USAID
FROM THE AMERICAN PEOPLE

IMPACT EVALUATION REPORT, July 2012

Rule of Law Stabilization Program – Informal Component (RLS-I)

Contract Number: DFD-1-00-04-00170-00
Task Order: DFD-I-05-04-00170

Interviews in Pashtun Kot, Faryab province

29 August 2012

This publication was produced for review by the United States Agency for International Development. It was prepared by Checchi and Company Consulting, Inc.

RLS-I Impact Evaluation Report,

July 2012

Rule of Law Stabilization Program – Informal Component

Contract Number: DFD-1-00-04-00170-00
Task Order: DFD-I-05-04-00170

Submitted 29 August 2012 by:

Michael Sinclair
Chief of Party
USAID/Afghanistan Rule of Law Stabilization Program – Informal Component
4th District, Ansari Square
2nd Street, House No. 149
msinclair@checchiconsulting.com

CONTENTS

Tables

GLOSSARY

AO	Assistance Objective
baad	customary practice of resolving a dispute by giving a girl from the offender's family in marriage to a male member of the victim's family
CPAU	Cooperation for Peace and Unity
DD or D-in-D	Difference in Differences design
GIRoA	Government of the Islamic Republic of Afghanistan
Hadith	Collection of scriptures detailing the actions, sayings, and tacit approvals or disapprovals of Islamic practices and beliefs of the Prophet Mohammad (PBUH) as documented by his companions and accompanied and verified by an authenticating record of the origin and lineage of each part of the collection, determining its authority as a source of Islamic law supplementing the Holy *Qur'an*
IDLG	Independent Directorate of Local Governance
IR	Intermediate Result
jirga (pl. *jirgee*)	a*d hoc* assembly of tribal elders convened to make specific decisions or resolve a specific dispute by consensus
jirgamar (pl. *jirgamaran*)	*jirga* member(s)
machalgha	cash deposit or bond posted by disputants; required under customary practice before a *shura* or *jirga* will consider a dispute, to ensure the disputant's compliance with the decision of the shura or jirga
OSDR	Organization for Sustainable Research and Development
RLS-I	Rule of Law Stabilization Program – Informal Component
Shari'ah	legal precepts found in the Holy *Qur'an* and the *Hadith*; sometimes used by non-scholars (and this report) to denote Islamic law or jurisprudence, which includes scholarly interpretations of the Holy *Qur'an* and the *Hadith*; *ijma* ("collective reasoning" or consensus among scholars); and *qiyas* or *ijtihad* ("individual reasoning" or deduction by analogy)
shura (pl. *shuragani*)	standing assembly of tribal elders and other community members that traditionally regulated community affairs and now also serves as a TDR forum
spinsary	(literally, feminine form of "white-headed") respected female elder(s) involved in dispute resolution
TDR	traditional dispute resolution
USAID	United States Agency for International Development
USG	United States Government

EXECUTIVE SUMMARY

The most recent phase ("Phase 2") of USAID's Rule of Law Stabilization Program – Informal Component (RLS-I) was implemented by Checchi and Company Consulting, Inc. (Checchi), in association with Management Systems International (MSI), and Cooperation for Unity and Peace (CPAU) the Checchi Team. The purpose of RLS-I was to contribute to USAID's sub-Intermediate Result "Strengthened traditional dispute resolution (TDR) and justice in contested areas". RLS-I activities included workshops, discussion groups, referral and case recording methods, encouraging women's participation in TDR processes, implementing a program of public outreach, and building networks of elders. Together, these activities were designed to enhance TDR mechanisms, improve linkages between TDR and the formal justice system, and provide avenues for resolution of long-standing and destabilizing conflicts.

Impact evaluation design

Program impact was to be measured through a panel design of a sample of elders, disputants, and citizens surveyed at program inception and again at its conclusion. Impact was defined as the difference in mean scores on various measures from baseline to endline (difference-in-differences design). This impact evaluation resulted in an in-depth dataset on the project districts and comparators, and utilized methods for testing the causal links of the RLS-I theory of change. RLS-I Phase 2 lasted only 10 months, with three to five months of program activities in the field. Due to the short program duration the research was not predicted to show impact for lagged effects in the perceptions and attitudes of TDR users and other citizens, or for changes in long-standing cultural practices. The RLS-I impact evaluation, while not conclusive, shed some light on these questions.

In spite of the short duration of the project, the RLS-I impact evaluation was designed to pilot a methodology to provide a credible estimate of program impact. This was difficult, however, as in addition to the short duration of the project there were other limitations to the validity of any claims made based upon the data collected.

- Neither districts nor individuals could be randomized, so comparison groups had to serve as a less desirable, but still serviceable, estimate of the counterfactual.
- The time between activity implementation and endline data collection was between two and four months, and the period for disputes to form and be settled was one month or less after treatment had ended. This limited period made the detection of impacts in the evaluation very unlikely.
- A different data collection firm was selected to conduct the endline data collection than had collected the baseline data. This had important consequences – both positive and negative – on the data and their analyses.

- Comparisons between districts in conflict areas are difficult at best. Ethnic and tribal overlays on state administrative boundaries, spillover, and localized insurgency, among other factors, limited the reliability of comparisons across boundaries.

Unfortunately, as a result of these challenges, the impact evaluation could not detect statistically valid change scores on many measures.

Gains in methodology for impact evaluations on sensitive topics in conflict areas

Fortunately, however, data collection and analyses were useful in other respects.

- First, the exercise provided documentation of TDR dynamics in Afghan communities, and is immensely rich in what the USAID Evaluation Policy refers to as "Learning for Effectiveness".
- Second, the exercise provided the basis for a monitoring system that can, with repeated application, robustly track changes in attitudes and practices relating to TDR and serve as a critical measure of development effectiveness.
- Third, design specifications for data collection allowed for a nuanced review of change, which is explored in this report. This report highlights the research findings – what the data do show – but always in light of the strength of the underlying comparison in making inferences.

Conclusions

While data quality issues adversely affected the validity of longitudinal measurements, examination of relationships between elder knowledge, disputant perception, and various program metrics within the treatment group suggested that knowledge was in fact important for improving disputants' assessments of TDR.

The data also suggested that the benefit of RLS-I was not transmitted through elder knowledge, but rather through continued elder attendance at RLS-I activities and the positive peer effects of RLS-I elder networks. These results posited a definite role for network and peer effects in program success, as is supposed by the development hypothesis and solicitation design.

Hypothesis I: The intervention will result in TDR decisions that better reflect and/or are based in Afghan law, Sharia, and human rights norms

As expected given the above, the overall treatment effect for knowledge was zero, but with some divergence according to the sub-topics of Afghan law and more *Shariah*-oriented topics of family, inheritance, and property. Citizen knowledge of Afghan law in communities that had received RLS-I outreach material increased by 6% relative to communities that had not.

Hypothesis 2: The intervention will result in TDR decisions and shura/jirga members being perceived as more impartial

For this hypothesis, there is more conclusive evidence that the data from baseline to endline are not comparable, as seen in item response patterns, respondent selection approaches, and differences in the enumerators. However, baseline and endline data on disputants proved useful in terms of learning for development effectiveness. The endline enumerators more successfully engaged female disputants for a more detailed measurement of the gender deficit in disputant perceptions.

Hypothesis 3: The intervention will result in a decrease in the number of TDR decisions that negatively impact women and children

RLS-I and USAID understood that this higher-order change was not likely in the short time frame of the Phase 2 intervention. From these data, indeed no effect on incidence of forced marriage or *baad* can be determined across the treatment population.

Hypothesis 4: The intervention will result in an increased role for women in TDR processes as disputants, witnesses or decision-makers

For the majority of questions on these topics, as predicted, there was little or no change from baseline to endline in attitudes or principles held by elders about women's participation in TDR. When elders were asked about specific cases of women's involvement in TDR, endline respondents were more than twice as likely to report a case with women's involvement (7% at baseline to 15% at endline). This may be a result of the women's (*spinsary)* groups in RLS-I activities.

Evaluation Research

Across the sample, the quantitative data show that documentation and registration of cases had improved substantially more for elders in treatment districts than those in comparison districts. These data were self-reported; RLS-I data on documentation and registration support the general pattern of self-reports, though with less incidence of documentation or registration.

In terms of lessons learned during the process of piloting the impact evaluation in a conflict-affected environment, it was found that using the same elders at baseline and endline helped withstand challenges to data quality – for example from using comparison units that may have been subject to different local dynamics or different sources of bias. Similarly, using disputant assessment data linked to a referring elder at baseline and endline allowed measurements of association between, for example,

elders' change in knowledge or extent of participation in RLS-I activities with disputant perceptions of procedural justice and equity of the decision. As a result of these efforts, and despite the challenges posed in taking longitudinal measurements in a conflict-affected environment, strong correlational evidence was found in support of the development hypothesis. Interestingly, the evidence also suggests that while the development hypothesis is well-formulated and possibly validated, the positive effect of RLS-I may not be transmitted exactly as the theory of change might suppose.

Key recommendations

Programming

1. ***Improve training and reinforce learning.***
   Modify workshop design to put additional emphasis on comprehension and retention by a low-literate, adult audience. This should include:
   - Sufficient time for multiple exposures to workshop content
   - Strengthened training of trainers, including performance monitoring and coaching
   - Increase use of adult learning principles for low-literate audiences
   - Increase active learning: role plays, case studies, and participants' own experiences used as discussion topics

2. ***Test assumptions on critical mass and saturation.***
   Programming choices should include discussion on best ways to support or scaffold elders to build critical mass at home.

3. ***Develop theoretical models for how inputs may affect disputant perception***
   The relationship between RLS-I and changes in disputant perceptions is not well-theorized. Research suggests that elders' attendance at RLS-I activities improves disputant assessment of adjudication; work to understand the process that leads to this apparently stabilizing effect.

4. ***Track specific applications of knowledge in events***
   Emphasize event monitoring to track how elders are asked to apply knowledge gained as a result of project activities. These measures will show the level of exposure necessary for elders to be able to adopt and use new knowledge and skills in their adjudication of disputes.

Impact evaluation research

1. ***Adopt a pipeline approach to program expansion***
   Comparison districts from Phase 2 should be short-listed for consideration as treatment districts in Phase 3, to assuage ethical challenges to their inclusion and take advantage of previous measurements to create stronger evaluation designs. For districts included in Phase 3 as comparison groups, selection should explicitly consider this pipeline approach – even if this is not communicated to comparison district respondents at the time of data collection.

2. ***Use the same data collection firm for the entire evaluation period***
   A major learning from the evaluation was that the disadvantages of using a different research partner at endline from baseline outweighed advantages. Switching partners may have corrected some bias in baseline data, but may also have invalidated evaluation measurements.

3. ***Ensure longer time periods between data collection***
   Conducting data collection in such a short time frame may distress some communities by asking disputants to recount negative, charged experiences very soon after they occurred. In addition, a longer time frame would allow the theory of change to be more fully tested, because the process of knowledge- and skills-building would have had more time to take root.

4. ***Integrate the research into an M&E system capable of robust inference***
   The RLS-I M&E system has hallmarks of industry-leading M&E: engaged and committed local leadership; effective and durable capacity building; grounded instrumentation and data capture; and off-the-shelf technology adapted for project use. M&E field work can support the impact evaluation in capturing contextual and secondary data for cross-fertilization of findings, and the impact evaluation is useful for the development of refined M&E processes and tools.

5. ***Consider disputant perceptions crucial for program context and learning***
   Disputant perceptions as impact measurements can be problematic. However, male and female disputants' experiences with the TDR system are critical for seeing change over time in perceptions of impartiality and justice. Shift from the large disputant sample size needed for making inference, to better connecting the contextual qualitative narratives with the numeric assessments.

I. INTRODUCTION

The development problem

The Afghan judicial system does not yet effectively reach many remote populations. Traditional dispute resolution (TDR) bodies, such as *jirgee* and *shuragani*, convene to resolve local level disputes. These systems provide useful and important services; however, TDR outcomes do not always reflect international human rights standards, or *Shari'ah* or Afghan law. The process can differ widely depending on the held beliefs of participating elders. Protections of human rights, guaranteed by the Afghan Constitution, are not universally known or enforced. Some decisions taken by *jirgee* and *shuragani*, and indeed some TDR patterns, reveal important gaps for women's legal rights.

The response

To address these issues, USAID developed the Rule of Law Stabilization Program – Informal Component (RLS-I), a program in Afghanistan that supports the USG's whole-of-government Rule of Law Strategy. RLS-I implemented by Checchi and Company Consulting, Inc. Under the overarching Assistance Objective, "Improved performance and accountability of governance," RLS-I contributes to USAID's first Intermediate Result under this AO, "Increased public confidence in the Rule of Law system." In turn, the sub-IR "Strengthened traditional dispute resolution and justice in contested areas" forms the purpose of the RLS-I intervention.

RLS-I began with Phase 1 (April 2010 to August 2011) and was followed by RLS-I Phase 2 (September 2011 to July 2012), which is the subject of this report. Phase 1 and Phase 2 were implemented in three regions of Afghanistan – north, south, and east.

RLS-I undertook a locally sensitive set of interventions designed to strengthen the informal justice sector. Activities included workshops, discussion groups, referral and case recording methods, encouraging women's participation in TDR processes, implementing a program of public outreach, and building networks of elders. Together, these activities were designed to enhance TDR mechanisms, improve linkages between TDR and the formal justice system, and provide avenues for resolution of long-standing and destabilizing conflicts.

The RLS-I development hypothesis is that skills- and knowledge-building of informal justice providers increases stability through increased access to justice and citizen confidence in TDR mechanisms. The if-then logic of programming presumes that these activities, combined with peer networking, will lead to greater knowledge of *Shari'ah* and Afghan law among jirga members, which will improve their

adjudication of disputes. A further program outcome will be strengthened linkages between the formal and informal justice systems, including the documentation and registration of TDR cases in the formal system. As these changes are institutionalized, community members using the TDR services of *jirgee* will perceive these processes as fair and impartial, and the effect will be improved performance and accountability of governance.

This impact evaluation of RLS-I Phase 2 activities resulted in an in-depth dataset on the targeted districts and comparators, to test the causal links of this theory of change. The RLS-I program lasted 10 months, and as such USAID and RLS-I understood that the comparison was likely to show no impact, much less the kinds and degree of changes that might be expected over a longer term. This is particularly true for indicators that are expected to lag, such as citizens' and disputants' perceptions and attitudes, and changes in long-standing cultural practices. However, the research process was also designed to gather in-depth qualitative data on process and change throughout the project life cycle, support improved programming, and result in useful lessons for development effectiveness – particularly in understanding how an impact evaluation can be carried out in conflict-affected environments, potential limitations to the research, and ways to link impact evaluation data to secondary data on villages and tribes, the formal justice sector, complementary stabilization efforts, and other features of the Afghan landscape.

Assumptions underlying this theory of change include the following:

- Participants are willing and able to change attitudes and practices that may conflict with Afghan law and *Shari'ah*
- The Afghan law, *Shari'ah*, and human rights workshop content are effectively imparted to participants
- Participants will be able to use their new knowledge effectively in context, upon returning to their communities
- Participation will generate a critical mass of *jirga* members in a given community sufficient to effect change in adjudication reflective of Afghan law, *Shari'ah* and human rights norms
- Improper influence and interference with informal dispute resolution from local power brokers will gradually lessen as a result of security and governance gains

That these assumptions prevail in the districts and villages where RLS-I is implementing the project is critical for success. Program staff and stakeholders looked for ways to mitigate circumstances outside the control of the activity, in an attempt to influence the risk factors underlying these assumptions. Utilizing Afghan staff and best practices in rural environments were important conditions with which RLS-I set the stage for success. The monitoring and evaluation ("M&E") feedback loop also brought real-time data back to project management to inform ongoing programming. Saturating districts with

training and networking coverage helped create the critical mass in villages that is necessary for change. But the difficult environment in Afghanistan continued to pose new challenges and threats to these assumptions.

