too difficult to decide in a consistent manner how to weight the importance of each attribute. However, if we allow our overall rating of stability for a region to be broken up into subcategories representing lines of effort such as 'Medical', 'Governance' and others as needed, we can begin to partition our data set into smaller clusters which become more and more closely related as the subdivision process continues.

We accomplish this partitioning by use of trees. The root node at the top of the tree represents stability. Stability can then be assigned any number of children. Those child nodes can then be broken up further and further as necessary until the analyst is left with only a small number of attributes as siblings whose relative importance can be determined more systematically.

Once the structure of the value system tree has been determined, the analyst must then decide what the minimum and maximum thresholds for the data are. Using infant mortality as an example, when measuring a region for stability, we may want to compare the infant mortality rate to that or the United States. The United States has an infant mortality rate of roughly 0.6%, which is extremely low when compared to second and third world nations. It must then be determined, with regard to stability, what a threshold is such that an infant mortality rate below the threshold no longer contributes to the region's stability. For the sake of this example we shall choose 6%. A similar decision must be made for the worst case value. Looking at data for Africa, we can see that no country in Africa has an infant mortality rate higher than roughly 11%. Therefore, an analyst may want choose this value as the worst case threshold.

Once the best case and worst case thresholds are established for all of the attributes in the data set, the analyst then chooses weights for each node in the tree relative only to its siblings. Thus, we only compare medical attributes with medical attributes, and governance attributes with other governance attributes, and so forth.

In the inner nodes of the tree, we compare between siblings as well. So, we need to choose the relative importance of medical stability to that of stability in governance. These broader categories are much easier to grasp with regard to relative importance than the more detailed measures taken from the data set. Once this has been completed for all nodes in the tree,

leaf nodes are attributes from the data set with best-case and worst-case values as well as a weight, all internal nodes are given weights, then the overall stability score for each region is computed using this fixed value system.

The value system itself is developed before it is applied to the data set and can be saved to an xml file for reuse on new data, or similar data sets. The data is then evaluated against the value system hierarchy tree and can be saved to a Microsoft Excel Workbook file where each region is printed out on its own worksheet and each region's worksheet has the structure of the tree preserved by noting the threshold values, parent node, and the evaluated value, by attribute. See Table 1 below for an example.

Node	Parent	Red	Green	Weight	CalculatedValue
Stability	null	0	1	1	0.541666667
Economy	Stability	0	1	4	0.541666667
Finance	Economy	0	1	2	0.5
Population.below.poverty.line	Finance	49	47	1	0.5
Production	Economy	0	1	1	0.625
GDP	Production	46.12	48.12	3	0.5
Per.capita.GDP	Production	2200	2300	1	1
Military	Stability	0	1	1	0.75
Population.fit.for.military.service	Military	2794997	2794999	1	0.5
Military.expenditures	Military	0.3	1.3	1	1
Health	Stability	0	1	2	0.5
Population.growth.rate	Health	3.082	1.082	1	0.5
Education	Stability	0	1	3	0.5
Education.expenditures	Education	2.7	4.7	1	0.5

Table 1. Cameroon value system evaluation output.

The capability to complete all of the actions required by users described above has been implemented in Java within DaViTo, and is available in the latest version of DaViTo, located here http://trac.nps.edu/davito. Also included at this link is a User's Guide that describes the functionality of DaViTo and assists the end user in getting started. All of the actions are completed within an intuitive graphical user interface similar to most computer applications in use today.

THIS PAGE INTENTIONALY LEFT BLANK

SECTION 3. DEVELOPING THE DATA SOURCES

3.1. FINDING THE DATA

The amount of data that needs to be analyzed at the tactical level to support strategic decisions can be overwhelming. Combine the vast amounts of data with the changes to the operations plan that state that coalition forces need to focus on applying counterinsurgency lines of effort simultaneously, rather than sequentially, and current methods used to support decision analysis are no longer suitable. For commanders to conduct resource allocation across all of the LOE simultaneously by the most effective means possible, new methods for collecting, analyzing and displaying data are required.

There is a growing recognition of the importance of data at all levels, from tactical to strategic levels, and increasing complexity for soldiers operating in full spectrum operations, beyond the COIN mission. New types of military engagements will require the Department of Defense to learn to access new data sets. For example, most recent Department of Defense policy and White House policy indicates data needs for topics that have been previously outside of the "comfort zone" for the U.S. Army. In collecting data for the framework, the team evaluated key policy guidance that has emerged from USG policy and Department of Defense doctrine, to include "Sustaining U.S. Global Leadership: Priorities for 21st Century Defense" (January 2012) indicating that the U.S. Army must be prepared to protect citizens and hand over security responsibilities to the HNSF. The data set includes points to consider for both overarching security sector and host nation armed forces, as in the recently published "The United States Army Operating Concept 2016-2028" (August 2010). The Department of Defense will be required to engage with host nation security forces in the prevention of mass atrocities as outlined in "Fact Sheet: A Comprehensive Strategy and New Tools to Prevent and Respond to Atrocities" (April 2012). Key concepts from these recent policies are included in the tool within the "Line of Effort" column, and in the Data Source section, the user is able to find key concepts that relate to these policies. For example, although it was not "Prevention of Mass Atrocity" is not included as a specific Line of Effort, concepts related to this foreign policy goal are included under the section titled "Establish Civil Security/Number of incidents of violence between people / groups of different communities." Data to support this issue includes texts, including

legal concepts related to prevention of mass atrocities is provided. Similarly, the tool allows the user to know that when looking at Prevention of Mass Atrocities, they should also consider the concept of "Responsibility to Protect" which is directly associated. In no way does this tool allow the user to instantly see what these concepts mean, but instead allows the user to know what data points and foundational research should be undertaken.

Although this initiative has been in production mode/draft mode over this fiscal year, it has already proved to be useful to various data requestors. For example, in support of the Protection of Civilians Working Group run out of the Peacekeeping and Special Operations Institute, the draft spread sheet allowed the Working Group to rapidly frame issues related to diverse topics, and allowed for the production of read-ahead data took less than five minutes to consolidate and send in an email, rather than an more extended research, analysis, and production effort. Time invested in advance to populate the tool paid dividends to the working group and prevented wasted time in searching for definitions of concepts and framing issues. For the purposes of the Protection of Civilians Working Group, the data provided is not in any way a complete guide, but rather starting points to help familiarize participants with key issues.

Additionally, the data set has helped influence scenario development for two efforts thus far. In the past, data collection for scenarios has been a time consuming undertaking, and triangulating data to determine whether the data is reliable and usable for scenario development has been a challenge. In scenario development conferences, often participants will spot a specific issue, or data gap, and then a researcher would need to go and find data points to fill the gap. At the previous Africa related conference, it was as simple as using the "find" function in the spreadsheet to rapidly point the scenario team to appropriate data sources for issues ranging from United Nations involvement in the area, potential concerns with refugee settlement, USG Interagency participation, and a host of other issues. The spreadsheet / data source tool allowed for definitive data on the wide range of issues, and allowed the team to avoid using opinion in lieu of research and analysis.

In the second scenario development, the geographic area was outside of Africa, but the data sources found for the Africa specific effort were still of utility to quickly access data on global issues related to PMESII-PT. For example, refugee resettlement was also an issue in the second scenario and the data points listed for the AFRICOM spreadsheet led to information on

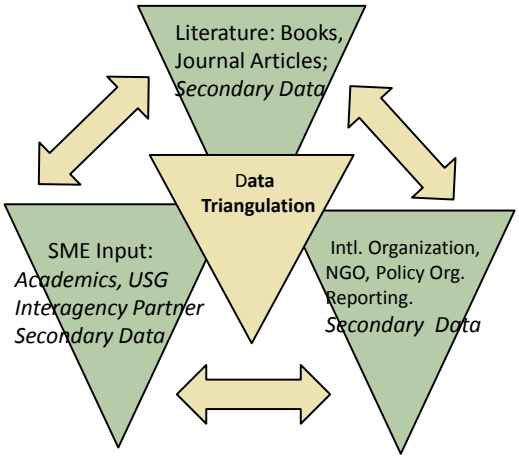
refugees in general and the organizations that are global responders. Although much of the tool would need to be adapted for areas outside of AFRICOM, a substantial number of the data points have applicability for other combatant commands.

As the data collection project progressed, the team endeavored to include emerging USG policy along with established doctrine. For example, as the Counterinsurgency (COIN) Center at Ft. Leavenworth undertook the rewrite of FM 3-24, potential changes to the COIN doctrine, not yet published, were included in the tool, so as to ensure that once new doctrine is published, the tool will be relevant to any changes. Similarly, as the Counterinsurgency Center trained Brigade Combat Teams for deployment to Afghanistan, the Complex Operations Date Development Activity (CODDA) provided quick turn data responses on issues related to rule of law, governance, human rights, interagency collaboration and other issues. While still in draft form, the tool allowed the CODDA to provide direct responses to the Counterinsurgency Center to inform pre-deployment training for a BCTs deploying to an area outside of AFRICOM. Commanders sought out data on larger conceptual issues in the pre-deployment phase, for example, Rule of Law, but once on the ground, the shift might turn to needing to know what USG agencies "do" Rule of Law and what Non-Governmental Organizations facilitate these programs. The tool allows the data requestor to get a quick familiarization with complicated issues, and allows the user to quickly identify leaders on the issue.

As a supplemental effort to this project, the project team did a data "deep dive" into available data sources in the Sahel region of Africa and more specifically in the country of Nigeria, as it represents a potential location for an upcoming IW TWG, and supports the "proof of principle" scenario methodology using survey data from the same region. The focus of this effort was finding strictly quantitative data, unlike some of the sources in the assessment framework, that can be used directly in future modeling or metric building works. Results of this effort are given in Appendix B with corresponding narrative as well as the data sets themselves in the form of outside links to related files. Also provided in Appendix C are the R scripts that were implemented for capturing and downloading the data. Only slight modifications to these scripts would be necessary in order to use them in order to find or download data sources from other areas of the world or similar websites.

3.2. GAP ANALYSIS

In a perfect world, it would be possible to consult with one data source, preferably an easy to access website, to access a reliable and up-to-date database to find information. The reality of the situation is that for the majority of the data needed, a user needs to consult with a variety of sources, compare results from data pulled from those sources, and triangulate the data in order to ensure that it is correct based on three types of sources: SME input from U.S. government and interagency partners, literature, books, and journal articles, as well as international and non-governmental organizations.



Goal: Identify three data points for each component of the Data Source section to allow user to draw data from multiple sources. For complicated data requests, best to refer to three data points, rather than one "go to site" to collect data.

Figure 1. Description of data sources

Although the initial goal for the data collection phase was to find one key source to mitigate the data gap, the team endeavored to find multiple data points for each line of the spreadsheet. Unfortunately, there is not uniformity in the level of detail provided for each section of the spreadsheet. For some needed data points, there was plethora of data points available, ranging from United Nations/World Bank Data, US Government, policy organizations, and academic literature. For other needed data points, only one, sometimes questionable data point was provided. When data was limited or questionable, the tool provides a full description of why

the data was possibly of diminished value/questionable in value. For example, although Wikipedia is not a definitive site, there were quite a few sites that had consolidated data sets on issues ranging from African Military Strength to Child Soldiers. In no way should one Wikipedia site be used as a sole source for data, but in some cases, it was of utility to include. For the majority of the data points requested, at least one data point was found to shed insight onto the issue. In some cases, one single report served to close the data gap. For example, USAID, "Africa Regional Rule of Law Status Review" (2009) included in depth/detailed information for the “Establish Civil Control” section and the user of the tool can see which specific chapter to consult in that report to find the needed data.

In more complicated data needs, numerous data points were included, so as to provide the user the opportunity to triangulate data. For example, an individual seeking data on Host Nation Security Forces will be provided with an open-source US Government link at the U.S. Department of State, on the Security Assistance landing page website. Additional data on Host Nation police inclusion in the Host Nation Security Forces is provided, via a contact at the U.S. Department of Justice. Additional data points for Host Nation Security Force include a text on Military Balance in Africa, a recently published report by the policy organization International Institute for Strategic Studies, and a link to SIPR NET that will provided data on integration of international forces with host nation forces. Additionally, two recent books on HNSF in the African context and one Wikipedia site are included. A distinct challenge for a data collector is that for most data needs, there is no single “go to” site or text that will provide the “right” answer. Instead, a skilled user of data will need to refer to multiple sites to collect relevant data for intended uses.

Not all data responses in the spreadsheet are as richly populated; but instead, provide the user with a pathway to find the needed data. For example, in the “Describing the Operational Environment” section, a required data point was “Political, Religious, Tribal Motivations of a Group.” Due to the broadness of this data requirement, the information provided is more of a general description of how to find the data, not a specific report or URL. The response to this requirement leads the user to consult with the US Embassy in the country of interest and provides the website of where to track down the appropriate personnel at the US Embassy with data on the issue. Additional data for this section directs the user to consult congressionally

mandated reports like the Annual Report to Congress on International Religious Freedom and provides the specific website for access. Additional texts are recommended to arm the researcher with foundational understanding of politics, religion, and tribal issues and understand associated issues, such as the role of land tenure and how it intersects with political, religious and tribal issues. It was not possible to draw in connected issues on all of the Data Source sections, but when possible, related issues were included to help the user frame the issue. Most of the books and journal articles listed in the spreadsheet are available on Google books, thus enabling quick download from remote locations. When possible, the spreadsheet provides the precise chapter to consult for a specific issue, thus allowing the user to rapidly sift through the text, without having to read the entire book.

The following are key data gaps that have not yet been completely mitigated:

1. Local Labor (19): We've found a proxy source for this in literature related to use of Chinese labor in lieu of local labor, but nothing substantive that speaks to quantitative data on local labor usage and impacts on local economies, etc. We project that this is a gap that may not be filled with research state side, but rather would require in country interviews/on site evaluation. Nothing within USAID or UN sites thus far has indicated that local labor data is systematically collected.

2. Insurgents / Violent Factions Operating in AO / Crime (20-28). Although we have found links for these data points, we would still consider this section to be a "gap." We are still looking to develop a spreadsheet on the SIPR side that will house more authoritative links for these issues from intelligence agencies, rather than foundational literature. This is a gap, but just needs more time to be closed by reaching out to sources at DIA and CIA. We've already made contact with Defense Intelligence Officer Martin Kindl (gatekeeper for all of DIA's AFRICOM threat data), but now need to drill down to the functional area and geographic area to fill gap.

3. Number and type of armed incursions by non-state actors from neighboring states, Kill ratio between foreign fighters and security forces, Amount of funding flowing from foreign states or transnational actors to violent factions, % of population that feels that they can travel safely within the country, outside of their own tribal area (58-61). Thus far, finding accurate reporting on NIPR side has been challenging. Issues such as incursions, kill ratios, and

perceptions of safety have been hard to find and should be considered gaps. Additional research on SIPR may provide some resources, and combined with literature review, may partially mitigate this gap. On the issue of perceptions of safety, this may be an ongoing gap which would be difficult to fill from a research position in the US. Data collection for perceptions on safety in tribal area may be a data question for Human Terrain Systems or other data collector. At best, we may only be able to find anecdotal data on this issue. There have been some efforts from JTF-HOA to implement the CIDNE database to track these issues and once this effort has been started, pulling data on these issues will be much easier.

4. Attrition rates (23). We have found academic resources, but not as of yet found satisfactory quantitative data for this component. In the end, we may only be able to get to SME data, not more detailed attrition rates.

5. Places of Religious Worship (36). Although we have been able to identify sources for discussion of major religions, actual locations of religious worship in the form of an eight digit grid is problematic.

6. School and Universities (40). We still need to submit an RFD to assessment team at USAID. It is possible that we may be able to get some assessments, but probably won't have a consolidated database of all educational facilities on continent without more time.

7. Level of governmental involvement with the people (43), Degree of reliance by the people on informal governmental structures for support (44). Although some academic and international organizations are listed for data points, this section needs much more work.

8. Influential figures publicly denounce acts of violence & other obstructionist behavior (51). We have found one USG data point and one academic source, but other data points are needed for this issue.

9. Number and severity of attacks of key HN facilities (to include both critical HN and privately owned facilities) (56). We are at this time uncertain at this time how to fill this and are reaching out to others to determine best approach. This might be compiled by various USG sources from a qualitative perspective to track trends, but not sure will find data that is accurate for number and severity.

10. Level of participation of ex-combatants in the political process & civil society (76). Gap. This is indicated as a gap, but can be closed with additional research. Although this data project has been country agnostic to a certain extent, the ex-combatant issue will be very country specific. Still trying to conceptualize how to find and display this data.

11. Recruitment rates vs. desertion rates (87), Identity groups are represented in the HNSF in the same proportion as they exist in the population as a whole (89), Locations of HNSF units (90), Ability to effectively plan, execute, and sustain operations (91), Number of autonomous HNSF operations and success rates. Thus far, the team has been unsuccessful in identifying data for these components and there is concern that accessing this data will be difficult if not impossible to access.

12. Government taxation vs. faction taxation vs. illegal extortion (101). We may only be able to get foundational data (literature review) here, and not exact/detailed quantitative data. With more time we can close the gap more with SME/reports from Transparency International, etc.

13. Perception among minority & majority identity groups of nepotism in the civil service (110). At present, this is still listed as a gap, due in part to the complexity of collecting data on this issue and will need additional time to at least partially mitigate with SME, literature review.

14. Prohibited political parties as a % of total (113). Initial research will allow for identification of political parties, but we are having more difficulty accessing information on prohibited parties.

15. New urban construction start rate (160). We are at the present time concerned that this data is not collected in a systematic way for most African countries. Will continue to look for data, but may only be able to find proxy sources.

For some data requests, there is no simple way to access the data. In these cases the spreadsheet guides the user through a potential path to find needed data. For example, there is not a specific database or site to consult for the presence of International Organizations in Africa. Instead, the response leads the user through various approaches to collect the data, from a mostly top down data collection methodology of checking United Nation umbrella sites, indicates other International and Regional Organizations with potential operations on the continent, and

provides recommended texts describing the mandates, strengths, and weaknesses of those entities. In sum, providing only information about the presence of the International Organization may not be effective and inclusion of academic studies of the effectiveness of the organization in the country will allow the researcher to compile a more robust response.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 4. SCENARIO METHODOLOGY & DEVELOPMENT

4.1 INTRODUCTION

The objective of this part of this effort is to develop a methodology and build a proof of principle scenario in a specific region or country in the AFRICOM AOR for use in future IW TWG's using Factor Analysis and Generalized Linear Models. This section will describe the survey data that was used, the recoding/imputation and factor analysis, and the subsequent linear and multiple logistic regression models that will allow us to predict future population Issue Stance Scores as well as Observed, Attitudes, and Behaviors. Additionally, a small "proof of principle" scenario will demonstrate how these models can be used to predict future population responses.

4.2 THE SURVEY DATA

This part of the project is based on survey data collected in six countries in the Western Trans-Sahel region of Africa. The surveys have been conducted over the past four years, though not every country was surveyed in each of the available years. The analysis for this portion of the effort focuses on the survey conducted in the country of Nigeria, during the year 2010. This particular country was chosen as it represents a possible and likely location for the upcoming Irregular Warfare Tactical Wargame scenario lead by the TRADOC Analysis Center – White Sands Missile Range.

These surveys were initially sponsored by AFRICOM, and conducted by a private contractor operating in the region with no discernible affiliation to the U. S. military or the U.S. government. AFRICOM's objective in conducting this project was to better ascertain how their actions affect the daily lives of the indigenous populations, while also looking to identify areas of the data that can be used when determining future courses of actions or allocations of resources (Kulzy, 2012).

The survey instrument for 2010 consists of 255 questions and 3,770 respondents for the country of Nigeria. However, of these questions, some are specific to only one or two countries. There are also questions to which a Likert scale value cannot be associated, so they are coded as nominal values. There are also a number of questions that were conditional on responses to other

questions. These conditional questions are, for example, specific to only one type of religion or are only answered if a previous question was affirmatively answered. These types of questions were omitted from the analyses, as they were deemed to bias the responses as they applied to only a subset of the population surveyed (Kulzy, 2012).

4.3 RECODING AND DATA IMPUTATION

Table 2 specifies the particular survey questions that were used in the analysis. All questions in the survey instrument that were asked of all respondents were included in the analysis. Conditional questions, based on skip questions, as discussed above, were not used in the analysis.

Source of Information	Q5
Quality of Life	Q6 – Q10
View of foreign countries	Q12, Q14, Q16, Q17, Q21 – Q23
Views of Nigeria	Q25
Trust and Religion	Q26 – Q34, Q36, Q37
Governance, Politics, and Security	Q40, Q45, Q48 – Q50, Q52
Acts of Violence	Q56 – Q59
U.S. Actions	Q60, Q62
Demographics	D12 – D17, D21 – D24, D26

Table 2. Related questions specific to the analysis

Crucial to any quantitative modeling of survey data is the appropriate preparation of the data. The first step in this process is re-coding the responses from the original Likert scale responses to numeric values. Various Likert scales were used in the survey and they differed both in terms of qualitative scales and response ranges. For example, a four-point Likert scale accounted for 66% of the total number of questions. Typically the response scales were in the

form of “always positive,” “somewhat positive,” “somewhat negative” and “always negative,” or some other similar positive to negative range. The survey also had questions with five-point Likert scale responses as well as binary responses. Recoding was done using the *CAR* package with the R statistical software program. Before the data was re-coded, it was important to determine how the response would be viewed. The factors used in this analysis were re-coded in a positive or negative direction depending on how a U.S. analyst would interpret the numeric variables loaded onto the factors. Consistency in the direction of the recoded variables does ease the burden of interpretation once the factors have been defined and the linear models fit. In general if a response was assumed to be positive to a U.S. analyst, then the response was given a positive value, and if it was assumed to be negative, then it was given a negative value. Numeric re-coding values range from a +2 to a -2. If the range was a four-point Likert scale, then the extreme positive and negative answers were given a +2 and -2 respectively. The moderate positive and negative were given a +1 and -1 respectively. The re-coding values for a fivepoint Likert scale is similar to a four point one, but with a 0 coded for neutral type responses such as “stayed the same.” Three-point Likert scales have a +2 and -2 for extremes and 0 coded to neutral responses, but there are no moderate values. There were also questions that offered binary responses, such as a general “Oppose” or “Support,” and a more formal choice of response as “Shari’ah reduces crime in society” or “Shari’ah does not affect the amount of crime in society.” These types of questions were given values of -2 or 2 (Kulzy, 2012).

