De-Radicalization and Reintegration from Violent Islamic Extremism*

— Research Design —

Robert Blair† *Brown Univ.*   Jun Goto‡ *Kobe Univ.*   Yosuke Nagai§ *Accept International*   Shunsuke Tsuda¶ *Brown Univ.*

Preliminary and incomplete — Please do not cite.

April, 2021

Abstract

The prevalence of violence by Islamic extremist organizations is a global threat. We collect novel data from ex-combatants of Al-Shabaab (a jihadist group responsible for the greatest volume of recent fatalities in Africa) imprisoned in Somalia. We identify extremist ideologies, preferences, and beliefs as key concepts and preconditions for achieving de-radicalization and reintegration. We implement an RCT intervention to evaluate the effectiveness of frequent rehabilitation counseling, including disillusionment management, and intergroup contact with community representatives in the form of a reconciliation dialogue. The degree of radicalization toward violent Islamic extremism is quantified by a natural language processing approach. A series of lab-in-the-field experiments is designed to capture other-regarding preferences, cooperative behaviors and subjective beliefs, and to test the role of information about ex-combatants' identities and statuses for identifying the barriers present in both civilians' and ex-combatants' sides.

Keywords: Conflicts, Ex-Combatants, Extremism, Islam, Peace Building, Prison, Lab-in-the-field Experiments, Text Mining, Somalia

*We thank comments from Dan Björkegren, Chris Blattman, Pedro Dal Bó, Emel Filiz-Ozbay, Andrew Foster, Jessica Goldberg, Ethan Kaplan, Nicola Limodio, Salma Mousa, Noah Nathan, Amma Panin, Eva Rios, Devesh Rustagi, Jesse Shapiro, Neil Thakral, and participants of Applied Micro Lunch (Brown), GPD Module on Experiments at Watson Institute (Brown), and Working Group in African Political Economy (WGAPE) Annual Meeting. We acknowledge support from the Japan Society for the Promotion of Science (19K01653), the Murata Science Foundation Research Grant, and the Bravo Center Research Funding.

†Joukowsky Family Assistant Professor of Political Science and International and Public Affairs. E-mail: Robert_Blair@brown.edu.

‡Associate Professor, Graduate School of Economics. E-mail: goto@econ.kobe-u.ac.jp

§CEO. E-mail: yosuke.nagai@accept-int.org

¶Ph.D. Candidate, Department of Economics. E-mail: Shunsuke_Tsuda@brown.edu.

1 Introduction

Violence by Islamic extremist organizations is prevalent all over the world, especially in the past decade in Africa (Figure 1). In Africa, the number of Islamic extremist organizations, the frequency of violent events involving them, and their geographical coverage have been increasing over the past decade. Finding ways to resolve conflicts involving violent Islamic extremist groups is a pressing task.

Promoting individual-level de-radicalization and reintegration is important to reduce violence by extremist groups for three reasons. First, peaceful negotiation with leaders of such groups is difficult and often has not been fruitful. In this environment, incentivizing individuals to abandon their affiliations with such groups may be a more feasible alternative. Second, if it is difficult for ex-combatants to reintegrate into society, then the risk of their re-radicalization and re-joining an extremist group is increased. Third, a re-integration process facilitated by civil society might incentivize combatants of violent groups to surrender at their own initiative.

This research attempts to uncover obstacles behind individual-level de-radicalization and reintegration from violent Islamic extremism. We will collect novel data from ex-combatants of a jihadist group imprisoned in Somalia. We implement an RCT intervention to evaluate the impact of a new generation of conflict resolution programs on radicalization and the pre-conditions for reintegration. The degree of radicalization is quantified by natural language processing approaches. We design a series of lab-in-the-field experiments to capture other-regarding preferences and subjective beliefs, which play key roles for reintegration after release.

We will attempt to collect information from all Al-Shabaab ex-combatants detained in the Mogadishu Central Prison in Somalia, in cooperation with the Somali government and the non-profit organization Accept International. It is presumed that there are about 700 former Al-Shabaab members in the prison. Al-Shabaab is an Islamic extremist organization, responsible for the greatest volume of recent fatalities among all militant groups in Africa[1]. Observing all members, including members not imprisoned, would be ideal, but is not possible. In this respect, our data is the second best option, but we believe that it will provide some of the most detailed information on violent extremists available in the research community.

Our conceptual framework identifies extremist ideologies, other-regarding preferences, and beliefs about others' cooperative behaviors as key concepts and pre-conditions for achieving de-radicalization and reintegration of the ex-combatants. We then propose three types of interventions (rehabilitation counseling; intergroup contact in the form of reconciliation dialogue; information provision) which are expected to shape these concepts. The frequent counseling (both at individual- and group-levels), including disillusionment management, and the reconciliation dialogue with community representatives are provided by our RCT intervention of a DRR (De-radicalization, Reinsertion, Reintegration) project, a new generation of conflict resolution programs. The information provision intervention is introduced in our lab-in-the-field experiments.

We quantify the degree of radicalization toward violent Islamic extremism. We define the degree of

[1]Indeed, more than 4,000 people were killed by Al-Shabaab in 2016. Al-Shabaab is under the umbrella of the global Al-Qaeda group. Al-Shabaab is active mainly in Somalia.

radicalization as the closeness of an ex-combatant's identity to violent Islamic extremist ideology relative to a general Muslim ideology. Directly eliciting and observing a person's internal ideology is difficult. We address this challenge in two ways. First, we adopt a natural language processing (NLP) approach to quantify individual ideological extremism. We conduct face-to-face interviews with each ex-combatant repeatedly in our counseling intervention. During these interviews, we record the entire conversation and transform it to text data. We rely on a deep-learning-based classification of the obtained text data into the extremist ideology. Second, we employ an implicit association test (IAT), which was developed in the field of psychology. The IAT has been widely applied to economics research to measure attitudes toward a specific object (e.g., Lowes et al. 2017).

A series of lab-in-the-field experiments is designed to quantify two key obstacles to reintegration.[2]. The first is ex-combatants' (civilians') other-regarding preferences toward civilians (ex-combatants). The second is cooperative behavior between ex-combatants and civilians, focusing on subjective beliefs about others' behaviors. (Sequential) self-other allocation games extended from Chen and Li (2009) capture other-regarding preferences. We design a "public investment game" (PIG) to capture subjective beliefs under cooperative actions. The lab-in-the-field experiments also allow us to investigate how information can be designed to update beliefs about others' behaviors and attitudes.[3]

This paper contributes to the literature on the economics of conflict from three perspectives.

First, this paper contributes to the literature on peacebuilding in fragile and post-conflict economies. Bauer et al. (2016) provide a comprehensive survey on this topic. This research is the first to study the reintegration process for ex-combatants of violent extremist organizations from its initial step. The motivation and context of our lab-in-the-field experiments is most similar to Bauer et al. (2017). Bauer et al. (2017) measure trust towards ex-combatants of the Lord's Resistance Army (LRA) in receiving communities in Northern Uganda. The ex-combatants in their sample have returned to, and are already living in, these communities. An important stage of the reintegration process that is missing from Bauer et al. (2017), and from most of the literature, is whether ex-combatants are willing to (re-)join these communities in the first place. We directly measure ex-combatants' attitudes and behaviors in advance of this process.

Second, this paper adds new evidence to the recently growing literature on individual-level interventions on high-risk populations, including non-cognitive skills (Blattman and Annan 2016; Blattman et al. 2017). We investigate whether ideologies, preferences, and subjective beliefs, all of which play key roles in reintegration and peacebuilding, can be updated through two distinct forms of external intervention,

[2]Ideally we could observe the entire process of reintegration beginning at the moment of release, but this might be infeasible for two reasons. First, it takes a long time to wait until ex-combatants finish their sentences. Second, different ex-combatants will reintegrate into geographically dispersed societies. The feasibility of tracking all of them is doubtful. We will, however, continue to explore the feasibility of tracking in the long run

[3]The influences of information on reintegration are also discussed in some recent research. Armand et al. (2020) study community acceptance message by FM radio on conflict outcomes involving LRA in Northern Uganda. Blouin and Mukand (2019) studies the role of radio propaganda for reintegration between Tutsi and Hutu in post-genocide Rwanda. In Armand et al. (2020), combatant-level behavior is a black box. In Blouin and Mukand (2019), what specific information has driven their result is a black box. We fill these knowledge gaps by providing ex-combatants with specific information in an experimental setting.

namely counseling and reconciliation dialogue.

Our reconciliation treatment is also related to recent research on the ways that intergroup contact shapes individual attitudes and behaviors towards out-groups (e.g., Lowe 2020; Mousa 2020; Scacco and Warren 2018). Among these papers, Mousa (2020) has the most similar motivation to our project. She provides evidence on the impact of intergroup contact in a post-conflict setting, focusing on the relationship between Muslims and Christians displaced by ISIS in Iraq. While she looks at prejudice toward Muslims in general, we focus on the relationship between civilians and ex-combatants from an Islamic extremist group. We examine how actual contact with a civilian representative shapes ex-combatants' preferences and beliefs about civilians and civil society.

Third, this study is broadly connected to research on the microfoundations of rebellion with primary data collection from the field (e.g., Beber and Blattman 2013; Sánchez De La Sierra 2020). To our knowledge, this paper is the first to directly observe the behavior of former members of Islamic extremist organizations. As there is little knowledge on violent Islamic extremism in economics both theoretically and empirically, this paper is expected to open this research area.