Learning for development effectiveness

The impact evaluation sought robust inference demonstrating the validity of the development hypothesis that skill- and knowledge-building of informal justice providers improves the adjudication of disputes resolved at the village level. However, this fundamental research question also depends on contextual questions that may need at least an exploratory evidentiary base to feed into the more fundamental research question. The more limited, but still crucial, research questions that remain unanswered include:

- What is the requisite amount of exposure to RLS-I activities before change in behavior might be effected?
- What is the time frame governing any treatment effect, and for how long does any treatment effect persist?
- What is the requisite number of participants from a given community needed to effect a change in dispute adjudication and outcomes in the community as a whole?
- Do RLS-I activities for women provide an indirect means of affecting dispute prevention, adjudication, and outcomes?
- Is the distinction between elders supported by the community and those imposed on it a meaningful one in the context of RLS-I treatment effect?

Two descriptive questions were also addressed in the research to support feedback into the RLS-I components that were not part of the impact evaluation hypotheses:

- What linkages exist between village *jirgee* and *shuragani* and their formal justice sector counterparts at the district level?
- What patterns among long-standing disputes are found in the respondent populations?

The RLS-I impact evaluation shed light on these questions while not being conclusive. Operating in highly fluid environments and piloting innovative methods conjectured to support the stabilization thesis, RLS-I learned crucial lessons about variability in TDR, the effects of insecurity on community resolutions to conflicts, and population readiness for women's roles and changes in harmful practices. The initial impact measurement generated from the first two data collection rounds provided valuable insights for the refinement of the evaluation questions that would best show whether RLS-I activities have a general effect on TDR adjudication and outcomes.

The RLS-I impact evaluation was part of a system of robust impact monitoring: tested instruments, a conflict-appropriate evaluation methodology and sensitive, detailed field data collection. The process allowed the program and M&E staff to better understand the opportunities and obstacles of an impact evaluation in a dynamic and conflict-affected environment – or, as termed in the USAID Evaluation Policy, "Learning for Effectiveness". Over time, with one to two data collections annually through 2014, this system sets the foundation for collecting quality data on this unique programming, providing increasingly reliable and valid impact findings. Evaluation methods and strategies emerging from this exercise as well as lessons learned are reported here for their future utility for other sites and programs as USAID implements its new Evaluation Policy.

II. METHODOLOGY

Summary of design

The evaluation was designed as a quasi-experimental, mixed methods study uniting data from quantitative and qualitative data streams both to show impact and to describe those elements that contribute to that impact. Elders, disputants, and citizens were queried from both treatment districts and a sample of non-equivalent comparison districts. These groups, queried in a longitudinal panel design, were compared through a "difference-in-differences" (D-in-D) design[1]. Scores on knowledge and attitudes of informal justice providers, dispute adjudication practiced by these providers, and disputant case assessments were contrasted from baseline to endline. By including a comparison group and testing both groups before and after the intervention period, D-in-D methods help control for unobserved characteristics that might otherwise explain outcomes. Please see the baseline report for additional detail and discussion on evaluation design issues.

Research partners

RLS-I subcontracted with local survey research firms to collect data in targeted treatment and comparison districts. At baseline, RLS-I selected Organization for Sustainable Development and Research (OSDR), an Afghan-owned firm with experience in the RLS-I intervention areas. EUREKA Research was selected as the research partner for endline data collection.

Changing research partners allowed RLS-I the opportunity to review prior assumptions and measurements, check the validity of baseline data patterns by replicating with different researchers, and

[1] Where the original respondents were not available or willing to be re-interviewed at endline, a cross-sectional sample of similarly treated or comparison elders was selected for the endline data collection. Both panel and cross-sectional matching were undertaken for analyses, as will be seen below.

explore new avenues for measuring access to justice. In this way, learning for development effectiveness became a priority, above and beyond impact measurement for its own sake. Prior to baseline, RLS-I discussed explicitly with USAID/Afghanistan (RLS-I Impact Evaluation Plan, 6 October 2011, p. 5) the improbability of change on the impact measurements given the short duration of the project. USAID, in both Afghanistan and Washington, underscored the utility of the research because of the unique intervention and the conflict-affected environment. In addition, it was thought that too little was still known about the environment in which the evaluation took place, whether the measurements were properly calibrated to the program being implemented, and whether the evaluation was unduly subject to hidden biases from the baseline research partner. The change in research partner was therefore consistent with these areas of learning for development effectiveness.

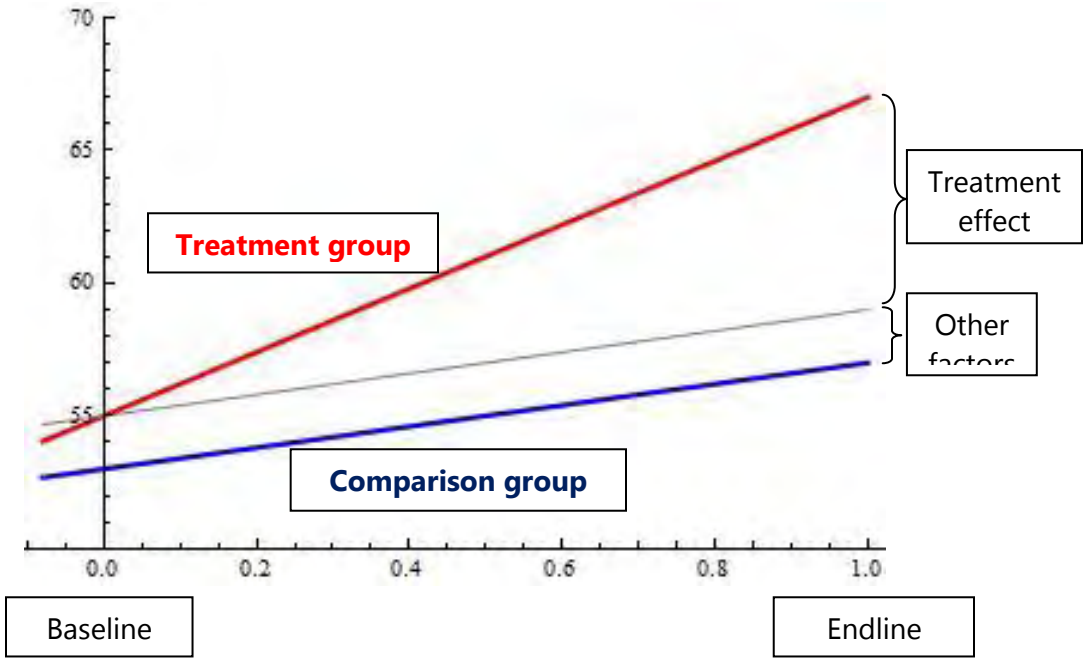
As is shown in this report, switching research partners at endline had both positive and negative results. The technical differences between the two partners in terms of their field work/field staff resulted in some incomparable data, as described in the section on "Limitations to design and measurement", below. Positive outcomes, however, include important tests of the validity of the baseline data on key measures of hypotheses such as incidence of forced marriage and *baad*. EUREKA Research was also somewhat more successful in identifying and interviewing female respondents than OSDR had been during the baseline data collection. As part of a process of learning for development effectiveness, the switch in research partner provided important lessons to incorporate in future evaluation efforts.

Impact evaluation

Key outcome scores on knowledge and attitudes of informal justice providers, dispute adjudication practiced by these providers, and disputant case assessments were contrasted from baseline to endline and the difference between these figures for treatment groups was compared to the difference between these figures for comparison groups. By including a comparison group and testing both groups before and after the intervention period, D-in-D methods helped control for unobserved district characteristics that might otherwise explain outcomes.

Non-random assignment reflects the fact that treatment districts were purposively selected and not randomly assigned while comparison districts were selected based on proximity and similarity to treatment districts and security considerations. This non-random assignment results in a lack of probabilistic equivalency between groups. In graphic form, the D-in-D model is as shown in Figure 2.1:

Figure 2.1: D-in-D model displayed graphically



Note that the gray line for the counterfactual is drawn parallel to the blue line for comparison group. This reflects a critical assumption that the treatment and comparison groups are similar enough that they mature at similar rates. If this assumption is violated, there will be additional variance introduced to the estimate of treatment effect.

Following the D-in-D model, the treatment and comparison groups were queried prior to and following the intervention. Treatment and comparison groups were compared geographically (district, province, region), demographically (socio-economic standing, tribe/ethnicity, education, population, security), and programmatically (features of the dispute resolution process such as case type, scope of dispute, etc.).

For disputants, the lowest unit of analysis was the village elder. For key informants, those interviewed at baseline were sought out at follow up, and the data from disputants referred by the interviewed elders at endline were compared to the data from disputants gathered from the same elders at baseline. This has the dual advantage of soliciting disputant data directly linked to program participants (elders) as well as removing variation in adjudication by elder. Where sampling disputants linked to the same elder was not feasible, similar key informants were identified for interview and referral of disputants, and the data were treated cross-sectionally rather than as a matched sample.

For the outreach component, a random cross-section of citizens was interviewed but those that accepted RLS-I outreach materials were specifically targeted at follow up. The outreach component thus included both a cross-sectional and panel design to measure potential effect at two levels – the movement in knowledge and attitude among those who accepted RLS-I outreach materials and the

movement in knowledge and attitude within the greater communities where the outreach materials were delivered.

Sample selection

Selection of districts

Treatment districts for the RLS-I program were selected in collaboration with USAID with the aim of improving stability in contested areas through improving traditional dispute resolution at the village and district level. Twelve new districts were selected for the RLS-I Phase 2 intervention, across nine provinces and across three regions of Afghanistan – east, north, and south.

For the purposes of the impact evaluation, proper sample selection was critical for the reliability of the resulting findings. Use of sampling greatly reduces costs for data collection, labor and transportation, while still providing a reasonable (though non-probability based) cross-section of the intervention sites. Sampling was done in this case by purposive selection, in which nine districts (from among eight provinces, as shown below) were chosen for the evaluation based on geographic coverage and on likely security factors and conditions, from the list of selected intervention districts.

In addition to the treatment/comparison sample for the impact evaluation, two additional samples were added to the research design.

- First, two RLS-I Phase 1 districts, Bihsud in Nangarhar province and Arghandab in Kandahar province, were included in the sample for the impact evaluation. This allowed estimation of program effect over a longer treatment period from the beginning of Phase 1 to the end of Phase 2. Though these districts lacked true baseline data and so could not provide estimates of treatment effect in their own right, they were analyzed in complementary fashion with the Phase 2 treatment districts to help shed light on the important research questions of (a) the requisite amount of exposure to treatment and duration of treatment period before there may be an accompanying shift in disputant perception, and (b) how treatment effects persist over time. For example, does disputant perception over time show an accelerating effect characteristic of increasing returns and a positive feedback loop? Or does disputant perception drop off over time similar to a model of decreasing returns to scale? Qualitative data from these sites provided useful lessons about the RLS-I program cycle and the life of the intervention to date, across RLS-I Phase 1, RLS-I Phase 2, and graduation (based on a set of criteria agreed on with USAID/Afghanistan).

- The second additional sample selection involved those districts where outreach activities were conducted under RLS-I Phase 2. In order to study the effects of the print materials produced and distributed by RLS-I, a sample of households in one district from each of the three program regions received these materials and was queried about the materials and their perceptions of rule of law in their communities. Those districts were Nurgal in the east, Puli Khumri in the north, and Chora in the south. With the citizen perceptions poll data, one district per region that received both the workshop activities and the public outreach was compared with another district receiving only the workshop activities but not the outreach component. This overlay on the sample was not planned to be large enough to generalize but rather to provide the opportunity to test instruments and procedures for citizen polling, learn how to sample women in districts effectively, and reach disputants for the main sample through random walk techniques. Most important, this work laid the groundwork for an ongoing monitoring system for capturing data on outreach efforts.

The treatment and comparison districts are presented in the table below; selection of comparison districts is described below. Blue shading denotes outreach districts and green shading denotes RLS-I Phase 1 districts.

Table 2.2: Treatment and comparison districts

Region	Province	Treatment	Comparison
East	Nangarhar	Bihsud	
	Laghman	Mihtarlam	Alingar
			Alishing
	Kunar	Nurgal	Chawkay
North	Baghlan	Puli Khumri	Aybak (Samangan)
	Faryab	Pashtun Kot	Shirin Tagab
South	Kandahar	Arghandab	Panjway
	Uruzgan	Chora	Khas Uruzgan
			Shahidi Hassas
	Zabul	Shahjoy	Shinkay

Phase 1 districts Outreach districts

Comparison district selection was first carried out based on logistical, geographic and security considerations. Program staff and stakeholder judgment were used to decide which districts were likely to be comparable, in the absence of complete district and village level data on matching characteristics. Baseline data on villages and districts (both treatment and comparison) were collected and analyzed for key characteristics.

During the analysis of endline data, treatment and comparison groups were compared geographically (district, province, region), demographically (socio-economic standing, tribe/ethnicity, education, population, security), and programmatically (features of the dispute resolution process such as case type, scope of dispute, etc). A final means to estimate treatment effect was by elder – by examining disputant perception linked to a specific elder who helped mediate the dispute. This had the advantage of removing variation in adjudication by elder and provided a more robust estimate of treatment effect with a comparative sample of disputant data linked to the same elders gathered at endline.

Selection of villages in districts

Village selection was directed by security conditions on the ground, representativeness of ethnicities and tribes, and geographic expediency. Data collection teams were composed of individuals from that district or nearby, and their local knowledge combined with district maps were used to seek a range of sites that were geographically and ethnically diverse across the whole of each sampled district. Supervisors accompanied each team to the districts. They presented an official authorization letter to the district governor and engaged local elders to introduce the study and negotiate entry to the district, explain the data collection scope and motivation, and identify key informants. As villages of different sizes were included in the sample, the field research team visited the number of villages necessary to reach the quota of individuals to be interviewed per district.

District and village authorities were also queried for data on the villages themselves; ethnic and tribal makeup of the district and its villages; the number of households; sources of income; education levels; number of female-headed households; and presence or absence of services and infrastructure. These data were used to give context to data on sub-provincial level strata. District and village identification used the same official numbering as the GIRoA system so that in the future national-level data could be further incorporated.

In the south, the data collection teams were very restricted by security concerns and getting to some remote villages proved to be impossible. The survey teams employed various techniques to minimize security risks while continuing to gather data as planned. In one case a team stayed in a village several extra days to avoid potential confrontations, employed local guides, and wrote the interview questions in regular notebooks (later transcribed into the survey forms) rather than bringing the forms to villages.

Selection of individuals in villages

Three types of individual respondents were sought for the evaluation: elders who serve on *jirgee* or *shuragani*,[2] disputants whose disputes were resolved by *jirgee* or *shuragani* in the recent past, and citizens to report their perceptions of rule of law and dispute resolution in their communities. Individual respondents were selected through a set of processes dictated by the study design as well as by conditions on the ground. The field research team was provided detailed instructions for identification of respondents within the district. They began in a district *shura* (if present) and with the district governor to gain access to villages in the district; the team supervisors carried an official authorization letter to introduce the study to selected individuals.[3] In parallel with the district saturation design of the intervention, a small number of key informants (elders) were sought in each village to participate in the study, rather than a large number of elders from one village.