The “Don’t know” and “No response” responses in this data were treated as unknown values that needed to be imputed. This is in contrast to the typical solution for handling missing data, which is to remove the associated entire observation from the data. This approach is often referred to as casewise deletion. In terms of survey analysis, casewise deletion means that if a respondent failed to respond to one question, then all of the rest of his or her information from the other 141 questions would be removed. For this data set, if casewise deletion was used in order to be able to first conduct a factor analysis and subsequently fit regression models, 2,240 of the 3,770 Nigerian observations (60%) would be removed from analysis. This is in spite of the fact that each question only had a very small percent of missing responses. Thus, imputation is crucial to this survey because imputing only 6% -8% of the data saves 60-72% of it for analysis. Missing data was handled using nearest neighbor hot-deck imputation, a more sophisticated

method than simple mean imputation, and was implemented using the state or region as a matching variable in order to account for spatial variation in the data (Kulzy, 2012). The hot-deck imputation method used in this effort is based on the *RANDwNND.hotdeck* function within *StatMatch* package of R. The imputation for the missing values, to include "Don't knows" and "No Response" responses, was done using the variables: region/state (the states of the country), "d5a" (religion), "d0" (gender), "urban/rural" (live in urban or rural area of state). The *RANDwNND.hotdeck* function initially subsets the data based on specific "donor class" variables. For this research, the donor class variable is the "state" variable. Basing the donor class on geographic state ensures that geographic heterogeneity is accounted for in the imputation. Within each state, then, the data is subset into two groups: the receivers and the donors. The receivers are those respondents who are missing the response to a particular question and the donors are those respondents who have answered the same particular question. For each receiver, a donor is then identified that is closest to the receiver in terms of Manhattan distance based on his or her religion, gender, and location (urban/rural). If there is more than one "closest" donor, then "one is picked at random" from among the tied group of the closest matches (D'Orazio, 2011).

Imputing all of the "Don't know" responses could have an impact on a few questions that loaded onto a factor with a minimal significant value of 0.4. Those questions loading as a 0.4 in one imputation would be considered significant. However, if the process was to be repeated, there is a chance that a minimal, loaded value question may now fall below the 0.4 threshold and be removed from the factor. It was determined to recode the "Don't know" responses in a manner that minimized the volatility of these few questions which rest on the cusp of the 0.4 threshold. It was assumed that a "Don't know" in the three and five point Likert scales would be equivalent to a "No Response" because a neutral, valued at zero, response was offered. Therefore, three and five point Likert scales of "Don't know" were imputed in the same manner as a "No Response." A more difficult question is how to best analyze "Don't know" responses in a two- and four-point Likert scales since these types of scales do not offer an explicit neutral response option. It is reasonable to assume that a "Don't know" response to a question with only "Strongly agree", "Agree," "Disagree," and "Strongly disagree" could, in fact, be using the "Don't know" response option to express neutrality, particularly when there was also a "No

response" option. Thus, in these cases a "Don't know" response was re-coded to a value of 0, a choice which seems conservative in the sense that without it imputing these responses would result in a potentially neutral person being given a positive or negative response (Kulzy, 2012). This assumption addresses over 60% of the missing data that would have otherwise required imputation. Roughly 6-8% of Nigeria's questions did not have a clear response, and eight of these questions are asked on either a two- or four-point Likert scale for Nigeria. Since there is no clear and definitive interpretation of the "Don't know" responses for these questions, and because of the large number of these questions, a closer analysis was performed. It is plausible to believe that without an option to be neutral, as in two- and four-point Likert scales, a logical interpretation of "Don't know" is neutral which would then result in re-coding it to zero. If this were to be the case then these questions would not be explicitly imputed. However, this is not necessarily true for other types of questions (Kulzy, 2012).

4.4 FACTOR ANALYSIS

One of the major challenges with large surveys is reducing the mass of data into useful information. Another challenge with surveys aimed at understanding the human terrain, particularly when applied to irregular warfare, is that the population characteristics of interest may not be directly measured via single questions. Factor analysis helps address both of these issues.

Critics of the factor analysis argue that its inherent subjectivity and flexibility allows analysts to manipulate the output. The non-unique solution of the factor loadings is often particular focus of this criticism. However, all mathematical and statistical models can be manipulated, and most involve making numerous subjective choices (choice of variables, model parameterization, etc). In this sense, factor analysis is no different. As with those methods, and research in general, it is incumbent on the researcher to ensure his or her results are not sensitive to, or dependent on, modeling choices. That said, remember that the goal of factor analysis is to create factors that are both statistically and substantively meaningful, and the latter implies -- perhaps requires -- a degree of subjectivity.

Factor analysis is a hybrid of social and statistical science. First conceived in the early 1900s, the goal was multivariate data reduction, but data reduction of a very specific type.

Essentially the idea is to explain the correlation structure observed in *p* dimensions via a linear combination of *r* factors, where the number of factors is smaller than the number of observed variables, and where the factors achieve both "statistical simplicity and scientific meaningfulness" (Harman, 1976).

Figure 2 illustrates the idea of factor analysis with six observed variables (i.e., survey question responses) that can be effectively summarized in terms of two latent variables (factors). Note that the survey question responses are observed with error (denoted by the ε_i terms) and the question responses are weighted linear combinations of the factors (where the weights are the λ_{ij}s). What factor analysis does is model the *p* observed variables as linear combinations of *r* factors, where the analyst has to pre-specify *r*, such that the model covariance matrix closely matches the sample covariance matrix of the observed variables.

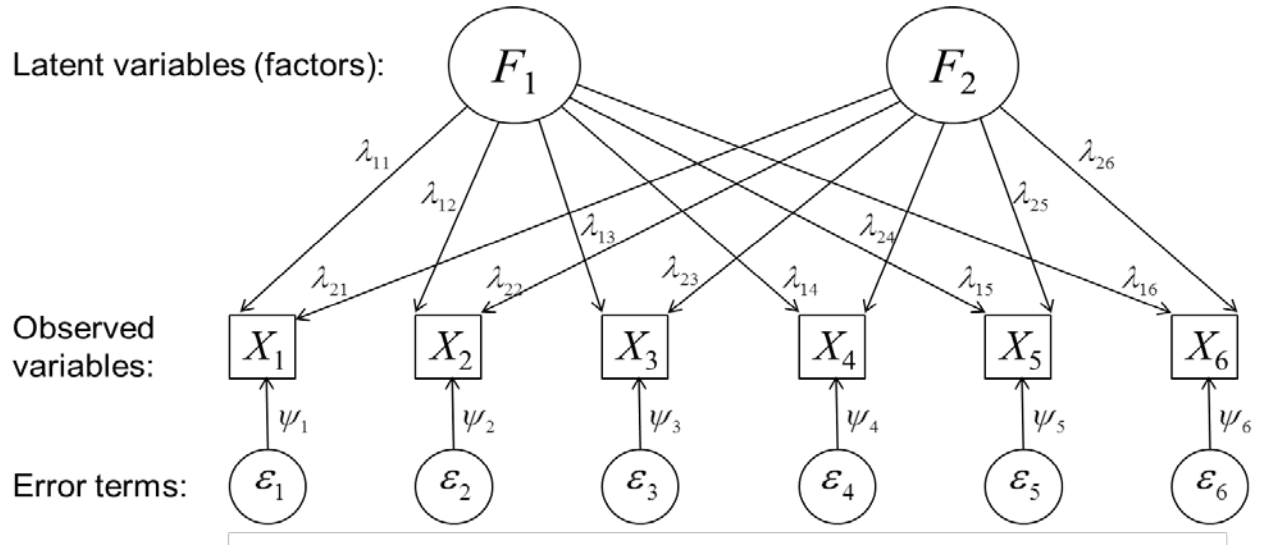


Figure 2. An illustrative example of factor analysis with six observed variables that can be effectively summarized in terms of two latent variables (factors).

An alternative to factor analysis is principal components which uses orthogonal transformations to convert a set of possibly correlated variables into a reduced set of uncorrelated variables that capture most of the variation in the original data. The transformation is defined so that the first principal component accounts for as much of the variability in the data as possible, and each succeeding component has the highest variance possible under the constraint that it be orthogonal to the preceding component or components. A principal components analysis, while

useful for efficiently summarizing data, does not necessarily result in factors with scientifically meaningful interpretations.

In contrast, factor analysis is specifically designed to look for meaningful commonality in a set of variables (DeCoster, 1998). There are two types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA looks to explore the data to find an acceptable set of factors. In this sense, it is much like exploratory data analysis. The goal is not so much to formally test hypotheses as it is to discover likely factors that will account for at least 50 percent of the common variation in the observed factors. CFA, on the other hand, begins with a theory or hypothesis about how the factors should be constructed and seeks to test whether the hypothesized structure adequately fits the observed data.

4.4.1 The Factor Analysis Model

Consider a survey consisting of *p* questions given to *n* respondents, where respondent *i*'s responses are denoted $\boldsymbol{y}_i = \{y_{i1}, \dots, y_{ip}\}$. From the data, a sample covariance matrix **S** is calculated in the usual way for the set of centered variables,

$$\boldsymbol{x}_i \triangleq \{y_{i1} - \bar{y}_1, \dots, y_{ip} - \bar{y}_p\},$$

where $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$. That is, the $j(k)^{\text{th}}$ entry of **S** is calculated as $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}x_{ik}$, $j \in \{1,2, \dots, p\}$ and $k \in \{1,2, \dots, p\}$.

The fundamental assumption of factor analysis is that, for some $r < p$, each of the *p* centered variables ($\boldsymbol{X} = \{X_1, \dots, X_p\}$) can be expressed as the sum of *r* common factors ($\boldsymbol{F} = \{F_1, \dots, F_r\}$) multiplied by their loadings ($\lambda_{i1}, \dots, \lambda_{ir}$) plus a unique factor ($\boldsymbol{E} = \{\varepsilon_1, \dots, \varepsilon_p\}$) multiplied by its associated loading ($\psi_1, \dots, \psi_p$),

$$\begin{aligned}
X_1 \triangleq Y_1 - \mu_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1r}F_r + \psi_1\varepsilon_1 \\
X_2 \triangleq Y_2 - \mu_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2r}F_r + \psi_2\varepsilon_2 \\
&\vdots \\
X_p \triangleq Y_p - \mu_p &= \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pr}F_r + \psi_p\varepsilon_p
\end{aligned} \tag{1}$$

where $\mu_j = \mathbb{E}(Y_j)$. Now, while the above formulation looks similar in many respects to a series of linear models, note that *everything* on the right-hand side of the *p* equations is *unobserved*. In spite of that, the goal is to estimate the loadings from the data so that the modeled covariance matrix **R** is "close to" the observed sample covariance matrix **S**.

Using matrix notation, Equation (1) can be expressed compactly as

$$\boldsymbol{X} = \boldsymbol{\Lambda F} + \boldsymbol{\Psi E}, \tag{2}$$

where $\boldsymbol{\Lambda}$ is the matrix of the loadings for the common factors of dimension $p \times r$ and $\boldsymbol{\Psi}$ is a matrix of dimension $p \times p$ with $\psi_1, \ldots, \psi_p$ on the diagonal and all off diagonal entries zero. Assuming $\mathbb{E}(\boldsymbol{E}) = \boldsymbol{0}$, we get to the whole point in fitting the factor analysis model, which is that we can use the estimated common factor loadings $\widehat{\boldsymbol{\Lambda}}$ to express the factors in terms of their constituent parts:

$$\mathbb{E}(\boldsymbol{F}) = \widehat{\boldsymbol{\Lambda}}^{-1}\mathbb{E}(\boldsymbol{X}). \tag{3}$$

One of the most common uses of exploratory factor analysis is to "determine what sets of items hang together in a questionnaire" (DeCoster, 1998). Thus, assuming Equation 1 is an appropriate model, via Equation 3 we can determine which of the survey questions are most related and, as desired, use them to estimate the underlying latent factor for any respondent as a linear combination of their responses to the survey questions. Furthermore, if the scientific meaningfulness goal is achieved, the latent variables will have useful and interpretable meanings that provide additional insight into the characteristics of the populations being studied.

Of course, at this point it should be evident that there will be no unique solution to this problem. There are simply too many degrees of freedom in the problem formulation and, even after some assumptions to make the problem solvable, there will still be an infinite set of solutions. This, along with the fact that the choice of solution is subjective, is one of the frequent criticisms of factor analysis. Nonetheless, as we will show, we have found the results to be quite informative and useful in our survey analyses, and there are ways to minimize the number of subjective modeling choices that must be made. There are three critical steps in fitting a factor analysis model: (1) Determining the number of factors, (2) fitting the model in order to estimate

the common factor loadings, and (3) rotating the loadings to find the preferred solution. We discuss each of these in turn.

4.4.2 Determining the Number of Factors

To conduct factor analysis, one must pre-specify the number of factors *r* to fit. In so doing, it is crucial not to underestimate or overestimate the number of factors. If too few factors are chosen then the fitted factors become overloaded with irrelevant variables. On the other hand, with an excessive number factors the variables may be spread out too much over the fitted factors. In either case, the result is likely to be that meaningful factors are never properly revealed.

This seems like a catch-22: To determine the correct factors, one must first know how many factors there are. However, over the years a number of solutions have been proposed, some that work better than others.

One early solution is the Kaiser rule which stipulates that the number of factors used in the model should equal the number of eigenvalues for the original data matrix that are greater than one. Another is to use a Scree plot to graph successive eigenvalues versus the number of factors and then setting *r* to the number of factors where the plotted line visually levels out (indicating that the remaining factors have little explanatory power).

The difficulty with the Kaiser rule and the Scree plot is they are heuristics. The Kaiser rule was designed to help the analyst of the early- to mid-1900s get “into the ballpark” with respect to an acceptable number of factors, but then the analyst was supposed to further refine the acceptable number of factors through trial and error. The Scree plot is also a heuristic because it allows for subjectivity in interpreting the plotted line where, to determine the number of factors, the analyst must visually determine when the line in the Scree plot levels out.

An alternative to these methods, which only became feasible with the widespread availability of significant computing power, is parallel analysis. Parallel analysis involves the construction of multiple correlation matrices from simulated data, where the average eigenvalues from the simulated correlation matrices are then compared to the eigenvalues from the real data correlation matrix. The idea of parallel analysis is that factors derived from the real data should

have larger eigenvalues than equivalent factors derived from repeatedly resampled or simulated data of the same sample size and number of variables. Then *r* is set to the number of factors in the actual data that are greater than the average of the equivalent simulated data factor eigenvalues (Hayton, Allen, & Scarpello, 2004).

4.4.3 Fitting the Model

Given that by definition $\mathbb{E}(\boldsymbol{X}) = \mathbf{0}$, and assuming that the common factors are independent of the unique factors, it is straightforward to show that the covariance matrix for **X** from Equation 2 is

$$\boldsymbol{R} = \boldsymbol{\Lambda R_F \Lambda'} + \boldsymbol{\Psi}^2, \tag{4}$$

where $\mathbf{R}_F$ is the covariance matrix of the factors (Mulaik, 2009). Further assuming that $\mathbb{E}(\boldsymbol{F}) = 0$ and $cov(\boldsymbol{X}) = \boldsymbol{I}$, where the former condition follows because the factors can always be rescaled and the latter because we assume the factors are independent, Equation 4 simplifies to

$$\boldsymbol{R} = \boldsymbol{\Lambda\Lambda'} + \boldsymbol{\Psi}^2. \tag{5}$$

Then from Equation 5, **Λ** and **Ψ** are estimated via maximum likelihood.

Note that the maximum likelihood estimators (MLEs) are not analytically derivable and must be solved for numerically using an iterative approach. Under the assumption that **F** and **E** are jointly normally distributed, the calculations essentially follow the usual estimation methods with an additional uniqueness condition added because of the indeterminacy of the factor analysis model.

4.4.4 Choosing the Preferred Rotation

Maximum likelihood estimation results in a non-unique solution for how the variables load onto the factors. That is, for any estimated common factor loading matrix $\widehat{\boldsymbol{\Lambda}}$ there are infinitely many other matrices that will fit the observed sample covariance matrix **S** equally well since

$$\widehat{\boldsymbol{\Lambda}}\boldsymbol{F} = \widehat{\boldsymbol{\Lambda}}\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{F} = \boldsymbol{\Lambda}^*\boldsymbol{F}^*, \tag{6}$$

where $\boldsymbol{\Lambda}^* = \widehat{\boldsymbol{\Lambda}}\boldsymbol{T}$ and $\boldsymbol{F}^* = \boldsymbol{T}^{-1}\boldsymbol{F}$ for some transformation matrix **T**.

Thus, after an initial solution is found, the final step in factor analysis is to rotate the variables to simplify their factor loadings. The rotation process is critical to factor analysis because it allows the analyst to identify the desired factor constructs, usually in terms of a simple structure of substantively interesting variables. However, this procedure is susceptible to criticism because all rotations are mathematically equivalent and thus the final choice is subjective.

There are two main types of rotation: (1) oblique, and (2) orthogonal. Orthogonal rotation is most commonly associated with what is called the "varimax" method, and oblique rotations are most commonly associated with what is called the "promax" method. The distinction between the two rotations is whether the factors are assumed to be correlated or not; orthogonal rotations are uncorrelated while oblique rotations may be correlated.

Kline says the most accepted method for creating factors with simple structure is varimax (Kline, 1994). On the other hand, the oblique method is recommended by Costello & Osborne because it can account for both correlated and uncorrelated factors (Costello & Osborne, 2005).

We used the varimax rotation on our survey data and found it to work well. As defined in Johnson & Wichern, the varimax procedure finds an orthogonal transformation matrix **T** that maximizes

$$V = \sum_{j=1}^{r} \left[\sum_{i=1}^{p} \tilde{\lambda}_{ij}^{4} - \frac{1}{p} \left(\sum_{i=1}^{p} \tilde{\lambda}_{ij}^{2} \right)^{2} \right], \tag{7}$$

where $\tilde{\lambda}_{ij} = \hat{\lambda}_{ij} / \sqrt{\sum_{j=1}^{r} \hat{\lambda}_{ij}^{2}}$ (Johnson & Wichern, 2002). Equation 7 is akin to calculating the sum of the variances of the factor loadings across the *r* factors. What varimax does is find the rotation that makes the high loadings as high as possible while simultaneously making the low loadings as low as possible on each factor.

4.4.5 Factor Analysis of the 2010 Nigeria Survey Data

As mentioned in Section 4.2, the Nigeria survey was fielded in 2010 to 3,770 respondents. A sample of sufficient size is an important consideration since the sample covariance matrix **S** is an estimate of some underlying true covariance matrix **Σ**. That is, since factor analysis focuses only on the sample covariance matrix, it is important that **S** is in fact a

good estimate of $\mathbf{\Sigma}$ in order to ensure the resulting factors represent underlying features of the population and not the noise or other artifacts of the sample.

The factor analysis models were fit using the R statistical package. In particular, the *factanal* function in the base package was used to fit the factor analysis model and rotate the loadings to get the final solution. And, we used the *fa.parallel* of the R *psych* package to do the parallel analyses (Revelle, 2011).

Prior to fitting the factor analysis models, we first cleaned and coded the data, and then we imputed a small number of missing values in order to prepare the data as described previously in detail in Section 4.3. The most important point to make here is that factor analysis can only be done with complete data and thus imputation is a critical step to complete prior to doing factor analysis. For our data, approximately six percent of the data was missing (due, for example, to respondents refusing or failing to answer one or more questions), but they were spread throughout the data set. Thus, if we had only used complete records, we would have eliminated 60 percent of the respondents. Imputation allowed us to use all the data and subsequent sensitivity analyses demonstrated that our imputation assumptions had no practical effect on the factor analysis results.

Returning to factor analysis, as discussed in Section 4.4.2, we first used parallel analysis to determine r, the number of factors. Figure 3 shows the results from *fa.parallel* for Nigeria, where the eigenvalues for 27 factors were greater than those from the simulated data (the blue line is greater than the dashed red line), so we set $r = 27$. Sensitivity analysis using other values of r subsequently confirmed that $r = 27$ was indeed appropriate. In the end, however, we only used 26 factors, as the last one contained low factor loadings, contained only two questions that were also repeated in another factor, and was therefore not used in this analysis. Of note, also is the fact that for this research, variables with loadings between 0.4 and -0.4 were removed.

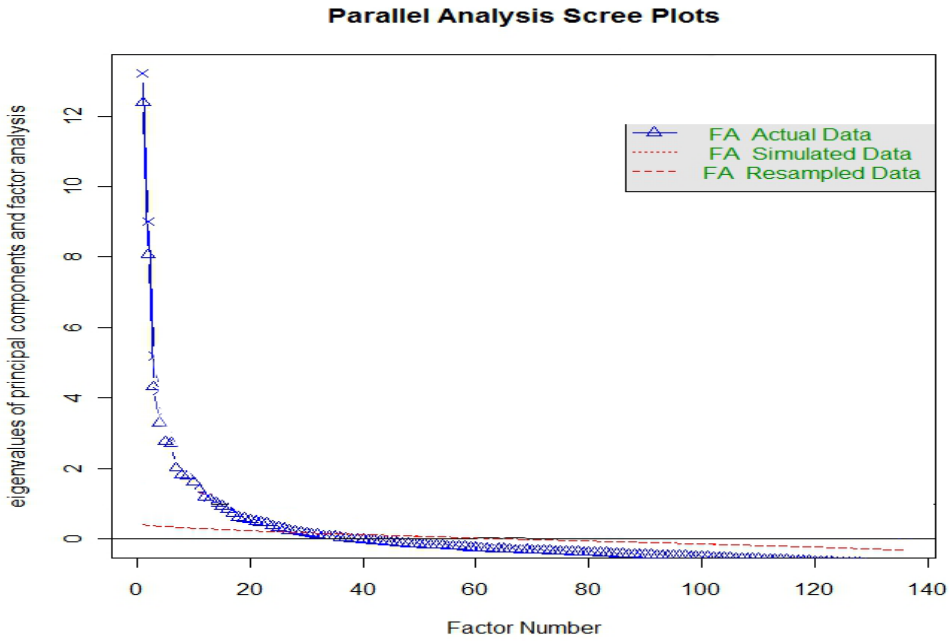


Figure 3. Parallel analysis where the eigenvalues for 27 factors were greater than those from the simulated data (the blue line is greater than the dashed red line).