In addition, this study is connected to the emerging literature using text data in economics (Gentzkow et al. 2019)[4]. Several studies in political science also use text data to investigate the political preferences of politicians (Laver and Benoit 2002; Laver et al. 2003).[5] However, empirical studies adopting text mining to identify religious extremism are scant, to the best of our knowledge.[6]

The reminder of this paper proceeds as follows. Section 2 describes the conflict environment in Somalia and our data collection. Section 3 describes the procedure of our RCT intervention. Section 4 describes our conceptual framework. Section 5 describes our approach to measuring the degree of radicalization. Section 6 describes the procedure of our lab experiments.

[4]For example, Gentzkow and Shapiro (2010) apply the simple NPL approach to analyze how newspapers incorporate slant to maximize their profits, in the face of local readers' demands. Hansen et al. (2018) use machine learning methods to study the impact of ensuring transparency on discussion for policymaking in the Federal Open Market Committee (FOMC).

[5]See Wilkerson and Casas (2017) which provide extensive review on the studies applying NLP approaches in political science.

[6]Note that subjects in our context who are ex-combatants of Al-Shabaab are fully different from ones in previous studies such as politicians: we cannot directly ask their ideology due to ethical reasons and thus we can only obtain noisy text data including topics and words which are totally unrelated to violent Islamic extremism. Even in this situation, machine learning approaches are still effective to capture unobservable ideology and personality as explained below. Moreover, these statistical approaches are necessary to assure scientific objectivity because it rules out econometrician's subjectivity as much as possible.

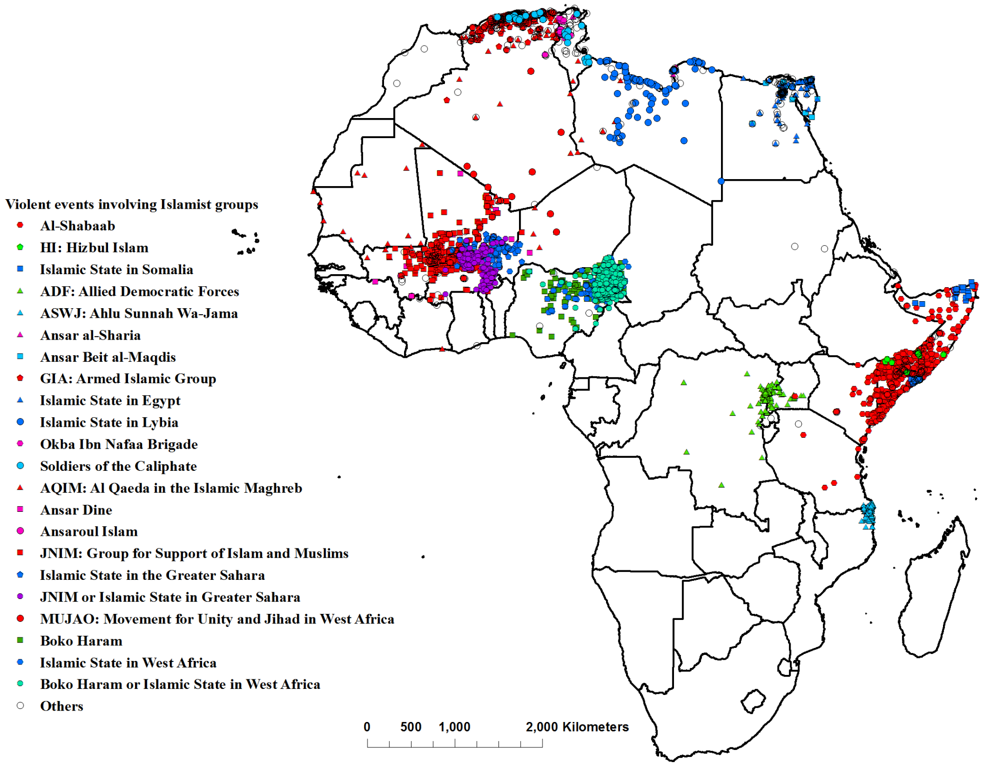

Figure 1: Islamist Violence in Africa from 2001-2019
(Source: Armed Conflict Location and Event Data Project)

2 Institutional Setting and Data

2.1 Al-Shabaab

Al-Shabaab is an Islamic extremist organization under the umbrella of the Al Qaeda network. It is also said that Al-Shabaab obtains financial support from the global Somali diaspora. Among militant groups, Al-Shabaab has caused the highest volume of fatalities in Africa over the past decade. Al-Shabaab is active in Somalia and beyond its borders (Kenya, Uganda, and Djibouti). Al-Shabaab adheres to an Islamist/Salafi/Jihadist ideology. Its stated goal is to "*topple the Somali government and establish an Islamic empire within the country guided by a strict reading of Shariah law*". Its estimated group size is about 7000-9000. Information in this subsection is partly based on Mapping Militant Organizations (2019).

2.2 Data Collection from Ex-Combatants at the Mogadishu Central Prison

We collect data from ex-combatants detained in the Mogadishu Central Prison (MCP henceforth; see Figure 2). The prison is administered by the Somali Custodial Corps and Prison Service, under the Somali government's Ministry of Justice. There are approximately 1,500 prisoners there, about half of whom are said to be Al-Shabaab ex-combatants. The MCP is the largest host of Al-Shabaab ex-combatants. We first define our sample and then discuss our target sample size.

2.2.1 Sample Definition and Characteristics

Our target sample is all Al-Shabaab ex-combatants detained in MCP who have less than two years until release and who agree to participate in our study. Our partner NGO (Accept International) contracts with the prison to conduct the DRR (De-radicalization, reinsertion, and reintegration; explained in Section 3) project for all such ex-combatants.

The ideal data for understanding violent Islamic extremism would be a representative sample of Al-Shabaab members. But this is infeasible. Ex-combatants detained in the MCP are not representative of Al-Shabaab members. We cannot observe Al-Shabaab members who died during fighting; those operating inside Al-Shabaab; those detained in other prisons; those with the highest profiles (such as those with very long prison sentences, or those sentenced to capital punishment); or defectors who voluntarily surrendered.

Nonetheless, our data is unique and valuable in its own way. Our sample consists of higher-risk individuals than are typically included in existing studies. For example, at least one recent study has collected data from Al-Shabaab defectors staying in rehabilitation centers.[7] But prisoners are different from defectors in key ways. In particular, defectors at rehabilitation centers are believed to be low-risk ex-combatants, while prisoners are regarded as a higher-risk population. This is because defectors have voluntarily abandoned their affiliations with Al-Shabaab, while prisoners were captured by state forces, and presumably did not want to reintegrate when they were captured. It is therefore plausible to assume that if these prisoners had not been captured, thay would have been more likely to be operating in Al-Shabaab. Defectors, in contrast, would have already been more de-radicalized and more willing to reintegrate into civilian life. As our research interest is to understand how to shift high-risk characteristics to low-risk ones, our newly-collected data is more suitable than existing data.

2.2.2 Variables

We collect three types of data. First, at the beginning of the study, we conduct a baseline survey to collect the ex-combatants' basic demographic information (e.g., hometown, age, clan, education, family information, etc), total prison terms, and remaining time until release.[8] Second, we collect recordings of conversations on non-sensitive topics during our counseling activities. The detailed procedure is described in Section 5. Third, several measures of other-regarding preferences and subjective expectations

[7] For example, Taylor et al. (2019) conducted interviews with 32 Al-Shabaab defectors at the Baidoa Rehabilitation Center in Southern Somalia. They point to the security concerns that defectors face, i.e., the risk of retaliation by Al-Shabaab, after graduating from the rehabilitation center as a key barrier to reintegration.

[8] It is ideal to obtain information on why they joined Al-Shabaab, their past activities when they belonged to Al-Shabaab, and how they were captured and ended up in prison as well. However, it is not realistic to ask these very sensitive questions in the prison. Eliciting true answers to such questions from the prisoners might also be infeasible. Therefore, instead of asking these sensitive questions, we use the total prison term as a proxy for the severity of past crimes that the ex-combatants committed.

are collected by a series of lab-in-the-field experiments. The detailed procedure is described in Section 6.

2.2.3 Sample Size and Power

There are about 700 Al-Shabaab ex-combatants in the MCP, with different lengths of time remaining until release. The DRR project by Accept International and our data collection will start with about 100-200 ex-combatants whose remaining time lengths until release are less than two or three years. As time passes, there will be new cohorts whose remaining periods become less than two or three years, to whom Accept International will start providing the program. We plan to collect data from them as well. Therefore, in theory, the minimum sample size is 100 and the maximum is about 700.[9]

We will calculate power and determine the target sample size after our pilot activities.

2.3 Data Collection from Civilians in Mogadishu

We also plan to conduct a survey targeting civilians in the city of Mogadishu. The main aim of the survey is to conduct a lab-in-the-field experiment to capture relationships between ex-combatants and civilians (see Section 6 for the experimental games). Our current plan to collect data from civilians via phone surveys, taking into account the security issue in Mogadishu.

As some of the games have sequential structures, we plan to use strategic methods to elicit preferences. Strategic methods allow us to randomly match prisoners and civilians and calculate payoffs *ex-post*. Then, we will pay the civilians through phone-based money transfer services after calculating payoffs. We are still considering the best ways to collect data in such an environment, and we may modify our plans through pilot studies.

3 RCT Intervention

3.1 De-radicalization, Reinsertion, and Reintegration (DRR) Project

We evaluate impacts of a new generation of conflict resolution programs, called **De-radicalization, Reinsertion, and Reintegration (DRR)** project. **Disarmament, demobilisation and reintegration (DDR)** is a major strategy for successful peacekeeping and has been commonly employed by UN Peacekeeping Operations following civil wars[10]. Yet, the feasibility to apply these methods for ongoing conflicts involv-

[9] These lower and upper bound sizes may also increase, because new captured ex-combatants may come to the prison subsequently as well. However, as it is almost impossible to estimate how many new ex-combatants will come later, we are ignoring this possibility for now. Moreover, there may be some high-profile ex-combatants whose prison lengths are very long, say, twenty years. It is not realistic to collect data from these prisoners. If there are many such prisoners, then the upper bound of the sample size would be significantly smaller than 700.