Female interviewers were employed to assist in locating and interviewing women involved in disputes as well as women citizens for the citizen perception survey. According to the data collection teams, it was difficult to locate and interview women with knowledge of a dispute. However, the data collectors at endline had somewhat more success with this process than did the data collectors at baseline.

The number of individuals sought in each category (elders, disputants, and citizens) and surveyed per district was based on the power analysis (described below) of sample size necessary for confidence in results, estimates of the existence of disputes dealt with in *jirgee* or *shuragani*, budget considerations, and the degree of stratification USAID required in the results.

For sampling of key informants (elders) in treatment districts, at baseline the data collectors used a list of known elders that RLS-I had compiled, so that in their selection of villages they could seek these individuals for interviews. Lists included the Independent IDLG *Shura* roster, the registered *malikan* roster, the *ulema shura* list, asking villagers whom they trusted to resolve disputes, or selection over the course of negotiating entry and welcome into a village. In the north, according to the data collection supervisors, many of the elders on the list could not be located; this also happened in certain provinces in the south, such as Uruzgan and Kandahar, while in the east, the lists were found to be very reliable. At endline, the data collectors used the list of baseline elder respondents in an attempt to assemble panel

[2] Though women do not serve on *jirgee* or *shuragani* in the target districts, RLS-I programming includes training for women in many of the same legal topics as those taught to male *jirga/shura* members, networking, discussion groups and development of *spinsary* groups. Nevertheless, as women elders do not traditionally have role in resolving the kinds of disputes brought before *jirgee*, the definition of "elders" for the purposes of this study was almost entirely male. For disputants, locating and gaining access to women who had been party to disputes proved difficult as well, but was carried out with more success at endline than at baseline. Sample numbers by gender are shown in the following section.

[3] It was not recommended for data collectors to enter a village without such an introduction, for reasons of security and respondents' potential willingness to receive them.

data on the experiences of these elders. When those elders could not be found or wished not to be interviewed again, a list of RLS-I participants was used to locate suitable respondents in treatment districts. In comparison districts, at endline the RLS-I-compiled lists of *jirga* and *shura* members were employed as during the baseline to identify appropriate respondents. In comparison districts, on occasions where the lists did not provide suitable elders, data collectors consulted citizens and leaders at the district and village levels to locate appropriate key informants.

Disputants were selected in one of several ways. Key informants (elders) were asked to identify disputants with whom to speak regarding disputes they themselves had settled. Additionally, key informants were asked if they could refer disputants in cases they were aware of even if the referring elder had not played any role in mediation. In some cases at endline, when a key informant identified a disputant the data collectors sought out the opposing party to that same dispute. Once a disputant was identified and interviewed, the disputant was queried whether they in turn knew of and could refer another disputant in the village; this method of snowball sampling made up 9% of the total sample. A final method of identification was soliciting at population collection centers such as the mosque, bazaar, transport depot, etc. The sample for the citizen perception survey was carried out in this way at baseline. At endline, the data collectors worked from the list of citizens surveyed at baseline and who agreed to accept the outreach materials, and supplemented the sample with other individuals sought through random walk and seeking respondents in public areas.

The numbers of respondents of each target group (elders, disputants and citizens) are shown in the table below, for both baseline and endline data collection. Comparison group samples are shown in parentheses beside the treatment group numbers for each type.

Table 2.3: Respondent sampling

Target group	Treatment *(Comparison)*	
	Baseline	Endline
Elders	190 *(130)*	210 *(227)*
Disputants	295 *(347)*	266 *(279)*
Citizens	958	891
Phase 1 districts		
Elders	55	68
Disputants	86	97

A group of respondents, shown in Table 2.3 under the heading Phase 1 districts, were part of the first cohort of elders trained in 2010 and early 2011. This sample was selected from two treated districts and

is shown in this report as a synthetic "time three" measurement – that is, what might be predicted if implementation and maintenance were to be continued for the current Phase 2 group of respondents.

Data collection instruments

The tools used to measure these hypotheses, as well as specific indicators measuring different components of the main hypotheses, are as follows:

- *Key informant interview*. Key measures included individual knowledge, attitudes, and practices, to the extent possible with specific examples of application of training content in local dispute resolution. There were also general questions on the structure/mapping of dispute resolution in a given village and district, as well as querying for the direct experience of the respondent in resolving disputes.

- *Disputant case assessment*. The disputant case assessment tool asked questions about specific cases resolved through the informal justice system and collected perceptual assessments of various aspects of the process of resolution and the case outcome.

- *Citizen perception survey*. Key questions were attitudes toward informal justice and the possible identification of disputants. The perception study also attempted to gauge any change in citizen perception of critical messaging from RLS-I outreach activities (distribution of print material in three districts).

In addition to these tools, the data collection protocol included an instrument to capture village characteristics such as groups, ethnicities and tribes; approximate number of households; income sources; education levels; female-headed households; presence or absence of mosques, paved roads, public parks, bazaars, link to the electrical grid, and water sources; and distance from the district center. These data are used to give context to data on sub-provincial level strata. District and village identification in this dataset included the official numbering from the GIRoA system so that in the future national-level data can be further incorporated.

Evaluation measurements and data types

The core measurement of this evaluation is that of difference-in-differences (D-in-D). First, the baseline measurement is subtracted from the measurement at endline for the treatment group, and again for the comparison group. Then the comparison group's difference is subtracted from that of the treatment group, to arrive at the estimate of the treatment effect. Mechanically, this measurement is presented in two ways. The first is linear as follows:

Table 2.4: Difference-in-differences design

Impact Measure	Baseline (T)	Baseline (C)	Endline (T)	Endline (C)	Difference (T)	Difference (C)	Treatment effect
Item	A	B	C	D	C-A	D-B	**(C-A) – (D-B)**

The second manner of presentation is in a more tabular format:

Table 2.5: Difference-in-differences, tabular format

Impact Measure	Pre	Post	Post - Pre
Treatment	A	C	C - A
Comparison	B	D	D - B
Treatment - Comparison	B – A	D - C	**(C-A) – (D-B)** or **(B-A) – (D-C)**

The linear presentation format will be used most often throughout the text that follows.

Matched sample (panel) data

The D-in-D measurements were constructed using three types of comparative data. Measuring change requires some level of equivalence between baseline and endline respondents: either the same individuals, similar individuals, or statistically similar individuals can be compared. In the case of the RLS-I impact evaluation, all three of these models were utilized and are described below.

Where possible, an actual match between baseline and endline respondents was sought, in what is called a panel design. Data collectors carried a list of those interviewed at baseline and attempted to interview the same individuals at endline, in both treatment and comparison districts. Just under half of elder respondents interviewed at baseline were again interviewed at endline; this provides the necessary conditions for evaluating change in those individuals and across the populations they represent. They are equivalent in that they are the same people, interviewed twice over time. Panel measurements provide the most power to detect a statistically significant treatment effect relative to other types of measurements.

In addition to the manual constructions above, D-in-D measurements may follow a linear regression format. This often facilitates analysis and also allows the inclusion of additional explanatory variables.

The D-in-D model using the same participants at baseline and endline may be expressed in regression notation as:

$$\Delta Y_i = \beta_0 + \beta_1 treatment$$

With ΔY_i denoting the change in a respondent's outcome score Y for a given unit of analysis *i*, and treatment = 1 if the given unit of analysis was subject to RLS-I programming, or treatment=0 if the unit of analysis is part of the comparison group. Thus the outcome variable Y is the change from baseline to endline, with β_0 signifying the change score of the outcome variable in the comparison group, and β_1 the change score of the outcome variable in the treatment group.

Pooled cross-sectional data

The panel data described above is based on successful matching of the same respondent at baseline and endline. Not only does this reduce the variance in the data, it also allows the inclusion of individual level characteristics to help explain varying levels of the treatment effect.

Where baseline respondents could not be found or did not wish to be interviewed again, other program participants were selected from lists provided by RLS-I (in the case of elders), or respondent selection followed the same steps as at baseline (in the case of citizens). They are equivalent to the individuals interviewed at baseline in their parallel characteristics, community roles, or status as a program participant. This is referred to as cross-sectional data. When independent samples of cross-sectional data (such as baseline and endline) are combined, it is referred to as pooled cross-sectional data. Pooled cross-sectional data is amenable to evaluation measurements given its longitudinal nature, but analysis typically takes place at the level of the entire group.

In regression format, the D-in-D measurement for pooled cross-sectional data is as follows:

$$y = \beta_0 + \delta_0 endline + \beta_1 treatment + \delta_1 endline \cdot treatment$$

In this format, δ_0 reflects the secular changes over time that are unrelated to treatment, β_1 reflects the change across the treatment and comparison group at endline, and δ_1 is the specific treatment effect of participation in RLS-I programming.

Propensity score matching

Given the lack of random sampling of respondents, even the use of evaluation designs for causal inference may break down in their intent of lending a causal interpretation to the phenomena being studied. When evaluations employ observational data, propensity score matching (PSM) has demonstrated good results in approximating the results from true experimental data in which treatment

status is randomized. Propensity scoring is based on the theorem that if an evaluation measure is independent of a participant's treatment status given a set of characteristics, then those characteristics of the respondent can be used to match a treatment participant with a comparison participant and approximate random assignment of treatment status (within that set of identified characteristics). PSM is thus a method of statistical matching that produces "as if" randomization of treatment status that allows a causal interpretation of the treatment effect.

To generate PSM measurements, first the matching characteristics were chosen and each data case assigned a probability of being in the treatment group (the propensity score) based on the chosen characteristics. Each case from the treatment group was then matched to another case from the treatment group with identical or near-identical propensity scores, with the same matching process applied to cases within the comparison group. Change scores on impact evaluation measures were then computed to produce the first difference. The second difference is achieved by comparing the propensity-matched change scores across treatment and control, again matching by the propensity score. This measurement is most conveniently presented in regression format, as follows:

$$\Delta Y_i = \beta_0 + \beta_1 treatment + \delta_0 propensity$$

In this format, the propensity variable serves to provide the second difference in the D-in-D score. It signifies that the treatment effect is generated by holding the propensity score fixed across treatment and comparison. Therefore, for whatever respondent characteristics were used to generate the propensity score, all evaluation measurements are conducted with treatment and comparison respondents who match on those characteristics. In this study, the characteristics used were region, district, age, level of education, and the respondent's assessment of the level of trust between citizens and government officials.

The measurements used in this study rely primarily on the pooled cross-sectional data type described above, with secondary measurements provided by the actual change scores of elders who agreed to be interviewed both at baseline and endline. The propensity-matched measurements provide a final corroboration of the pooled cross-section and matched sample measurements and in some cases provide useful insight as to the quality of the measures.

Mixed methods: analysis and treatment of data

Rigorous impact evaluation on quantitative measures was combined in this study with qualitative data collection methods that elicited rich narrative data from respondents on their experiences with and perceptions of dispute resolution. The sensitive topics covered in the research benefited from allowing respondents to tell their stories, and the resulting data was used to understand the meaning of simple numbers. There are times when numbers and narratives contradict, as well, when respondents provided

a socially acceptable response on a scale or simple yes/no answer, compared to the detail and nuance achieved through in-depth interviewing. It is important to understand how these different data streams can be used to triangulate findings and support or refute the hypotheses underpinning the research.

Qualitative data are collected and scanned for content; for the current study, these responses serve as background to the quantitative research. This mixed-methods research, while labor intensive in the data collection, entry, and analysis, allows the matching of respondent narrative with fixed categories of description. These fixed categories include case types and subtypes, the scope of the dispute, its duration and costs, and so on, and provide a crucial backdrop against which to interpret the dispute narratives and organize the pathways of resolution according to the pre-coded categories.

Limitations to design and measurement

While the RLS-I impact evaluation was designed to provide a credible estimate of program impact, there are still threats to the validity of any claim. Under RLS-I, neither districts nor individuals could be randomized to treatment or control groups. As a substitute, the impact evaluation identifies comparison groups to serve as a less robust, but still serviceable, estimate of the counterfactual[4]. Observable characteristics can be controlled for with the proper data collection and inclusion within a regression analysis. However, unobserved variables still differentially affect treatment and comparison to an unknown extent and in unknown directions and are therefore persistent threats to the internal validity of the estimate of treatment effect. This caution is especially relevant to the question of the extent to which the estimate of treatment effect can be generalized to a wider population of interest. Additional limitations are described below.

Insufficient time for program implementation

RLS-I Phase 2 was initially designed as a 10-month program with six months allocated to implementation. Interventions were to begin immediately after baseline data collection and district assessments were completed. This allowed not only for the time for the intervention to take place but also for an additional period of time between the end of the intervention and the beginning of endline data collection. In other words, elders who participated in the training and other activities would have time to take the new skills back to their villages and practice them prior to the endline data collection. During that additional period of time disputants would have their disputes heard and resolved by elders with new skills and tools. The evaluation design was consistent with this time frame for intervention, practical experience, and disputant reactions to the modified TDR bodies.

[4] The counterfactual is an estimate of what would have happened in the absence of the program.

In reality, activity implementation was between two and four months after baseline data collection, and the period for disputes to be settled was within one month after treatment had ended. Some disputes in the sample come from the time period during elders' training, and the extent to which these conflicts benefited from initial or partial RLS-I intervention is likely to be minimal. In the north, with the CPAU intervention model, far fewer weeks of intervention were completed before data collection began.

This extremely limited period for RLS-I intervention and elders' practical experience with their new skills made the detection of impacts in the evaluation much less likely. RLS-I and USAID agreed at the evaluation's outset that impacts were unlikely in the planned short time frame, and that any impacts found were likely only at the first level of the logic chain – elders' knowledge gain. However, RLS-I and USAID also agreed on the evaluation's goals of learning for development effectiveness. These lessons learned, then, are a significant focus of this report.

Insufficient time allocated for disputes to be resolved and reconciled

The endline study was carried out two to four months after the treatment intervention began, depending on district schedules. As a result there are measurement limitations because of the short duration of treatment. The program inputs were likely not substantive or sustained enough to be able to detect treatment effects among elders and, even less so, all the way to the adjudication of disputes. It is also important to note that change in disputant perceptions may be a heavily lagged variable: though adjudication itself may change, disputant perceptions of changes in adjudication may take longer to filter through the population and be detected by such a study. The short time frame of implementation exacerbates this effect.

Re-interviewing respondents so soon after baseline data collection was also problematic, as respondents tended to refuse re-interviews or particular questions more frequently, remembered knowledge questions and responses from the first iteration, confounded results with angry or resistant responses, "hid" behind long-ago disputes, or were simply annoyed by the burden of two long interviews in a short period. Disputants sought out in the endline data collection were asked to report on disputes occurring within that short time frame. This limited the possible pool of disputes but also meant that the subtler social and personal feelings about those disputes were more raw and immediate. In Arghandab, for instance, respondents from some villages had recently been involved in the resolution of cases of murder or manslaughter but were not ready to recount such experiences. In the future RLS-I will need to consider more carefully the appropriate time to collect data after the resolution of a dispute such that it remains timely and relevant, but also that recounting the dispute would not distress respondents or risk re-igniting tensions within the community.