The list of the factors and the questions that load onto each is given below in Figure 4. Each factor name was chosen subjectively based on the content of the questions that loaded onto each particular factor. These 26 factors, in addition to 4 other survey questions that were not used in the factor analysis, will become the variables used in the next part of the project where we build regression models in order to predict future population issue stance scores and observed attitudes and behaviors.

The final step in this factor analysis is to compute a factor score for each respondent. This is a necessary step if we wish to conduct further analysis with the factors or to use them in any kind of model building. The score for a given factor is simply the linear combination of each measure or question, weighted by the corresponding factor loading (DeCoster, 1998). We can further refine this by rescaling the resulting factor score by dividing by the column (factor score) sums, thereby obtaining a factor score of between -2 and 2, the same as our recoded scale as described in Section 4.3.

Factor Name	Factor No.	Questions							
Shari'a Law	X1	q32a	q32b	q32c	q32d	q32e	q33	q57	
U.S. Assistance to Nigeria	X2	q21a	q21b	q21c	q21d	q21e	q21f	q21g	q21h
Chinese Assistance to Nigeria	X3	q22a	q22b	q22c	q22d	q22e	q22g	q22h	
Social & Essential Services	X4	q8edu	q8hea	q8wat	q9edu	q9hea	q9wat		
Trust in Government Agencies	X5	q49na	q49pp	q49af	q49cj	q49lp	q49lg		
External Security	X6	q23b	q23c	q23d	q23e	q23f			
General Trust	X7	q26a	q26b	q26c	q26d	q26e			
Non-Western Countries	X8	q12ni	q12ir	q16so	q16li	q16sa			
Local & National Freedom	X9	q48a	q48b	q48c	q48d	q48e	q48f		
Democracy	X10	q40	q42	q44	q45				
Other's Values	X11	q17sa	q17fr	q17ch	q17ir	q17us			
Daily Life Acceptance	X12	q27a	q27b	q29a	q29b				
Use of Violence	X13	q25a	q25b	q25c					
Terrorism Enablers	X14	q23a	q59d						
Family & Friends	X15	q27c	q27d	q29c	q29d				
Civic Duty	X16	d24a	d24b						
Attacks on U.S.	X17	q58a	q58b	q58c					
Discussion of U.S.	X18	q62a	q62b	q62c					
Electricity	X19	q8ele	q9ele						
Western Countries	X20	q12uk	q12fr	q14usa					
Trust in Policy Makers	X21	q49pr	q49pm	q50					
Religious Freedom in the West	X22	q37c	q37d						
Religious Intolerance	X23	q36a	q36b						
Civility	X24	q28	q30						
Policy and Law	X25	q31a	q31b						
Roads	X26	q8roa	q9roa						

Figure 4. List of factors and factor names

4.5 PREDICTIVE MODELS

We now move on to use what we have done with the data through the recoding, imputation, and factor analysis to building regression models that will enable us to predict a population's response in light of future events within the context of the TRAC IW TWG.

In short, the IW TWG seeks to stimulate a player such that he/she are forced to make the "best" decisions and develop appropriate courses of action in a given location and scenario. In order to do this, the game model must be able to provide feedback from the local populace to the player on how player decisions effect population perceptions. The subsequent linear and multinomial logistic regression models that predict population responses were built specifically with this functionality in mind, to stimulate player action and decision making in a simpler, and more traceable way than is currently being used with TRAC's Cultural Geography model.

4.5.1 Predicting Issue Stance Scores Using Linear Regression

The first step in building linear regression models used to predict future issue stance scores and the subsequent OABs (though using different model), is to determine what issues are most important to the population. That is, of all of the factors that we identified during the factor analysis, which ones matter most to the people as well as providing the most predictive power? To do this, we take the 26 factors and 4 other survey questions (q6, q7, q10, d23) that were not used in the factor analysis (this will avoid multi-colinearity problems), and regress each against all the other ones, thereby creating 30 linear regression models all with 29 predictor variables (no interaction terms were used). In order to create the simplest predictive model that minimizes over-fitting, we use a stepwise deletion process, specifically the *stepAIC* function in R. This function, in order to find the statistically significant predictor variables, deletes the term with the highest *p*-value (greater than 0.05), re-runs the model, and continues this process until all the remaining variables have *p*-values that are less than 0.05. The 30 models, now simplified with only significant predictor terms remaining, are then compared based on their adjusted R^2 value. Those models with an adjusted R^2 of greater than 0.4, and that do not violate any of the usual linear regression modeling assumptions, are chosen as the "best" ones, and in this context represent the key issues that matter most to the population as well as those with the most predictive influence. Each of the four factors X2, X4, X5, and X10, also account for a large proportion of the total variance, again indicating that these four are the key issues to the population. We get four that meet these criteria: models with X2, "U.S. Assistance to Nigeria", X4, "Social & Essential Services", X5, "Trust in Government Agencies", and X10, "Democracy", as the response variables. Since we don't want any one of the four response variables being predictor variables in one of the other four's regression equation, we re-build each of the four models, taking out the other three response variables if they were present as predictors. Our four issue stance / linear regression equations are then given by:

$$X_2 = -0.19 + 0.03X_1 + 0.38X_3 + 0.07X_6 - 0.08X_8 + 0.05X_9 + 0.09X_{14} + 0.03X_{17} + 0.12X_{18} + 0.24X_{20} + 0.08X_{21} + 0.03X_{22} + 0.02X_{23} + 0.03X_{26} + 0.05q_7$$

$$X_4 = 0.09 + 0.09X_3 - 0.04X_6 - 0.05X_{13} + 0.09X_{14} - 0.04X_{15} - 0.01X_{16} - 0.02X_{17} + 0.11X_{19} + 0.04X_{20} + 0.04X_{21} + 0.04X_{22} - 0.02X_{23} + 0.12X_{24} + 0.05X_{25} + 0.3X_{26} + 0.05d_{23} + 0.09q_6 + 0.13q_7 + 0.02q_{10}$$

$$X_5 = -0.51 - 0.03X_1 + 0.1X_3 + 0.16X_7 + 0.03X_8 + 0.16X_9 + 0.02X_{11} + 0.08X_{12} - 0.04X_{14} - 0.02X_{15} + 0.02X_{16} + 0.05X_{18} + 0.06X_{19} - 0.03X_{20} + 0.35X_{21} - 0.02X_{22} - 0.04X_{23} - 0.03X_{25} + 0.05X_{26} - 0.02q_6 + 0.07q_7 + 0.02q_{10}$$

$$X_{10} = -0.06 + 0.03X_3 + 0.09X_7 - 0.13X_8 + 0.18X_9 + 0.1X_{11} - 0.06X_{13} - 0.04X_{14} + 0.04X_{15} + 0.05X_{16} - 0.03X_{20} + 0.31X_{21} + 0.02X_{22} - 0.03X_{23} + 0.05X_{24} + 0.05X_{26} + 0.05d_{23} + 0.02q_6 + 0.2q_7 + 0.05q_{10}$$

These regression equations will now allow us to predict future issue stance scores, which will be demonstrated through a small use case in Section 4.5.3.

4.5.2 Predicting Future OABs Using Multinomial Logistic Regression

In the previous section, we showed how a linear regression model can be used to predict future issue stance scores from a given population. We now move on to the next step, predicting future observed attitudes and behaviors (OABs) using a different type a model, the multinomial logistic regression.

A simple logistic regression model can be used in situations where the response variable is dichotomous or binary, that is, the response measurement for each subject is a "success" or "failure". This model type can be modified to handle cases where the outcome variable is nominal with more than two levels (Hosmer & Lemeshow, 2000). For instance, we could employ a multinomial logistic regression if we wanted to model the choice of a meal plan from among three offered to students at a university. If the meal plans are represented by "A", "B", and "C", we could model, based on whatever predictor variables we have chosen, the probability of a student choosing one of the three meal plans as a function of those covariates. We must, however, pay attention to the scale that is used, as different methods can be employed if the scale is nominal or ordinal (Hosmer & Lemeshow, 2000). For our purposes here, we will use a nominal scale. To develop the model, assume we have *p* covariates and a constant term, denoted by the vector **x**, of length *p + 1*. Since we have three outcome variables in our meal plan

example, we will need two logit functions, and we will compare the baseline outcome, meal plan "A" (or $P(Y = 0)$), against the others. We denote the two logit functions as:

$$g_1(\mathbf{x}) = ln\left[\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p \text{ , and}$$

$$g_2(\mathbf{x}) = ln\left[\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p \text{ .}$$

Notice that there are separate parameters for each logit function, meaning that the effects vary according to the response category paired with the baseline (Agresti, 1996). The conditional probabilities of each of the three outcome variables given **x** are then:

$$P(Y=0|\mathbf{x}) = \frac{1}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}},$$

$$P(Y=1|\mathbf{x}) = \frac{e^{g_1(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}} \text{ , and}$$

$$P(Y=2|\mathbf{x}) = \frac{e^{g_2(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}}.$$

A general expression for the conditional probability in an *n* category model is:

$$P(Y=j|\mathbf{x}) = \frac{e^{g_j(\mathbf{x})}}{\sum_{k=0}^{n-1} e^{g_k(\mathbf{x})}}.$$

We can estimate the value of the parameters by first constructing a likelihood function for a sample of *n* independent observations, given by:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}[\pi_0(\boldsymbol{x}_i)^{y_{0i}}\pi_1(\boldsymbol{x}_i)^{y_{1i}}\pi_2(\boldsymbol{x}_i)^{y_{2i}}] \text{ .}$$

By taking the log of this likelihood function we get:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i}g_1(\boldsymbol{x}_i) + y_{2i}g_2(\boldsymbol{x}_i) - ln\left(1 + e^{g_1(\boldsymbol{x}_i)} + e^{g_2(\boldsymbol{x}_i)}\right) \text{ .}$$

The likelihood equations are constructed by taking the first partial derivatives of $L(\boldsymbol{\beta})$ with respect to each of the unknown parameters. The general form of these equations is:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{jk}} = \sum_{i=1}^{n} \boldsymbol{x}_{ki}\left(y_{ji} - \pi_{ji}\right).$$

The maximum likelihood estimator is then obtained by setting these likelihood equations equal to zero and solving. The solution requires the same type of iterative computation that is used in the simpler binary outcome case (Hosmer & Lemeshow, 2000). For a more detailed discussion, see *Applied Logistic Regression* by Hosmer & Lemeshow.

With now a basic understanding of the multinomial logistic regression model, we can move on to a description of the methodology that we used in order to predict future OAB scores. The goal here is to determine with what probability, after a game event occurs, the population will blame an actor for that event happening, and to see over time with a small use case that follows from section 4.5.1, how these probabilities change. As our response variable, we used question 47 of the survey described earlier in section 4.2. The question asked: "In your opinion, which of the following groups is most to blame for ongoing violence in your country today?" The response options were: "Rebel Groups", "International Terrorists", "Common Criminals", "The Military", "Government Officials", or "Foreign Countries". This particular question was chosen because it was the only one that asked about the specific actors that we felt were most relevant in an IW TWG scenario. Since we wanted a samples' issue stance score to have some influence over their OAB towards an actor, we built a multinomial logistic regression model with question 47 as the six category response variable, and the four key issues, X2, X4, X5, and X10, as the predictor variables. The *mlogit* library in the R statistical package gives us the following five logit functions, using "Rebel Groups" as the baseline, where $\mathbf{x} = \langle X_2, X_4, X_5, X_{10} \rangle$:

$$g_1(\mathbf{x}) = ln\left[\frac{P(Y = Terrorists|\mathbf{x})}{P(Y = Rebels|\mathbf{x})}\right] = -0.51 - 0.31X_2 + 0.06X_4 + 0.01X_5 + 0.23X_{10}\,,$$

$$g_2(\mathbf{x}) = ln\left[\frac{P(Y = Criminals|\mathbf{x})}{P(Y = Rebels|\mathbf{x})}\right] = 0.85 - 0.34X_2 - 0.01X_4 + 0.11X_5 + 0.08X_{10}\,,$$

$$g_3(\mathbf{x}) = ln\left[\frac{P(Y = Military|\mathbf{x})}{P(Y = Rebels|\mathbf{x})}\right] = -0.21 - 0.12X_2 + 0.06X_4 - 0.09X_5 + 0.04X_{10}\,,$$

$$g_4(\mathbf{x}) = ln\left[\frac{P(Y = Government|\mathbf{x})}{P(Y = Rebels|\mathbf{x})}\right] = 1.8 - 0.14X_2 + 0.02X_4 - 0.16X_5 - 0.2X_{10}\,,\text{and}$$

$$g_5(\mathbf{x}) = ln\left[\frac{P(Y = Foreign|\mathbf{x})}{P(Y = Rebels|\mathbf{x})}\right] = -1.2 - 0.07X_2 - 0.02X_4 - 0.01X_5 - 0.3X_{10}\,.$$

The six conditional probability models are then given as:

$$P(Y = Rebels|\mathbf{x}) = \frac{1}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})} + e^{g_3(\mathbf{x})} + e^{g_4(\mathbf{x})} + e^{g_5(\mathbf{x})}},$$

$$P(Y = Terrorists|\mathbf{x}) = \frac{e^{g_1(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})} + e^{g_3(\mathbf{x})} + e^{g_4(\mathbf{x})} + e^{g_5(\mathbf{x})}},$$

$$P(Y = Criminals|\mathbf{x}) = \frac{e^{g_2(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})} + e^{g_3(\mathbf{x})} + e^{g_4(\mathbf{x})} + e^{g_5(\mathbf{x})}},$$

$$P(Y = Military|\mathbf{x}) = \frac{e^{g_3(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})} + e^{g_3(\mathbf{x})} + e^{g_4(\mathbf{x})} + e^{g_5(\mathbf{x})}},$$

$$P(Y = Government|\mathbf{x}) = \frac{e^{g_4(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})} + e^{g_3(\mathbf{x})} + e^{g_4(\mathbf{x})} + e^{g_5(\mathbf{x})}}, \text{and}$$

$$P(Y = Foreign|\mathbf{x}) = \frac{e^{g_5(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})} + e^{g_3(\mathbf{x})} + e^{g_4(\mathbf{x})} + e^{g_5(\mathbf{x})}}.$$

These multinomial logistic regression equations will be used in our use case to determine future observed attitudes and behaviors of the population towards each actor in the proof of principle scenario that follows. In order to determine how well our models fit the data, we could subset our data into a training set as well as a test set, re-build our models on the training set, apply these to our test set, and see how well our models predict our response variable. Ideally, our test set would be next year's survey, assuming of course the same questions are asked, enabling us to determine the predictive power of our models.

4.5.3 Proof of Principle Scenario

In order to predict future issue stance scores, we would require a certain amount of subject matter expert (SME) input. That is, for each event scheduled to happen during our small use case, we would need to solicit SME input in order to determine how these would affect population views with respect to the 26 factors and 4 additional survey questions. Each of the 30 variables would get a score between -2 and 2 for each event, with -2 corresponding to a highly negative impact, -1 to a slightly negative impact, 0 to no impact, 1 to a slightly positive impact, and 2 to a highly positive impact. For our purposes in this project, as it is only a "proof of principle", SME input was notional and generated in a random fashion using an Excel

spreadsheet and input into the models from there (see Appendix D). Additionally, if we should use this methodology during an actual IW TWG, we would probably want to subset the data into different population stereotypes before building our models, and then use those models and SME input as described above for each separate stereotype. This would enable us to more effectively model the population. But again, as this was only a "proof of principle", we built one set of models for the entire population. We first need to calculate the initial issue stance score and OAB probabilities in order to instantiate our model. The initial issue stance score will result in a number between -2 and 2 (the same range as the re-scaled factor scores), and is accomplished by using the mean score for each factor as input for each of the four separate equations. The initial issue stance scores are given in Table 3.

Response Variable	**Initial Issue Stance Score**
X2. U.S. Assistance to Nigeria	0.178
X4. Social & Essential Services	0.151
X5. Trust in Government Agencies	-0.145
X10. Democracy	0.272

Table 3. Initial issue stance ccores by key issue

The initial OABs are calculated similarly, using the mean factor scores for X2, X4, X5, and X10 as inputs for our conditional six probability models. The initial OAB probabilities are given in Table 4.

Actor	**Initial OAB Probability**
Rebel Groups	0.093
International Terrorists	0.057
Common Criminals	0.206
Military	0.076
Government Officials	0.541
Foreign Countries	0.027

Table 4. Initial OAB probabilities by actor

Once our initial issue stance scores are determined, we can now use our linear regression equations in order to predict, with SME input, future scores. Given below in Figure 5 are the results of a small "proof of principle" example consisting of only 20 events with randomly generated scores, each occurring randomly over 200 time steps. These graphs show the cumulative change for each of the four issue stances over time.

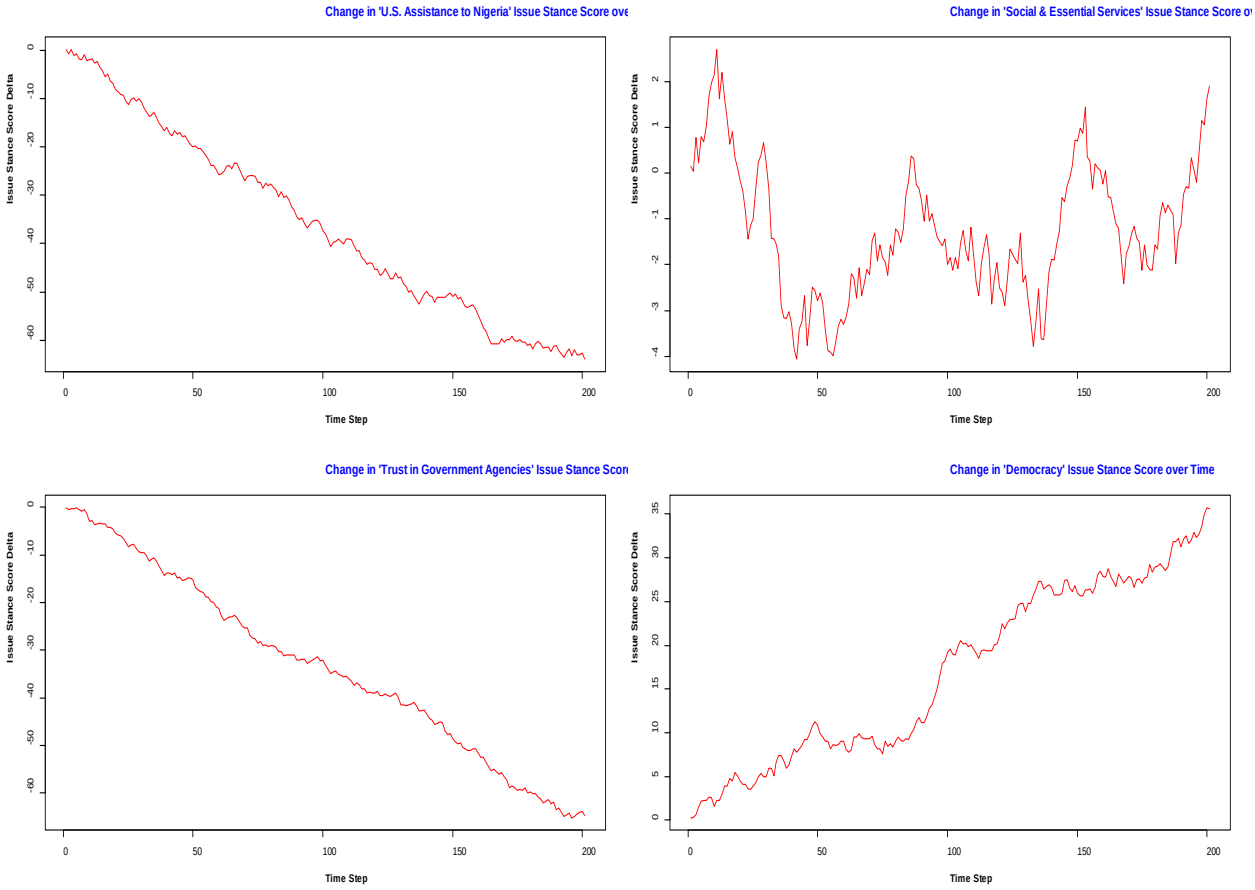


Figure 5. Cumulative issue stance score over time for the 4 key issues.

We can see from the graphs that our randomly generated events have made the population's issue stance concerning "U.S. Assistance to Nigeria" and "Trust in Government Agencies" both decrease over time, while "Social & Essential Services" and "Democracy" see an upward trend. Shown below in Figure 6 is a brief listing of events (including the first and last 25) by time step and the change in each issue stance score.