[10] Impacts of DDR are recently analyzed in the field of political science (e.g., Humphreys and Weinstein 2007, which find little evidence of DDR on demobilization and reintegration in Sierra Leone). As noted in Matanock (2020), most evaluations on DDR programs relied on case studies and results are mixed. Gilligan et al. (2013) is an exception utilizing quasi-experimental situation in Burundi, showing large decline in poverty and improvement in livelihood due to DDR.

Figure 2: Mogadishu Central Prison

ing Islamic extremist organizations, including Al-Shabaab, is limited (Muggah and O'Donnell 2015). It is unrealistic to satisfy the preconditions for initializing the DDR, like a peace agreement. Indeed, peaceful negotiation with leaders of such radical terrorist organizations is difficult and has not been fruitful in most cases.

In this environment, incentivizing individuals to abandon their affiliations with such groups may be a more feasible alternative. Promoting individual-level de-radicalization and reintegration is key for reducing violence by extremist groups for the following reasons. If it is difficult for ex-combatants to reintegrate into a society, then the risk of their re-radicalization and re-joining an extremist group is increased. Moreover, a civil society facilitating re-integration process might also incentivize combatants of a violent groups to surrender at their own initiatives.

With this motivation, our study partner, Accept International [11] , implements a new program in Somalia: **De-radicalization, reinsertion, and reintegration (DRR)** project. Some components of its program are constructed reflecting the qualitative findings in Nagai (Forthcoming)[12]. The detailed program of the DRR project consists of the following components:

(I) Rehabilitation Counseling

We conduct frequent counseling activities, consisting of individual- and group-level counselings:

- Care Counseling (1 on 1):

  The NGO will listen to ex-combatants' voices and then build strong relationship with them to understand their needs and to increase their motivation for their actions for reintegration and long life after their release from the prison. Moreover, in this counseling, the NGO discusses with ex-combatants about the difference between general Islamic doctrine and Islamic extremism and about the life with general Islamic doctorine[13]. Recordings, whose text data will be used to estimate detainees' ideologies, are obtained from this form of counseling.

- Disillusionment Management (group session):

  The risk of re-radicalization can increase when ex-members face the reality that things did not work out as they had expected. In this session, we will discuss problems that they will face after their release from the prison and consider solution ideas together so that the risk of re-radicalization can be minimized.

[11] https://accept-int.org/en/. Accept International specializes in eradicating terrorism and resolving armed conflicts. The NGO has nearly 10 years' experience of conducting their projects (in Somalia, Kenya, and Indonesia) and building trust with local governments to de-radicalize criminal gangs and ex-combatants from extremist groups. They also collaborate with the UN for some projects. It is noteworthy that they have contributed to de-radicalize and peacefully demobilize Somali gangs in Nairobi.

[12] Nagai (Forthcoming) conducted qualitative interviews with disengaged Al-Shabaab combatants and ordinary civilians in Mogadishu for examining necessary conditions of successful reintegration.

[13] The NGO is not in charge of the whole set of formal religious education. Rather, this discussion aims for complementing the formal religious education with an imam, which is provided to all prisoners.

Counseling sessions are going to be conducted twice a week from the beginning of the DRR program until the time of release from the prison. In the current plan, all of these three components are cyclically and regularly implemented.

(II) Reconciliation Dialogue (group session with society representatives)

Lack of mutual understanding is always one of obstacles for reconciliation. We will invite some community leaders and hold dialogue sessions so that ex-members can understand their ideas and what is needed for reconciliation, and of course, that community leaders also deepen understanding of the reality of ex-members of Al-Shabaab. This dialogue session is going to be held about once a month.

(III) Follow-up for Reinsertion and Reintegration (1 on 1) (release phase)

One month before release, we will implement job management counseling (1 on 1). Considering 67% the youth unemployment rate, we support various things to deal with their economic difficulty. We start with job counseling to learn the skills they have now and how to develop them. We follow by making their CVs. We also provide a reference letter with them and endeavor negotiations with a potential employer. Additionally, we create an opportunity to develop essential communication skills for job hunting.

At the time of release, we will provide a reintegration kit, including mobile phones and emergency contact lists, and establish a system for communicating with them after they have completed the program or released. We will accept them back in the case of an emergency, and they may use this center as a shelter that they may visit and stay at anytime.

We are currently discussing what additional experimental variations we will create at the release phase for our future research projects.

3.2 Procedure of the RCT Intervention

Due to the mission of Accept International and their contract with the Somali Government, Accept International provides the DRR program to all the ex-combatants before they are released. In order to obtain causal effects of each component of the DDR program, we randomize the order of providing treatments to the ex-combatants. In addition to the recordings in each 1 on 1 care counseling, several outcome measures are obtained by lab-in-the-field experiments. Some randomly-chosen ex-combatants receive treatments before the lab experiments, while others receive them after the experiments.

Figure 3 provides a simple illustration of the RCT intervention design. We take the following steps to sample the ex-combatants and to assign treatment status to them with stratified randomization:

Step 0. At the beginning of the study, we list up ex-combatants whose remaining times are less than two years until their releases from the prison.

Step 1. We construct strata of ex-combatants based on their remaining times until releases from the prison within a range of one month.

Step 2. We prepare three phases from the beginning of the study. Some outcome measures by lab-in-the-field experiments will be obtained between Phase 1 and Phase 2. If capacity allows, we collect outcome measures by lab-in-the-field experiments again after all phases.

Step 3. We randomly divide ex-combatants in the same stratum into three groups, C, T1, and T2. Different groups obtain different interventions in each phase. Refer to Table 1 for detail.

Using the first set of outcome measures between Phase 1 and Phase 2, our RCT design allows us to obtain the following causal effects. We obtain the impact of the counseling treatment by comparing outcomes between C and T1. We obtain how that impact is enhanced with the reconciliation dialogue by adding the comparison of outcomes between C and T2. We obtain the impact of reconciliation dialogue among counseling takers by comparing outcomes between T1 and T2.

Using the second set of outcome measures after Phase 3, our RCT design allows us to conduct the following additional investigations. We can test the persistence of the impacts of the counseling and reconciliation dialogue treatments by comparing outcomes between C and T1. We can test the effectiveness of the reconciliation dialogue at the later stage of conseling relative to that at the earlier stage by comparing outcomes between T1 and T2.

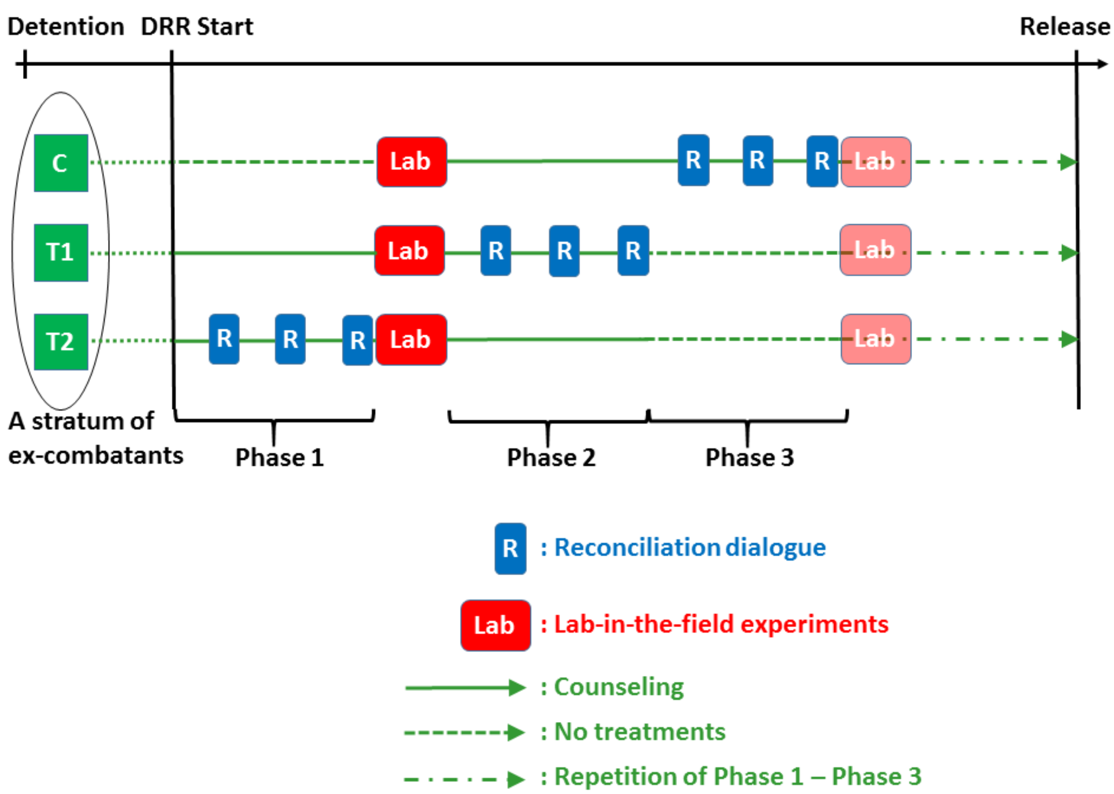

Figure 3: Simple Illustration of RCT Intervention

Table 1: Procedure of Randomizing Orders of Treatments

Groups	Phases		
	Phase 1	**Phase 2**	**Phase 3**
C	No treatments	(I) Counseling	(I) Counseling + (II) Reconciliation dialogue
T1	(I) Counseling	(I) Counseling + (II) Reconciliation dialogue	No treatments
T2	(I) Counseling + (II) Reconciliation dialogue	(I) Counseling	No treatments

4 Conceptual Framework: Inducing De-Radicalization and Reintegration from Violent Islamic Extremism

Promoting individual-level de-radicalization and reintegration is important to reduce violence by extremist groups for three reasons. First of all, peaceful negotiation with leaders of such groups is difficult and in many cases has not proven fruitful[14]. In this environment, incentivizing individuals to abandon their affiliations with such groups may be a more feasible alternative. Second, if it is difficult for ex-combatants to reintegrate into civil society, then the risk of them re-radicalizing and re-joining an extremist group is increased. Third, a re-integration process facilitated by civil society might also incentivize combatants of violent groups to surrender at their own initiatives.