Survey fatigue

Many interviews were one hour in length. According to data collectors, some respondents were impatient to finish and return to their work or other activities. Survey fatigue is a threat in development environments of all types; in Afghanistan, however, the prevalence of survey research is very high in a number of regions where conflicts are incipient or recent. The likelihood of respondent refusals to participate, or refusals to answer particular questions, is higher in these areas. There is also a risk of negative reactions when community members or leaders have expectations that surveys will result in development activities. For comparison districts in particular, their participation is subject to goodwill; the lack of development investment after baseline data collection can erode that goodwill prior to endline data collection.

Dearth of female respondents

Few respondents among the key informants and, to a lesser degree, disputant surveys were female. As the program emphasized reducing harmful practices toward women and children, this lack of data could potentially have missed any harm reduction accomplished by RLS-I. The endline data collectors were able to interview a greater number of women, allowing the first robust measurement of the gender deficit in disputant perception. While this first evaluation round cannot measure change in female disputant perception, RLS-I now has a baseline of female disputants to serve as the baseline measurement in the event of future evaluation research in a subsequent program phase.

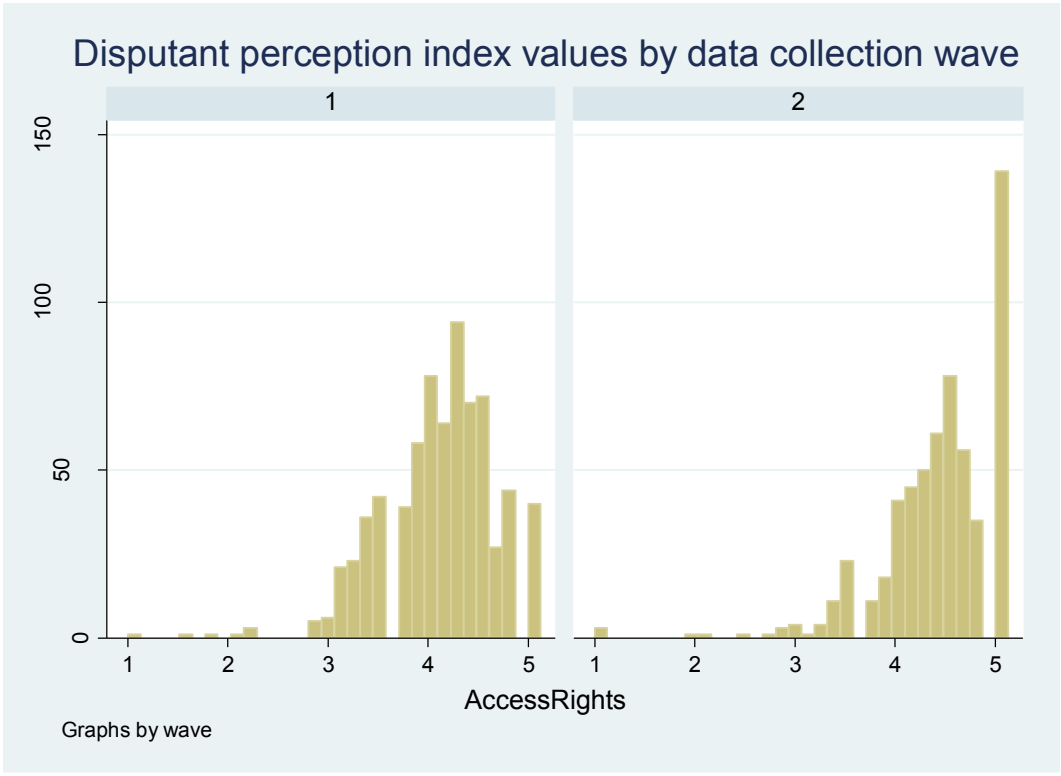
Change in research partner

As described above, a different firm was selected to conduct the endline data collection. This change had important effects – both positive and negative – on the resulting data and their analyses. In terms of field work techniques, the differences in data from baseline to endline as a result of switching research partners were greater than anticipated. There were in fact extreme differences in the measures from baseline to endline, most contrary to expectations based on the theory of change. Differences in the demographics and experience of the data collectors may have played a role, with one firm employing older and more experienced enumerators for whom rapport was easier, but for whom probing for deeper or more sensitive responses seems to have proved more difficult. These differences highlight the idiosyncratic data collection methods of each of the subcontractors, rather than any treatment effect or lack thereof.

As an example, Figure 2.6 below illustrates the differences in baseline and endline data in the disputant perception index of Access Rights, which consists of seven questions asked of respondents on how they perceive their access to justice. The average rating given by respondents to the seven items is listed on

the horizontal axis, on a scale of one to five. The length of the bars reflects the number of respondents who provided those average ratings.

Figure 2.6: Response differences from baseline to endline, access to justice index



Baseline values show a distribution of responses across the population, with slightly more frequent responses at the high range of the scale, but on balance following a roughly normal distribution. Endline values, on the other hand, show a severely skewed population with an excessive listing for values of the maximum scale value of 5, resulting in a nearly bimodal rather than normal distribution. A 5 on the index value requires that all 7 items making up the scale were marked as a 5. Endline enumerators therefore appear to have used the scale more as binary yes-no type of question, rather than probing for the nuance afforded by the 5-point strength of response scale.

The point may be illustrated more generally by isolating all disputant index values that scored at the maximum scale value of 5. This is shown in Table 2.7.

Table 2.7: Percent of cases with the maximum index value (5)

Time	Access Rights	Decision Subverted	Free Forum	Outcome Just
Baseline	5%	27%	23%	21%
Endline	22%	41%	38%	32%

The values for endline differ greatly from the baseline values, suggesting a misapplication of the response scale at endline and a resulting difference in the data unrelated to treatment. This confounds any attempt to detect movement in the data attributable to RLS-I. Such a systematic difference in applying the response scale is thought to be a major factor behind the likely incomparability between disputant perception scores from baseline to endline.

At the same time, advantages of the choice of a different survey firm were also apparent in the endline data, confirming some weaknesses in the baseline data collection. For example, a greater number of women were surveyed at endline, including among women who were party to disputes. This allowed for greater understanding of the differences in women's experiences of TDR in their communities, and the inequalities inherent in many *jirgee*. In addition, these data can serve as comparison for ongoing evaluation research on the deficit in women's participation and access to justice.

Additionally, the endline research partner appears to have conducted a more competent survey at the level of citizen or household. At baseline, the data showed very low incidence of *baad* and forced marriage, far lower than what RLS-I anecdotal reporting had suggested. At endline, the incidence was higher and closer to expected levels of incidence of *baad* and forced marriage. Differences in the way enumerators asked the questions and probed for understanding appears to account for this variation, as opposed to a multi-fold increase in incidences of *baad* in the few months since baseline data were collected.

Individually, the baseline and endline data sets are valid and useful. However, it is believed that measurements from baseline to endline are in many cases invalidated due to the greater than anticipated differences in data collection methods.

Challenges in the use of comparison groups

Conflict-affected environments pose unique challenges for evaluation research. Targeted interventions often, as in the case of RLS-I, do not permit randomization and, therefore, the use of experimental design. This is because the districts selected for intervention often have particular characteristics that draw donors' attention and intervention, such as districts that have recently been "cleared" of insurgents. The inability to randomize leaves quasi-experimental design options, including the use of comparison districts, to identify the counterfactual – that is, what would have happened had the intervention not taken place.

Comparisons between districts under Afghanistan's conflict-affected conditions are also difficult. Ethnic and tribal overlays on state administrative boundaries (such as districts) translate into variability that is not easily matched with nearby districts. Spillover is likely among networked elders, among whom knowledge and/or a seat in the training venue can be an influential good. Networking as a part of RLS-I

programming provides benefits to participants; but for the purpose of comparisons across districts it often presents favorable conditions for spillover. Localized insurgency or other power-related events within a district also limit the reliability of comparisons across boundaries. The fluidity of political and security dynamics are highly unpredictable and variable across districts, as individuals, militant groups, tribes or others assert claims over smaller and larger areas. If a district in which this is occurring is matched in the evaluation design with a district that is more stable, research findings can mask unobserved variables. Each of these factors threatens the critical assumption in the evaluation design that in the absence of RLS-I programming treatment and comparison districts would share the same rate of maturation.

Interpreting findings in light of challenges

As a result of the aforementioned challenges, the impact evaluation was not able to detect statistically valid change scores on many measures. However, the data collection and analyses were extremely useful in three other respects.

First, the exercise provided thorough documentation of the dynamics of dispute resolution in Afghan communities and is immensely rich in what the USAID Evaluation Policy refers to as "Learning for Effectiveness," one of the two primary motivations behind investing in impact evaluations. Responses to questions asked in the field provide a wealth of data on the strength of the instruments and items created and allow for improvements for a future longer-term study designed to detect effects. The data collected for the evaluation serve as an effective baseline on TDR interventions for future programming use for those populations.

Second, the exercise provided the basis for an impact monitoring system that can, with repeated application, robustly track changes in attitudes and practices relating to informal dispute resolution and serve as a critical measure of development effectiveness leading to transition to full Afghan control by 2014.

Third, specifications for data collection in both the panel and cross-sectional design will allow for reviewing the data in different ways – by exact match of individuals at baseline and endline (for those individuals willing to be re-interviewed), by propensity score matching of individuals with parallel characteristics, and across the entire dataset, at different points in time. These different technical methods allow for a more nuanced view of change and are explored in this report. As a result of these challenges in the data, RLS-I has been diligent in examining the data for the findings that are detectable, but also in weighing those findings against data weaknesses before drawing conclusions. As a result, this report highlights the findings of the report – what the data do show – but always in light of the strength of the underlying comparison in making inferences about programming.

III. FINDINGS

This section presents the data from baseline and endline data collections, including the D-in-D change scores. The section is organized by hypothesis, followed by findings on secondary research questions and extensions to the core analyses and measurements. Findings differ from conclusions in that the former show the data at "face value", including some of the contextual issues around the data that may affect their validity or reliability. Conclusions, by contrast, involve reflection on what the data actually mean, after contextualizing them with data quality, their fit with prior knowledge, and their utility for explaining existing theory.

The RLS-I development hypothesis is that skills- and knowledge-building of informal justice providers increases stability through increased access to justice and citizen confidence in TDR mechanisms. While longitudinal measurements were clouded by issues of incomparability from baseline to endline, as described in the section on "Methodology" above, examination of relationships between elder knowledge, disputant perception, and various program metrics within only the treatment group suggested important lessons for improving or refining programming.

Hypothesis 1

The intervention will result in TDR decisions that better reflect and/or are based in Afghan law, Shari'ah, and human rights norms

Hypothesis I is measured by elder self-reports of changes in adjudication of disputes, unsuccessful attempts to change adjudication so as to better reflect Afghan law and *Shari'ah*, and gains in specific knowledge points of Afghan law and *Shari'ah*. Citizen gains in knowledge of Afghan law are included here as a complement to elder knowledge, and to demonstrate the possible change in citizens where outreach efforts are active. Finally, knowledge change scores are cross-referenced against a variety of contextual factors to help identify the environmental determinants, if any, that facilitate knowledge gain.

Knowledge questions at baseline and endline

The theory of change underlying RLS-I presumes that improving elders' knowledge of Afghan law, *Shari'ah*, and international human rights norms will strengthen their dispute resolution practice. The impact evaluation baseline instrument for elders included twenty-two questions designed to elicit elders' knowledge on four legal themes from the workshops. The short time frame of the intervention made significant impacts unlikely, so these items were designed to serve as a pilot, to understand item

response rates and procedures, and to allow the incorporation of lessons learned for a longer intervention.

Baseline results were extremely high on five of the knowledge questions (more than 89% of respondents answered these questions correctly), indicating that these were too simple for respondents. With such high scores, the comparison with endline would likely be no effect or negative effect. These outcomes were most likely because of the scant room for improvement above the high scores and because the phenomenon of "regression toward the mean" would likely result in a reduction in scores. Because of these factors, these items would not be useful for the long-term impact evaluation design. Refining these questions and adding greater difficulty was necessary. The new items were to be tested in the endline data collection and, at the same time, included in multiple-item indices by topic. For purposes of both the Phase 2 impact evaluation and the longer term, adapting these questions was an imperative prior to endline data collection. In line with standard questionnaire design methodology for longitudinal studies, to refine the instrumentation and improve the tool's ability to test knowledge, new and somewhat more difficult questions were proposed.

RLS-I developed replacement questions on the same themes that were open-ended, requiring a short answer instead of a simple true-false response. The baseline and replacement questions are shown below, categorized into the workshop legal themes and the percentage of respondents answering these questions correctly at baseline.

Table 4.1: Baseline and replacement knowledge questions, by topic and baseline score

Baseline knowledge items dropped for endline (true-false)	**Topic**	**Baseline score**	**Replacement questions** (short answer)
According to *Shari'ah*, if a husband dies without children, his wife shall receive ¼ of the inheritance.	Inheritance	89%	According to *Shari'ah*, if a husband dies without children, what share of the inheritance shall his wife receive?
According to *Shari'ah*, if someone accesses a stream to irrigate his land, but the irrigation channel crosses other people's lands, the owners of those lands cannot prevent the person from digging this irrigation canal.	Property	90%	There are two pre-emptors: one person is a shareholder in the land being sold, while the other person pre-empts the boundaries of the land. Which person has precedence in invoking the right of pre-emption?
According to *Shari'ah*, it is better for the guardian of a female to wait until she is adult age before marriage and then seek her consent.	Family	94%	According to *Shari'ah*, what are the two conditions a sane and mature woman must meet in order to enter into Nikah without permission of her guardian?

Baseline knowledge items dropped for endline (true-false)	**Topic**	**Baseline score**	**Replacement questions** (short answer)
According to *Shari'ah*, if a husband dies, his wife shall receive 1/8 of the inheritance, while the rest will be distributed to the sons and daughters, with two portions going to the sons for each portion going to the daughter.	Inheritance	97%	According to *Shari'ah*, if a man with children dies how much does his wife inherit from the legacy?
According to *Shari'ah*, a guardian of a female should not allow her to enter into a marriage she is not satisfied with.	Family	93%	In *Shari'ah*, if it is known that a proposed marriage will lead to suffering and still proceeds, is the marriage agreement valid?

Unfortunately, after translation, the research partner staff formatted the answer space for the new questions (where open-ended responses were to be written in) as if they were again true-false questions. This invalidated the measures, since data collectors checked a box instead of answering the question as it was asked and translated. In terms of measuring change in knowledge over time, this left the remaining questions – the ones that baseline data show to have been more difficult for the elders. The following table shows the number of these questions by their respective topics at baseline and endline.

Table 4.2: Number of knowledge items, by topic, baseline and endline

Topic	**Number of items**	
	Baseline	**Endline**
Afghan law	8	8
Family	6	4
Property/Deeds	4	4
Inheritance	3	0
Total	**21**	**16**

Two knowledge items from inheritance, two from family, and one from property were lost from baseline to endline. Only one inheritance question remained; this question was included with the other property questions, because it was closest to those questions. All knowledge measurements for the endline report follow the arrangement at endline; that is, the 16 questions available for measurement at endline are the ones used for the sake of comparison from the baseline data.

Change in knowledge

The knowledge measures show change between baseline and endline on each of the questions; change is positive for treated respondents on seven of the knowledge items, and negative on nine items. Comparison respondents generally had the same change reaction in the data; this pattern supports RLS-I's analysis that the endline data collection subcontractors asked these questions very differently and accepted responses differently (with more or less strict interpretation of respondents' answers, for example from the baseline data collectors). There is also evidence in narrative responses of spillover from treatment to comparison districts.