Time	Event	X2	X4	X5	X10
0	0	0.178	0.151	-0.145	0.272
1	16	-0.859	0.046	-0.482	0.331
2	2	0.131	0.79	-0.4	0.578
3	20	-1.193	0.227	-0.354	1.51
4	19	-0.859	0.793	-0.194	2.178
5	16	-1.896	0.688	-0.531	2.237
6	12	-2.005	1.035	-0.77	2.284
7	11	-0.928	1.696	-0.435	2.631
8	10	-2.203	1.983	-1.326	2.567
9	17	-2.081	2.149	-2.951	1.597
10	19	-1.747	2.715	-2.791	2.265
11	15	-2.661	1.635	-3.632	2.256
12	19	-2.327	2.201	-3.472	2.924
13	20	-3.651	1.638	-3.426	3.856
14	4	-4.224	1.194	-3.502	3.878
15	20	-5.548	0.631	-3.456	4.81
16	13	-5.018	0.906	-4.224	4.486
17	20	-6.342	0.343	-4.178	5.418
18	7	-6.854	0.125	-4.561	5.043
19	5	-8.141	-0.16	-5.498	4.452
20	7	-8.653	-0.378	-5.881	4.077
21	4	-9.226	-0.822	-5.957	4.099
22	9	-9.27	-1.43	-6.456	3.602
23	10	-10.545	-1.143	-7.347	3.538
24	8	-11.291	-0.985	-8.252	3.918
25	11	-10.214	-0.324	-7.917	4.265
176	4	-60.405	-2.004	-59.322	27.751
177	1	-60.328	-2.098	-58.959	29.19
178	3	-61.104	-2.126	-59.968	28.321
179	19	-60.77	-1.56	-59.808	28.989
180	16	-61.807	-1.665	-60.145	29.048
181	2	-60.817	-0.921	-60.063	29.295
182	13	-60.287	-0.646	-60.831	28.971
183	7	-60.799	-0.864	-61.214	28.596
184	8	-61.545	-0.706	-62.119	28.976
185	1	-61.468	-0.8	-61.756	30.415
186	1	-61.391	-0.894	-61.393	31.854
187	15	-62.305	-1.974	-62.234	31.845
188	11	-61.228	-1.313	-61.899	32.192
189	17	-61.106	-1.147	-63.524	31.222
190	18	-62.096	-0.45	-63.047	32.145
191	8	-62.842	-0.292	-63.952	32.525
192	3	-63.618	-0.32	-64.961	31.656
193	11	-62.541	0.341	-64.626	32.003
194	6	-61.831	0.084	-64.246	32.925
195	5	-63.118	-0.201	-65.183	32.334
196	11	-62.041	0.46	-64.848	32.681
197	18	-63.031	1.157	-64.371	33.604
198	1	-62.954	1.063	-64.008	35.043
199	19	-62.62	1.629	-63.848	35.711
200	10	-63.895	1.916	-64.739	35.647

Figure 6. Partial listing of cumulative issue stance changes over time

Turning our attention now to predicting future OAB probabilities using the multinomial logistic regression equations developed in the previous section, and using the same 20 events across 200 time steps as described above, we can look at how the OABs toward each actor change over time as seen in Figure 7.

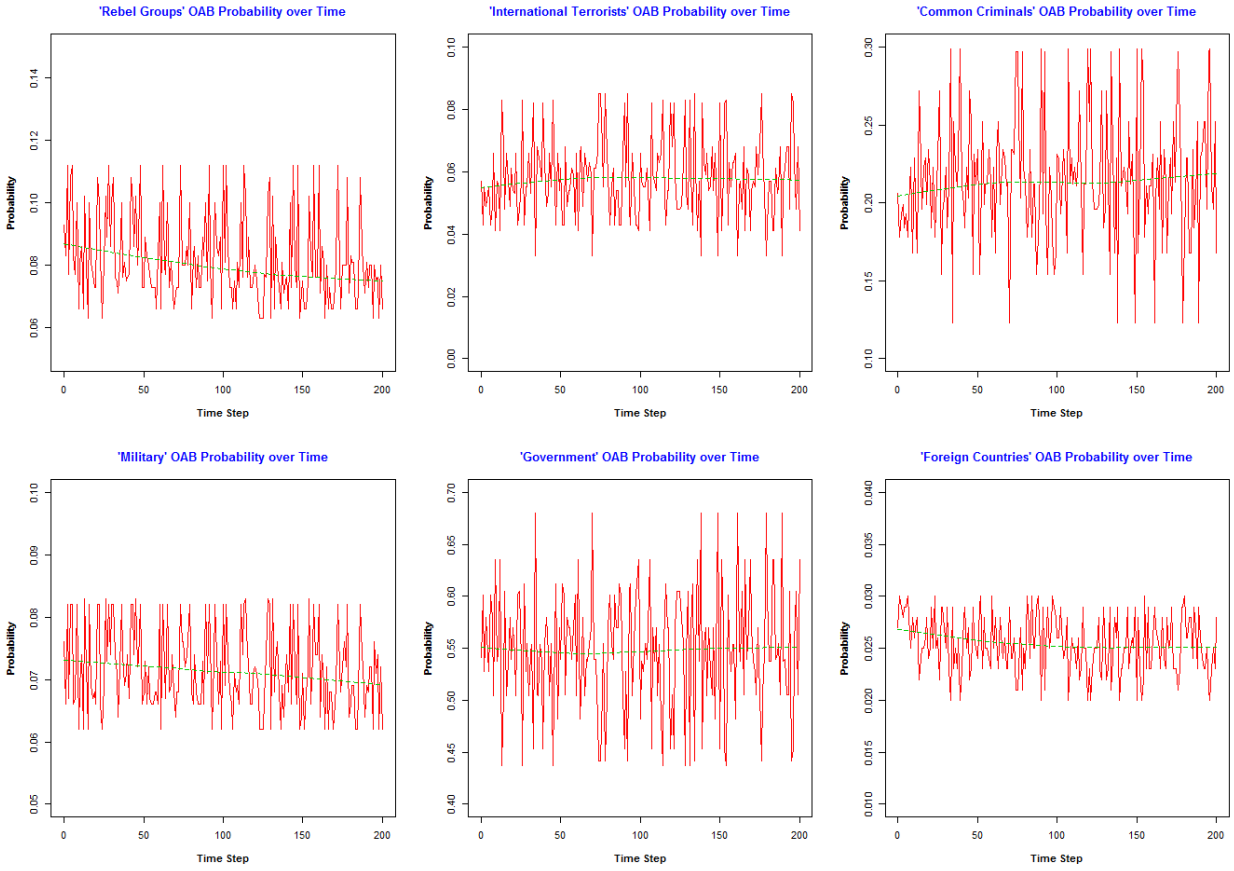


Figure 7. Observed attitude and behavior probabilities over time.

The "Government" OAB has the most variation over time, while the others tended to stay relatively close to their initial value. This is primarily due to the fact that an overwhelming majority of survey respondents had selected "Government Officials" as the primary source of blame for the ongoing violence in their country. Shown below in Figure 8 is a listing of events (including the first and last 25) by time step and the change in each OAB.

Time	Event	Rebel_Groups	International_Terrorists	Common_Criminals	Military	Government	Foreign_Countries
0	0	0.093	0.057	0.206	0.076	0.541	0.027
1	16	0.073	0.061	0.234	0.068	0.539	0.024
2	2	0.108	0.053	0.184	0.082	0.545	0.029
3	20	0.075	0.082	0.299	0.071	0.453	0.02
4	19	0.1	0.066	0.223	0.082	0.504	0.025
5	16	0.073	0.061	0.234	0.068	0.539	0.024
6	12	0.086	0.055	0.203	0.074	0.555	0.027
7	11	0.112	0.055	0.193	0.082	0.528	0.029
8	10	0.066	0.059	0.213	0.068	0.57	0.023
9	17	0.071	0.033	0.123	0.064	0.68	0.029
10	19	0.1	0.066	0.223	0.082	0.504	0.025
11	15	0.073	0.054	0.212	0.066	0.57	0.025
12	19	0.1	0.066	0.223	0.082	0.504	0.025
13	20	0.075	0.082	0.299	0.071	0.453	0.02
14	4	0.081	0.057	0.229	0.069	0.538	0.026
15	20	0.075	0.082	0.299	0.071	0.453	0.02
16	13	0.089	0.043	0.154	0.073	0.612	0.029
17	20	0.075	0.082	0.299	0.071	0.453	0.02
18	7	0.077	0.049	0.199	0.067	0.58	0.028
19	5	0.063	0.048	0.196	0.062	0.605	0.025
20	7	0.077	0.049	0.199	0.067	0.58	0.028
21	4	0.081	0.057	0.229	0.069	0.538	0.026
22	9	0.083	0.043	0.178	0.066	0.601	0.03
23	10	0.066	0.059	0.213	0.068	0.57	0.023
24	8	0.076	0.063	0.211	0.074	0.554	0.023
25	11	0.112	0.055	0.193	0.082	0.528	0.029
176	4	0.081	0.057	0.229	0.069	0.538	0.026
177	1	0.102	0.083	0.272	0.083	0.437	0.022
178	3	0.066	0.041	0.168	0.062	0.635	0.028
179	19	0.1	0.066	0.223	0.082	0.504	0.025
180	16	0.073	0.061	0.234	0.068	0.539	0.024
181	2	0.108	0.053	0.184	0.082	0.545	0.029
182	13	0.089	0.043	0.154	0.073	0.612	0.029
183	7	0.077	0.049	0.199	0.067	0.58	0.028
184	8	0.076	0.063	0.211	0.074	0.554	0.023
185	1	0.102	0.083	0.272	0.083	0.437	0.022
186	1	0.102	0.083	0.272	0.083	0.437	0.022
187	15	0.073	0.054	0.212	0.066	0.57	0.025
188	11	0.112	0.055	0.193	0.082	0.528	0.029
189	17	0.071	0.033	0.123	0.064	0.68	0.029
190	18	0.08	0.085	0.297	0.076	0.442	0.021
191	8	0.076	0.063	0.211	0.074	0.554	0.023
192	3	0.066	0.041	0.168	0.062	0.635	0.028
193	11	0.112	0.055	0.193	0.082	0.528	0.029
194	6	0.112	0.066	0.231	0.082	0.482	0.026
195	5	0.063	0.048	0.196	0.062	0.605	0.025
196	11	0.112	0.055	0.193	0.082	0.528	0.029
197	18	0.08	0.085	0.297	0.076	0.442	0.021
198	1	0.102	0.083	0.272	0.083	0.437	0.022
199	19	0.1	0.066	0.223	0.082	0.504	0.025
200	10	0.066	0.059	0.213	0.068	0.57	0.023

Figure 8. Partial listing of observed attitude and behavior probabilities over time

For both the predicted issue stance scores as well as the OAB probabilities, the event driven values in the form of a look-up table are available in Appendix D.

THIS PAGE INTENTIONALLY LEFT BLANK

SECTION 5. CONCLUSION

The assessment framework built for this project was designed with two things in mind: one, to give an analyst operating in the AFRICOM AOR a starting point or list of "good ideas" from which they could then refine based on specific locations or circumstances, and two, its broad and rather generic construct allowed the project team to gather as many data sources as possible to inform the framework, and subsequently, the end user. Finding the actual sources of data can at times be difficult. While the project team endeavored to discover as many quantitative, sources as we could, in some cases we fell short, and were only able to identify ones of a qualitative nature, and sometimes none at all. These areas were identified as gaps in our data collection efforts, and should give the user an idea of what data is and is not available. The additional functionality added to DaViTo completed during this effort allows the user to in a sense automate the assessments process, making for easier data input as well as the capability to display analytic results. Finally, this effort sought to create a methodology for building a scenario that can predict future population responses to issue stance and OABs in direct support of TRAC's IW TWG. Part of the methodology, to include the recoding/imputation and factor analysis relied heavily on work done by LCDR W. Kulzy (see References). His thesis compared factors between countries in the Sahel region of Africa, and not, as this effort did, on building predictive models, but his work was nonetheless critical to this project. Additionally, while this effort developed a scenario in Nigeria, the same methodology could be used anywhere in the world, provided good survey data is available. The recoding and imputation, factor analysis, and model building are universally applicable regardless of scenario location.

Future work along the lines of this project would undoubtedly involve the refinement of the assessment framework and further development of data sources as they become more available. The scenario methodology also has room for growth. Instead of utilizing a single set of models for the issue stances and OABs, a nice follow-on to this effort would be to develop sets of dynamic and continually updating models having the ability to adjust during a wargame according to events and the shifting views of the population. This would undoubtedly mirror reality better than a single set of static equations that do not change as the population and players do.

APPENDIX A. ASSESSMENT FRAMEWORK USER'S GUIDE

At present, the spreadsheet tool contains 158 lines of data sources and reading through the document is a time consuming endeavor. When time is of the essence, the most efficient way to use the tool is to start with a brief familiarization of topics included in the data set by looking at the Operational Environment section, and reviewing which Lines of Effort are included for the user. Next, the user should look to the "Find" tool to rapidly identify where in the spreadsheet a certain type of data can be found. If a response is found, go to that specific section of the spreadsheet and look at responses, usually found in Column G, "Data Source." In addition to the specific data source in Column G, the user of this tool should look to the left columns to determine if overarching data sources are available. When an overarching product related to many data sources, it is listed in Column A. For example, rather than listing the US Army Africa G2 has in depth Operational Environment briefs that can be accessed via http://www.intelink.sgov.gov/wiki/usaraf_G2_PMESII-JIPOE in numerous components of Column G, we have chosen to put it only once in Column A, under "Describing the Operational Environment." When a source appears on Column A, it should consider this as an overarching source with general applicability to numerous lines in Column G-Data Source.

Similarly, as the user moves to the right of the spreadsheet, in the Area and Indicator column, there are similar websites/texts that have general applicability to multiple sections of Column G-Data Source. If there is an annotation in Area and Indicator section of a data source, the tool user should assume that this source has diverse applicability to numerous components of Column G. In summary, general sources appear to the left of the spreadsheet, while specific data points appear to the right.

In some cases, there are conceptual similarities and similar data sources for certain issues. In these cases, Column D simply refers the user to a separate section of the spreadsheet, rather than copying the entire list of data sources. For example, line 21 of the tool requires data on Political / religious / tribal motivations by group and line 22 indicates a need for Size of the Groups. The data sources found for both sections were the same, thus the user is referred to the line above, rather than copying all again.

Finally, the user of the tool should exercise caution in evaluating the data sources. Ultimately, data should be pulled from numerous data points and not a single source. For example, in the event that a researcher is looking the issue of civilian casualties in an African Conflict, it would be of utility to look at the text that appears which describes the issue of Civilian Casualties generally, then look to the Africa landing page of the International Committee of the Red Cross for high profile cases, drill down to the country office to collect additional environmental, and finally look to advocacy organization listed for additional data. Keep in mind that the list of data sources is not rank ordered. If one book is listed second, it does not mean that the first listed book is considered to be “better.”

APPENDIX B. NIGERIA FOCUSED DATA COLLECTION EFFORTS

Background:

Nigeria is one of the countries in the Sahel region of Africa. For the purposes of this document, we have taken "Sahel" to include Nigeria and also Benin, Burkina Faso, Chad, Côte d'Ivoire, Ghana, Guinea, Guinea-Bisseau, Mali, Mauritania, Niger, Senegal, Sierra Leone, and Togo, plus Nigeria's neighbor of Cameroon.

Administratively, Nigeria is divided into 36 states plus the Federal Capital Territory (FCT). We will include the FCT when we refer to "states." States are divided into "local government areas," or LGAs. There are 787 LGAs in Nigeria, and it may be worth noting that some names are duplicated across states (there are also variations in spelling from source to source).

Population by LGA:

We extracted the population data from a table we found at this link. The link is currently broken, but the file, named "Vol 03 Table DSx LGAPop by SDistrict-PDF.pdf," is available here. Each of the 37 population tables was highlighted in the PDF file and copied to the clipboard, and then the R function `pdf.grabber()`(see Appendix D) was used to produce an R object reflecting the contents of that table. Each table gives the LGA name, the land area of the LGA, and the population in that LGA (by gender, and in total). Observe that the "District Number" field as provided is not always sequential. The state-specific tables were combined into one object with 787 rows, which can be found in `state.lga.csv`. We ignored the "Disputed Areas" counts in Taraba State (p. 32 of the PDF file), which accounted for 10,106 men and 9,858 women. Therefore our grand total is smaller than the one in the PDF by these numbers. Although we do have population data by LGA, most data available from major organizations like the United Nations and the CIA is at the country level only.

United Nations Data:

The United Nations has "summary data" for each member country, at pages with addresses like, for example, http://data.un.org/CountryProfile.aspx ?crName=CAMEROON. For

each country, the data between the header “Summary Statistics” and the “trade profile” was highlighted and copied to the clipboard; then the R function `handle.un.stats()`(see Appendix D) was used to produce a country-specific data set. In addition to Nigeria, we did this for each of the Sahel countries. The resulting combined data set is provided as `un.data.xlsx`.

CIA Word Factbook:

The CIA World Factbook is available for download at this link. The directory called “fields” holds a set of files, each file holding all the values of one “fact” for all countries. Our function `fact.grabber()`grabs a single fact for a single country; `fact.looper()`(see Appendix D) grabs all the facts for a single country; the script file `cia.script.R` (see Appendix D) runs the loop to extract all facts for all relevant countries. The full set of facts for the Sahel countries plus Nigeria is found in `Ciafactsfull.xls`. Many times the facts contains several years’ worth of data, separated by semi-colons. A second output file, `Ciafacts.xls`, contains only the most recent data for any fact (this workbook has two sheets. The second, “Xposed,” presents the data transposed, so that there is one row per country and one column per fact).

World Bank Data:

The World Bank’s web site contains a substantial amount of data, organized by 18 broad topic areas like Agriculture and Rural Development, Education, and Poverty. Each topic is addressed by a number of indicators (though some indicators pertain to more than one topic). There are 318 distinct indicators; the list of indicators organized by topic can be seen at this link. Our `WorldBank.R` (see Appendix D) script serves to read all of these indicators directly from the web (these steps require the use of R’s `RCurl` and `XML` packages to read in and decode the data, and `RODBC` for an easy way to write them out as separate worksheets in an Excel workbook). This script requires the `wb.format.one()`function (see Appendix D), which writes a particular indicator to disc or returns it (some tables, however, are empty and those are skipped). Many entries are missing. Most indicators stop at the year 2010; in order to ensure that all output has the same number of columns we have omitted 2011 data when it appears. The data set is seen as `WB.datadump.xls`. Note that, because of operator error, the set of countries

included here omits Sierra Leone but includes Liberia, and each worksheet starts with two column headers.

World Bank ADI:

The World Bank also makes available its Africa Development Indicator (ADI) data. This can be downloaded from the "Databank" link in the top-right of this page. This has a similar flavor to the data above but it extends back as far as 1960 and includes 2,409 indicators. We found the spreadsheet exported from the databank to seem some manipulation, which was performed by the `ADI.handler.R` script (see Appendix D). The data set resulting from this analysis is found here.

African Development Bank Group: The ADBG's Data Query page can be found at this link. This data is organized into 16 topics, which at the site are called "Indicators," but we will use "indicators" as in "World Bank Data" above to refer to the individual values. There are 708 unique indicators in this data. Because the topics are of quite different sizes, we have downloaded this data into a group containing topics 1-6, one containing topics 7-12, one with topic 13 and another with topic 14. The workbooks containing this data can be found in this directory.

Piracy Data:

We extracted piracy data from the ICC Commercial Crimer Services database. An example report can be found at this site and others can be retrieved by varying the last three digits of that address. Our `pirate.grabber()` function (see Appendix D) extracts one such report; `pirate.looper()` (see Appendix D) extracts a set; and the `Piracy.R` script (see Appendix D) assembles these into a database, extracting only those for which the reported country was one of Ghana, Benin, Togo, Nigeria, Cameroon, or Equitorial Guinea. The data set is located here.

Other "Official" Data Sources:

The United Nations Economic Commission for Africa (UNECA) maintains a substantial library of publications at this site. These include the African Statistical Yearbook, the 2011 version of which can also be found here. The UN also maintains its Data Mart which has a large

and perhaps slightly unfocused collection of data some of which is detailed at the national level. The Nigerian National Bureau of Statistics website is in the process of being upgraded. The "data portal" holds great promise for extracting official government data by state. The National Population Commission site seems to be not quite as useful. Some data on energy production can be found at the U.S. Energy Information Administration's web site. The Nigeria page can be found at this link.

Other Unofficial Data Sources:

Detailed trade information can be found at the "African Growth and Opportunity Act" website maintained by the Trade Law Centre for Southern Africa. Nigeria's page can be found at this link. This website hosted at Columbia University delivers poverty and food security data for many countries. However, the selection for Nigeria is limited. The Cleen foundation reports what it says are crime statistics for Nigeria (see this link), but no source appears to be given.

Shape Files:

Nigeria's site at maplibrary.org (at this page) contains shape files giving a satellite image of Nigeria, and the polygons that make up the states and LGAs. Some general notes on sources are available. In particular the satellite imagery is said to have come from NASA. The website for the GIS software DIVA-GIS (found at this page) provides many shape files of different sorts. The software's focus is on natural and biological resources and we have downloaded plausible-looking shape files giving water areas, rivers, and roads. The sources for the shape files are rarely revealed, however.