However, measuring the actual reintegration process is a challenging task in our context, as we would need to wait a long time to observe the release of these ex-combatants, and even if we could wait, it might be difficult to track all of them. Instead, in this paper we propose to measure key *pre-conditions* for achieving successful de-radicalization and reintegration when ex-combatants are released from the prison.[15]

In the first subsection, we describe key concepts regarding the pre-conditions of de-radicalization and reintegration. In the second subsection, we discuss potential forces to shape these pre-conditions. In the third subsection, we briefly discuss the interlinkage and complementarity between the key concepts specified in the first subsection.

4.1 Key Concepts

Of the five concepts introduced below, we regard (A) – (C) as especially key to this study. Depending on the local situation, we may also investigate (D) – (E).

(A) Radicalization toward violent Islamic extremism

It is difficult to explain conflicts involving Islamic extremist organizations using existing economic models of conventional warfare in which monetary incentives play significant roles. Indoctrination of members in violent Islamist organizations toward jihadist ideology is prevalent to enhance organizational power. If the ex-combatants are radicalized toward violent extremism or Islamic extremism ideology,

[14] The signing of the U.S.-Taliban Agreement on February 2020 is an exceptional case.

[15] As a separate research project, we will also attempt to track ex-combatants and observe actual reintegration process in the long run.

their de-radicalization is thus key for mitigating the risk of recidivism after release from prison. Section 5 describes our methods to quantify the degree of radicalization.

(B) Ex-combatants' (civilians') other-regarding preferences toward civilians (ex-combatants)

This is a key pre-condition for successful reintegration of ex-combatants into society. If ex-combatants have negative preferences towards civilians such that they are not willing to cooperate or reciprocate, building peaceful relationships between these two groups is going to be challenging. A utility function which incorporates other-regarding preferences can be expressed as $U(y_i, y_{-i}) = u(y_i) + v(y_i, y_{-i})$ where y_i is the own resource allocation and y_{-i} is the others' allocation. One tractable way of capturing these relationships is a well-known functional form proposed by Fehr and Schmidt (1999):

$$U(y_i, y_{-i}) = y_i + \sum_{j \neq i} (-\alpha \max\{y_j - y_i, 0\} - \beta \max\{y_i - y_j, 0\}) \tag{1}$$

Applying Iriberri and Rey-Biel (2013)'s method enables us to classify social preferences into four categories summarized in Table 2.

Table 2: Summary paramters and social preferences

Types	Selfish	Inequality averse	Positive altruistic	Competitive
Parameters	$\alpha = 0$ and $\beta = 0$	$\alpha > 0$ and $\beta > 0$	$\alpha < 0$ and $\beta > 0$	$\alpha > 0$ and $\beta < 0$

If an individual is selfish ($\alpha = 0$ and $\beta = 0$), he never cares about others' wealth. If an individual is inequality averse ($\alpha > 0$ and $\beta > 0$), his utility increases as the wealth gap shrinks. In this case, α and β are parameters for envy and guilt, respectively. If an individual is positive altruistic $\alpha < 0$ and $\beta > 0$), others' gains always make him happy. If an individual is competitive ($\alpha > 0$ and $\beta < 0$), he always feels happy when he hurts others. Classifying other-regarding preferences is crucial in our context to capture the relationship between ex-combatants and civilians. For instance, if ex-combatants prefer to hurt civilians and prefer to benefit those sharing a similar ideology, we would observe $\alpha > 0$ and $\beta < 0$ when they interact with civilians, and $\alpha < 0$ and $\beta > 0$ when they interact with other ex-combatants in the prison. Section 6.1 describes how to estimate these preference parameters.

(C) Beliefs and cooperative behavior

Cooperative behavior between ex-combatants and civilians is another key precondition for reintegration. Though several factors might affect cooperative behavior in general,[16] we focus on subjective beliefs about others' behaviors and other-regarding preferences (introduced above) as primary mechanisms driving cooperation (Fischbacher and Gachter 2010). Our argument partly aligns with a theoretical argument

[16]Other factors outside this paper's focus include, for example, strong reciprocity (e.g., Palfrey and Prisbrey 1997; Andreoni 1995) and cultural norms (e.g., Henrich 2004; Gächter and Herrmann 2009).

by Rohner et al. (2013), which discusses how cooperation is hindered by incomplete information and inaccurate beliefs about others, thereby driving a cycle of war.

Even if both ex-combatants and civilians are willing to cooperate with each other to engage in productive economic activities, if (for instance) ex-combatants have negative beliefs that civilians are not cooperative, ex-combatants may give up on cooperation. Or, if ex-combatants have overconfidence that civilians will be cooperative, then the risk of re-radicalization after release might increase due to a gap between their expectations and the reality in a society.[17] Since ex-combatants are located in the prison, opportunities to interact with civilians are limited, as is information about the societies into which they will reintegrate. Therefore, both ex-combatants and civilians engage in highly internal processes of belief formation without actually observing the objects of those beliefs (at least while ex-combatants are detained in the prison). As a consequence, they might have inaccurate beliefs about the attitudes of others. Focusing on subjective beliefs about others will thus allow us to observe a potential obstacle to reintegration. Section 6.2 describes how we plan to elicit subjective beliefs about others.

(D) Ex-combatants' subjective expectations about income

Perceptions about future income also play an important role in the reintegration process. Excessively low subjective expectations about income might create a risk of re-radicalization. Excessively high subjective expectations might foster disillusion upon release. Both cases could become obstacles to reintegration. Section 6.4 describes how we plan to elicit subjective expectations about income.

(E) Preferences for anti-social behavior

Violent activities are specific cases of criminal activities. Obviously we cannot measure actual violent activities in the lab. We instead measure preferences for anti-social behavior as a rough proxy for the underlying preferences that may be conducive to violence. Section 6.5 describes potential lab-in-the-field games to proxy these concepts.

4.2 Primary Mechanisms toward De-Radicalization and Reintegration

We propose three types of interventions (rehabilitation counseling; intergroup contact in the form of reconciliation dialogue; information provision) to *directly* shape ideologies, preferences (especially other-regarding preferenes), and beliefs, which correspond to concepts (A) – (C) above.

[17] Indeed, from qualitative interviews with disengaged combatants of Al-Shabaab in Mogadishu, Nagai (Forthcoming) reports a significant gap between ex-combatants' expectations and community members' attitudes toward them as a major obstacle behind reintegration.

4.2.1 Shaping ideologies, preferences, and beliefs by the rehabilitation counseling

One intervention is repetitive and frequent rehabilitation counseling, consisting of both individual- and group-level activities, over time. This first intervention corresponds to treatment (I) in section 3[18]. This intervention aims to directly shape ex-combatants' intrinsic motivations and re-define new identities to help them reintegrate into civil society. Rehabilitation counseling could directly influence the degree of radicalization and subjective beliefs on others' behavior.

Radicalization toward violent Islamic extremism.—Individual-level care counseling addresses each prisoner's needs and helps him to re-define a mindset for transitioning from life with a violent group to life in a civil society. The concrete discussion about the life in a civil society after release from the prison could shape one's ideology away from violent Islamic extremism. Moreover, the discussion about the ideological difference between general Muslim doctrine and Islamic extremism could directly facilitate this re-definition of one's identity toward de-radicalization from extremism.

Other-regarding preferences.—Throughout the counseling sessions, ex-combatants discuss life in civil societies. Acquiring new information regarding an outside society in general may change preferences of interacting with civilians, away from competitive. Literature mainly in social psychology discusses debiasing in the absence of direct contacts. This stream of literature includes listening stories on role models, experts' opinions, perspective taking, etc (surveyed in Bertrand and Duflo 2017). Since the counseling intervention involves various types of informaiton, we do not focus on particular mechanisms. Yet, as a principle aim of the counseling, counselors are most likely to provide information which should increase familiarity with a civil society.

Beliefs about cooperative behaviors.—Group-level disillusionment management aims to narrow the gap between ex-combatants' expectations about life after release and the reality that they will be likely to face. This intervention is expected to directly influence ex-combatants' subjective beliefs about civilians' cooperative behaviors (measured by our lab-in-the-field experiments, described in section 6). Whether ex-combatants' expectations about the propensity of cooperation by civilians should increase or decrease is not clear. The direction of change may depend on initial expectations, which may differ heterogeneously across prisoners with different severities of past activities. We could instead expect that the accuracy of beliefs (the gap between subjective beliefs and the actual propensity to cooperate by civilians measured in the lab game) improves by this intervention. This argument also implies that the variance of ex-combatants' subjective beliefs about civilians cooperative behaviors could decrease by this intervention.

4.2.2 Shaping ideologies, preferences, and beliefs, reinforced by the reconciliation dialogue

Another intervention is to enhance mutual understandings between civilians and ex-combatants through direct interactions. This corresponds to treatment (II) in section 3. This intervention is distinct from the previous counseling intervention in that it involves the direct interaction with civilians from the outside

[18]This approach has similar motivations and characteristics as the Cognitive Behavioral Therapy (CBT) recently conducted by Blattman et al. (2017) with high-risk men in Liberia.

society, and is designed to disentangle its additional effects. Reconciliation dialogue could reinforce the effects of the rehabilitation counseling on the degree of radicalization, other-regarding preferences, and subjective beliefs about others' cooperative behaviors.