Table 4.3 demonstrates how difference-in-differences (D-in-D) data are arrayed in this report (see Evaluation Measurements for more detailed discussion). Tables regarding this topic first show the treatment and comparison scores for baseline, then treatment and comparison scores for endline, followed by each baseline and endline change score (the difference between treatment and comparison for each time period). The final score is the second differencing – the difference between the baseline change score and the endline change score.

Table 4.3: Presentation of Difference-in-Differences (D-in-D) findings

Knowledge question	Baseline (T)	Baseline (C)	Endline (T)	Endline (C)	Difference (T)	Difference (C)	Treatment effect
Item	A	B	C	D	C-A	D-B	**(C-A) – (D-B)**

While there is variation among the D-in-D results (both positive and negative gains are seen in elders' knowledge scores), neither USAID nor the evaluation team expected to see impact from baseline to endline. The time period of the intervention was too short to expect movement on the knowledge measures. The challenges and complexity of this capacity building intervention also precluded any expectations of change after only two to four months of intervention. The target audience is particularly hard to reach: largely illiterate and remote from even district centers, the elders face significant obstacles to sustained learning. The insecure environments in the selected project districts affect attendance at training events. Qualitative responses from event evaluations indicated that there was more content in each one-day workshop than elders felt they could absorb, understand, and utilize once they had returned home. Importantly, capacity building efforts often have an initial effect of challenging participants' understanding, leading to a sense of uncertainty about their own knowledge and skills. One effect of measuring impact before participants have had sufficient time to assimilate knowledge is that their responses evince this uncertainty and confusion. Significant spillover also appears to contribute to the variability in scoring, in which comparison district respondents have taken part in RLS-I events or and/or learned about them from their peers in treatment districts.

For these reasons, the expected treatment effect was zero for the RLS-I Phase 2. The actual changes by knowledge question are presented in Table 4.4 below. The sixteen questions included are those from baseline that remained on the endline instrument.

Table 4.4: Percent of correct responses to knowledge questions, elders

Knowledge question	Baseline (T)	Baseline (C)	Endline (T)	Endline (C)	Difference (T)	Difference (C)	D-in-D
If someone is being held in police custody the elders can negotiate his or her release	87%	88%	90%	95%	2.9%	7.1%	**-4.2%**
Under *Shari'ah*, if a witness signs a deed where the claimant asserts false information, the witness is responsible for this false act even if he was not aware	93%	87%	51%	47%	-42.3%	-39.3%	**-3.0%**
If the police imprison you, you do not have the right to receive visits from your family	89%	86%	59%	51%	-29.3%	-34.4%	**5.0%**
If you are accused of a crime before a court of law, the government is required to provide you with a defense lawyer even if you cannot afford to hire one	41%	19%	88%	86%	46.5%	66.9%	**-20.3%**
If police accuse you of a crime before a court, the court assumes that you are guilty and you must prove that you are innocent based on evidence	8%	23%	14%	17%	5.7%	-6.5%	**12.2%**
According to *Shari'ah*, women do not have the right to own property	73%	79%	58%	49%	-14.5%	-30.0%	**15.4%**
If police detain you for any reason, they are allowed to hold you for a maximum of 72 hours. After this time, they must either bring a formal charge, or set you free	97%	87%	83%	82%	-13.3%	-4.9%	**-8.5%**
According to *Shari'ah*, if someone revives useless and unowned land, for example by constructing a building or planting crops, the revived land shall belong to the person who revived it	51%	53%	62%	61%	11.9%	7.5%	**4.3%**
According to Afghan law, the government courts are the only recognized body for handling criminal cases	38%	28%	67%	57%	29.4%	28.6%	**0.8%**

Knowledge question	Baseline (T)	Baseline (C)	Endline (T)	Endline (C)	Difference (T)	Difference (C)	D-in-D
If a woman is unhappy in her marriage and goes to stay with her parents, she has broken Afghan law for the crime of running away	56%	40%	43%	41%	-12.6%	1.0%	**-13.6%**
Under Afghan law, a woman has a right to request a divorce her husband	6%	6%	36%	48%	30.5%	41.5%	**-11.1%**
If one party has a valid deed to land and elders split the land with someone without a valid deed, the elders have violated the owner's property rights	84%	88%	50%	60%	-33.8%	-28.3%	**-5.4%**
If a person is tried in the government courts and convicted of a crime, under Afghan law the elders may negotiate his or her release.	20%	16%	57%	45%	36.6%	28.6%	**8.0%**
What does *Shari'ah* demand for conditions of request and acceptance for *nikah*?	93%	85%	49%	26%	-43.6%	-59.3%	**15.8%**
Is the practice of *baad* consistent with Islam and the Holy *Qur'an*?	97%	93%	71%	74%	-25.9%	-19.0%	**-6.9%**
Why is the practice of *baad* considered contrary to Afghan law?	85%	80%	45%	26%	-39.5%	-53.5%	**14.0%**

The response patterns from baseline to endline are roughly parallel between treatment and comparison respondents: when treatment responses rise or fall, comparison responses do in most cases as well. There are mixed responses on four questions. In addition, the questions with highest scores at baseline are those that fall in endline data collection, and vice versa. Such instances could be interpreted as regression to the mean, in which the endline scores on a given measure revert closer to their actual values in the population. This will at first glance appear to be a gain or loss in terms of treatment effect, but would in actuality be a normal correction for an initially uncommon value drawn from the population at baseline.

The knowledge measures also reflect differences in subcontractor field methods as mentioned above, in particular because the direction of change is the same for both treatment and comparison respondents. Enumerators at endline appear to have judged responses with a stricter standard, compared to the enumerators at baseline. The different methods result in data that appear to show the elders losing knowledge, but since that tendency is across both treatment and comparison samples this may be interpreted as an artifact of the data collection.

The individual knowledge questions were organized according to the topics of Afghan law, family law, and property law[5]. In addition to an overall law score, there is also a measure for all questions that were not related specifically to Afghan law. For lack of a better term, the set of questions not dealing explicitly with matters of Afghan law will be referred to as non-Afghan law items. Given that the workshop content on family, inheritance, and property law are more oriented toward basic *Shari'ah* content, this set of questions may also be considered a crude proxy for *Shari'ah*.

Table 4.5: D-in-D measurement of knowledge questions, by topic of law and overall

Topic	Baseline (T)	Baseline (C)	Endline (T)	Endline (C)	Difference (T)	Difference (C)	Treatment effect
Family	74%	70%	55%	46%	-19.0%	-24.0%	**5.0%**
Property	76%	77%	54%	54%	-21.6%	-23.0%	**1.4%**
non-Afghan	75%	73%	52%	48%	-22.5%	-25.5%	**2.9%**
Afghan Law	56%	49%	63%	61%	7.6%	11.6%	**-4.0%***
Overall	65%	61%	59%	55%	-7.0%	-6.4%	**-0.9%**

* Significant at 10%

[5] Inheritance-related questions were two of those that were invalidated by the error in instrumentation described above. As such, that category of questions from the baseline was lacking data at endline, and the remaining questions that the evaluation team could use were divided among family and property categories so that the number of questions would be sufficient in each category to make comparisons.

For Afghan law questions, baseline respondents were correct less frequently at baseline; at endline, the D-in-D measure shows a further reduction in knowledge relative to the comparison respondents. For property and family law, there are statistically insignificant improvements in knowledge among treatment respondents relative to comparison. Overall, there was a net negative effect, at a statistically insignificant level, resulting in an overall finding of no net effect from RLS-I activities.

A clearer picture emerges when the change scores are disaggregated by region, as shown in the table below. The south region is driving the negative change scores, while in the east there is a slight gain across respondents. The north region shows an interesting divergence between a strong gain in all items not related to Afghan law, and an almost equally negative performance on items pertaining to Afghan law.

Table 4.6: Change scores on knowledge questions, by topic of law and region

Region	Family	Property	Non-Afghan law	Afghan law	Overall
East	4.3%	1.2%	2.2%	3.9%	**2.5%**
North	17.5%	5.5%	12.2%	-8.8%	**2.3%**
South	-14.7%	-5.8%	-10.2%	-7.0%	**-9.8%**
Overall	4.3%	1.4%	2.9%	-4.0%	**-0.9%**

Media outreach and TDR: citizens' knowledge change

More citizen respondents (14%) reported receiving some messaging about legal rights or *jirgee/shuragani* in the past three months at endline than at baseline (3%). Over two-thirds (70%) of those receiving such material said that it was either somewhat or extremely beneficial for such information to be shared through public materials such as booklets, calendars, television, and radio.

Citizens were also asked six questions on Afghan law and their rights. Across these six questions, the resulting change score shows treatment group citizens gaining 6% more knowledge than did comparison group citizens. The questions and responses for treatment districts (those with the outreach component) and comparison districts are as follows:

Table 4.7: Percentage of correct responses to knowledge questions, citizens

Questions	Baseline (T)	Baseline (C)	Endline (T)	Endline (C)	Difference (T)	Difference (C)	Treatment effect
If a woman is unhappy in her marriage and goes to stay with her parents, she has broken Afghan law for the crime of running away	52%	46%	24%	52%	-28%	6%	**-34%**
If you are accused of a crime before a court of law, the government is required to provide you with a defense lawyer even if you cannot afford to hire one	43%	16%	75%	70%	32%	54%	**-22%**
If police accuse you of a crime before a court, the court assumes that you are guilty and you must prove that you are innocent based on evidence	16%	21%	21%	27%	5%	6%	**-1%**
If police detain you for any reason, they are allowed to hold you for a maximum of 72 hours. After this time, they must either bring a formal charge, or set you free	90%	91%	76%	51%	-14%	-40%	**26%**
If the police imprison you, you do not have the right to receive visits from your family	68%	71%	65%	61%	-3%	-10%	**7%**
According to Afghan law, the government courts are the only recognized body for handling criminal cases	43%	56%	78%	53%	35%	-3%	**38%**

The response patterns from baseline to endline are again parallel between treatment and comparison respondents: when treatment responses rise or fall, comparison responses do as well, for four of the six questions. In addition, the questions with highest scores at baseline are those that fall in endline data collection; this can be interpreted as a regression to the mean, in which the endline sample of individuals answers incorrectly more often than did the baseline sample.

The knowledge measures also reflect differences in subcontractor field methods just as in the case of the elders' knowledge responses. Again, the direction of change is the generally same for both

treatment and comparison respondents[6]. Enumerators at endline may have judged responses with a stricter standard, compared to enumerator behavior at baseline.

Elders' change in adjudication

Elders were asked at baseline and endline whether or not there was anything different in the way they or their communities resolved disputes since the time of the RLS-I intervention. Ten percent more treatment respondents said that something had changed in their adjudication of disputes than did comparison group respondents. From baseline to endline, treatment respondents went from 20% saying there had been a change, to 49%. Comparison group respondents also were more likely to say there had been a change, but at a lesser rate of increase (16% to 34%).

Respondents were asked to specify what they did differently. The three most frequently cited changes were that they are now documenting cases (10%), that they are now making just, impartial decisions that are accepted by the community (20%), and that they are now following Afghan law, *Shari'ah*, or human rights norms (19%). Other responses are shown in the table below with frequencies and the percentage of respondents who cited those changes. Responses also mentioned increased cooperation between *jirgee* and district government and no longer allowing forced marriage.

Table 4.8: Changes cited in resolution of disputes, by program participation

Changes to adjudication	Endline respondents *n=243*	
	#	Gross %
Now seeking alternatives to *baad*	10	4%
Now documenting cases	25	10%
Now not demanding *machalgha*	4	2%
Now making just, impartial decisions that are accepted by the community	49	20%
Now women's disputes are also resolved	3	1%
Now following Afghan law, *Shari'ah*, or human rights norms	45	19%

[6] With only six questions in this item bank compared to sixteen for elders, the evidence for subcontractor differences among citizens is somewhat weaker; however, the pattern does still appear to hold.

Hypothesis 2:

The intervention will result in TDR decisions and* shura/jirga *members being perceived as more impartial

Measuring access to justice

The core impact evaluation measurements for the disputant case assessment consisted of a battery of attitudinal items on case resolution process and outcome. The attitudinal items were adapted from a methodology of measuring the costs and access to pathways of justice established by the Tilburg Institute for Interdisciplinary Studies in Civil Law and Conflict Resolution Systems (TISCO). The TISCO *Measuring Access to Justice Handbook* establishes ten dimensions of measurement according to a five-point Likert scale capturing the extent to which the disputant believes a given statement to be true.

RLS-I adapted the TISCO methodology and pathways to justice by establishing four paths of inquiry specific to USAID Rule of Law objectives and evaluation hypotheses:

- Access rights: procedural justice regardless of decision
- Decision subverted: lack of corruption from external actors or within *jirga*, or both
- Freedom of forum: lack of coercion in choosing forum
- Outcome just: fairness and equity of decision[7]

Following the assessment scale in the TISCO *Measuring Access to Justice Handbook*, each item was evaluated along a five-point Likert scale: To no extent (1), To little extent (2), To some extent (3), To great extent (4), and Completely (5). Eighteen questionnaire items are included among these four indices. Respondents express their perceptions of how well these aspects of justice performed in their own cases.

The summary values of the four index items are tabled below. For convenience, the degree of support for a given statement is collapsed into a binary value, with a positive value assigned to any selection of To some extent (3), To great extent (4), or Completely (5).

[7] The International Development Law Organization (IDLO) followed a similar process in conducting exploratory research comparing the measurements of formal and informal pathways to justice, which was cited as a possible model to follow in the USAID Request for Task Order Proposals for RLS-I. In the specific context of RLS-I baseline evaluation data, the intent is not to compare formal and informal paths to justice (though gain and loss data relative to other pathways could be envisioned for future data collection), but rather to test directly the RLS-I development hypothesis that program inputs will improve disputant perceptions of the *jirga* decision-making process and outcomes.

Table 4.9: Value of four indices, overall

Index	Index value
Access rights	4.12
Decision subverted	1.43
Freedom of forum	4.35
Outcome just	4.41

Partly for ease of presentation and computation, it is these mean index values that serve as the primary measurements of Hypothesis 2 and other outcome- and impact-level measures.

In the table below, the D-in-D for index scores is shown. This is the degree to which disputants' perceptions from treated districts changed compared to the changes in comparison districts. For the three positive-scaled indices, the change in disputants' perceptions from treated districts was less than that in comparison districts. The Decision Subverted scale is reversed, in that a lower score is better (i.e., disputants perceive less that their decision was somehow negatively influenced by some type of subversion.) On that index, the direction of difference is positive, but the finding was not large enough to be statistically significant.