APPENDIX C. R CODE FOR CAPTURING / DOWNLOADING DATA SOURCES

```
> pdf.grabber
function () {

# This function turns some data copied from the document at
# http://population.gov.ng/images/stories/Vol%2003%20Table%20DSx%20LGAPop%20by%20SDistrict-PDF.pdf into a table. Copy the state-specific table to the clipboard and this formats the table.

# The key delimiters are integers that name the LGAs (in at least one
# case the compilers of the table skipped an integer, so we use length()
# rather than max()) and senatorial districts (apparently, each states
# gets three senators, except for Abuja FCT, which gets one).

# Between the row number and the senator is the name of the district,
# which can have multiple words. The four columns after the senator give
# the land size, male pop'n, female pop'n, and total.

# Read in the data; remove everything before the first "1". If there
# are zero "1", or two or more, stop.

str <- scan ("clipboard", what="")
one <- str == "1"
if (sum (one) != 1) stop (paste ("Ones in positions", which (one)))
str <- str[-(1:(which(one) - 1))]

# Kill the commas, because at least one state has 'em.

str <- gsub (",", "", str)

# You're a row number if you're numeric and you're < 50 (because no state
# has more than 44 LGAs!) and you're an integer.

nums <- as.numeric (str)
rownums <- which (!is.na (nums) & nums < 50 & nums - trunc(nums) < .00001)

# In Akwa Ibom, rows 234 and 256 should be 23 and 25.

rownums <- sort (c(rownums, which ( (nums == 234 | nums == 256) & nums - trunc(nums) < .00001)))

senator <- which (is.element (str, c("A", "B", "C")))

# Construct the "out" matrix; put row numbers and senators in.

out <- matrix ("", length (nums[rownums]), 7)
out[,1] <- nums[rownums]
out[,3] <- str[senator]

# For each row, paste the name parts together and move the
# populations into "out".

for (i in 1:nrow(out)) {
if ((rownums[i] + 1) == (senator[i] - 1))
    out[i,2] <- str[rownums[i] + 1]
else
    out[i,2] <- paste (str[(rownums[i] + 1):(senator[i] - 1)], collapse=" ")
out[i, 4:7] <- str[(senator[i] + 1):(senator[i] +4)]
}

# The result is a matrix. Make it a data frame, converting those last four columns to numeric.

out <- as.data.frame (out, stringsAsFactors=FALSE)
names (out) <- c("Dist", "Name", "Sen", "Land", "Male", "Female", "Total")
out[,4:7] <- matrix (as.numeric (unlist (out[,4:7])), ncol=4)
return (out)
```

```
}
# HANDLE.UN.STATS

> handle.un.stats
function ()
{

# Read data in. Ignore quotes, or you'll be abused by Cote d'Ivoire!

a <- scan ("clipboard", what="", sep="\t", quote=NULL)
a <- a[a != "Summary statistics" & a != "Economic indicators"]
a <- a[a != "Environment" & a != "Social indicators"]
a <- a[a != "Top"]
a
}

# FACT.GRABBER

> fact.grabber
function (item = "2089.html", country)
{
fname <- paste ("fields/", item, sep="")
txt <- scan (fname, what="", sep="\n", quiet=T)
hdr <- txt[grep ("::", txt)][1]

# The header is between two colons and a ">". There should only be one
# "::", but there were two, for example, in 2033.html for Venezuela.
# So take the first instance.

colons <- regexpr ("::", hdr)
hdr <- substring (hdr, colons  + 3) # allow for a space
lt <- regexpr ("<", hdr)
hdr <- substring (hdr, 1, lt - 1)
death <- readHTMLTable (fname)[[country]]
if (length (death) == 0)
   d2 <- "Missing"
else
    d2 <- as.character(death[2,2])
return (c(hdr, d2))
}

# FACT.LOOPER

> fact.looper
function (country = "ni")
{

# Grab all the CIA facts for one country. The relevant files have names that start with "2."

f <- list.files ("fields", pattern="^2")

out <- matrix ("", length (f), 3)
out[,1] <- f
for (i in 1:length(f)) {
    out[i, 2:3] <- fact.grabber (f[i], country)
}
out
}

# cia.script.R

# CIA Fact Book stuff

# We dump and unzip the CIA fact book.

library (XML)
fact.grabber (2003, "ni")

# extracts the particular report about Nigeria in document <...> field/2003.html.
```

```
fact.looper (, "ni")

# gets (by default) all the facts about Nigeria.

countries <- c("Nigeria", "Mauritania", "Senegal", "Guinea-Bissau",
"Guinea", "Sierra Leone", "Liberia", "Cote d'Ivoire", "Mali",
"Burkina Faso", "Ghana", "Togo", "Benin", "Niger", "Chad", "Cameroon")
countries <- cbind (countries, c("ni", "mr", "sg", "pu", "gv",
"sl", "li", "iv", "ml", "uv", "gh", "to", "bn", "ng", "cd", "cm"))

# "Countries" is 16x2, each row giving a name and a two-letter identifier
# for one of the countries of the Sahel (plus Cameroon).
# So we get the whole set like this:
for (i in 2:16) {
assign (countries[i,2], fact.looper (countries[i,2]), pos = 1, immediate=TRUE)
}

# Ensure all the first columns match

for (i in 2:16) print (all (ni[,1] == get (countries[i,2])[,1])
)

# Now let's assemble them. The embedded new-line characters turn out to get in the way
# (who saw that coming?). So let's change them to, I don't know, semi-colons for the moment.
# Also kill any spaces that precede or follow new lines.

cia <- eval (parse (text = paste ("data.frame (", paste (countries[,2], "[,3,drop=F]",
collapse=","), ", stringsAsFactors=FALSE)")))
cia <- data.frame (ni[,2], cia, stringsAsFactors=FALSE)
names (cia) <- make.names (c("Measure", countries[,1]))
for (i in 1:ncol (cia))
{

# The outer ()+ thing means "match lazily, that is, keep matching until you come to an end
#(instead of stopping as soon as you see a match." The inside part describes space(s), new-line,
# space(s). Finding the maximal set of those, turn them into one semi-colon. The asterisks in
#front of the [[:blank:]] part says "zero or more blanks."

# match zero or more blanks, new line, zero or more blanks; change all of those to ;
    cia[,i] <- gsub ("(*[[:blank:]]\n*[[:blank:]])+", ";", cia[,i])
}

write.table (cia, "../cia.tsv", sep="\t", row.names=FALSE, col.names=TRUE, quote=FALSE)
#
# This got turned into Ciafactsfull.xls.
#
# Here we remove everything in any entry starting at the first semi-colon.
for (i in 1:ncol (cia))
{
    semicolon <- regexpr (";", cia[,i])
    cia[semicolon > 1,i] <- substring (cia[semicolon > 1,i], 1, semicolon[semicolon > 1] - 1)
}
write.table (cia, "../cia2.tsv", sep="\t", row.names=FALSE, col.names=TRUE, quote=FALSE)

# This got turned into Ciafacts.xls. Page 2 has the transposed version, which might be easier to
read.

# WORLDBANK.R

# Grab World Bank stuff

# The World Bank Indicators are held here:

# http://data.worldbank.org/indicator

wb.indic.out <- getURI ("http://data.worldbank.org/indicator")

# This item has 18 tables, each with a set of indicators.
```

```
wb.indicators <- vector ("list", 18)
for (i in 1:18) {
    this.set <- readHTMLTable (wb.indic.out)[[i]]
    this.set <- as.character(unlist (this.set))
    this.set <- this.set[this.set != ""]
    index <- character (length (this.set))
    for (j in 1:length (this.set)) {
        start <- regexpr (this.set[j], wb.indic.out, fixed=TRUE)
        mychars <- substring (wb.indic.out, start - 100, start)
        mychars <- substring (mychars, regexpr ("indicator/", mychars) + nchar ("indicator/"))
        index[j] <- substring (mychars, 1, regexpr ("\"", mychars) - 1)
    }
    wb.indicators[[i]]   <-   data.frame   (Description   =   this.set,   Indicator   =   index,
stringsAsFactors=FALSE)
}

# Some weird one we fix by hand

wb.indicators[[3]][18,2] <- "NY.GNP.PCAP.CD"
wb.indicators[[4]][15,2] <- "NY.GNP.PCAP.CD"
wb.indicators[[4]][38,2] <- "BX.TRF.PWKR.CD.DT"
wb.indicators[[8]][22,2] <- "CM.MKT.INDX.ZG"
wb.indicators[[8]][27,2] <- "BX.TRF.PWKR.CD.DT"
wb.indicators[[16]][6,2] <- "SP.POP.SCIE.RD.P6"
wb.indicators[[16]][10,2] <- "SP.POP.TECH.RD.P6"

names (wb.indicators) <- c("Agriculture & Rural Development", "Aid Effectiveness",
"Climate Change", "Economic Policy &  External Debt", "Education",
"Energy & Mining", "Environment", "Financial Sector", "Gender", "Health",
"Infrastructure", "Labor & Social Protection", "Poverty", "Private Sector",
"Public Sector", "Science & Technology", "Social Development",
"Urban Development")

# For each indicator, hit the web and extract table 1.

for (i in 13:18) {
    cat ("We're at the top, and i is ", i, "\n")
    newlist <- vector ("list", nrow (wb.indicators[[i]]))
    names (newlist) <- wb.indicators[[i]][,2]
    for (j in 1:nrow (wb.indicators[[i]])) {
        cat ("We're in the loop, and j is ", j, "\n")
        wb.tbl.out      <-      getURI      (paste      ("http://data.worldbank.org/indicator/",
wb.indicators[[i]][j,2], sep=""))

# Issues: "page could not be found" and no columns with dates (that is, "2" in the #name).
#
if (regexpr ("page could not be found", wb.tbl.out) <= 0) {
     wb.countries <- readHTMLTable (wb.tbl.out, stringsAsFactors=FALSE)[[1]]
      if (!all (regexpr ("2", wb.countries[,-1]) < 0)) {

# Clean up column names, just because. Some have nothing in them, so in particular, they don't
have a 2 or a "C."

                names (wb.countries) <- gsub ("\n", "", names (wb.countries))
                wb.countries <- wb.countries[,regexpr ("C", names(wb.countries)) > 0 | regexpr
("0", names (wb.countries)) > 0]
                names (wb.countries) <- gsub ("[[:space:]]+$", "", names (wb.countries))
                names (wb.countries) <- gsub ("^[[:space:]]+", "", names (wb.countries))
                names (wb.countries)[1] <- "Country.name" # for neatness
                result   <-   wb.countries[is.element   (wb.countries$Country.name,   c("Nigeria",
"Benin", "Burkina Faso",
                "Cameroon", "Chad",  "Cote d'Ivoire",
                "Ghana",  "Guinea",  "Guinea-Bisseau",  "Liberia",  "Mali",  "Mauritania",  "Niger",
"Senegal", "Togo") ),]
                newlist[[j]] <- result
                }
             else
                 cat ("Sadly, that table was empty.\n")
            }
```

```
        else
            cat ("Sadly, nothing was found there.\n")
    }
    assign (paste ("wb.indic.", ifelse (i < 10, "0", ""), i, sep = ""), newlist)
}

# Now that they're all built, let's write them out.

for (i in 1:length (wb.indicators))
    for (j in 1:nrow (wb.indicators[[i]]))
        wb.format.one (i, j)

# Now check this out
library (RODBC)
myod <- odbcConnectExcel2007 ("q:/africom/Africa EDA/WorldBank/WB.Datadump.xls",
                 readOnly=FALSE)
for (i in 1:length (wb.indicators)) {
    for (j in 1:nrow (wb.indicators[[i]])) {
        tbl <- wb.format.one (i, j, write.out=F)
        if (is.null (tbl) || all (tbl[-1,-1] == ""))
            cat ("Skipping empty table ", i, j, "\n")
        else {
            cat ("About to try to write table ", i, j, "\n")
            sqlSave (myod, tbl, tablename = wb.indicators.shortnms[i],
                rownames=FALSE, colnames=TRUE, safer=FALSE, fast=FALSE, append=T)
# insert blank line
            tbl <- as.data.frame (matrix ("", 1, 5))
            sqlSave (myod, tbl, tablename = wb.indicators.shortnms[i],
                rownames=FALSE, colnames=FALSE, safer=FALSE, fast=FALSE, append=T)

        }
    }
}
odbcCloseAll ()

# wb.format.one

> wb.format.one
function (topic, indic, write.out=TRUE, dir = "q:/africom/africa eda/WorldBank/")
{
     fname <- paste ("wb.topic.", ifelse (topic <= 9, "0", ""), topic, ".tsv", sep="")
     fname <- paste (dir, fname, sep="")
     str <- paste ("wb.indic.", ifelse (topic <= 9, "0", ""), topic, "[[", indic, "]]", sep="")
     tbl <- eval (parse (text = str))
     topic.name <- names (wb.indicators)[topic]
     indic.name <- wb.indicators[[topic]][indic,1]
     indic.short <- wb.indicators[[topic]][indic,2]
     if (write.out == FALSE) {

# For this purpose we require exactly the columns "Country",
# "2007", "2008", "2009", and "2010." If some of these don't exist,
# create them. We do this by creating a full-size item like this
# and keeping the relevant columns. If 2011 exists, delete it.

         if (is.null (tbl))
             return (tbl)
         if (any (names (tbl) == "2011"))
             tbl <- tbl[,names(tbl) != "2011"]
         bigtbl <- as.data.frame (matrix ("", nrow(tbl), ncol=5), stringsAsFactors=FALSE)
         names(bigtbl) <- c("Country.name", "2007", "2008", "2009", "2010")
         bigtbl[,match (names(tbl), names(bigtbl))] <- tbl
         hdr.1 <- c(topic.name, "", "", indic.short, "")
         hdr.2 <- c(indic.name, "", "", "", "")
         bigtbl <- rbind (hdr.1, hdr.2, bigtbl)
         return (bigtbl)
     }
     if (indic == 1) append <- FALSE else append <- TRUE
##   cat (paste (topic.name, "\t\t", indic.short, "\t\n", sep=""))
     cat (paste (indic.name, "\t\t\t\t\n", sep=""), file = fname,
```

```
          append=ifelse (indic == 1, FALSE, TRUE))
     cat (paste (indic.short, "\t\t\t\t\n", sep=""), file = fname, append=T)
     write.table (tbl, file = fname, append=TRUE, sep="\t",
                  col.names=TRUE, row.names=FALSE, quote=FALSE)
     cat ("\n", file = fname, append=TRUE)
}

# ADI handler


# Handle the funky ADI data. Highlight A1:BC84062.

adi <- scan ("clipboard", what="", sep="\t", quote=NULL)
adi <- adi[adi != "\""]
adi.nms <- adi[1:55]
adi <- as.data.frame (matrix (adi[-(1:55)], nrow = 84061, ncol=55, byrow=T))
names (adi) <- make.names (adi.nms)
# They spell "Guinea-Bissau" without the second "e"
keepers <- c("Nigeria", "Benin", "Burkina Faso", "Chad", "Cote d'Ivoire",
"Ghana", "Guinea", "Guinea-Bissau", "Mali", "Mauritania", "Niger", "Senegal",
"Sierra Leone", "Togo", "Cameroon")

# Save those and write out.

adi <- adi[is.element (adi$Country.Name, keepers),]
write.table (adi, "Q:\\Africom\\Africa EDA\\DeliverToTom\\adi.csv",
sep="\t", row.names=FALSE)

# Pirate Grabber

> pirate.grabber
function (input = 130)
{
base <- "http://www.icc-ccs.org/piracy-reporting-centre/live-piracy-report/details/116/"
uri <- paste (base, input, sep="")
pirates <- getURI (uri)
cat (pirates, file = "h:/temp/killme.txt")
p2 <- scan ("h:/temp/killme.txt", sep="\n", what="", quote=NULL)
p3 <- gsub ("<.*?>", "", p2) # remove HTML between brackets
p3 <- gsub ("\t", "", p3) # remove tabs
p3 <- gsub("^ .", "", p3) # remove leading white space
p3 <- gsub(" +$", "", p3) # remove trailing white space
p3 <- p3[p3 != ""]
p3 <- p3[-(1:which (p3 == "IMB Live Piracy Map 2012")[3])]
p3 <- p3[p3 != "Ports Anchorages"] # that just gets in the way!

# Here we fix some of their broken (colon-less) stuff.

p3[p3 == "CPA in nm"] <- "CPA in nm:"
p3[p3 == "crew threatened"] <- "Crew threatened:"
p3[p3 == "Date Hijacked Vessel Released"] <- "Date Hijacked Vessel Released:"

p3 <- p3[1: (grep ("WNI", p3) - 1)]
p3 <- p3[-(1:grep (":", p3)[1] - 1)]
p3.colons <- grep (":", p3)

# Everything between "Narrations:" and "Date Hijacked Vessel Released"
# constitutes narration.

narr.start <- which (regexpr ("Narrations:", p3) > 0) + 1
narr.end <- which (regexpr ("Date Hijacked Vessel Released:", p3) > 0) - 1
date.released <- narr.end + 2
p3 <- p3[1:(date.released + 3)] # cut off everything else
narration <- paste (p3[narr.start:narr.end], collapse="\n")

# Don't include narrations or date of release in the "nms" part.

p3.colons <- p3.colons [!is.element (p3.colons, narr.start:narr.end)]
p3.colons <- p3.colons[!is.element (p3.colons, date.released)]
```

```
# Some are missing, so here's a weak plan. Go down the
# vector. If there's a colon, it's a header, else it's a value.

nms <- p3[p3.colons]
vals <- character (length (nms))
ctr <- 0
for (i in 1:length (p3)) {
    if (i %in% p3.colons)
        ctr <- ctr + 1
    else {
        if (i == narr.start)
            vals[ctr] <- narration
        else
            vals[ctr] <- p3[i]
    }
}
return (cbind (nms, vals))
}

# Pirate looper

function ()
{

# 173 incidents as of 5/22/2012. Let's grab them all and put their output into a list.

keepsters <- vector ("list", 173)
for (i in 1:173) {
    cat ("Let's try number", i, "\n")
    keepsters[[i]] <- pirate.grabber (i)
}
keepsters
}

# Piracy.R

the.big.out <- pirate.looper ()

# Some of these have 48 rows, some have 50.

country.grabber <- function (x) {
    x[x[,1] == "Country:",2]
}
where <- sapply (the.big.out, country.grabber)
the.big.out <- the.big.out[which (is.element (where,
c("Ghana", "Benin", "Togo", "Nigeria","Cameroon", "Equitorial Guinea")))]

# All of these have 50 rows...

sapply (the.big.out, nrow)

# ... and they all have the same entries in the first column

sapply (the.big.out, function (x) all (x[,1] == the.big.out[[1]][,1]))

# So put all the second columns together...

pirates <- sapply (the.big.out, function(x) x[,2])

# Transpose, so that the attacks are in rows...
#
pirates <- t(pirates)

# ...and add labels. These may have colons and/or trailing spaces.

colname <- the.big.out[[1]][,1]
colname <- sub (" +$", "", colname)
colname <- sub (":", "", colname)
colname <- sub (" +$", "", colname)
```

```
dimnames(pirates) <- list (NULL, colname)
pirates <- as.data.frame (pirates, stringsAsFactors=FALSE)

# Fix lat and long. These are four columns each.

lat <- pirates[,grep ("LAT",  names (pirates))]
newlat <- apply (lat, 1, function (x) {
    xx <- as.numeric (x[1:3])
    amt <- xx[1] + xx[2]/60 + xx[3] / 3600
    return (ifelse (x[4] == "N", amt, -amt))
})
pirates[, grep ("LAT", names (pirates))[1]] <- newlat # replace the first
pirates <- pirates[, -grep ("LAT", names (pirates))[-1]] # delete the others
names (pirates)[grep ("LAT", names (pirates))] <- "LAT"

long <- pirates[,grep ("LONG",  names (pirates))]
newlong <- apply (long, 1, function (x) {
    xx <- as.numeric (x[1:3])
    amt <- xx[1] + xx[2]/60 + xx[3] / 3600
    return (ifelse (x[4] == "E", amt, -amt))
})
pirates[, grep ("LONG", names (pirates))[1]] <- newlong # replace the first
pirates <- pirates[, -grep ("LONG", names (pirates))[-1]] # delete the others
names (pirates)[grep ("LONG", names (pirates))] <- "LONG"

rm (lat, long, newlat, newlong, colname)


write.table (pirates, "clipboard", sep="\t", quote=FALSE, col.names=TRUE, row.names=FALSE)
```

APPENDIX D. SME INPUT & LOOK-UP TABLE

Event	X1	X3	X6	X7	X8	X9	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	Safety	Goals	Services	Equality
1	-2	0	-1	0	-2	1	1	0	-2	1	1	0	-2	2	-1	-1	2	1	-2	0	0	-2	1	0	1	-1
2	0	2	2	0	0	0	-2	2	-2	-1	2	1	2	-2	0	2	1	-1	-1	0	0	2	-1	2	-1	0
3	0	-1	1	-2	2	-1	2	2	-1	0	-1	0	0	0	0	0	0	1	0	0	1	0	0	1	-2	0
4	0	-1	1	-2	2	1	0	2	1	1	0	-1	2	-2	1	0	2	1	0	-1	-2	-2	0	2	0	0
5	2	-2	-2	2	-2	-2	-2	2	1	-2	0	1	-2	0	-2	0	0	0	1	2	2	1	0	-2	-2	-2
6	0	2	2	-1	-1	1	2	0	-1	-2	1	0	1	1	-2	-1	2	-2	1	0	0	1	0	-1	-1	-2
7	0	-1	0	2	0	2	-2	0	0	-2	-1	1	2	2	2	0	-2	0	0	-1	2	0	-2	-1	0	2
8	-2	-2	-1	0	-1	-1	0	-2	2	2	0	1	0	-2	0	1	1	0	1	2	2	-2	1	0	2	0
9	2	0	-2	0	0	-1	-1	2	2	0	1	1	0	1	1	1	0	-1	0	1	-1	-2	-1	-2	0	1
10	0	-2	-1	0	1	0	-1	0	0	-2	0	-1	1	-2	0	1	-1	1	-2	1	-1	0	2	0	2	0
11	-1	2	0	0	0	0	0	2	0	2	2	0	0	2	2	0	1	0	0	0	-1	0	0	1	0	0
12	1	1	2	1	-1	0	-2	0	-2	-2	-2	-1	1	-1	0	-1	0	-1	0	1	0	1	0	0	-1	2
13	-2	2	2	-1	1	-2	-2	-2	0	0	2	0	2	-1	2	0	0	1	0	1	-2	-1	0	0	2	-1
14	-2	0	-1	1	0	1	0	0	-2	-2	2	0	-2	-2	-1	0	1	-2	0	2	1	0	1	0	-2	-1
15	2	-2	1	2	0	2	1	-2	2	-1	2	2	1	-2	-2	1	-1	-2	1	-1	0	0	-1	-1	0	-2
16	0	-1	0	0	0	-2	-2	-1	0	0	0	2	-2	-2	2	-1	2	2	0	0	0	-1	0	-2	0	0
17	-2	0	1	1	-1	-2	0	-1	0	-2	-1	-2	-1	1	-1	2	-2	2	1	0	2	0	0	2	0	-1
18	1	-2	2	2	0	0	1	0	1	0	0	0	0	1	0	-2	2	-2	0	2	0	2	0	2	0	-1
19	0	2	-1	-2	-2	1	-2	1	-2	-1	-1	0	1	0	1	-2	2	0	2	2	-1	0	1	1	-1	-2
20	2	-2	1	1	-2	0	-2	-1	0	-2	-2	1	2	-1	0	-2	2	1	-2	2	-1	-2	0	0	0	2

Figure 9. Notional SME input values for each factor by event

Event	X2	X4	X5	X10	Rebels	Terrorists	Criminals	Military	Government	Foreign
1	0.077	-0.094	0.363	1.439	0.102	0.083	0.272	0.083	0.437	0.022
2	0.99	0.744	0.082	0.247	0.108	0.053	0.184	0.082	0.545	0.029
3	-0.776	-0.028	-1.009	-0.869	0.066	0.041	0.168	0.062	0.635	0.028
4	-0.573	-0.444	-0.076	0.022	0.081	0.057	0.229	0.069	0.538	0.026
5	-1.287	-0.285	-0.937	-0.591	0.063	0.048	0.196	0.062	0.605	0.025
6	0.71	-0.257	0.38	0.922	0.112	0.066	0.231	0.082	0.482	0.026
7	-0.512	-0.218	-0.383	-0.375	0.077	0.049	0.199	0.067	0.58	0.028
8	-0.746	0.158	-0.905	0.38	0.076	0.063	0.211	0.074	0.554	0.023
9	-0.044	-0.608	-0.499	-0.497	0.083	0.043	0.178	0.066	0.601	0.03
10	-1.275	0.287	-0.891	-0.064	0.066	0.059	0.213	0.068	0.57	0.023
11	1.077	0.661	0.335	0.347	0.112	0.055	0.193	0.082	0.528	0.029
12	-0.109	0.347	-0.239	0.047	0.086	0.055	0.203	0.074	0.555	0.027
13	0.53	0.275	-0.768	-0.324	0.089	0.043	0.154	0.073	0.612	0.029
14	-0.804	-0.069	-0.057	0.446	0.08	0.068	0.252	0.072	0.505	0.023
15	-0.914	-1.08	-0.841	-0.009	0.073	0.054	0.212	0.066	0.57	0.025
16	-1.037	-0.105	-0.337	0.059	0.073	0.061	0.234	0.068	0.539	0.024
17	0.122	0.166	-1.625	-0.97	0.071	0.033	0.123	0.064	0.68	0.029
18	-0.99	0.697	0.477	0.923	0.08	0.085	0.297	0.076	0.442	0.021
19	0.334	0.566	0.16	0.668	0.1	0.066	0.223	0.082	0.504	0.025
20	-1.324	-0.563	0.046	0.932	0.075	0.082	0.299	0.071	0.453	0.02

Figure 10. Look-up table for issue stance and OAB by event

APPENDIX E. R CODE FOR FACTOR ANALYSIS AND REGRESSION MODELS

There are five distinct pieces of R code that follow, one each for the recoding/imputation of the data, the factor analysis, recoding the response variables, building the models, and a script that will manipulate the data as well as generate plots for the use case implementation.