Radicalization toward violent Islamic extremism.—Given that the ex-combatants live in the prison, this intervention is the only opportunity for them to interact with civilians (except with the NGO and prison workers). That is, the ex-combatants interact with the identity close to what the rehabilitation counseling attempts to make them have. This force could thus work toward de-radicalization by inducing ideological *conformity* as an additional channel.

Other-regarding preferences.—The ex-combatants have opportunities to meet community representatives several times. Different people are invited to the prison across sessions. Beyond preferences toward specific community representatives, this sequence of interacting with community representatives is expected to shift the ex-combatants' other-regarding preferences toward civilians *in general* away from the competitive category. This expected changes are in line with the contact hypothesis (Allport 1954) in social psychology, stating that positive interpersonal contact is an effective method to reduce prejudice[19]. This may further enhance cooperative behaviors between ex-combatants and civilians.

Beliefs about cooperative behaviors.—Reconciliation dialogue aims to promote mutual understandings between ex-combatants and civilians. One dimension of the mutual understanding is that the ex-combatabnts understand how general civilians view them. This understanding could reform beliefs on civilians' cooperative behaviors. Therefore, the reconciliation dialogue could reinforce the effects of the rehabilitation counseling on beliefs in the same direction through the similar mechanism.

4.2.3 Shaping preferences and beliefs by information provision

We have been focusing on ex-combatants' behaviors so far. At the same time, civilians' attitudes toward ex-combatants also play important roles for reintegration. Indeed, qualitative interviews by Nagai (Forthcoming) reveal that many civilians have fears about accepting ex-combatants in their communities. In some cases civilians express less fear when they know that ex-combatants have undergone any sort of de-radicalization or reintegration programming. We propose information provision to both civilians and ex-combatants for removing each barrier specific to each side. These interventions correspond to information treatments in our lab-in-the-field experiments described in section 6.3.

Civilians' other-regarding preferences toward ex-combatants.—First, in order to investigate whether civilians' attitudes towards ex-combatants differ from their attitudes towards other civilians, we will vary the information about the identity of civilians' partners in our lab games. Civilians do not know that their partners are ex-combatants in one arm, while they know that information in another arm. Second, we provide civilians with information that their partner ex-combatants have completed the two interventions (rehabilitation counseling and reconciliation dialogue). This will allow us to test whether civilians respond positively to the knowledge that ex-combatants have undergone de-radicalization and reintegration training.

[19]See Pettigrew and Tropp (2000) for the meta-analysis of testing contact hypothesis. Among 515 studies reviewed in the paper, 94% of them found positive impacts of inter-group contacts on prejudice.

Civilians' beliefs about ex-combatants' cooperative behaviors.—A similar argument as above applies here.

Ex-combatants' beliefs about civilians' cooperative behaviors.—First, we randomly inform ex-combatants that their paired civilians know that civilians are paired with prisoners. For another arm, ex-combatants do not recieve such information. We examine ex-combatants' perceived barriers stemming from their identity. Second, among those who recieve the first information, we randomly inform ex-combatants that thier paired civilians know that prisoners completed the two treatments (rehabilitation counseling and reconciliation dialogue). We eximine whether the perceived barrier as an ex-combatants is mitigated by knowing civilians' knowledge on treatments.

4.2.4 Shaping preferences and beliefs by empathy effect

This intervention corresponds to arm (d) in 6.3, narratives of civilians and ex-combatants. The detailed argument may be added later.

5 Quantifying Radicalization toward Violent Islamic Extremism by Natural Language Processing Approaches

This section describes natural language processing (NLP) approaches to quantify[20]:

(A) Degree of radicalization toward Islamic extremism.

The degree of radicalization toward Islamic extremism is defined as the similarity of word usage patterns with actual Islamic extremists. NLP approaches aim to estimate to what extent each subject supports the extreme ideology or terrorism using non-sensitive text information which is fully unrelated to terrorism or Islamic extremism. Explicitly measuring such ideological extremism from unrelated text information is a challenging task. We employ three approaches to overcome this challenge. In the first and second approaches, we rely on a deep-learning-based classification while the third approach employs a non-supervised machine learning method.

5.1 The First Approach: An Application of Non-censored Conversation Data

Our first approach takes three steps: model construction, a validation test, and similarity calculation. First, we construct a deep-learning-based model to statistically calculate how similar a given statement is to the statements made by an extremist in Twitter. Second, we check the validity of this model by investigating whether it works for statements which are unrelated to extremism: does it precisely detect the extremists' statements even when they talk about non-sensitive topics such as family issues or entertainments? Finally, we put our subject's conversation data collected in the counseling treatment into the model and obtain the statistical similarity of statements between each subject and extremists.

[20] Other approaches for measuring radicalization are described in Appendix B.

5.1.1 Model Construction to Identify Extremists

Preparation of Datasets. We prepare two datasets: the Twitter dataset and the Dark Web dataset. The Twitter dataset is used for construction of deep-learning-based model. The Dark Web dataset is complementarily used to assign a tag, i.e., an extremist or a non-extremist, to the Twitter dataset. First, using Twitter streaming API, we scrape tweets containing one or more extremism-related keywords (ISIS, bomb, suicide etc) and acquire the Twitter dataset[21]. Note that this dataset potentially contains statements made by both *true* extremists and *false* extremists (or non-extremists) and each tweet does not have any explicit label or tag disentangling extremists from non-extremists. We also collect text information from Dark Web forums using the same keywords adopted in construction of the Twitter dataset and obtain the Dark Web dataset. It is worth noting that the Dark Web dataset surely consists of statements made by *true* extremists only.

Training Dataset Construction. We will assign a label of either an extremist or a non-extremist on each tweet in the Twitter dataset by applying computational text analysis methods. More specifically, we use probability latent semantic indexing (PLSI) to calculate the statistical consistency between each tweet and statements in the Dark Web dataset[22]. If a given tweet is closely similar to statements in the Dark Web dataset which are surely made by extremists, we assign an extremist-tag for that tweet and otherwise we assign a non-extremist-tag.

Model Construction and Tuning of Parameters. We use the tagged Twitter dataset as a training data and apply the long short-term memory with Convolutional Neural Network (LSTM-CNN) model to obtain the parameters that classify unknown statements into an extremist and a non-extremist. In other words, the model is shaped to quantify the probability that each target's statement is made by extremists.

A Difference with Previous Literature. These procedures are identical with the sentiment analysis technique developed by Ahmad et al. (2019) except for the way to assign a tag on either an extremist or a non-extremist in the construction of training dataset. More precisely, Ahmad et al. (2019) manually assigns a tag for each tweet while we use the PLSI method to do that. We automate the tag assignment process by the computational text analysis method to improve the practical feasibility of checking how sensitive the parameters are to choices of statements from Dark Web forums in construction of the Dark Web dataset. For example, we will investigate how parameters in the model change when we strictly focus on Islamic extremism rather than other religious extremism. To implement such sensitivity analyses, manual assignment of an extremist tag entails a tremendous time cost.

5.1.2 A Validation Test

The LSTM-CNN model constructed by the above method is designed to classify the statement which is related to extremism or terrorism into an extremist tag or a non-extremist tag. Therefore, we need to check whether or not the model also works for statements which are related to neither extremism nor terrorism. Remember that, in the tagged Twitter dataset, we have an extremist tag and a non-extremist

[21] We will use the words' list which is same with Ahmad et al. (2019)

[22] A detailed statistical model on the PLSI is available in Appendix A.

tag. Focusing on an extremist tag only, we can extract past tweets which are fully unrelated to extremism or terrorism but surely made by extremists. Now we put these past posts into the model and calculate the probability that an extremist tag is appropriately assigned to these posts. If the probability is plausibly high, then we consider that the model can also work for statements unrelated to extremism.

5.1.3 Similarity Calculation between Subjects and Extremists

We apply the LSTM-CNN model with non-censored conversation data obtained from counseling treatments. Note that the model is designed to calculate the probability that an input (i.e., a set of statements) is made by extremists. Namely, we can get a continuous variable on how likely each subject is to have extreme ideology in a sense that each subject's statements are consistent with ones held by extremists in Twitter.

5.2 The Second Approach: An Application of Censored Conversation Data with Pre-Identified Non-Sensitive Topics

In the first approach, we directly apply the LSTM-CNN model to non-censored conversation data. This approach might not work if such data is noisy and it does not have sufficient variations to be classified into either an extremist tag or a non-extremist tag by the model. We take a different method to deal with this problem. Our second approach is identical with the first approach except for the final step of similarity calculation: we will use censored conversation data rather than non-censored one collected from the counseling treatment. Specifically, in order to increase the model precision, we identify non-sensitive topics in which extremists and non-extremists in the tagged Twitter dataset use statistically different word patterns, and then ask questions related to these topics in the counseling treatment. This censored data contains sufficient variations such that the model precisely classifies each conversation into an extremist-tag and a non-extremist tag.

5.2.1 Identifying Non-Sensitive Topics

This subsection explains how we identify the non-sensitive topics which are fully unrelated to Islamic extremism in the tagged Twitter dataset. First, we apply the Latent Dirichlet allocation (LDA) model or topic model to obtain topic distributions in all tweets in this dataset. Technically, we characterize each text into simple topic structures by reducing its dimension. These topics should consist of sensitive topics which are directly related to extremism and non-sensitive topics which are not related to it such as family issues, entertainments, and general social issues etc. Second, we pick up specific topics among non-sensitive ones in which a pattern of word usage is statistically different between extremists and non-extremists. For instance, suppose that the data leads us to choose the topic related to family issues. This topic selection means that when Twitter users discuss about this specific topic, a pattern of word usage is clearly different between extremists and non-extremists such that the former use aggressive words to protect his or her family while the latter use more moderate expressions. Therefore, we can more easily quantify the similarity of a given text information collected from our subject to tweets posted by either extremists or non-extremists, compared to the first approach. Finally, if a set of statements made by a subject is closer to those made by extremists than other subjects, the model should predict that the probability that this subject holds the same ideology with extremists is higher than others.