Table 4.10: Disputant perceptions on index measures

Index value	D-in-D change score	p-value	Effect size
Access rights	-0.16	0.019	-0.11
Decision subverted	-0.06	0.580	-0.03
Freedom of forum	-0.40	0.000	-0.26
Outcome just	-0.23	0.001	-0.16

Previous discussion has already established the likely invalidation of the D-in-D measures for disputant assessment data. However, other possible explanations should not be ruled out. One such explanation is that RLS-I programming may have raised awareness among citizens of their rights and protections under the law, regardless of what forum they chose to resolve a dispute. This awareness could have had the effect of changing the respondent's assessment of the response scale itself from baseline to endline, thus making the negative treatment effect an actual program result, as disputants' assessments became more demanding.[8]

[8] A recent impact evaluation of a peacebuilding program in Liberia that targeted traditional justice structures reported a similar result in that the program had the short-run effect of exacerbating existing or past conflicts as a result of opening dialogue about healing conflict. (Blattman, Christopher; Alexandra Hartman; and Robert Blair. 2011. "Can we teach peace and conflict resolution? Results from a randomized evaluation of the Community Empowerment Program (CEP) in Liberia: A program to

However, in the absence of a stronger theoretical explanation, incomparable data is the most likely explanation of the negative treatment effects detected in the data. See the Extensions section for more promising data analysis within only the treatment group, linking improvement in disputant perception with the number of activities an elder attends and with the overall number of elders per district passing through the RLS-I core curriculum.

Regardless of the change scores, disputant perceptions may be cross-referenced against various background characteristics relevant to dispute resolution for better understanding of case dynamics. In the table below, a selection of background factors is correlated with each index. The cells indicate both the direction (positive or negative) and magnitude (strong, weak, or nonexistent) of the relationship.

Table 4.11: Disputant perception scores correlated with possible explanatory factors

	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
Female	Strong negative relationship	Positive relationship	Positive relationship	Strong negative relationship
Government trust	Strong positive relationship	Strong positive relationship	Weak positive relationship	Strong positive relationship
South	Moderate negative relationship	Strong positive relationship	Weak negative relationship	Strong negative relationship
Duration	Strong negative relationship	Positive relationship	Negative relationship	Weak negative relationship
% of Tashkil (District court staffing)	Negative relationship	Weak positive relationship	Weak positive relationship	Positive relationship
District court caseload (1389)[9]	Positive relationship	No relationship	Positive relationship	Weak positive relationship

build peace, human rights, and civic participation." Accessed on August 24, 2012 at Innovations for Poverty Action website at https://www.poverty-action.org/sites/default/files/blattman_hartman_blair_can_we_teach_peace_ipa_liberia_0.pdf.

[9] The Islamic Year 1389 covers the period from on/about March 21, 2010 to March 20, 2011.

The sample of disputants at endline included more female respondents (11%) compared with the baseline sample (4%). This allows for more detailed analysis of responses from females involved in disputes. Women rated access to justice and a just outcome more negatively than did male respondents; they also were more likely to perceive subversion of the decision, though the trend is not as strong as with the other two indices. Similarly, women perceived more freedom in choice of forum than did men; the relationship between being female and a higher score on that index was positive, though not as strong as the negative relationship on access to justice and the just outcome.

Respondents who indicated stronger trust in government had higher perceptions on all four indices. On the other hand, the longer the duration of the process to resolve a respondent's dispute, the poorer the perceptions on all four indices. Similarly, respondents from the south tended to perceive lower scores for all four indices, with stronger negative correlation for both the just outcome and the possible subversion of the decision.

Impartiality – the elders' perspective

The elder interview included a battery of questions on impartiality, using a set of statements with a scaled response: Strongly Agree, Somewhat Agree, Somewhat Disagree, and Strongly Disagree. For purposes of the analysis, the responses are collapsed into Agree and Disagree, to test for changes from baseline to endline. The table below shows "agree" responses to the seven questions.

Table 4.12 Elder perceptions of impartiality, by region, percentage of agreement

	Treatment Baseline *n=190*	Treatment Endline *n=208*	Comparison Baseline *n=229*	Comparison Endline *n=227*	D-in-D score
Sometimes *jirga/shura* members base their decisions on which party is more powerful in the community	11	41	14	40	4
Jirgee/shuragani are often unwillingly influenced by people with their own interest in a case	11	35	11	38	-3
Decisions made by *jirgee* or *shuragani* are usually consistent with Afghan law	71	90	67	90	-4
Decisions made by *jirgee* or *shuragani* are usually consistent with *Shari'ah*	88	90	94	90	6
Decisions made by *jirgee* or *shuragani* are fair	95	93	94	92	0
Decisions made by *jirgee* or *shuragani* find justice for the community	95	89	95	89	0
Decisions made by *jirgee* or *shuragani* consult all relevant parties or witnesses to a case, including women	75	78	60	84	-21***

*** *Significant at .01 level*

The differences between treatment and comparison groups were usually quite small, while the difference from baseline to endline for both groups is large. This again appears to be an artifact of the change in data collection partner, since the changes from baseline to endline are often very significant. The idiosyncrasies of data collection between the two groups have more power to explain these results than would the assumption that RLS-I has had a negative or neutral effect. Spillover is also highly possible. With the exception of the last difference, the differences are not statistically significant; the final question about consulting all relevant parties is statistically significant at the .01 level (designated throughout this report with the customary *** notation.)

The question of whether or not the elders are influenced by outside interests is also asked, in a set of questions about whether people who are not part of a case ever attempt to influence its outcome. More than half (57%) of treatment district elders at baseline said this did happen; at endline, only 28% said this happened. Those who said it did happen also noted that it happened about three-quarters of the time, consistent with the figure at baseline. Nearly two-thirds of respondents at baseline said that these attempts were "never" successful, compared to 38% at endline. At baseline fewer than 8% said that these attempts were successful somewhat or very often, while at endline nearly half said these attempts were successful somewhat or very often. Idiosyncratic data collection is likely to be the cause of these discrepant patterns in the data.

Impartiality – citizens' perceptions

The citizen perception interview also includes the battery of questions on impartiality, using the same set of statements with the scaled response: Strongly Agree, Somewhat Agree, Somewhat Disagree, and Strongly Disagree. For purposes of the analysis, the responses are collapsed into Agree and Disagree, to examine the D-in-D change scores. Respondents from treated districts rated their *jirgee* or *shuragani* more impartial, and in five of these response rates, the difference was statistically significant.

Table 4.13 Citizen perceptions of impartiality, by region

	Treatment Baseline *n=454*	Treatment Endline *n=257*	Comparison Baseline *n=273*	Comparison Endline *n=319*	D-in-D score
Sometimes *jirga/shura* members base their decisions on which party is more powerful in the community	29%	58%	13%	44%	-2%
Jirgee/shuragani are often unwillingly influenced by people with their own interest in a case	30%	55%	18%	40%	3%

	Treatment Baseline *n=454*	Treatment Endline *n=257*	Comparison Baseline *n=273*	Comparison Endline *n=319*	D-in-D score
Decisions made by *jirgee* or *shuragani* are usually consistent with Afghan law	64%	84%	80%	86%	14%***
Decisions made by *jirgee* or *shuragani* are usually consistent with *Shari'ah*	80%	85%	93%	83%	15%***
Decisions made by *jirgee* or *shuragani* are fair	86%	83%	98%	73%	22%***
Decisions made by *jirgee* or *shuragani* find justice for the community	87%	80%	96%	73%	16%***
Decisions made by *jirgee* or *shuragani* consult all relevant parties or witnesses to a case, including women	53%	57%	57%	48%	13%***

*** *Significant at 1%*

The differences between treatment and comparison groups were larger for citizens, while the difference from baseline to endline for both groups remains generally large. This may also be an artifact of the change in data collection partner. All but the first two differences are statistically significant at the .01 level. However, the idiosyncrasies of data collection between the baseline and endline subcontractors cannot be discounted in this analysis, as they may explain these results better than would the assumption that RLS-I caused these generally positive effects. Spillover is also possible.

Hypothesis 3

The intervention will result in a decrease in the number of TDR decisions that negatively impact women and children

Elders' views on forced marriage and *baad*

The RLS-I Impact Evaluation Plan did not predict higher order effects such as social change at this level to occur after only three months of project implementation. From these data, no effect on *baad* or forced marriage can be determined. That being said, treated elders were asked about changes in their *jirgee* as a result of RLS-I interventions and ten elders specifically mentioned that their *jirgee* no longer

used *baad* to resolve disputes, but rather some compensation in the form of labor, land or money. Three respondents said that forced marriages were no longer acceptable in their communities as a result of the messages brought back from the training.

The elder interview included a set of questions about dispute resolution practices that are harmful to women and children. Forced marriage, or marriage against the will of one or both parties, was the topic of one set of questions. Only five elders responded that they knew of any such cases at baseline, less than 5% of the sample, across both treatment and comparison samples. At endline, 22 elders said they knew of such cases, over 10% of the sample.

Similarly, when asked how many instances of forced marriage or marriage against the will of one or both parties they could recall, 8.5 percent of baseline respondents cited figures totaling only eight cases in total. At endline, 53 respondents reported knowing of a case in the last three months, which represents a significantly larger 27.5% of the sample.

The degree of difference in these findings is most likely related to the ways the two data collection subcontractors asked questions of respondents. At baseline, RLS-I surmised that, to some extent, the low reporting of *baad* and forced marriage could be tied to the sensitivity of the topic for respondents, and in the case of *baad*, the relative infrequency of major cases (such as for murder) being resolved through *jirgee*. The endline enumerators may have been better able to establish rapport with respondents, or there may be other factors at work. Survey fatigue or annoyance at being queried again after such a short period may have resulted in some respondents inflating responses or providing confounding responses. The exact circumstances of the data collection and the differences compared to baseline may be unknowable; the figures collected at endline, however, are more concordant with other published studies.

With respect to forced marriage, elders were asked their opinions on whether those marriages were more likely to face disputes. On average, at baseline the elders in treatment districts felt that such marriages later resulted in disputes between the husband and wife or between families about 33% of the time; at endline, the treatment district elders said this was true about 28% of the time. However, at endline, enumerators asked this question of all or nearly all respondents, while at baseline the enumerators only asked this question when the answer to the question of existence of forced marriage was yes; the measure is likely invalidated by this difference.

Hypothesis 4

The intervention will result in an increased role for women as disputants, witnesses or decision-makers

Elders' views on women's roles in TDR

This section of the interview included questions on women's roles in traditional dispute resolution – about their participation as decision-makers or witnesses and their ability to present their cases directly rather than through an intermediary.

The elders were asked whether it was possible for a woman to sit as a member of a *jirga* or *shura*. Across the baseline sample, 46% said this was possible, while 49% said it was not possible. At endline, just under 40% of treatment district respondents said this was possible.

Women's ability to testify in a *jirga* was also queried: if a case before the jirga directly involved a woman, how she would present her case? A slight majority of baseline respondents (56%) said that a family member or other representative would present the case on her behalf, while at endline the response was less than half of treatment respondents (46%). A small percentage at baseline (12.6%) said she would present her case personally before a member of the *jirga/shura*, but at endline this was over a third (38%). For comparison group respondents, the change was similar, indicating a difference in the way questions were asked rather than a program effect.

When asked, "If a woman had critical knowledge concerning a case, would she be called upon to testify before the *jirga* or *shura*?" over three-quarters of baseline respondents in treatment districts (78%) said that she would, and at endline the figure was roughly the same (77%). Further, elders were asked if, generally speaking, the testimony of women is accepted by the *jirga* or *shura*; the great majority at baseline (93%) and endline (90%) said that yes, women's testimony would be accepted at a *jirga*.

The foregoing questions queried elders about their principles regarding women's participation in *jirga* proceedings. The following question was asked about concrete cases involving women as disputants, witnesses or members of a *jirga* or *shura*, of which they might be aware. Seven percent of elders interviewed at baseline knew of a case, whereas at endline nearly 15% said they knew of a case where women participated in *jirga* proceedings, from among the treatment districts. This may be a result of the opening of *spinsary* groups as part of RLS-I activities in their districts; however, comparison district respondents also showed an increase for this response.

Elders' opinions about women's involvement in *jirgee or shuragani* were queried in three attitudinal questions as shown in the table below.

Table 4.15: Elders' opinions on acceptability of women's participation in TDR, D-in-D

	Treatment Baseline *n=188*	Treatment Endline *n=208*	Comparison Baseline *n=229*	Comparison Endline *n=223*	D-in-D score
If I were part of a *jirga* or *shura*, I would not accept having a woman sit on the *jirga/shura*	48	49	59	62	-2
The wives of the Prophet Muhammad (PBUH) played a role in resolving disputes	95	95	98	96	2
My community would not accept women as members of a *jirga* or *shura*	85	60	82	59	2

The degree to which treatment and comparison respondents changed from baseline to endline for the first two questions was very small, and the change scores are statistically insignificant. Change on the third measure, however, was larger for both groups (resulting in a small D-in-D score). This may reflect some effect from RLS-I and its spillover with the advent of *spinsary* groups in which women are participating to adjudicate disputes within families or it could be interference from the differences between the two survey firms carrying out data collection at baseline and endline.

Citizens' views on women's roles in TDR

As with elders, citizen respondents were asked about women's ability to testify in a *jirga*: if a case before the *jirga* directly involved a woman, how she would present her case? A majority of baseline respondents in treatment districts (70%) said that a family member or other representative would present the case on her behalf, while at endline the response was around half that figure (38%). A small percentage at baseline (7%) said she would present her case personally before a member of the *jirga/shura*, but at endline this was over a quarter (29%). For comparison group respondents, the change was in the other direction (that is, fewer respondents said a woman could present her own case), which may indicate a treatment effect, but caution must be used in interpretation due to the uncertainties in the data collection.

The degree to which women would be called upon to testify before a *jirga* dropped for both treatment and comparison groups from baseline to endline, in about the same measure (21%). This is likely a result of differences in the data collection, because of the size of the change and because there is no differentiation between treatment and comparison change. However, there was more nuance in the set of responses from treatment districts, in which respondents indicated conditionality: whether a woman

was called to testify might depend on the situation. In contrast, two-thirds of the comparison respondents simply answered that a woman would not be called to testify. Treatment district citizens were 25% more likely to say that it was possible for a woman to sit on a *jirga* or *shura* than were comparison district respondents.

Respondents were asked whether *jirgee* would, in principle, accept women's testimony. More than three-quarters of comparison respondents said a woman's testimony would not be accepted at all, while only 30% of treatment respondents agreed, and there were more variant answers from among treatment respondents. Using the D-in-D calculation, a 9% gain was seen in women's testimony accepted compared to respondents in comparison districts.

Secondary research questions

For the additional descriptive questions noted below, the information found in qualitative and quantitative data from the current dataset is presented here. Understanding these topics will allow feedback into the RLS-I components that are not part of the impact evaluation hypotheses:

- What linkages exist between village *jirgee* and *shuragani* and their formal justice sector counterparts at the district level? Have linkages been strengthened by the RLS-I intervention?
- What patterns among long-standing disputes are found in the respondent populations?

Linkages with the formal sector

Elders were asked about the documentation and registration of their *jirga* decisions to understand the linkage, if any, between village dispute resolution and the formal justice sector. In many qualitative replies elders reported documenting and registering decisions as one of the most important changes in their *jirgee* since participating in RLS-I. Across the sample, the quantitative data show that documentation and registration of cases had improved substantially more for elders in treatment districts than those in comparison districts. Table 4.16 shows the D-in-D calculation for documentation and registration of cases with districts:

Table 4.16: D-in-D change scores for documentation and registration of cases

	Baseline Treatment	Baseline Comparison	Endline Treatment	Endline Comparison	Change score (Treatment)	Change score (Comparison)	D-in-D

Are cases documented in your community?	42%	36%	63%	40%	21%	4%	**17%**
If yes, what percentage of cases are documented?	51%	35%	64%	69%	13%	34%	**-21%**
Are cases registered in your community?	12%	10%	50%	28%	38%	18%	**20%**
If yes, what percentage of cases are registered?	22%	36%	58%	54%	36%	18%	**18%**

The p-value for documented cases is 0.008, while the other three change scores are significant at the 0.001 level. There was progress in documentation and registration for both the treatment and comparison groups, but here were statistically significant changes for the better for treatment group respondents than for comparison group respondents. For comparison groups, the increase in documentation may have come from general development, alternative programming, spillover from RLS-I, or some other source, but the change among treatment districts is most likely purely the positive effect of RLS-I program treatment.