I. Data Recode and Imputation

```
## Script for recoding and imputing the 2010 Sahel (Nigeria) Survey Data
## This program will output 3 files:
## 1. A recoded data set according to the recode functions listed below
## 2. A recoded and imputed data set using hot decking
## 3. A final data set with only the variables (questions) necessary for factor analysis

## Read in .sav file
library(foreign)

nig10 <- read.spss("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Nigeria_2010_weights.sav",use.value.labels=TRUE,max.value.labels=Inf, data.frame=TRUE)

# Replace ":" in d17 with "-"
nig10$d17 <- gsub(":","-",nig10$d17)

# Delete BLANK1-BLANK15,"hh_1-7_1-12","reas1-12","length_int",and "sexagewgt"
data <- nig10[,-
c(11,18:101,151,196,199,228,230,235,237,240,242,265,288,333,345,347,358:369,370,375)]

## Recoding scheme based off of number of points in Likert Scale (-2 to 2, "Don't Know = 0")

library(car) # package for recoding

# 1. Two point questions where "Yes" is most positive (Don't Know = 0)
recodeTwoPos <- function(x){
       recode(x,
              '"Yes"= 2;
              "No"= -2;
              "The government serves the interests of all people equally" = 2;
              "The government favors certain groups over others"= -2;
              "Does not affect the amount of crime in society"= 2;
              "Reduces crime in society"= -2;
              "Promotes harsh criminal punishments"= 2;
              "Promotes fair criminal punishments"= -2;
              "Does not affect the amount of corruption in society"= 2;
              "Reduces corruption in society"= -2;
              "Denies women rights"= 2;
              "Denies women\'s rights"= 2;
              "Protects women"= -2;
              "Protects women\'s rights"= -2;
              "Does not treat women as equals to men"= 2;
              "Does not treat women  as equals to men"= 2;
              "Treats women as equals to men"= -2;
              "Treats women as equals to men"= -2;
              "Non-Muslims in Nigeria should be free to worship in their own way"= 2;
              "Non-Muslims in [COUNTRY] should be free to worship in their own way."= 2;
              "Non-Muslims in Nigeria should not be free to worship in their own way"= -2;
              "Non-Muslims in [COUNTRY] should not be able to worship in their own way"= -2;
       "Islam teaches people to deal with non-believers with cooperation and understanding"= 2;
       "Islam teaches people to deal with nonbelievers with cooperation and understanding"= 2;
       "Islam teaches people to deal with non-believers with confrontation and struggle"= -2;
       "Islam teaches people to deal with nonbelievers with confrontation and struggle."= -2;
              "Non-Muslim and Muslim cultures can peacefully exist together" = 2;
```

```
                "Non-Muslim and Muslim cultures can peacefully exist together."= 2;
                "War between Non-Muslim and Muslim cultures is inevitable"= -2;
                "War between Non-Muslim and Muslim cultures is inevitable."= -2;
                "European/American culture is not a threat to traditional Muslim values"= 2;
                "European/American culture is not a threat to traditional Muslim values."= 2;
                "European/American culture is a threat to traditional Muslim values"= -2;
                "European/American culture is a threat to traditional Muslim values."= -2;
                "Muslims who live in France are free to practice Islam"= 2;
                "Muslims who live in France are free to practice Islam."= 2;
                "Muslims who live in France cannot freely practice Islam"= -2;
                "Muslims who live in France cannot freely practice Islam."= -2;
                "Muslims who live in the United States are free to practice Islam"= 2;
                "Muslims who live in the United States of America are free to practice Islam."= 2;
                "Muslims who live in the United States cannot freely practice Islam"= -2;
        "Muslims who live in the United States of America cannot freely practice Islam."= -2;
                "Muslims are treated fairly in the world today"= 2;
                "Muslims are treated fairly in the world today."= 2;
                "Muslims are being oppressed in the world today"= -2;
                "Muslims are being oppressed in the world today."= -2;
"The office of the president should be held by the person most capable regardless of their
regional origin"= 2;
"The office of the president should be alternately held by a notherner and a southerner"= -2;
                "Marabouts sending young boys into the street is a form of exploitation."= 2;
"Marabouts sending young boys into the street is a necessary part of their religious education."=
-2;
                "Don\'t know" = 0;
                "Don\'t Know" = 0;
                "Dont know"= 0;
                "DK"= 0;
                "No answer"= NA;
                "No response"= NA;
                "No repsonse"= NA;
                "No Response"= NA;
                "NR"= NA; ',
                        as.factor.result=FALSE)
}

# 2. Two point questions where "Yes" is most positive (Don't Know = NA)
recodeTwoPos1 <- function(x){
        recode(x,
                '"Yes"= 2;
                "No"= -2;
                "The government serves the interests of all people equally" = 2;
                "The government favors certain groups over others"= -2;
                "Does not affect the amount of crime in society"= 2;
                "Reduces crime in society"= -2;
                "Promotes harsh criminal punishments"= 2;
                "Promotes fair criminal punishments"= -2;
                "Does not affect the amount of corruption in society"= 2;
                "Reduces corruption in society"= -2;
                "Denies women rights"= 2;
                "Denies women\'s rights"= 2;
                "Protects women"= -2;
                "Protects women\'s rights"= -2;
                "Does not treat women as equals to men"= 2;
                "Does not treat women  as equals to men"= 2;
                "Treats women as equals to men"= -2;
                "Treats women as equals to men"= -2;
                "Non-Muslims in Nigeria should be free to worship in their own way"= 2;
                "Non-Muslims in [COUNTRY] should be free to worship in their own way."= 2;
                "Non-Muslims in Nigeria should not be free to worship in their own way"= -2;
                "Non-Muslims in [COUNTRY] should not be able to worship in their own way"= -2;
        "Islam teaches people to deal with non-believers with cooperation and understanding"= 2;
        "Islam teaches people to deal with nonbelievers with cooperation and understanding"= 2;
        "Islam teaches people to deal with non-believers with confrontation and struggle"= -2;
        "Islam teaches people to deal with nonbelievers with confrontation and struggle."= -2;
                "Non-Muslim and Muslim cultures can peacefully exist together" = 2;
                "Non-Muslim and Muslim cultures can peacefully exist together."= 2;
                "War between Non-Muslim and Muslim cultures is inevitable"= -2;
```

```
                "War between Non-Muslim and Muslim cultures is inevitable."= -2;
                "European/American culture is not a threat to traditional Muslim values"= 2;
                "European/American culture is not a threat to traditional Muslim values."= 2;
                "European/American culture is a threat to traditional Muslim values"= -2;
                "European/American culture is a threat to traditional Muslim values."= -2;
                "Muslims who live in France are free to practice Islam"= 2;
                "Muslims who live in France are free to practice Islam."= 2;
                "Muslims who live in France cannot freely practice Islam"= -2;
                "Muslims who live in France cannot freely practice Islam."= -2;
                "Muslims who live in the United States are free to practice Islam"= 2;
                "Muslims who live in the United States of America are free to practice Islam."= 2;
                "Muslims who live in the United States cannot freely practice Islam"= -2;
        "Muslims who live in the United States of America cannot freely practice Islam."= -2;
                "Muslims are treated fairly in the world today"= 2;
                "Muslims are treated fairly in the world today."= 2;
                "Muslims are being oppressed in the world today"= -2;
                "Muslims are being oppressed in the world today."= -2;
"The office of the president should be held by the person most capable regardless of their
regional origin"= 2;
"The office of the president should be alternately held by a notherner and a southerner"= -2;
                "Marabouts sending young boys into the street is a form of exploitation."= 2;
"Marabouts sending young boys into the street is a necessary part of their religious education."=
-2;
                "Don\'t know" = NA;
                "Don\'t Know" = NA;
                "Dont know"= NA;
                "DK"= NA;
                "No answer"= NA;
                "No response"= NA;
                "No repsonse"= NA;
                "No Response"= NA;
                "NR"= NA; ',
                        as.factor.result=FALSE)
}

# 3. Two point questions where "Yes" is most negative ("No" and "Oppose" is positive, Don't Know
= 0)
recodeTwoNeg <- function(x){
        recode(x,
                '"Yes"= -2;
                "No"= 2;
                "Oppose"= 2;
                "Support"= -2;
                "Justified"= -2;
                "Not Justified"= 2;
                "Don\'t know" = 0;
                "Don\'t Know" = 0;
                "Dont know"= 0;
                "DK"= 0;
                "No answer"= NA;
                "No response"= NA;
                "No Response"= NA;
                "No repsonse"= NA;
                "NR"= NA; ',
                        as.factor.result=FALSE)
}

# 4. Two point questions where "Yes" is most negative ("No" and "Oppose" is positive, Don't Know
= NA)
recodeTwoNeg1<- function(x){
        recode(x,
                '"Yes"= -2;
                "No"= 2;
                "Oppose"= 2;
                "Support"= -2;
                "Justified"= -2;
                "Not Justified"= 2;
                "Don\'t know" = NA;
                "Don\'t Know" = NA;
```

```
                "Dont know"= NA;
                "DK"= NA;
                "No answer"= NA;
                "No response"= NA;
                "No Response"= NA;
                "No repsonse"= NA;
                "NR"= NA; ',
                        as.factor.result=FALSE)
}

# 5. Three point questions where "Most" is preferred (positive)
recodeThreePos <- function(x){
        recode(x,
                '"Improved"= 2;
                "Stayed the same"=0;
                "Gotten worse"=-2;
                "Worsened"=-2;
                "Government and religion should be kept separate" = 2;
                "Government and religion should be kept separate."= 2;
"Our country should remain a secular democracy, but religion should play a greater role in govt"=
0;
"Our country should remain a secular democracy, but religion should play a greater role in
government."=0;
                "Our country should be governed by religious leaders"= -2;
                "Our country should be governed by religious leaders." = -2;
                "Our country should be governed by civil law" =2;
                "Our country should be governed by civil law." =2;
                "Our country should be gvoerned by a combination of civil and religious law"= 0;
                "Our country should be governed by a combination of civil and religious law."=0;
                "Religious laws should govern all spheres of life"= -2;
                "Jihad is an inward personal and moral struggle"= 2;
                "Jihad is both"= 0;
                "Jihad is taking up arms against enemies of Islam"= -2;
                "The U.S is engaged to fight terrorism"= 2;
                "The U.S. is engaged to fight terrorism"= 2;
                "Both"= 0;
                "None"= 0;
                "The U.S. is engaged to fight Islam"= -2;
                "Dont know"= NA;
                "No response"= NA;
                "No repsonse"= NA;
                "No answer"= NA;
                "Dont know"= NA;
                "Don\'t know"= NA;
                "Don\'t Know"= NA;
                "No Response"= NA;
                "No response"= NA;
                "No Response"= NA;
                "No answer"= NA;',
                        as.factor.result=FALSE)
}

# 6. Three point questions where "Most" is least preferred (negative)
recodeThreeNeg <- function(x){
        recode(x,
                '"Positive influence"= -2;
                "Neutral influence"= 0;
                "Negative influence"= 2;
                "No influence"= 0;
                "Don\'t know"= NA;
                "Dont Know"= NA;
                "No Response"= NA;
                "No repsonse"= NA;
                "No response"= NA; ',
                        as.factor.result=FALSE)
}

# 7. Four point questions where "Most" is preferred (positive)
recodeFourPos <- function(x){
```

```
recode(x,
        '"Never"=-2;
        "Several times a year"= -1;
        "Several times a month" = 1;
        "Several times a week"= 2;
        "Very safe"= 2;
        "Fairly safe"= 1;
        "Not very safe"= -1;
        "Not safe at all"= -2;
        "Very satisified"= 2;
        "Very satisfied"= 2;
        "Somewhat satisfied"= 1;
        "Not very satisfied"= -1;
        "Somewhat frustrated"= -1;
        "Not at all satisfied"= -2;
        "Very frustrated"= -2;
        "Nigeria is not a democracy"= -2;
        "Very favorable"= 2;
        "Somewhat favorable"= 1;
        "Somewhat unfavorable"= -1;
        "Very unfavorable"= -2;
        "Very similar"= 2;
        "Somewhat similar"= 1;
        "Only a little similar"= -1;
        "Not similar at all"= -2;
        "A lot"= 2;
        "A Lot"= 2;
        "A fair amount"= 1;
        "A Fair amount"= 1;
        "Fair amount"= 1;
        "A little"= -1;
        "Not at all"= -2;
        "No trust at all"= -2;
        "A lot of confidence"= 2;
        "A fair amount of confidence"= 1;
        "Only little confidence"= -1;
        "Only a little confidence"= -1;
        "No confidence at all"= -2;
        "Very stable"= 2;
        "Somewhat stable"= 1;
        "Somewhat fragile"= -1;
        "Very fragile"= -2;
        "Strongly agree"= 2;
        "Somewhat agree"= 1;
        "Somewhat disagree"= -1;
        "Strongly disagree"= -2;
        "Strongly approve"= 2;
        "Somewhat approve"= 1;
        "Somewhat disapprove"= -1;
        "Strongly disapprove"= -2;
        "Very good"= 2;
        "Somewhat good"= 1;
        "Somewhat poor"= -1;
        "Very poor"= -2;
        "Very good"= 2;
        "Good"= 1;
        "Fair"= -1;
        "Poor"= -2;
        "Often"= 2;
        "Sometimes"= 1;
        "Rarely"= -1;
        "Never"= -2;
        "Very easy"= 2;
        "Somewhat easy"= 1;
        "Somewhat hard"= -1;
        "Very hard"= -2;
        "Very often"= 2;
        "Fairly often"= 1;
        "Not very often"= -1;
```

```
                "Not at all"= -2;
                "Very important"= 2;
                "Fairly important"= 1;
                "Not very important"= -1;
                "Not at all important"= -2;
                "Very responsive"= 2;
                "Somewhat responsive"= 1;
                "Not very responsive"= -1;
                "Not at all responsive"= -2;
                "[COUNTRY]is not a democracy"= NA;
                "NA"= NA;
                "DK"= 0;
                "NR"= NA;
                "No answer"= NA;
                "Don\'t know"= 0;
                "Don\'t Know"= 0;
                "Dont know"= 0;
                "No response"= NA;
                "No response"= NA;
                "No repsonse"= NA;
                "No repsonse"= NA;
                "No Response"= NA; ',
                       as.factor.result=FALSE)
}

# 8. Four point questions where "Most" is least preferred (negative)
recodeFourNeg <- function(x){
        recode(x,
                '"Always justified"= -2;
                "Sometimes justified"= -1;
                "Rarely justified"= 1;
                "Never justified"= 2;
                "Strongly agree"= -2;
                "Somewhat agree"= -1;
                "Somewhat disagree"= 1;
                "Strongly disagree"= 2;
                "Often"= -2;
                "Sometimes"= -1;
                "Rarely"= 1;
                "Never"= 2;
                "DK"= 0;
                "NR"= NA;
                "No answer"= NA;
                "Don\'t know"= 0;
                "Don\'t Know"= 0;
                "Dont know"= 0;
                "No response"= NA;
                "No repsonse"= NA;
                "No Response"= NA; ',
                       as.factor.result=FALSE)
}

# 9. Five point questions where "Most" is preferred (positive)
recodeFivePos <- function(x){
        recode(x,
                '"Always"= 2;
                "Most of every day"= 1;
                "Only a few hours a day"= 0;
                "Only a few hours a week"= -1;
                "Never"= -2;
                "Upper- Plenty of disposable money"= 2;
                "Upper middle- Able to purchase most essential goods"= 1;
                "Lower middle- Able to meet basic needs with some non-essential goods"= -1;
                "Poor- Able to meet basic needs"= -1;
                "Very poor- Unable to meet basic needs without charity"= -2;
                "Plenty of disposable money"= 2;
                "Able to purchase most non-essential goods"= 1;
                "Able to meet basic needs with some non-essential goods"= 0;
                "Able to meet basic needs" = -1;
```

```
                "Unable to meet basic needs without charity"= -2;
                "DK"= NA;
                "NR"= NA;
                "No answer"= NA;
                "Don\'t know"= NA;
                "Don\'t Know"= NA;
                "Dont know"= NA;
                "No response"= NA;
                "No repsonse"= NA;
                "No Response"= NA; ',
                        as.factor.result=FALSE)
}

# 10. Five point questions where "Most" is the least preferred (negative)
recodeFiveNeg <- function(x){
        recode(x,
                '"Increased a lot"= -2;
                "Increased a little"= -1;
                "Stayed the same"= 0;
                "Decreased a little"= 1;
                "Decrease a lot"= 2;
                "Increased dramatically"= -2;
                "Increased slightly" = -1;
                "Stayed the same"= 0;
                "Decreased slightly"= 1;
                "DK"= NA;
                "NR"= NA;
                "No answer"= NA;
                "Don\'t know"= NA;
                "Don\'t Know"= NA;
                "Dont know"= NA;
                "No response"= NA;
                "No repsonse"= NA;
                "No Response"= NA; ',
                        as.factor.result=FALSE)
}

# These are recoded for imputation purposes as the Match.var variable. Others may be included.
recodeDem <- function(x){
        recode(x,
                '"Christianity"= 1;
                "Christianity (Catholic, Protestant, Evangelical, etc)"= 1;
                "Islam"= -1;
                "Traditional"= 0;
                "No religion"= 0;
                "Others"= 0;
                "Other"= 0;
                "Judaism"= 0;
                "Animism"= 0;
                "Missing" = 0;
                "No Response"= 0;
                "No response"= 0;
                "Don\'t know"= 0;
                "Male"= 1;
                "Female"= -1;
                "Rural"= 1;
                "Urban"= -1; ',
                        as.factor.result=FALSE)
}

# Recode Question 47 for model building purposes
recodeQ47 <- function(x){
        recode(x,
                '"Rebel groups"= 0;
                "International terrorists"= 1;
                "Common criminals"= 2;
                "The military"= 3;
                "Government officials"= 4;
                "Foreign country"= 5;
```

```
                "Other"= NA;
                "Don\'t know"= NA;
                "No Response"= NA; ',
                       as.factor.result=FALSE)
}

# Link each question to specific recode functions and recode

data$urbanrural <- as.numeric(recodeDem(data$urbanrural))
data$q5 <- as.numeric(recodeFourPos(data$q5))
data$q6 <- as.numeric(recodeFourPos(data$q6))
data$q7 <- as.numeric(recodeFourPos(data$q7))
data$q8edu <- as.numeric(recodeFourPos(data$q8edu))
data$q8hea <- as.numeric(recodeFourPos(data$q8hea))
data$q8wat <- as.numeric(recodeFourPos(data$q8wat))
data$q8roa <- as.numeric(recodeFourPos(data$q8roa))
data$q8ele <- as.numeric(recodeFourPos(data$q8ele))
data$q9edu <- as.numeric(recodeThreePos(data$q9edu))
data$q9hea <- as.numeric(recodeThreePos(data$q9hea))
data$q9wat <- as.numeric(recodeThreePos(data$q9wat))
data$q9roa <- as.numeric(recodeThreePos(data$q9roa))
data$q9ele <- as.numeric(recodeThreePos(data$q9ele))
data$q10 <- as.numeric(recodeTwoPos(data$q10))
data$q12uk <- as.numeric(recodeFourPos(data$q12uk))
data$q12fr <- as.numeric(recodeFourPos(data$q12fr))
data$q12ni <- as.numeric(recodeFourPos(data$q12ni))
data$q12ir <- as.numeric(recodeFourPos(data$q12ir))
data$q12ch <- as.numeric(recodeFourPos(data$q12ch))
data$q14usa <- as.numeric(recodeFourPos(data$q14usa))
data$q16so <- as.numeric(recodeFourPos(data$q16so))
data$q16li <- as.numeric(recodeFourPos(data$q16li))
data$q16sa <- as.numeric(recodeFourPos(data$q16sa))
data$q17sa <- as.numeric(recodeFourPos(data$q17sa))
data$q17fr <- as.numeric(recodeFourPos(data$q17fr))
data$q17ch <- as.numeric(recodeFourPos(data$q17ch))
data$q17ir <- as.numeric(recodeFourPos(data$q17ir))
data$q17us <- as.numeric(recodeFourPos(data$q17us))
data$q21a <- as.numeric(recodeFourPos(data$q21a))
data$q21b <- as.numeric(recodeFourPos(data$q21b))
data$q21c <- as.numeric(recodeFourPos(data$q21c))
data$q21d <- as.numeric(recodeFourPos(data$q21d))
data$q21e <- as.numeric(recodeFourPos(data$q21e))
data$q21f <- as.numeric(recodeFourPos(data$q21f))
data$q21g <- as.numeric(recodeFourPos(data$q21g))
data$q21h <- as.numeric(recodeFourPos(data$q21h))
data$q22a <- as.numeric(recodeFourPos(data$q22a))
data$q22b <- as.numeric(recodeFourPos(data$q22b))
data$q22c <- as.numeric(recodeFourPos(data$q22c))
data$q22d <- as.numeric(recodeFourPos(data$q22d))
data$q22e <- as.numeric(recodeFourPos(data$q22e))
data$q22f <- as.numeric(recodeFourPos(data$q22f))
data$q22g <- as.numeric(recodeFourPos(data$q22g))
data$q22h <- as.numeric(recodeFourPos(data$q22h))
data$q23a <- as.numeric(recodeFourPos(data$q23a))
data$q23b <- as.numeric(recodeFourPos(data$q23b))
data$q23c <- as.numeric(recodeFourPos(data$q23c))
data$q23d <- as.numeric(recodeFourPos(data$q23d))
data$q23e <- as.numeric(recodeFourPos(data$q23e))
data$q23f <- as.numeric(recodeFourPos(data$q23f))
data$q25a <- as.numeric(recodeFourNeg(data$q25a))
data$q25b <- as.numeric(recodeFourNeg(data$q25b))
data$q25c <- as.numeric(recodeFourNeg(data$q25c))
data$d5a <- as.numeric(recodeDem(data$d5a))
data$q26a <- as.numeric(recodeFourPos(data$q26a))
data$q26b <- as.numeric(recodeFourPos(data$q26b))
data$q26c <- as.numeric(recodeFourPos(data$q26c))
data$q26d <- as.numeric(recodeFourPos(data$q26d))
data$q26e <- as.numeric(recodeFourPos(data$q26e))
data$q27a <- as.numeric(recodeTwoPos1(data$q27a))
```

```
data$q27b <- as.numeric(recodeTwoPos1(data$q27b))
data$q27c <- as.numeric(recodeTwoPos1(data$q27c))
data$q27d <- as.numeric(recodeTwoPos1(data$q27d))
data$q28 <- as.numeric(recodeThreePos(data$q28))
data$q29a <- as.numeric(recodeTwoPos(data$q29a))
data$q29b <- as.numeric(recodeTwoPos(data$q29b))
data$q29c <- as.numeric(recodeTwoPos(data$q29c))
data$q29d <- as.numeric(recodeTwoPos(data$q29d))
data$q30 <- as.numeric(recodeThreePos(data$q30))
data$q31a <- as.numeric(recodeThreePos(data$q31a))
data$q31b <- as.numeric(recodeThreePos(data$q31b))
data$q32a <- as.numeric(recodeTwoPos(data$q32a))
data$q32b <- as.numeric(recodeTwoPos(data$q32b))
data$q32c <- as.numeric(recodeTwoPos(data$q32c))
data$q32d <- as.numeric(recodeTwoPos(data$q32d))
data$q32e <- as.numeric(recodeTwoPos(data$q32e))
data$q33 <- as.numeric(recodeTwoNeg(data$q33))
data$q34a <- as.numeric(recodeTwoPos(data$q34a))
data$q34b <- as.numeric(recodeTwoPos(data$q34b))
data$q36a <- as.numeric(recodeFourNeg(data$q36a))
data$q36b <- as.numeric(recodeFourNeg(data$q36b))
data$q37a <- as.numeric(recodeTwoPos(data$q37a))
data$q37b <- as.numeric(recodeTwoPos(data$q37b))
data$q37c <- as.numeric(recodeTwoPos(data$q37c))
data$q37d <- as.numeric(recodeTwoPos(data$q37d))
data$q37e <- as.numeric(recodeTwoPos(data$q37e))
data$q40 <- as.numeric(recodeFourPos(data$q40))
data$q41a <- as.numeric(recodeTwoPos1(data$q41a))# Don't know=0 here because it is not an opinion
data$q42 <- as.numeric(recodeFourPos(data$q42))
data$q44 <- as.numeric(recodeFourPos(data$q44))
data$q44na <- as.numeric(recodeTwoPos(data$q44na))
data$q45 <- as.numeric(recodeFourPos(data$q45))
data$q47 <- as.numeric(recodeQ47(data$q47))
data$q48a <- as.numeric(recodeFourPos(data$q48a))
data$q48b <- as.numeric(recodeFourPos(data$q48b))
data$q48c <- as.numeric(recodeFourPos(data$q48c))
data$q48d <- as.numeric(recodeFourPos(data$q48d))
data$q48e <- as.numeric(recodeFourPos(data$q48e))
data$q48f <- as.numeric(recodeFourPos(data$q48f))
data$q49pr <- as.numeric(recodeFourPos(data$q49pr))
data$q49pm <- as.numeric(recodeFourPos(data$q49pm))
data$q49na <- as.numeric(recodeFourPos(data$q49na))
data$q49pp <- as.numeric(recodeFourPos(data$q49pp))
data$q49af <- as.numeric(recodeFourPos(data$q49af))
data$q49cj <- as.numeric(recodeFourPos(data$q49cj))
data$q49rl <- as.numeric(recodeFourPos(data$q49rl))
data$q49lp <- as.numeric(recodeFourPos(data$q49lp))
data$q49lg <- as.numeric(recodeFourPos(data$q49lg))
data$q50 <- as.numeric(recodeFourPos(data$q50))
data$q52 <- as.numeric(recodeFourPos(data$q52))
data$q56b <- as.numeric(recodeThreePos(data$q56b))
data$q57 <- as.numeric(recodeTwoNeg(data$q57))
data$q58a <- as.numeric(recodeTwoNeg(data$q58a))
data$q58b <- as.numeric(recodeTwoNeg(data$q58b))
data$q58c <- as.numeric(recodeTwoNeg(data$q58c))
data$q59a <- as.numeric(recodeFourPos(data$q59a))
data$q59b <- as.numeric(recodeFourNeg(data$q59b))
data$q59c <- as.numeric(recodeFourNeg(data$q59c))
data$q59d <- as.numeric(recodeFourPos(data$q59d))
data$q60 <- as.numeric(recodeThreePos(data$q60))
data$q62a <- as.numeric(recodeFourPos(data$q62a)) #It’s opinions; can't determine a pos or neg
data$q62b <- as.numeric(recodeFourPos(data$q62b))
data$q62c <- as.numeric(recodeFourPos(data$q62c))
data$q62d <- as.numeric(recodeFourPos(data$q62d))
data$d0 <- as.numeric(recodeDem(data$d0))
data$d13 <- as.numeric(recodeFourPos(data$d13))  # conditional question
data$d15 <- as.numeric(recodeThreePos(data$d15))
data$d16 <- as.numeric(recodeFiveNeg(data$d16))
data$d17 <- as.numeric(recodeFivePos(data$d17))
```

```
data$d21 <- as.numeric(recodeFivePos(data$d21))
data$d22 <- as.numeric(recodeTwoPos1(data$d22)) # Don't Know = NA. No cell phone?
data$d23 <- as.numeric(recodeFourPos(data$d23))
data$d24a <- as.numeric(recodeTwoPos1(data$d24a)) # Don't Know = NA
data$d24b <- as.numeric(recodeTwoPos1(data$d24b))
data$d24c <- as.numeric(recodeTwoPos1(data$d24c))
data$d24d <- as.numeric(recodeTwoPos1(data$d24d))
data$d24e <- as.numeric(recodeTwoNeg1(data$d24e))
data$d26a <- as.numeric(recodeFourPos(data$d26a))
data$d26b <- as.numeric(recodeFourPos(data$d26b))
data$d30 <- as.numeric(recodeFourPos(data$d30))

write.table(data,"C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Recode_10.csv",sep=",",col.names=TRUE,row.names=FALSE,quote=TRUE,na="NA")


## Imputation Using Hotdeck Method

library(StatMatch)

imputeHD <- function(Question,Dframe,Donor.Class,Match.vars){

Data.rec <- Dframe[is.na(Dframe[,Question])==TRUE,]
Data.rec <- subset(Data.rec,select=-get(Question))

Data.don <- Dframe[is.na(Dframe[,Question])==FALSE,]

imp.RAND <- RANDwNND.hotdeck(data.rec=Data.rec,data.don=Data.don,match.vars=Match.vars,
               don.class=Donor.Class,dist.fun="Manhattan")

Data.rec.imp <-
create.fused(data.rec=Data.rec,data.don=Data.don,mtc.ids=imp.RAND$mtc.ids,z.vars=Question)

final <- rbind(Data.rec.imp,Data.don)
return(final)
}

HD.loop <- function (Dframe, Donor.Class, Match.vars, Question) {
       empty <- "False"
       while (empty == "False"){
              final <- imputeHD (Question[1], Dframe, Donor.Class, Match.vars)
              Question <- Question[-1]  # remove that question FIFO
              Dframe <- final  # update Dframe with new data
              if (length(Question) < 1){
                     empty <- "True"
              }
       }
       final
}

Match.vars <- c("d5a","d0","urbanrural")
data$state <- as.factor(data$state) # state must be a factor
Donor.Class <- c("state") #state is the donor class
Dframe <- data
Question <- c("q5","q6","q7","q8edu", "q8hea", "q8wat", "q8roa", "q8ele", "q9edu", "q9hea",
"q9wat","q9roa","q9ele", "q10", "q12uk", "q12fr", "q12ni", "q12ir", "q12ch", "q14usa",
"q16so", "q16li", "q16sa", "q17sa", "q17fr", "q17ch", "q17ir", "q17us", "q21a", "q21b", "q21c",
"q21d", "q21e", "q21f", "q21g", "q21h", "q22a","q22b", "q22c", "q22d", "q22e", "q22f", "q22g",
"q22h","q23a", "q23b", "q23c", "q23d", "q23e", "q23f", "q25a", "q25b", "q25c", "q26a","q26b",
"q26c","q26d", "q26e", "q27a", "q27b", "q27c", "q27d", "q28", "q29a", "q29b", "q29c", "q29d",
"q30","q31a", "q31b", "q32a", "q32b", "q32c", "q32d", "q32e", "q33", "q34a", "q34b", "q36a",
"q36b","q37a", "q37b", "q37c", "q37d", "q40", "q41a", "q42", "q44", "q44na", "q45", "q47",
"q48a", "q48b", "q48c","q48d", "q48e", "q48f", "q49pr", "q49pm", "q49na", "q49pp", "q49af",
"q49cj","q49rl", "q49lp", "q49lg","q50", "q52", "q56b", "q57", "q58a", "q58b", "q58c", "q59a",
"q59b", "q59c", "q59d", "q60", "q62a","q62b", "q62c", "q62d", "d13", "d15", "d16", "d17", "d21",
"d22", "d23", "d24a", "d24b", "d24c", "d24d","d24e", "d26a", "d26b", "d30")

rec.imp.data <- HD.loop(Dframe,Donor.Class,Match.vars,Question)
```

```
write.table(rec.imp.data,"C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Rec_Imp_10.csv",sep=",",col.names=TRUE,row.names=FALSE,quote=TRUE,na="NA")

# Delete all variables except those we want to create factors with (taken from the Questions)
final.data <- rec.imp.data[,c("q5","q6","q7","q8edu", "q8hea", "q8wat", "q8roa", "q8ele",
"q9edu", "q9hea","q9wat","q9roa","q9ele", "q10", "q12uk", "q12fr", "q12ni", "q12ir", "q12ch",
"q14usa","q16so", "q16li", "q16sa", "q17sa", "q17fr", "q17ch", "q17ir", "q17us", "q21a", "q21b",
"q21c","q21d", "q21e", "q21f", "q21g", "q21h", "q22a","q22b", "q22c", "q22d", "q22e", "q22f",
"q22g", "q22h","q23a", "q23b", "q23c", "q23d", "q23e", "q23f", "q25a", "q25b", "q25c",
"q26a","q26b", "q26c","q26d", "q26e", "q27a", "q27b", "q27c", "q27d", "q28", "q29a", "q29b",
"q29c", "q29d", "q30","q31a", "q31b", "q32a", "q32b", "q32c", "q32d", "q32e", "q33", "q34a",
"q34b", "q36a", "q36b","q37a", "q37b", "q37c", "q37d", "q40", "q41a", "q42", "q44", "q44na",
"q45", "q48a", "q48b", "q48c","q48d", "q48e", "q48f", "q49pr", "q49pm", "q49na", "q49pp",
"q49af", "q49cj","q49rl", "q49lp", "q49lg","q50", "q52", "q56b", "q57", "q58a", "q58b", "q58c",
"q59a", "q59b", "q59c", "q59d", "q60", "q62a","q62b", "q62c", "q62d", "d13", "d15", "d16", "d17",
"d21", "d22", "d23", "d24a", "d24b", "d24c", "d24d","d24e", "d26a", "d26b", "d30")]

# Check to see if there are any missing values remaining
for (i in 1:ncol(final.data)) {
       check <- sum(is.na(final.data[,i]))
       # show(check)
}
sum(check)

write.table(final.data,"C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Final_10.csv",sep=",",col.names=TRUE,row.names=FALSE,quote=TRUE,na="NA")
```