5.2.2 Obtaining Text Data about Non-Sensitive Topics

Next, we proceed to acquire conversation data for each subject which is recorded during the counseling treatments. During these treatments, we ask general questions related to the non-sensitive topics such as family issues or entertainments etc. In order to estimate the degree of radicalization and its change, these questions should satisfy two conditions: they should be identical over time and they should also induce a set of words which are enough to identify the relative closeness to tweets contents posted by extremists.

5.2.3 Similarity Calculation between Subjects and Extremists

Finally, we apply the LSTM-CNN model with the censored conversation data obtained from the counseling treatment to estimate the statistical similarity of word usage patterns between subjects and extremists. This process is exactly same with the first approach except for using censored data instead of non-censored one.

5.3 The Third Approach: A Non-Deep-Learning-Based Model

The first and second approach might not work if the deep-learning process suffers from overfitting problems. In such cases, the third approach has an advantage over the two approaches. The third approach relies on a non-supervised machine learning algorithm instead of the deep-learning-based model. Its procedure is exactly same with the second approach until the step of obtaining text data using questions related to non-sensitive topics in the counseling treatment.

The difference between this approach and the second approach is that the former calculates similarity in the sense of consistency in topic distributions estimated by the LDA model through the cosine similarity estimation, while the latter calculates it through the deep-learning-based model. In other words, in the third approach, we compare the topic distributions of a given subject's statement with the topic distributions of non-extreme tweets made by extremists. If topic distributions overlap each other, it implies that a subject's statement is statistically similar to extremists' tweets because they elicit same word usage patterns in a given topic[23].

[23]This approach is widely used in the field of political science. For instance, Laver et al. (2003) adopt non-supervised machine learning methods to estimate politician's positions and their similarity to others using the manifesto data in Britain and Ireland. Statistical methods to assess the similarity are described in Appendix A.

5.4 Model Selection

So far, we propose three different approaches to estimate the degree of radicalization toward Islamic extremism using text data. This subsection explains the way of evaluating relative performances of the three approaches and the criteria for our model selection. Essentially, we create "pseudo" censored conversation data, relying on Twitter only. This pseudo data consists of tweets which are related to non-sensitive topics identified in the second approach and surely made by extremists. This pseudo data enables us to evaluate each approach's performance: as long as we put exactly same data into the different algorithms developed by the three approaches, respectively, the estimated probabilities are comparable such that the algorithm returning the highest probability that these tweets are made by extremism should have the best performance.

Exact procedure is as follows. First, we construct a new dataset that contains tweets which are fully unrelated to extremism but surely posted by extremists. Technically, using Twitter API, we randomly scrape 10,000 tweets if each tweet includes at least one name of the leaders of extremist organizations and it contains positive semantic words in neighborhood of this figure's name. We assume that a Twitter user who made such a tweet can be considered as supporting extremism. Second, we need to replicate the censored conversation data, which is used in the second and third approaches, from these 10,000 tweets. We extract past tweets posted by each user only if they include specific words which are frequently used in the non-sensitive topics. This set of past tweets composes the pseudo censored conversation data. Finally, we calculate the probability that these tweets are made by extremists using three different approaches proposed above. The approach returning the highest probability is considered as the best approach in this context.

5.5 Empirical Concerns

One empirical concern about our NLP approaches for measuring radicalization would be that there are alternative explanations for similarity in speech patterns that have nothing to do with radicalization. For example, other factors affecting similarity in speech include vernacular expressions.

We avoid such confounding effects in the following three ways. First, we will collect the data from a broader geographical range of Dark Web forums which should entail highly general text information. Second, when we examine the treatment impacts on radicalization, we will focus on how the similarity between a subject and extremists changes over time. Third, the deep-learning based model can control for fixed effects as in reduced forms. These fixed effects absorb confounding factors and thus we only rely on within-pair (a subject and extremists) time-variant variations.

6 Lab-in-the-Field Experiments

Lab-in-the-field experiments are designed to capture the following four components:

(B) Ex-combatants' (civilians') other-regarding preferences toward civilians (ex-combatants)

(C) Cooperative behavior, especially ex-combatants' (civilians') subjective beliefs about civilians' (ex-combatants') behavior

(D) Subjective expectations on income after releasing from the prison

(E) Preferences for anti-social behavior

Games in subsections 6.1 attempts to capture (B) and some aspects of (E). Modified version of the public goods game presented in subsection 6.2 captures (C). Subsection 6.4 describe the way to capture (D). In subsection 6.5, we also list additional games that may be able to capture important aspects of (E). These divisions also facilitate our adjustments depending on implementing constraints in the field. If it is difficult to conduct games for general civilians outside the prison in Mogadishu (due to a security reason etc), then we will implement games only in subsections 6.1 and 6.5 with the prisoners.

6.1 Capturing Social Preferences: (Sequential) Self-Other Allocation Games

To capture social preferences, we follow and extend games conducted in Chen and Li (2009). Self-other allocation games consist of two-person dictator games and two-person response games. Table 3 illustrates the payoff structure of these games drawing from Chen and Li (2009). Individuals are notified that they are matched with an ex-Al-Shbaab prisoners or an citizen living in Mogadishu. Matching partner is reshuffled game by game. Current payoff structures are just a duplication of Chen and Li (2009), and we may modify payoffs and number of games depending on the environment.

In the two-person dictator games (Panel A), individuals assigned as "person B" allocate resources between own and "person A" from two options. These dictator games allow us to estimate parameters representing other-regarding preferences. We plan to add choices to capture **destructive behaviors**. B may hurt or benefit himself and destruct all the A's profits. "Dict new 1" and "Dict new 2" in Table 3 are examples that B benefits himself by destructing A's profit, and "Dict new 3" and "Dict new 4" are examples that B hurts himself by destructing A's profit.

For a few dictator games in Panel A, we also ask prisoners (both those who are assigned to person A and B) to hide their identities by giving up some of their payoffs. By doing so, they can replace their identity to "someone living in Mogadishu" or complete anonymous. This treatment aims to measure **perceived value** of their identity as prisoners. We use Becker-DeGroot-Marschak mechanism (BDM mechanism; Becker et al. 1964), which is frequently used to elicit truth-telling WTP. In the BDM mechanism, one can bid for an option or an item (in our case, hiding identities). Then, the price of the option is randomly drawn by an experimenter, and the bidder will pay the drawn price if the drawn price is lower than the bidding price. Theoretically speaking, prisoners may hide their identity in a dictator game if expected payoffs by hiding their identity are higher than what they pay. By realizing their "cost" of their identity through treatments, they may adjust their WTP for hiding their identity. If prisoners are over-confident on their identity, believing that they would be accepted by a society, their WTP are expected be higher after treatments.

Table 3: Payoff tables of self-other allocation game

	A stays out	If A enters, B chooses
Panel A: Dictator games		
Dict 1		(400,400) vs (750,400)
Dict 2		(400,400) vs (750,375)
Dict 3		(300,600) vs (700,500)
Dict 4		(200,700) vs (600,600)
Dict 5		(0,800) vs (400,400)
Dict new 1		(400,400) vs (0,410)
Dict new 2		(400,400) vs (0,600)
Dict new 3		(400,400) vs (0,390)
Dict new 4		(400,400) vs (0,200)
Panel B: Response games: B's payoffs identical		
Resp 1a	(750,0)	(400,400) vs (750,400)
Resp 1b	(550,550)	(400,400) vs (750,400)
Resp 6	(100,1000)	(75,125) vs (125,125)
Resp 7	(450,900)	(200,400) vs (400,400)
Panel C: Response games: B's sacrifice helps A		
Resp 2a	(750,0)	(400,400) vs (750,375)
Resp 2b	(550,550)	(400,400) vs (750,375)
Resp 3	(750,100)	(300,600) vs (700,500)
Resp 4	(700,200)	(200,700) vs (600,600)
Resp 5a	(800,0)	(0,800) vs (400,400)
Resp 5b	(0,800)	(0,800) vs(400,400)
Panel D: Response games: B's sacrifice hurts A		
Resp 10	(375,1000)	(400,400) vs (350,350)
Resp 11	(400,1200)	(400,200) vs (0,0)
Resp 12	(375,1000)	(400,400) vs (250,350)
Resp 13a	(750,750)	(800,200) vs (0,0)
Resp 13b	(750,750)	(800,200) vs (0,50)
Resp 13c	(750,750)	(800,200) vs (0,100)
Resp 13d	(750,750)	(800,200) vs (0,150)

Payoff tables are from Chen and Li (2009). We adjust some numbers and add new variations. Indicating "new" for the game labels (e.g. "Dict new 1") are newly added variations. Payoffs are in experimental currency unit, converted to real currency in the end of our experiment.

Response games (Panel B to D) capture additional behaviors. The game has two stages. Person A is a first mover, who decide to stay out the game or let Person B to choose allocations. If player A stays out, then payoffs for both players are determined at that point. For games in Column B, a person B's payoffs are identical conditional on a person A decides to enter. Person B needs to sacrifice his payoffs to benefit person A in panel C, and person B needs to sacrifice his payoffs to hurt person A in panel D. From these person B's response, we can analyze preferences for rewards and punishment. Importantly, response games with punishment can be also interpreted as **retaliatory behavior** in our context. Choice Resp 13a is a good example. Person B could gain 750 if A stays out, but if A decides to enter, B could get 200 or 0. B may "retaliate" A's payoff by hurting himself to punish A's unfriendly behavior.