In fact, given anecdotal information gathered through RLS-I M&E activities, it is reasonable to assume that some spillover is in fact part of the effect in comparison districts. The D-in-D change score might actually underestimate the general treatment effect where spillover of program effects is occurring. For example, a 38% increase in registration is a program effect (as opposed to the 20% D-in-D change score), while at least part of the 18% increase in comparison district is a spillover. The negative treatment effect found by D-in-D for the extent of documentation could in fact be a spillover as well, given the pattern of gain for both treatment and comparison and the overall pattern of gain for these measures.

RLS-I access to registration books in treatment districts brings additional findings for the east and south. While the previous table showed what elders report for themselves, the record books show actual cases that are documented and those that are registered with district courts. According to a sample of registration books reviewed from treatment districts, 22% of elders have documented their cases, while this figure was closer to 40% in the self-reported data. Some 14% of elders register cases, slightly higher than the 10%-12% reported at baseline.

Table 4.17: Registration book data on documentation and registration, from a sample of treatment districts

Region	Elders sampled	% elders document	# disputes documented	% elders register	# disputes registered	% disputes registered

East	52	21.2%	18	9.6%	6	33%
South	84	22.6%	26	16.7%	14	54%
Overall	136	22%	44	14%	20	45%

Receiving the RLS-I decision books appeared to increase the likelihood of documentation. Of 180 respondents who did not receive an RLS-I decision book, 59% reported that decisions in their community were documented. For the 88 respondents who did receive an RLS-I decision book, 66% reported that decisions in their community were documented, a statistically significant difference of 7%. Distributing an RLS-I decision book did not have an effect on the self-reported rate of dispute registration.

Long-standing disputes

Respondents were asked about long-standing disputes, if any, with which the respondent might be familiar. As with questions about cases of forced marriage and *baad*, few elders knew of such long-standing disputes in their villages and communities but more respondents at endline provided information on such disputes. At baseline, only five respondents from across the baseline sample (treatment and comparison districts) cited and described long-standing disputes, fewer than 3%. At endline, 33 such disputes were described, about evenly split between treatment (16) and comparison respondents (17). Most respondents said there was just one such dispute rather than multiple long-standing disputes in their areas, though at endline there were seven respondents who said they knew of "two or three" such cases.

The interviewer asked the elder to describe a specific "difficult dispute," in which he or she has been involved. In 24 such descriptions the elder reported that this was the result of a prior or long-standing dispute (11.4%). Elders reported that their decisions were not accepted or implemented in fewer than 3% of these cases; 5% of elders wished they could resolve some point or aspect of the case differently.

IV. EXTENSIONS TO CORE ANALYSIS

Given the likely invalidation of scores based on change from baseline to endline across treatment and comparison groups, exploration of the determinants of movement within only the treatment group or examining relationships only at endline takes on greater relevance. This section examines dynamics between exposure to RLS-I activities, absolute levels of elder knowledge and disputant perception, and changes in knowledge and disputant perception as a function of exposure to RLS-I activities. The

cautious top-line finding is that exposure to RLS-I activities does positively affect disputant perceptions, but the transfer and application of knowledge is not the mechanism driving this change.

Validating the development hypothesis in the absence of longitudinal measures

The impact evaluation for RLS-I Phase 2 tests the development hypothesis that skills- and knowledge-building of informal justice providers improves citizen access to justice. This was to be documented through longitudinal measurements of both treatment and comparison groups that would demonstrate a treatment effect from RLS-I and by extension would validate the hypothesis. In the absence of validating the development hypothesis through a statistically significant treatment effect in elder knowledge and disputant perception, it remains to be seen whether the data might produce evidence of a statistical relationship between the degree of exposure to RLS-I activities and the impact evaluation measures. If such a link could be established or at least suggested it would be evidence in support of the theory of change even if it could not by itself validate the development hypothesis.

To help shed light on a potential relationship between degree of exposure to RLS-I treatment and degree of movement within treatment group outcome measures from baseline to endline, RLS-I first did a manual linking of disputants to the referring elder who helped adjudicate the dispute. In this way disputant scores could be cross-referenced against elder data in a search for correlation. Then elders were cross-referenced against the program activity database in order to access program metrics such as activities attended and size of district cohorts progressing through the curriculum. Program metrics for linked elder and disputant data were then cross-referenced against elder knowledge and disputant perception within two dimensions: the *level* of such knowledge and perception at endline, and the *change* in knowledge and perception from baseline to endline as a function of program metrics. The final analysis juxtaposed the two outcome measures of elder knowledge and disputant perception as linear functions of each other, with knowledge (or gain in knowledge) seen as a determinant of disputant perception (or positive improvement in perception). This is the most direct expression of the RLS-I development hypothesis and program results framework, in which training leads to improved knowledge, skills, attitude, and behavior followed by a corresponding shift in the perceptions of the users of informal justice.

Elder knowledge and disputant perception

Operationalizing the development hypothesis through the program results framework follows a standard theoretical progression from program inputs realizing change in participant knowledge, skills, and attitude, with subsequent change in behavior in their home environment leading to eventual social

change in the target communities. This modality of change can be explicitly tested due to the fact that a majority of disputants were identified by the same elders who were interviewed for knowledge and attitude. Examining disputant perceptions as a function of elder knowledge may help shed light on the question of whether knowledge of Afghan law and *Shari'ah* matters for disputant perception of procedural justice and equitable outcomes.

A simple table of correlations between elder knowledge and disputant perception identifies potential connections. Cases are limited to endline to guard against any unrelated changes from baseline to endline that might obscure or confound a connection between elder knowledge and disputant perception.[10]

Table 4.18: Endline correlations between elder knowledge and disputant perceptions

	Afghan law	Non-Afghan law
Access Rights	0.071	0.114**
Decision Subverted	-0.140***	-0.179***
Free Forum	0.037	0.068
Outcome Just	0.143***	0.089*

Afghan law is correlated with procedural justice and equity but not with subversion of decision or freedom of forum. Questions on non-Afghan law are correlated with subversion only. It is also important to note the need to unpack the overall set of questions into its constituent topics of Afghan law and non-Afghan law, as correlations in the overall topic are commonly driven by either Afghan law or non-Afghan law but seldom both. Regardless of the specific connections, the most important potential finding is that there is a connection between elder knowledge and disputant perception, as would be predicted by the development hypothesis.

The next step is to generate specific associations from the data to assess model fit. When knowledge scores are modeled as predictive values of disputant perception scores (including control variables for the north and south regions) the only significant finding is that between knowledge of Afghan law and disputant assessments of the equity of informal decisions. The mean knowledge score in Afghan law at endline (61%) is associated with an increase in disputant perception on the equity index by 0.21 on the five-point assessment scale (p=.021). See Annex Table 1 for the table of coefficients generating the predicted values.

[10] For example, scores on Afghan law increased from 53% to 61% from baseline to endline across both treatment and comparison groups, while scores on non-Afghan law fell from 75% to 50% from baseline to endline across both treatment and comparison groups. Limiting analysis to endline prevents finding an apparent correlation of programmatic relevance that is merely reflecting the secular change in time.

Elder knowledge and change in disputant perception

Whether a causal link may be suggested between RLS-I and the knowledge/perception relationship can be investigated more carefully through the relationship between a *change* in elder knowledge and levels of disputant perception scores at endline. Note that while the previous analysis has relied on the entire sample of elder knowledge scores cross-referenced against the entire sample of disputant perception scores that could be linked to a referring elder (n=980 for elders, n=579 for disputants), examining change scores severely limits the sample size to only those elders who could be explicitly linked from baseline to endline (n=189). The measures remain valuable, however, in that they can guard against misleading conclusions based only on relationships between levels of knowledge, which might reflect general trends in data rather than an actual connection of programmatic relevance.

District level means in knowledge change scores for Afghan and non-Afghan law are as follows:

Table 4.19: Mean change scores in Afghan law and topics outside Afghan law, by district

Province	District	Status	Average change in Afghan law	Average change in non-Afghan law	n
Nangarhar	Bihsud	Treatment	-3.4%	-34.2%	6
Laghman	Mihtarlam	Treatment	9.1%	-11.4%	17
Laghman	Alingar	Comparison	0.1%	-7%	15
Laghman	Alishing	Comparison	12.1%	-19.5%	16
Kunar	Nurgal	Treatment	24.7%	-22%	25
Kunar	Chawkay	Comparison	21.2%	-26.5%	12
Baghlan	Puli Khumri	Treatment	-33.9%	-37.5%	4
Faryab	Pashtun Kot	Treatment	2.1%	-23.2%	10
Faryab	Shirin Tagab	Comparison	13.6%	-14.5%	10
Zabul	Shahjoy	Treatment	-4.7%	-27.3%	9
Province	**District**	**Status**	**Average change in Afghan law**	**Average change in non-Afghan law**	**n**
Uruzgan	Chora	Treatment	16%	-43.8%	24
Uruzgan	Khas Uruzgan	Comparison	4.2%	-38.7%	8
Uruzgan	Shahidi Hassas	Comparison	32.3%	-26%	8
Kandahar	Arghandab	Treatment	-7.2%	-3.2%	7

Kandahar	Panjway	Comparison	0%	-29%	8
Overall			**10.2%**	**-21.9%**	**189**

As before, a table of correlations provides an initial snapshot of potential connections.

Table 4.20: Correlations between elder knowledge and disputant perceptions, change scores

	Δ Afghan law	Δ non-Afghan law
Access rights	0.195	-0.100
Decision subverted	-0.333***	-0.071
Free forum	0.218*	0.004
Outcome Just	0.216*	0.000

A change in elder score on questions pertaining to Afghan law has low to moderate correlation with the disputant assessment scores, but there is no correlation between an elder change in knowledge of non-Afghan law and the same disputant assessment scores. Improving elders' knowledge of Afghan law appears closely connected to disputant perception.

When disputant perception is modeled as a function of elder knowledge, the link between improvement in Afghan law and disputant perception persists for subversion of decision and equity of outcome, but falls away for procedural justice and freedom of forum. The mean change in knowledge of Afghan law from baseline to endline (10.2%) is associated with a fall in disputant assessment on the subversion of decision index by 0.14 on the five-point scale. See Annex Table 2 for the table of coefficients generating the predicted values.

Another illustration of the occasionally disparate relationship between knowledge of Afghan law and knowledge of the other training topics may be seen by cross-referencing them with each other. The correlation is -0.213 (p=0.003), suggesting a trade-off in learning one sort of knowledge or the other[11]. This can be seen intuitively by examining the above table of mean change scores by district. The districts of Mihtarlam, Alishing, Nurgal, Chawkay, Shirin Tagab, Chora, and Shahidi Hassas show extreme disparities in knowledge gain scores by the topics of Afghan or non-Afghan law. Another five districts show little or no gain in one topic followed by strong gain in the other. In only one district, Puli Khumri is there strong gain or loss across both topics.

[11] See Annex Table 3 for the relationship between change in Afghan and non-Afghan law in regression format.

This has direct implications for the stabilization thesis. While the development hypothesis is formulated and measured in terms of access to justice, the environment in which RLS-I operates is very much affected by counterinsurgency doctrine and stabilization. While not attempting to measure stabilization directly, RLS-I does incorporate elements of the stabilization thesis; i.e., that strengthening informal justice systems ultimately leads to greater linkages between the formal and informal justice sectors, thereby contributing to the GIRoA state-building process.

The trade-off between gain in knowledge in Afghan law and non-Afghan law points to a possible situation in which the Afghan formal and informal justice sectors are seen as opposed to one another rather than complementary, making the informal justice sector part of the battle space for hearts and minds in contested areas. The RLS-I role in such a battle space would be to shift knowledge and perception such that informal justice is seen as a complement to a functioning, transparent, and effective state justice system. One target perception is whether *Shari'ah* is thought to be the primary source of jurisprudence behind a given resolution to a dispute. RLS-I addresses this directly through its workshops introducing Afghan criminal and Constitutional law, as well as *Shari'ah* as it is reflected in Afghan law and the Constitution.

Change in elder knowledge and change in disputant perception

The final issue to be examined regarding the relationship between elder knowledge and disputant perception is whether a gain or loss in one is associated with a corresponding gain or loss in the other. Whereas previous analysis had examined all disputant assessment scores that had been directly referred by an elder against the overall knowledge scores of those elders, here the analysis is more exact. Cases are limited to only those elders who are in both the baseline and endline data, as well as who directly referred a party to a dispute the elder helped mediate, again both at baseline and endline. Given these strict parameters, sample size for change in disputant scores is only 60-70. Therefore, any findings using this analysis may provide evidence in support of the development hypothesis and help develop lines of inquiry for further research, but cannot by itself validate the development hypothesis.

The table of correlations is as follows:

Table 4.21: Correlations between change in knowledge and change in disputant perception

	Δ Afghan law	Δ non-Afghan law
Δ Access Rights	0.000	-0.107
Δ Decision Subverted	-0.239***	-0.112
Δ Free Forum	-0.023	-0.128

Δ Outcome Just	0.232*	-0.131

The correlations largely follow the pattern established by the change in knowledge and the level of disputant perception. The relationship between change in knowledge of Afghan law and change in disputant assessment of subversion of decision is strong. The mean change in knowledge of Afghan law from baseline to endline (10.2%) is associated with a fall in disputant assessment on the subversion of decision index by 0.13 on the five-point scale, while a knowledge change score at the 75th percentile (25%) is associated with a fall in disputant perception on the subversion of decision index by 0.32 on the five-point scale. What is notable here is that exact disputant scores are linked to their exact referring elders. See Annex Table 4 for the table of coefficients generating the predicted values.[12]

In conclusion, cross-referencing elder knowledge against disputant perception offers some evidence that knowledge does in fact matter for disputant perception, and that knowledge of Afghan law is more important for disputant perception than knowledge of non-Afghan law as it is defined by RLS-I. Elder knowledge of Afghan law is associated with subversion of decision and equity of outcome, while elder knowledge of topics outside Afghan law may be weakly connected to procedural justice. Elders tended to learn one sort of law at the expense of the other, with possible implications for the stabilization thesis – especially as the relationship was slightly stronger in the south than elsewhere.

Identifying critical mass

The findings for each hypothesis presented in Section III, above, explicitly looked at treatment and comparison groups across baseline and endline in a search for a treatment effect to validate the development hypothesis. The preceding sections looked at correlations and simple regressions only at endline in a search for a statistical relation between elder knowledge and disputant perception. This section combines these two lines of inquiry by returning to the search for a treatment effect, this time using two program metrics that allow for the intensity of treatment to vary, and cross-referenced against elder knowledge and disputant perception. While data quality issues have likely confounded attempts to look across both time and group, identifying a statistical relationship between intensity of treatment and RLS-I outcome and impact measures would provide additional evidence in support of the development hypothesis even if it could not fully validate it. Note further that examining intensity of treatment looks only within the treatment group to identify heterogeneous responses by program participants.

[12] Also note that the measurements based on overall values and individual change scores largely cohere, lending additional validity to the estimates.