II. Factor Analysis

```
## Script for conducting Factor Analysis on the 2010 Sahel (Nigeria) Survey Data
# Function finds optimal number of factors, forms a matrix of the factor loadings as the output.
# Prints out the optimal number of factors used based off of eigenvalues.
# Prints out the factor matrix with loadings > 0.4 or < -0.4.
# Prints out the variable names by factor as well as the factor names.
# Prints the % of variance the factor will explain via eigenvalues.
# Modifies the loading matrix by deleting factors that are n/a.
# Calculates the matrix of factor scores.
# Scales the factor score matrix appropriately to values between -2 and 2.

final.data <- read.csv("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Final_10.csv")

factorNames <- c("1. Sharia Law", "2. U.S. Assist to Nigeria", "3. China Assist to Nigeria", "4.
Social & Essential Services","5. Trust in Gov Agencies", "6. External Security", "7. General
Trust", "8. Non-West Countries", "9. Local and National Freedom","10. Democracy", "11. Others
Values", "12. Daily Life Acceptance", "13. Use of Violence","14. Terrorism Enablers", "15. Family
and Friends", "16. Civic Duty", "17. Attacks on U.S.","18. Discussion of U.S.", "19.
Electricity", "20. Western Countries", "21. Trust in Policy Makers","22. Religious Freedom in the
West", "23. Religious Intolerance", "24. Civility", "25. Policy and Law","26. Roads", "27. None",
"28. None", "29. None")

initial.factor.analysis <- function(data,num){

## Find the optimal number of factors for a field of data
       ev <- eigen(cor(data))
       if(num!=0) {
              num <- num
       }
       else {
              num <- length(ev$values[ev$values > 1])
       }

## Conduct factor analysis
       fact <- factanal(data,factors=num,rotation="varimax")

## Convert the factor loadings to a matrix and name the factors
       fa.mat <- numeric(0)
       for(i in 1:num){
```

```
                fake.fac.load <- fact$loadings[,i]
                fake.fac.load[fact$loadings[,i] < 0.4 & (fact$loadings[,i] > -0.4)] <- 0
                fa.mat <- cbind(fa.mat, fake.fac.load)
        }
        colnames(fa.mat) <- c()
        rownames(fa.mat) <- c()
        rownames(fa.mat) <- c(colnames(data))
        colnames(fa.mat) <- colnames(fa.mat, do.NULL= FALSE, prefix = "Factor.")
        fa.mat # matrix with loadings > 0.4 or < -0.4

## Calculate the variance of each variable

        i.j.MatrixLoc <- which(fa.mat!=0, arr.ind=TRUE)
        z <- tapply (i.j.MatrixLoc[,1], i.j.MatrixLoc[,2],
                        function (x) sum (ev$values[x]))/length(ev$values)
        z <- as.matrix(z)
        dim(z) <- length(z)
        rownames(z) <- rownames(z, do.NULL= FALSE, prefix = "Factor.")

## Print the Output

cat("The number of factors (based off of eigen values or given) is: ", num, "\n",
sep="",file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Data10FactorOutput.txt",
append=FALSE)
cat("\n","The number of relevent factors is: ",length(z),"\n", sep="",
file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Data10FactorOutput.txt", append=TRUE)
cat("\n","The variables per factor are: ", "\n","================================",
sep="",file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Data10FactorOutput.txt",
append=TRUE)

        x <- numeric(0)
        for(i in 1:ncol(fa.mat)){
                f <- rownames(fa.mat)[which(fa.mat[,i]!=0)]
                x <- fa.mat[which(fa.mat[,i]!=0),i]
                x <- as.matrix(x)
                rownames(f) <- c(colnames(fa.mat[,i]))
                colnames(x) <- c(colnames(fa.mat[,i]))

cat("\n","Factor",i,"= ", sep=" ",
file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Data10FactorOutput.txt", append=TRUE)
cat(round(x,4), sep=", ", file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Data10FactorOutput.txt", append=TRUE)
cat("\n","Factor",i,"= ", sep=" ",file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Data10FactorOutput.txt", append=TRUE)
cat(f, sep=", ", file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Data10FactorOutput.txt", append=TRUE)
cat("\n","----------------------------------------------------------------", "\n", sep="",
file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Data10FactorOutput.txt", append=TRUE)
}
cat("\n","----------------------------------------------------------------","\n","\n","The
variance impact of each factor is in % : ", "\n",
"================================================","\n",
sep="",file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel Survey/Data10FactorOutput.txt",
append=TRUE)

write.table(round(z,4)*100,"C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Data10FactorOutput.txt", append=TRUE,sep="= ", col.names=FALSE, row.names=TRUE,
quote=FALSE, na="NA")
}

initial.factor.analysis(final.data,29)

factor.analysis <- function(data,num,name){

        fact <- factanal(data,factors=num,rotation="varimax")

## Convert the factor loadings to a matrix and name the factors
        fa.mat <- numeric(0)
        for(i in 1:num){
```

```
                fake.fac.load <- fact$loadings[,i]
                fake.fac.load[fact$loadings[,i] < 0.4 & (fact$loadings[,i] > -0.4)] <- 0
                fa.mat <- cbind(fa.mat, fake.fac.load) # builds a matrix of factors
        }
        colnames(fa.mat) <- c()
        rownames(fa.mat) <- c()
        rownames(fa.mat) <- c(colnames(data))
        colnames(fa.mat) <- colnames(fa.mat, do.NULL= FALSE, prefix = "Factor.")
        fa.mat # matrix with loadings > 0.4 or < -0.4

        if (is.na(name)==FALSE){
                colnames(fa.mat)<- c(name)
                return(fa.mat)
        }
        else{
        return(fa.mat)
        }
}

Nig.factors <- factor.analysis(final.data,29,factorNames)

## Modify factors & Create Matrix of Factor Scores

Nig.factors <- Nig.factors[,-c(27,28,29)] # delete factors 27, 28, 29
Nig.factors[24,8] <- 0 # delete q17sa in factor 8
Nig.factors[27,8] <- 0 # delete q17ir in factor 8

final.data <- as.matrix(final.data)

factor.scores <- data.frame(final.data%*%Nig.factors)

## Scale factor scores by dividing by factor loading sums to get scores between -2 and 2

loadSum <- colSums(data.frame(Nig.factors))
factor.scores <- apply(factor.scores,1,function(x)x/loadSum)
factor.scores <- data.frame(t(factor.scores))

write.table(factor.scores,"C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/FactorScores_10.csv",sep=",",col.names=TRUE,row.names=FALSE,quote=TRUE,na="NA")
```

III. Recode Response Variables

```
## Code for recoding response variables

library(car) # package for recoding

demoVar <- read.csv("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/Rec_Imp_10.csv",header=TRUE)

## Questions to add in the model and corresponding recoding

Actor <- as.factor(demoVar[,"q47"])
Safety <- demoVar[,"d23"]
Goals <- demoVar[,"q6"]
Services <- demoVar[,"q7"]
Equality <- demoVar[,"q10"]

## Combine the data sets into initial states for modeling

model.data <- na.omit(data.frame(cbind(factor.scores,Safety,Goals,Services,Equality)))
```