6.2 Capturing Cooperative Behavior and Beliefs: Public Investment Game (PIG)

Setup

We design a modified version of public goods game. We name it "public investment game" (PIG). The main goal of this game is to analyze cooperative behaviors, with a particular focus on subjective expectations over others' behavior. In addition, we also analyze cooperative behavior through the game. Note that, in order to focus on subjective belief as the primary mechanism driving cooperation, we explicitly eliminate free-riding problem of standard public goods game.

We create groups, each of which consists of 5 players. Each game is played with these 5 players by each group. Among 5 players in each group, there are N_d are ex-combatants and N_c ordinary civilians. N_d and N_c *exclude* the player himself, and vary by groups. Players are endowed with 100 tokens (that can be transformed to cash or some other forms of compensations). Players observe other members' identities (i.e. ex-combatants or citizens), but do not know any additional information. Players do not interact in person[24].

Experimental Procedure

The game is played in two steps.

Step 1. Each participant decides whether he keeps a token or invests it to the group.

If more than or equal to three participants have invested, the participants who have invested gains the three-times as high amount of money as the amount that they have invested. If less than three participants have invested, the money invested will be lost.

Step 2. After each participant has decided his investment decision, the enumerator uses a visual method to identify the player's subjective expectation on others' investment decision. The visual method is explained in following section (Section 6.2).

As a summary, Table 4 shows the payoff for an individual:

[24] To make the experimental design as simple as possible, we also consider to make the game played by two players. Yet, games with multiple agents is more interesting to study in the sense that players strategically consider others' beliefs.

Table 4: Own payoffs for PIG by own and others' behavior

Own behavior \Others	$n_d + n_c \geq 2$	$n_d + n_c < 2$
Invest	300	0
Not invest	100	100

Visual method to identify subjective expectation

To backup each individual's subjective expectation over other players' investment, we plan to use a visual method. The method is designed to be simple enough that ex-combatants who are likely to have less knowledge on probabilities could intuitively reveal their preferences.

First, the enumerator prepares 10 beans. Second, to identify subjective probabilities on other ex-combatants in the game, the enumerator asks a player to allocate those beans to the numbers $n_d \in \{0, N_d\}$. The enumerator explains that those numbers reflect the number of ex-combatants in the same game who invested in the experiment. The player is told to allocate more beans to an event that is more likely to occur, and less beans to an even that is less likely to occur. Figure 4 illustrates this activity. Similarly, he also allocates 10 beans to the number of civilians, $n_c \in \{0, N_c\}$.

6.3 Randomizing Information Provision

Randomly adding some treatment arms to the benchmark experiment in Section 6.1 and 6.2 enable us to evaluate impacts of potential policy analogues on enhancing reintegration. For example, We may randomly distribute information towards civlians and ex-combatants as follows.

Randomizing information toward the civilians

(a) Inform civilians that their partners in their group are random persons living in Mogadishu

(b) Inform civilians that their partners in their group are ex-combatants detained in the prison

(c) Inform civilians that ex-combatants in their group took treatments (I) and (II) in section 3

((d) Inform civilians of some narratives about ex-combatants' backgrounds)

Randomizing information toward the ex-combatants

(a) Inform ex-combatants that civilians in their group do not know that you are prisoners.

(b) Inform ex-combatants that civilians in their group know that you are prisoners.

(c) Inform ex-combatants that civilians in their group know that ex-combatants in the group took treatments (I) and (II) in section 3

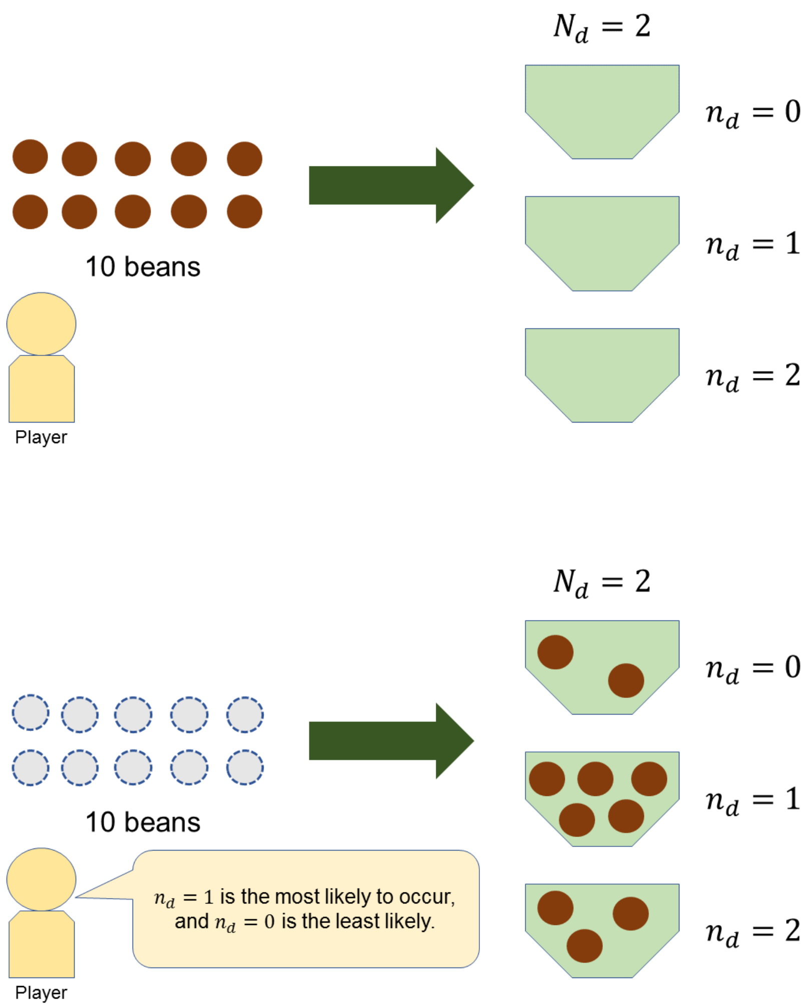

Figure 4: Visual Method in the Public Investment Game

((d) Inform ex-combatants of some narratives about victimizations of civilians by violent extremist groups)

6.4 Capturing Subjective Expectations on Income

Again, we plan to use visual aids to elicit subjective expectation on income after release. We follow De Mel et al. (2008)'s method: we ask ex-combatants to consider a fixed number of individuals just like theirs, and ask them to allocate beans for future income. More detailed explanation and alternative methods for visualizations are summarized in Delavande et al. (2011).

6.5 Other Games: Capturing Anti-Social Behaviors

Depending on the capacity of the prison and our constraints, we may introduce additional games to capture other important features of deradicalization and reintegration. Followings are our candidates:

- Adding **third-party punishment** (e.g., Fehr and Fischbacher 2004; Nese et al. 2013) & **risk** to the destruction game: Note that what we are observing is very likely to be lower-ranked ex-members of Al-Shabaab. That is, they are likely to be agents controlled and dictated by higher-ranked bosses (principals), from the perspective of a standard principal-agent model. The lower-ranked workers' failures of or escapes from destructive activities might entail risks of being punished by higher-ranked bosses.

- **Bomb Risk Elicitation Task** (BRET) (Crosetto and Filippin 2013): Participants decide how many boxes to open out of 100. There is a bomb in one of the 100 boxes. Payoffs increase with the number of boxes opened, but become zero if the box with the bombed is also opened.

- **Cheating behavior** (e.g., Kajackaite and Gneezy 2017): Measuring lies. This framework applies to general criminal behaviors, but we are not very sure how it is relevant or of first-order importance for capturing potential activities of terrorist organizations.

7 Empirical Strategies

7.1 Estimating Other-Regarding Preferenecs

We use results from the dictator games presented in Section 6.1 to estimate other-regarding preferences. Binary-response data in the dictator games can be estimated with following formula:

$$Pr(x_B = Left) = \frac{exp(\gamma U_B(y_B, y_A | x_B = Left))}{exp(\gamma U_B(y_B, y_A | x_B = Right)) + exp(\gamma U_B(y_B, yA | x_B = Left))} \tag{2}$$

where $x_B \in \{Left, Right\}$ is a choice of person B (a dictator), y_A and y_B are final allocations for person A (a recipient) and B, and γ reflects the sensitivity of the choices to utility differences. $\gamma = 0$ implies the model is random choice with equal probability (McFadden 1981). $U_B(y_B, y_A)$ is an individual's utility function explicitly written as Equation 1.

We interpret our results in the following two ways.

Assuming homogeneity in the baseline preference: Assuming everyone has the same preference without treatments and anonymity, we estimate impact of treatments and person A's identity on α and β on Equation 1. Formally, we can write

$$\alpha_{st} = \alpha_{0st} + \alpha_{1st} D_{ex} + \alpha_{2st} D_1 + \alpha_{3st} D_1 D_2 + \alpha_{4st} D_{ex} D_1 + \alpha_{5st} D_{ex} D_1 D_2 \tag{3}$$

where α_{0st} is a baseline preference toward civilians without treatments for dictator (person B)'s identity $s \in \{civilian, prisoner\}$ and information treatments $t \in \{a, b, c\}$ described in Section 6.3. D_{ex} is a dummy variable taking 1 if person A is ex-combatants, and D_1 and D_2 are dummy variables of treatments from the RCT. Note that one of our treatment (D_2, reconciliation dialogue) is always treated together with D_1 (counseling), so that we do not include non-interaction terms of D_2. Obviously, we can characterize β in the same way.

Assuming heterogeneity in the baseline preference: The homogeneity assumption above may be strong. We also interpret our results by relaxing this assumption and applying Iriberri and Rey-Biel (2013)'s method to sort individual's social preferences into four categories (Selfish; Inequality averse; Positive altruistic; Competitive: See Table 2). We then compare fractions of each type by recipients' identities and treatment statuses. We propose to estimate in this way because our current experimental procedure cannot capture shifts from one preference to another [25]. Comparing fractions of each preference at least enables us to capture net shifts from one preference to another by others' identities and the treatments.

7.2 Estimating Treatment Effects on Other Outcomes

For the other outcomes, we estimate the following equation:

$$y_i = \theta_{0st} + \theta_{1st} D_{1i} + \theta_{2st} D_{1i} D_{2i} + \boldsymbol{X_i \theta} + \varepsilon_i \tag{4}$$

where y_i is an outcome variable for an individual i. We again allow differential treatment effects by individuals' indentity $s \in \{civilian, prisoner\}$ and information treatments $t \in \{a, b, c\}$. We consider the dependent variables that we have discussed: the degrees of radicalization measured in several ways; subjective beliefs about others' behaviors; subjective expectations on future income. D_{1i} is the counseling treatment dummy and D_{2i} is the reconciliation dialogue treatment dummy. We include additional controls such as age, clan, hometown and prison term. We cluster standard erros by the stratum of randomization (i.e. remaining time length until release).

Furthermore, we investigate heterogeneity of treatment effects by important characteristics of prisoners' profiles. The first is prison length as a proxy of severeness of activities at Al-Shabaab. The second is remaining time length until release.

References

Ahmad, Shakeel, Muhammad Zubair Asghar, Fahad M Alotaibi, and Irfanullah Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques,"

[25] For example, suppose an individual is classified as "positive altruistic" when a recipient was a civilian and he had a treatment. Since we do not conduct our lab experiment at the baseline before the RCT intervention, we cannot observe his preference without the treatment: he might have changed his preference from, say, selfish to positive altruistic due to the treatment, or he was positive altristic without the treatment but he just become more altruistic after the treatment.

Human-centric Computing and Information Sciences, 2019, *9* (1), 24.

Allport, Gordon Willard, *The nature of prejudice*, Cambridge, MA: Addison-Wesley, 1954.

Andreoni, James, "Cooperation in public-goods experiments: kindness or confusion?," *The American Economic Review*, 1995, pp. 891–904.

Armand, Alex, Paul Atwell, and Joseph F Gomes, "The reach of radio: Ending civil conflict through rebel demobilization," *American Economic Review*, 2020, *110* (5), 1395–1429.

Bauer, Michal, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts, "Can war foster cooperation?," *Journal of Economic Perspectives*, 2016, *30* (3), 249–74.

_ , Nathan Fiala, and Ian Levely, "Trusting former rebels: An experimental approach to understanding reintegration after civil war," *The Economic Journal*, 2017, *128* (613), 1786–1819.

Beber, Bernd and Christopher Blattman, "The logic of child soldiering and coercion," *International Organization*, 2013, *67* (1), 65–104.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak, "Measuring utility by a single-response sequential method," *Behavioral science*, 1964, *9* (3), 226–232.

Bertrand, Marianne and Esther Duflo, "Field experiments on discrimination," in "Handbook of economic field experiments," Vol. 1, Elsevier, 2017, pp. 309–393.

Blattman, Christopher and Jeannie Annan, "Can employment reduce lawlessness and rebellion? A field experiment with high-risk men in a fragile state," *American Political Science Review*, 2016, *110* (1), 1–17.

_ , Julian C Jamison, and Margaret Sheridan, "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia," *American Economic Review*, 2017, *107* (4), 1165–1206.

Blouin, Arthur and Sharun W Mukand, "Erasing Ethnicity? Propaganda, Nation Building, and Identity in Rwanda," *Journal of Political Economy*, 2019, *127* (3), 1008–1062.

Chen, Yan and Sherry Xin Li, "Group identity and social preferences," *American Economic Review*, 2009, *99* (1), 431–57.

Crosetto, Paolo and Antonio Filippin, "The "bomb" risk elicitation task," *Journal of Risk and Uncertainty*, 2013, *47* (1), 31–65.

Delavande, Adeline, Xavier Giné, and David McKenzie, "Measuring subjective expectations in developing countries: A critical review and new evidence," *Journal of development economics*, 2011, *94* (2), 151–163.

Fehr, Ernst and Klaus M Schmidt, "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

_ **and Urs Fischbacher**, "Third-party punishment and social norms," *Evolution and human behavior*, 2004, *25* (2), 63–87.

Fischbacher, Urs and Simon Gachter, "Social preferences, beliefs, and the dynamics of free riding in public goods experiments," *American economic review*, 2010, *100* (1), 541–56.

Gächter, Simon and Benedikt Herrmann, "Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2009, *364* (1518), 791–806.

Gentzkow, Matthew and Jesse M Shapiro, "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 2010, *78* (1), 35–71.

_ **, Bryan Kelly, and Matt Taddy**, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–74.

Gilligan, Michael J, Eric N Mvukiyehe, and Cyrus Samii, "Reintegrating rebels into civilian life: Quasi-experimental evidence from Burundi," *Journal of Conflict Resolution*, 2013, *57* (4), 598–626.

Hansen, Stephen, Michael McMahon, and Andrea Prat, "Transparency and deliberation within the FOMC: a computational linguistics approach," *The Quarterly Journal of Economics*, 2018, *133* (2), 801–870.

Henrich, Joseph, "Cultural group selection, coevolutionary processes and large-scale cooperation," *Journal of Economic Behavior & Organization*, 2004, *53* (1), 3–35.

Humphreys, Macartan and Jeremy M Weinstein, "Demobilization and reintegration," *Journal of conflict resolution*, 2007, *51* (4), 531–567.

Iriberri, Nagore and Pedro Rey-Biel, "Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do?," *Quantitative Economics*, 2013, *4* (3), 515–547.

Kajackaite, Agne and Uri Gneezy, "Incentives and cheating," *Games and Economic Behavior*, 2017, *102*, 433–444.

Laver, Michael and Kenneth Benoit, "Locating TDs in policy spaces: the computational text analysis of Dáil speeches," *Irish Political Studies*, 2002, *17* (1), 59–73.

_ **, _ , and John Garry**, "Extracting policy positions from political texts using words as data," *American political science review*, 2003, *97* (2), 311–331.

Lowe, Matt, "Types of contact: A field experiment on collaborative and adversarial caste integration," 2020.

Lowes, Sara, Nathan Nunn, James A Robinson, and Jonathan L Weigel, "The evolution of culture and institutions: Evidence from the Kuba kingdom," *Econometrica*, 2017, *85* (4), 1065–1091.

Mapping Militant Organizations, "Al Shabab," *Stanford University*, 2019.

Matanock, Aila M, "Experiments in Post-Conflict Contexts," *Advances in Experimental Political Science. Eds. James N. Druckman and Donald P. Green. New York: Cambridge University Press, forthcoming*, 2020.

McFadden, Daniel, "Econometric models of probabilistic choice," *Structural analysis of discrete data with econometric applications*, 1981, pp. 198–272.

Mel, Suresh De, David McKenzie, and Christopher Woodruff, "Returns to capital: Results from a randomized experiment," *The Quarterly Journal of Economics*, 2008, *123* (4), 1329–72.

Mousa, Salma, "Creating Coexistence: Intergroup Contact and Soccer in Post-ISIS Iraq," 2020.

Muggah, Robert and Chris O'Donnell, "Next generation disarmament, demobilization and reintegration," *Stability: International Journal of Security and Development*, 2015, *4* (1).

Nagai, Yosuke, "Reintegration of Al-Shabaab's Defectors in Somalia: An Examination of Conditions for Successful Reintegration," *Peace and Conflict Studies*, Forthcoming.

Nese, Annamaria, Arturo Palomba, Patrizia Sbriglia, Maurizio Scudiero et al., "Third party punishment and criminal behavior: An experiment with the Italian Camorra prison inmates," *Economics Bulletin*, 2013, *33* (3), 1875–1884.

Palfrey, Thomas R and Jeffrey E Prisbrey, "Anomalous behavior in public goods experiments: How much and why?," *The American Economic Review*, 1997, pp. 829–846.

Pettigrew, Thomas F and Linda R Tropp, "Does intergroup contact reduce prejudice? Recent meta-analytic findings," *Reducing prejudice and discrimination*, 2000, *93*, 114.

Rohner, Dominic, Mathias Thoenig, and Fabrizio Zilibotti, "War signals: A theory of trade, trust, and conflict," *Review of Economic Studies*, 2013, *80* (3), 1114–1147.

Sánchez De La Sierra, Raúl, "On the origins of the state: Stationary bandits and taxation in eastern congo," *Journal of Political Economy*, 2020, *128* (1), 000–000.

Scacco, Alexandra and Shana S Warren, "Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria," *American Political Science Review*, 2018, *112* (3), 654–677.

Taylor, Christian, Tanner Semmelrock, and Alexandra McDermott, "The Cost of Defection: The Consequences of Quitting Al-Shabaab," *International Journal of Conflict and Violence (IJCV)*, 2019, *13*, a657–a657.

Wilkerson, John and Andreu Casas, "Large-scale computerized text analysis in political science: Opportunities and challenges," *Annual Review of Political Science*, 2017, *20*, 529–544.

— Appendix —

A Data Construction and Machine Learning Methods for Natural Language Processing Approaches

To be added later.

B Other Approaches for Measuring Radicalization toward Violent Islamic Extremism

We will also use an implicit association test (IAT) and other behavioral measures for obtaining information on radicalization.