IV. Model Building

```
#### Function to iterate regression models IOT pick the best ones
```

```
library(MASS)
data.best <-
data.frame(matrix(rep(0,nrow(model.data)*ncol(model.data)),nrow(model.data),ncol(model.data)))
names(data.best) <- names(model.data)
for (i in 1:ncol(model.data)){
        reg <- lm(model.data[,i] ~ .,data=model.data[,-c(i)])
        reg.step <- stepAIC(reg,scope = list(upper = ~ ., lower = ~ 1),trace=FALSE)
                if (summary(reg.step)$adj.r.squared > 0.39){
                        data.best[,i] <- model.data[,i]
        }
}

which(colSums(data.best)!=0)

### Building, initializing,& predicting future Issue Stance Scores

## Model Build

rx2 <- lm(X2..U.S..Assist.to.Nigeria ~ . - X4..Social...Essential.Services -
X5..Trust.in.Gov.Agencies - X10..Democracy,data=model.data)
rx2.step <- stepAIC(rx2,scope = list(upper = ~ . - X4..Social...Essential.Services -
X5..Trust.in.Gov.Agencies - X10..Democracy, lower = ~ 1),trace=FALSE)
summary(rx2.step)

rx4 <- lm(X4..Social...Essential.Services ~ . - X2..U.S..Assist.to.Nigeria -
X5..Trust.in.Gov.Agencies - X10..Democracy,data=model.data)
rx4.step <- stepAIC(rx4,scope = list(upper = ~ . - X2..U.S..Assist.to.Nigeria -
X5..Trust.in.Gov.Agencies - X10..Democracy, lower = ~ 1),trace=FALSE)
summary(rx4.step)

rx5 <- lm(X5..Trust.in.Gov.Agencies ~ . - X2..U.S..Assist.to.Nigeria -
X4..Social...Essential.Services - X10..Democracy,data=model.data)
rx5.step <- stepAIC(rx5,scope = list(upper = ~ . - X2..U.S..Assist.to.Nigeria -
X4..Social...Essential.Services - X10..Democracy, lower = ~ 1),trace=FALSE)
summary(rx5.step)

rx10 <- lm(X10..Democracy ~ . - X2..U.S..Assist.to.Nigeria - X4..Social...Essential.Services
- X5..Trust.in.Gov.Agencies,data=model.data)
rx10.step <- stepAIC(rx10,scope = list(upper = ~ . - X2..U.S..Assist.to.Nigeria -
X4..Social...Essential.Services - X5..Trust.in.Gov.Agencies, lower = ~ 1),trace=FALSE)
summary(rx10.step)

## Generate initial Issue Stance Scores using mean factor scores

intx2 <- intersect(names(coef(rx2.step)),names(model.data))
intx4 <- intersect(names(coef(rx4.step)),names(model.data))
intx5 <- intersect(names(coef(rx5.step)),names(model.data))
intx10 <- intersect(names(coef(rx10.step)),names(model.data))
ndx2 <- data.frame(matrix(round(colMeans(model.data[,c(intx2)]),3),1,NROW(intx2),byrow=TRUE))
names(ndx2) <- c(intx2)
ndx4 <- data.frame(matrix(round(colMeans(model.data[,c(intx4)]),3),1,NROW(intx4),byrow=TRUE))
names(ndx4) <- c(intx4)
ndx5 <- data.frame(matrix(round(colMeans(model.data[,c(intx5)]),3),1,NROW(intx5),byrow=TRUE))
names(ndx5) <- c(intx5)
ndx10 <- data.frame(matrix(round(colMeans(model.data[,c(intx10)]),3),1,NROW(intx10),byrow=TRUE))
names(ndx10) <- c(intx10)

## Predict inital Issue Stance Scores

nx2 <- data.frame(round(predict(rx2.step,ndx2,type="response"),3))
nx4 <- data.frame(round(predict(rx4.step,ndx4,type="response"),3))
nx5 <- data.frame(round(predict(rx5.step,ndx5,type="response"),3))
nx10 <- data.frame(round(predict(rx10.step,ndx10,type="response"),3))

## Output initial Issue Stance Score files to excel

library(xlsx)
names(nx2) <- c("X2_Predict")
names(nx4) <- c("X4_Predict")
```

```
names(nx5) <- c("X5_Predict")
names(nx10) <- c("X10_Predict")
write.xlsx(nx2,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X2_Initial_Issue",row.names=FALSE,append=TRUE)
write.xlsx(nx4,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X4_Initial_Issue",row.names=FALSE,append=TRUE)
write.xlsx(nx5,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X5_Initial_Issue",row.names=FALSE,append=TRUE)
write.xlsx(nx10,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X10_Initial_Issue",row.names=FALSE,append=TRUE)

## Read-in SME input files

pdx2 <- read.xlsx("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/SME.xlsx",sheetIndex=1,sheetName="X2",as.data.frame=TRUE,header=TRUE,keepFormulas=FALSE)
pdx2 <- pdx2[,-c(1,2)]
pdx4 <- read.xlsx("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/SME.xlsx",sheetIndex=2,sheetName="X4",as.data.frame=TRUE,header=TRUE,keepFormulas=FALSE)
pdx4 <- pdx4[,-c(1,2)]
pdx5 <- read.xlsx("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/SME.xlsx",sheetIndex=3,sheetName="X5",as.data.frame=TRUE,header=TRUE,keepFormulas=FALSE)
pdx5 <- pdx5[,-c(1,2)]
pdx10 <- read.xlsx("C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/SME.xlsx",sheetIndex=4,sheetName="X10",as.data.frame=TRUE,header=TRUE,keepFormulas=FALSE)
pdx10 <- pdx10[,-c(1,2)]

## Predict future Issue Stance Scores based on events
event <- c(1:20)
p2 <- data.frame(round(predict(rx2.step,pdx2,type="response"),3))
p4 <- data.frame(round(predict(rx4.step,pdx4,type="response"),3))
p5 <- data.frame(round(predict(rx5.step,pdx5,type="response"),3))
p10 <- data.frame(round(predict(rx10.step,pdx10,type="response"),3))

px2 <- cbind(event,p2)
px4 <- cbind(event,p4)
px5 <- cbind(event,p5)
px10 <- cbind(event,p10)

## Output predicted Issue Stance Score files to excel

names(px2) <- c("Event","X2_Predict")
names(px4) <- c("Event","X4_Predict")
names(px5) <- c("Event","X5_Predict")
names(px10) <- c("Event","X10_Predict")
write.xlsx(px2,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X2_Predict_Issue",row.names=FALSE,append=TRUE)
write.xlsx(px4,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X4_Predict_Issue",row.names=FALSE,append=TRUE)
write.xlsx(px5,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X5_Predict_Issue",row.names=FALSE,append=TRUE)
write.xlsx(px10,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="X10_Predict_Issue",row.names=FALSE,append=TRUE)

### Building, initializing, and predicting future OABs

## Model Build

library(mlogit)

wr.data <- data.frame(cbind(Actor,factor.scores))
wr.data <- wr.data[,c(1,3,5,6,11)]

WR <- mlogit.data(wr.data,varying=NULL,choice="Actor",shape="wide")

weight.reg <- mlogit(Actor ~ 1 | X2..U.S..Assist.to.Nigeria + X4..Social...Essential.Services +
X5..Trust.in.Gov.Agencies + X10..Democracy,data=WR,reflevel="0")
wsum <- summary(weight.reg)

## Predict Initial OAB Probabilities
```

```
oab.data <- wr.data[,-c(1)]
wr <- data.frame(matrix(round(colMeans(oab.data),3),1,4,byrow=TRUE))
names(wr) <- names(oab.data)

log0 <- rep(0,1)
log1 <- wsum$coef[["1:(intercept)"]] +
wsum$coef[["1:X2..U.S..Assist.to.Nigeria"]]*wr$X2..U.S..Assist.to.Nigeria +
wsum$coef[["1:X4..Social...Essential.Services"]]*wr$X4..Social...Essential.Services +
wsum$coef[["1:X5..Trust.in.Gov.Agencies"]]*wr$X5..Trust.in.Gov.Agencies +
wsum$coef[["1:X10..Democracy"]]*wr$X10..Democracy
log2 <- wsum$coef[["2:(intercept)"]] +
wsum$coef[["2:X2..U.S..Assist.to.Nigeria"]]*wr$X2..U.S..Assist.to.Nigeria +
wsum$coef[["2:X4..Social...Essential.Services"]]*wr$X4..Social...Essential.Services +
wsum$coef[["2:X5..Trust.in.Gov.Agencies"]]*wr$X5..Trust.in.Gov.Agencies +
wsum$coef[["2:X10..Democracy"]]*wr$X10..Democracy
log3 <- wsum$coef[["3:(intercept)"]] +
wsum$coef[["3:X2..U.S..Assist.to.Nigeria"]]*wr$X2..U.S..Assist.to.Nigeria +
wsum$coef[["3:X4..Social...Essential.Services"]]*wr$X4..Social...Essential.Services +
wsum$coef[["3:X5..Trust.in.Gov.Agencies"]]*wr$X5..Trust.in.Gov.Agencies +
wsum$coef[["3:X10..Democracy"]]*wr$X10..Democracy
log4 <- wsum$coef[["4:(intercept)"]] +
wsum$coef[["4:X2..U.S..Assist.to.Nigeria"]]*wr$X2..U.S..Assist.to.Nigeria +
wsum$coef[["4:X4..Social...Essential.Services"]]*wr$X4..Social...Essential.Services +
wsum$coef[["4:X5..Trust.in.Gov.Agencies"]]*wr$X5..Trust.in.Gov.Agencies +
wsum$coef[["4:X10..Democracy"]]*wr$X10..Democracy
log5 <- wsum$coef[["5:(intercept)"]] +
wsum$coef[["5:X2..U.S..Assist.to.Nigeria"]]*wr$X2..U.S..Assist.to.Nigeria +
wsum$coef[["5:X4..Social...Essential.Services"]]*wr$X4..Social...Essential.Services +
wsum$coef[["5:X5..Trust.in.Gov.Agencies"]]*wr$X5..Trust.in.Gov.Agencies +
wsum$coef[["5:X10..Democracy"]]*wr$X10..Democracy

logits <- cbind(log0,log1,log2,log3,log4,log5)
prob <- data.frame(round(exp(logits)/rowSums(exp(logits)),3)) # This is the data frame of
probabilities
colnames(prob) <-
c("Rebel_Groups_Predict","International_Terrorists_Predict","Common_Criminals_Predict",
"Military_Predict","Government_Predict","Foreign_Countries_Predict")

## Output initial OAB Probability files to excel

names(prob[1]) <- c("Rebel_Groups_Predict")
names(prob[2]) <- c("International_Terrorists_Predict")
names(prob[3]) <- c("Common_Criminals_Predict")
names(prob[4]) <- c("Military_Predict")
names(prob[5]) <- c("Government_Predict")
names(prob[6]) <- c("Foreign_Countries_Predict")
write.xlsx(prob[1],file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Rebels_Initial_OAB",row.names=FALSE,append=TRUE)
write.xlsx(prob[2],file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Terrorists_Initial_OAB",row.names=FALSE,append=TRUE)
write.xlsx(prob[3],file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Criminals_Initial_OAB",row.names=FALSE,append=TRUE)
write.xlsx(prob[4],file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Military_Initial_OAB",row.names=FALSE,append=TRUE)
write.xlsx(prob[5],file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Government_Initial_OAB",row.names=FALSE,append=TRUE)
write.xlsx(prob[6],file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="ForiegnCountries_Initial_OAB",row.names=FALSE,append=TRUE)

## Predict future OAB Probabilities based on events

pd <- cbind(px2,px4,px5,px10)[,c(2,4,6,8)]

log00 <- rep(0,20)
log11 <- wsum$coef[["1:(intercept)"]] + wsum$coef[["1:X2..U.S..Assist.to.Nigeria"]]*pd$X2_Predict
+ wsum$coef[["1:X4..Social...Essential.Services"]]*pd$X4_Predict +
wsum$coef[["1:X5..Trust.in.Gov.Agencies"]]*pd$X5_Predict +
wsum$coef[["1:X10..Democracy"]]*pd$X10_Predict
```

```
log22 <- wsum$coef[["2:(intercept)"]] + wsum$coef[["2:X2..U.S..Assist.to.Nigeria"]]*pd$X2_Predict
+ wsum$coef[["2:X4..Social...Essential.Services"]]*pd$X4_Predict +
wsum$coef[["2:X5..Trust.in.Gov.Agencies"]]*pd$X5_Predict +
wsum$coef[["2:X10..Democracy"]]*pd$X10_Predict
log33 <- wsum$coef[["3:(intercept)"]] + wsum$coef[["3:X2..U.S..Assist.to.Nigeria"]]*pd$X2_Predict
+ wsum$coef[["3:X4..Social...Essential.Services"]]*pd$X4_Predict +
wsum$coef[["3:X5..Trust.in.Gov.Agencies"]]*pd$X5_Predict +
wsum$coef[["3:X10..Democracy"]]*pd$X10_Predict
log44 <- wsum$coef[["4:(intercept)"]] + wsum$coef[["4:X2..U.S..Assist.to.Nigeria"]]*pd$X2_Predict
+ wsum$coef[["4:X4..Social...Essential.Services"]]*pd$X4_Predict +
wsum$coef[["4:X5..Trust.in.Gov.Agencies"]]*pd$X5_Predict +
wsum$coef[["4:X10..Democracy"]]*pd$X10_Predict
log55 <- wsum$coef[["5:(intercept)"]] + wsum$coef[["5:X2..U.S..Assist.to.Nigeria"]]*pd$X2_Predict
+ wsum$coef[["5:X4..Social...Essential.Services"]]*pd$X4_Predict +
wsum$coef[["5:X5..Trust.in.Gov.Agencies"]]*pd$X5_Predict +
wsum$coef[["5:X10..Democracy"]]*pd$X10_Predict

logits1 <- cbind(log00,log11,log22,log33,log44,log55)
prob1 <- data.frame(round(exp(logits1)/rowSums(exp(logits1)),3))
colnames(prob1) <- c("Rebel Groups","International Terrorists","Common
Criminals","Military","Government","Foreign Countries")

## Output predicted OAB Probability files to excel

poab0 <- data.frame(cbind(event,prob1[,1]))
poab1 <- data.frame(cbind(event,prob1[,2]))
poab2 <- data.frame(cbind(event,prob1[,3]))
poab3 <- data.frame(cbind(event,prob1[,4]))
poab4 <- data.frame(cbind(event,prob1[,5]))
poab5 <- data.frame(cbind(event,prob1[,6]))
names(poab0) <- c("Event","Rebel_Groups_Predict")
names(poab1) <- c("Event","International_Terrorists_Predict")
names(poab2) <- c("Event","Common_Criminals_Predict")
names(poab3) <- c("Event","Military_Predict")
names(poab4) <- c("Event","Government_Predict")
names(poab5) <- c("Event","Foreign_Countries_Predict")
write.xlsx(poab0,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Rebels_Predict_OAB",row.names=FALSE,append=TRUE)
write.xlsx(poab1,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Terrorists_Predict_OAB",row.names=FALSE,append=TRUE)
write.xlsx(poab2,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Criminals_Predict_OAB",row.names=FALSE,append=TRUE)
write.xlsx(poab3,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Military_Predict_OAB",row.names=FALSE,append=TRUE)
write.xlsx(poab4,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="Government_Predict_OAB",row.names=FALSE,append=TRUE)
write.xlsx(poab5,file="C:/Users/tmdevean/Desktop/IW TWG/2010 Sahel
Survey/ALL.xlsx",sheetName="ForeignCountries_Predict_OAB",row.names=FALSE,append=TRUE)
```

V. Use Case

```
### Example Use Case

time.step <- data.frame(c(1:200))
names(time.step) <- c("Time")
events <- data.frame(sample(1:20,200,replace=T))
names(events) <- c("Event")

event.list1 <- merge(cbind(time.step,events),px2)
event.list2 <- merge(cbind(time.step,events),px4)
event.list3 <- merge(cbind(time.step,events),px5)
event.list4 <- merge(cbind(time.step,events),px10)
event.list5 <- merge(cbind(time.step,events),poab0)
event.list6 <- merge(cbind(time.step,events),poab1)
event.list7 <- merge(cbind(time.step,events),poab2)
event.list8 <- merge(cbind(time.step,events),poab3)
event.list9 <- merge(cbind(time.step,events),poab4)
event.list10 <- merge(cbind(time.step,events),poab5)
```

```
event.list <-
cbind(event.list1,event.list2,event.list3,event.list4,event.list5,event.list6,event.list7,event.l
ist8,event.list9,event.list10)

event.list <- event.list[,c(1,2,3,6,9,12,15,18,21,24,27,30)]

event.list <- event.list[order(event.list[,"Time"]),]
event.list <- event.list[,c(2,1,3,4,5,6,7,8,9,10,11,12)]

in.time <- data.frame(c(0))
names(in.time) <- c("Time")
in.event <- data.frame(c(0))
names(in.event) <- c("Event")

event.list <-
rbind(cbind(in.time,in.event,nx2,nx4,nx5,nx10,prob[1],prob[2],prob[3],prob[4],prob[5],prob[6]),ev
ent.list)

## Issue Stance Score Plots
par(mfrow=c(2,2))
plot(event.list$Time,event.list$X2_Predict,type="l",xlab="Time Step",ylim=c(-2,2),ylab="Issue
Stance Score",main="'U.S. Assistance to Nigeria' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$X2_Predict,iter=10),lty="dashed",col="139")

plot(event.list$Time,event.list$X4_Predict,type="l",xlab="Time Step",ylim=c(-2,2),ylab="Issue
Stance Score",main="'Social & Essential Services' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$X4_Predict,iter=10),lty="dashed",col="139")

plot(event.list$Time,event.list$X5_Predict,type="l",xlab="Time Step",ylim=c(-2,2),ylab="Issue
Stance Score",main="'Trust in Government Agencies' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$X5_Predict,iter=10),lty="dashed",col="139")

plot(event.list$Time,event.list$X10_Predict,type="l",xlab="Time Step",ylim=c(-2,2),ylab="Issue
Stance Score",main="'Democracy' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$X10_Predict,iter=10),lty="dashed",col="139")

## OAB Probability Plots
par(mfrow=c(2,3))
plot(event.list$Time,event.list$Rebel_Groups_Predict,type="l",xlab="Time
Step",ylim=c(0.05,0.15),ylab="Probability",main="'Rebel Groups' OAB Probability over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$Rebel_Groups_Predict,iter=10),lty="dashed",col="139")

plot(event.list$Time,event.list$International_Terrorists_Predict,type="l",xlab="Time
Step",ylim=c(0,0.1),ylab="Probability",main="'International Terrorists' OAB Probability over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$International_Terrorists_Predict,iter=10),lty="dashed",co
l="139")

plot(event.list$Time,event.list$Common_Criminals_Predict,type="l",xlab="Time
Step",ylim=c(0.1,0.3),ylab="Probability",main="'Common Criminals' OAB Probability over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$Common_Criminals_Predict,iter=10),lty="dashed",col="139")

plot(event.list$Time,event.list$Military_Predict,type="l",xlab="Time
Step",ylim=c(0.05,0.1),ylab="Probability",main="'Military' OAB Probability over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$Military_Predict,iter=10),lty="dashed",col="139")

plot(event.list$Time,event.list$Government_Predict,type="l",xlab="Time
Step",ylim=c(0.4,0.7),ylab="Probability",main="'Government' OAB Probability over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$Government_Predict,iter=10),lty="dashed",col="139")
```

```
plot(event.list$Time,event.list$Foreign_Countries_Predict,type="l",xlab="Time
Step",ylim=c(0.01,0.04),ylab="Probability",main="'Foreign Countries' OAB Probability over
Time",col="2",col.main="4",font.lab="2",font.main="2")
lines(lowess(event.list$Time,event.list$Foreign_Countries_Predict,iter=10),lty="dashed",col="139"
)

sh.elist <- event.list[,-c(1,2)]
delta.event <- cumsum(sh.elist)
delta.event <- cbind(data.frame(c(1:201)),delta.event)
names(delta.event) <-
c("Time","X2_Delta","X4_Delta","X5_Delta","X10_Delta","Rebel_Delta","Terrorist_Delta",
"Criminal_Delta","Military_Delta","Government_Delta","Foreign_Delta")

## Issue Stance Score Cumulative Plots
par(mfrow=c(2,2))
plot(delta.event$Time,delta.event$X2_Delta,type="l",xlab="Time Step",ylab="Issue Stance Score
Delta",main="Change in 'U.S. Assistance to Nigeria' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")

plot(delta.event$Time,delta.event$X4_Delta,type="l",xlab="Time Step",ylab="Issue Stance Score
Delta",main="Change in 'Social & Essential Services' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")

plot(delta.event$Time,delta.event$X5_Delta,type="l",xlab="Time Step",ylab="Issue Stance Score
Delta",main="Change in 'Trust in Government Agencies' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")

plot(delta.event$Time,delta.event$X10_Delta,type="l",xlab="Time Step",ylab="Issue Stance Score
Delta",main="Change in 'Democracy' Issue Stance Score over
Time",col="2",col.main="4",font.lab="2",font.main="2")
```

LIST OF REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis* (1st ed., p. 206). John Wiley & Sons, Inc.

Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research and Evaluation, 10*(7), 1-9.

DeCoster, J. (1998). Overview of Factor Analysis. Retrieved July 25, 2012, from http://www.stat-help.com/notes.html

Dziedzic, M., Sotirin, B., & Agoglia, J. (Eds.). (2008). *Measuring Progress in Conflict Enviroments (MPICE)* (1st ed.).

D'Orazio, M. (2011). StatMatch: Statistical Matching.

Harman, H. H. (1976). *Modern Factor Analysis* (3rd ed.). University of Chicago Press.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. *Organizational Research Methods, 7*(2), 191-205.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed., pp. 260-264). John Wiley & Sons, Inc.

Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5th ed., p. 504). Prentice-Hall.

Kilcullen, D. J. (2009). Measuring Progress in Afghanistan.

Kline, P. (1994). *An Easy Guide to Factor Analysis* (1st ed.). Routledge.

Kulzy, W. W. (2012). *Modeling Indigenous Population Attitudes in Support of Irregular Warfare Analysis*. Naval Postgraduate School.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous Conjoint Measurement: A New Scale of Fundamental Measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Mulaik, S. A. (2009). *Foundations of Factor Analysis* (2nd ed., p. 136). CRC Press.

Obama, B. (2009). Speech by President Barack Obama to the Ghanian Parliament.

Ploch, L. (2011). Africa Command: U.S. Strategic Interests and the Role of the U.S. Military in Africa. *CRS Report for Congress*. Retrieved July 24, 2012, from http://www.fas.org/sgp/crs/natsec/RL34003.pdf

Revelle, W. (2011). psych: Procedures for Psychological, Psychometric, and Personality Research.

The White House. (2002). *The National Security Strategy of the United States of America*. Washington, DC.

United States. (2006). *Counterinsurgency: Field Manual 3-24*. Washington, DC: Headquarters, Dept. of the Army.

United States. (2009). *Tactics in Counterinsurgency: Field Manual 3-24.2*. Washington, DC: Headquarters, Dept. of the Army.

Upshur, W. P., Roginski, J. W., & Kilcullen, D. J. (2012). Recognizing Systems in Afghanistan. *Prism, 3*(3), 87-104.

Varner, M. A. (2007). The Strategic Importance of Africa Command. Retrieved June 29, 2012, from http://www.thepresidency.org/storage/documents/Vater/Varner.pdf

THIS PAGE INTENTIONALLY LEFT BLANK