Finally, note that analysis within a dose-response framework has a direct link to programmatic interest in critical mass for social change, conditions allowing for consolidation of gains/prevention of regress, and ultimately establishing benchmarks for graduation such that donors would have some indication as to when resources could be re-directed to new treatment areas.

Elder knowledge and exposure to RLS-I treatment

The first metric on the exposure to RLS-I treatment is simply a count of the number of activities attended by RLS-I participants from the evaluation data. One hundred and five elders[13] could be identified within the program activity database across six treatment districts[14], with an overall average of 6.3 activities attended per elder for Phase 2. A district breakdown is as follows:

Table 4.22: Activities attended and participating elders, by district

District	Number of RLS-I activities attended	Number of participating elders
Mihtarlam	6.7	20
Nurgal	7.1	27
Puli Khumri	3.5	16
Pashtun Kot	3.6	12
Shahjoy	5.6	5
Chora	8.6	25
Overall	**6.3**	**105**

The higher figures for Mihtarlam and Nurgal reflect the relatively more permissive operating environment, while the high participation in Chora reflects uncommonly strong management capacity in the Uruzgan office. In Shahjoy, on the other hand, RLS-I experienced significant difficulties in enrolling elders who had participated in the baseline data collection. Puli Khumri and Pashtun Kot reflect participation in CPAU's Broad-Based Education curriculum with a model that features a dedicated cohort passing through a fixed curriculum. Less is known, however, about the specifics and fidelity of implementation of the CPAU program.

The first question of interest is whether the degree of exposure to RLS-I treatment affects either the level of or change in knowledge within the treatment group from baseline to endline.

[13] One elder from the comparison district of Chawkay attended the east regional network meeting but is not included in the analysis.
[14] Arghandab and Bihsud elders are not included here as baseline data from Phase 1 were not available for analysis.

Attending ten RLS-I activities, after controlling for region, is associated with 8% less knowledge among elders in Afghan law, 3% less in non-Afghan law, and 5% less overall. Substituting in the district-level means of knowledge scores for each topic provides an actual estimate of the effect of RLS-I activities on elder knowledge for the majority of program participants:

Table 4.23: Predicted knowledge scores based on number of RLS-I activities attended, by district

District	Average of RLS-I activities attended	Afghan law	non-Afghan law
Mihtarlam	6.7	-5.3%	-2.0%
Nurgal	7.1	-5.7%	-2.1%
Puli Khumri	3.5	-2.8%	-1.1%
Pashtun Kot	3.6	-2.9%	-1.1%
Shahjoy	5.6	-4.5%	-1.7%
Chora	8.6	-6.8%	-2.6%
Overall	**6.3**	**-5.1%**	**-1.9%**

See Annex Table 5 for the table of coefficients generating these predicted values.

Recall that since the analysis is examining the level of Afghan knowledge, these measures are vulnerable to the incomparability problem between baseline and endline data sets. For elders who could be matched from baseline and endline, one may shift the analysis to the rate of change in knowledge as a function of exposure to RLS-I activities. However, the coefficients on the change in knowledge, rather than the endline knowledge level, do not substantively differ[15].

The finding from examining the degree of exposure to RLS-I activities corroborates the original knowledge change score reporting. There is a possible negative treatment effect in some cases, but more likely a zero treatment effect or an inability to make a determination after considering data collection issues. In some sub-analyses there may be some knowledge gain, namely in the east region in Afghan law, the north region in non-Afghan law, and Arghandab in non-Afghan law.

[15] One example of where it is helpful to examine the change in knowledge rather than level of knowledge at endline is the topic of non-Afghan law in the north, where the endline treatment score fell drastically relative to baseline treatment (from 73% to 58%), and yet simultaneously realized a 12% knowledge gain relative to the comparison group, whose mean score fell from 76% to 48%. In this case, cross-referencing activities with the level of knowledge predicted that attending 10 RLS-I activities would result in an 11% knowledge loss in absolute terms, while cross-referencing activities with the knowledge change score predicted that attending ten RLS-I activities would result in a 9% knowledge gain.

Disputant perception and exposure to RLS-I treatment

The question of interest is whether elder attendance at RLS-I activities affects the perception of parties to disputes the elders help mediate, irrespective of the elder's knowledge. Mean values by district are as follows:

Table 4.24: Mean number of activities attended and disputant assessments, by district

District	Average # of RLS-I activities attended	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
Mihtarlam	6.7	4.79	1.11	4.88	4.79
Nurgal	7.1	5.00	1.03	5.00	4.83
Puli Khumri	3.5	3.96	2.90	3.87	3.92
Pashtun Kot	3.6	4.10	2.18	4.08	4.38
Shahjoy	5.6	4.40	2.93	4.45	4.54
Chora	8.6	4.37	2.93	4.34	4.49
Overall	**6.3**	**4.44**	**2.30**	**4.48**	**4.50**

A table of correlations helps identify potential connections.

Table 4.25: Correlations between activities attended and index scores

	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
# RLS-I activities attended	0.327***	-0.019	0.174	0.315***

Exposure to RLS-I activities is positively associated with disputant perceptions of procedural justice and equity of decisions. When disputant perception is modeled as a function of the number of RLS-I activities attended by the adjudicating elder (with control variables for the north and south regions included), the connection between equity of decision and activities attended falls away while the connection with procedural justice remains strong. The mean value of RLS-I activities attended (6.3) is associated with an increase on the procedural justice index by 0.13 on the five-point scale. The 25th percentile (3 activities) is associated with an increase of 0.06, while the 75th percentile (9 activities) is associated with an increase of 0.18. See Annex Table 6 for the table of coefficients generating the predicted values.

Correlations between elder attendance at RLS-I activities and the change in a disputant's assessment show still stronger patterns:

Table 4.26: Correlations between elder attendance and disputant assessment change scores

	ΔAccess Rights	ΔDecision Subverted	ΔFreedom of Forum	ΔOutcome Just
# RLS-I activities attended	0.414**	-0.165	0.385*	0.499**

Modeling change in disputant perception as a function of the adjudicating elders' attendance at RLS-I activities indicates strong relationships. The mean value of 6.3 activities attended is associated with improvement in procedural justice, freedom of forum, and equity of decision by 0.68, 0.59, and 0.70 respectively on a five-point scale. These values amount to 0.39 – 0.50 of their respective standard deviations, indicating a substantive effect. However, one must also be aware that given the combination of partial data collected for activities attended and the limited data available for changes in disputant perception, the sample size for this analysis is 20-25. Due to a sample size just large enough for statistical analysis, regional control variables are not included. The reader should note the inability to extrapolate from such a small sample as well as the potential of other variables affecting the relationship but not included in the estimating equation. Annex Table 7 has the coefficients generating the predicted values.

Elder knowledge and network effects

Attendance at RLS-I activities is an individual measure. Not to be forgotten are considerations of peer effects and critical mass within a district or village contributing to shared knowledge gain and collective improvement in adjudication of disputes. RLS-I targets from 40 to 120 informal justice providers per district for passage through its core curriculum of learning workshops, depending upon management capacity, the security environment, and the size and population of a given district. It is of interest to identify whether there is some level of collective progress through the RLS-I program cycle that engenders learning and improved adjudication. To measure this, RLS-I added the number of elders passing through all five of its core curriculum workshops as a variable to analyze in relation to elder knowledge and disputant perception. Like the attendance metric, size of the district cohort is a variable that affects only the treatment group with intensity varying by district. The size of each district cohort is as follows:

Table 4.27: Size of district cohorts

District	Number of elders passing	1390 population	Cohort proportion of

	through core curriculum	estimate[16]	population
Mihtarlam	84	126,000	0.067%
Nurgal	136	30,800	0.44%
Puli Khumri	38	203,600	0.019%
Pashtun Kot	12	183,500	0.007%
Shahjoy	24	56,800	0.042%
Chora	76	49,700	0.15%
Overall	**63**	**108,400**	**0.12%**

The first question of interest is whether the size of the district cohort helps engender learning. The correlation between cohort and knowledge of Afghan law is0 .271 (p=0.000) while the correlation with non-Afghan law is 0.068. However, when elder knowledge is modeled as a function of the size of the district cohort, both Afghan and non-Afghan law show strong relationships. The mean cohort size of 63 is associated with a 13% improvement in knowledge of Afghan law, while the same cohort size is associated with a 19% drop in knowledge of non-Afghan law. The effect is the same for change in knowledge rather than level of knowledge at endline.

The data so far suggest that knowledge of Afghan law is sensitive to peer effects, as measured by the size of the district cohort successfully passing through the RLS-I core curriculum. There is also a possible negative relationship between size of cohort and knowledge loss in non-Afghan law. This needs further investigation but one line of inquiry is that even when knowledge is absorbed erroneously or not absorbed at all, peer effects are a significant factor in the learning dynamic at work.

Disputant perception and network effects

While a mixed picture emerged in examining size of district cohort and elder knowledge, the correlations between size of district cohort and disputant perception are consistent.

Table 4.28: Correlations between size of district cohort and disputant perceptions

	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
Cohort	0.637***	-0.473***	0.498***	0.389***

[16] Source: Central Statistics Office (CSO)

When disputant index values are modeled as a function of the size of the district cohort (including control variables for north and south regions), only the connection with procedural justice remains strong. The mean value of 63 elders passing through the full curriculum is associated with a 0.25 increase in disputant perception on the procedural justice index. At the 25th percentile of 24 elders the associated value is 0.10 and the 75th percentile of 84 elders is associated with a 0.34 improvement. See Annex Table 10 for the coefficients generating the predicted values.

The correlations between the size of district cohort are even stronger, as are the modeled relationships generating predicted values.

Table 4.29: Correlations between size of district cohort and disputant perception change scores

	Δ Access Rights	Δ Decision Subverted	Δ Freedom of Forum	Δ Outcome Just
Cohort	0.544***	-0.516***	0.670***	0.357*

Again however, one should be cautious in interpreting data with change in disputant score, due to the small sample size. See Annex Table 11 for the coefficients generating the predicted values.

Discussion of findings

The data indicate that elder attendance at RLS-I activities is not associated with their level of knowledge or gains in knowledge, as predicted because of the short duration of Phase 2. However, elder attendance does have a relationship with disputant perception irrespective of elder knowledge, most notably the index values of subversion of decision and equity of outcome. At the same time, the size of the district cohort did have a strong relationship with knowledge of Afghan law (and a possible negative relationship with knowledge of non-Afghan law).

The correlations between size of the district cohort and disputant perception are strong, though only procedural justice stands out when the data is modeled in regression format. It is interesting to note that in the previous section knowledge of Afghan law was linked to the disputant index values of subversion of decision and equity of outcome, while procedural justice seems to be affected more by the size of the district cohort.

The initial motivating question behind these additional findings was whether knowledge actually mattered as an intervening process leading to change in disputant perception. The answer seems to be yes in general, and yet the current data cannot answer the question whether RLS-I is affecting disputant perception through the mechanism of improved elder knowledge.

On the other hand, the data do provide evidence that RLS-I is positively affecting disputant perception irrespective of knowledge. This is seen through the program metrics of activity attendance and size of the district cohort, neither of which depend on elder knowledge. One interpretation could be that RLS-I activities are filling a crucial governance gap in contested areas, which help establish a normative framework of both statutory law and *Shari'ah* that elders and citizens may follow in the absence of a functioning state justice system and as an alternative to Taliban justice.

V. CONCLUSIONS

The RLS-I Impact Evaluation Plan did not predict finding a treatment impact from baseline to endline in the RLS-I impact evaluation. Rather the exercise was seen as a test of items, instruments, procedures and analyses under the dynamic conditions in a conflict-affected state. Moreover, the theory of change posits a progression from workshop and elder activity inputs that would only later affect more important lagged variables such as disputants' perception of the impartiality of TDR. Changes in social order – such as encouraging women's participation in *jirgee* or *shuragani* and reducing the incidence of practices harmful to women and girls – represent higher-order changes as well, that could not be expected after intervention periods of only two to four months.

Nevertheless, patterns in the data from baseline and from endline are valuable for understanding the challenges of data collection in dynamic, conflict-affected environments. This section presents those conclusions that are supported by the data for each hypothesis below, followed by conclusions on secondary research questions and extended analyses of data on elders' and disputants' experiences in RLS-I.

Hypothesis 1: The intervention will result in TDR decisions that better reflect and/or are based in Afghan law, *Shari'ah*, and human rights norms

Measurement: Direct determinations of decisions within prescribed law were not considered feasible. Therefore, the evaluation used proxy measurements of knowledge gain in Afghan law and *Shari'ah* as a marker for adjudication that more closely followed these standards.

Findings: Respondents improved their knowledge of Afghan law approximately 10% against baseline. However, respondents in comparison districts improved more, resulting in a treatment effect of -4% (p=0.074). On the other hand, scores on the topics of family, property, and inheritance law, which were more geared towards *Shari'ah* as reflected in Afghan Constitutional and statutory law, fell for participants – a result that is thought to be driven by idiosyncratic differences in the data collection methods of the different research partners at baseline and endline. Respondent scores in the treatment

group fell less than those in the comparison group, yielding a positive treatment effect of 1.4%, but one that was not statistically different from zero. On balance, the overall treatment effect for knowledge was zero, but with some divergence according to the sub-topics of Afghan law and more *Shari'ah*-oriented topics of family, inheritance, and property. There are also regional differences driving the results, with slight knowledge gains across the board in the eastern region, mixed results in the northern region, and negative gains across the board in the southern region.

On a more encouraging note, citizen knowledge of Afghan law in communities that had received RLS-I outreach material increased by 6% relative to communities that had not – a result statistically different from zero.

Conclusion: Apart from the slightly more positive results for citizens, the knowledge gains that were theorized to have foundational import as part of the RLS-I theory of change were not found at endline. Two interpretations can be posited: there was in fact zero treatment effect, or the data were simply incomparable. If it is assumed that idiosyncratic differences in data collection methods were applied uniformly by each research partner across all data collection districts, then the D-in-D measurements retain their validity and one may conclude a zero treatment effect for knowledge. This is a reasonable assumption given that each survey research firm typically provides uniform training to its enumerators specific to a given data collection effort. On the other hand, differences in enumerator quality and practice by region were both observed by RLS-I monitors and reported by the research firms. Idiosyncratic differences in data collection that varied by region, province, or district would invalidate the change measurements and the conclusion would be that no insight may be taken from the D-in-D measurements. This study cannot make a definitive determination as to which of these scenarios may be true. The evidence suggests the latter given the parallel trends for treatment and comparison groups and given that the response trends show high variation from baseline to endline. However, given the expectation of zero treatment effect due to the short time span of intervention and the concordant findings from treatment data, this explanation is highly credible.

If a zero treatment effect for knowledge were posited, explanations are available. In workshop event reports, elders said the workshops were too short to absorb and practice the new skills. Scholars were said to have included too much content in the workshops, and too little hands-on practice. Participants reported feeling overwhelmed, which could have limited their ability to either share the learning with colleagues or to change their adjudication practices, in addition to limiting their ability to retain the workshop content.

Evaluation evidence from other training interventions shows that such dynamics are not uncommon. Training effectiveness is limited by trainers' abilities to impart information in a way that participants can assimilate, particularly with low-literacy audiences. Adult learners also learn best when such workshops

include both theory and practice, and the scholars chosen to lead RLS-I workshops have that skill set in varying degrees, according to RLS-I event reports. Training can essentially push participants from a beginning state of certainty about their knowledge and skills into a state of uncertainty about what they know and can do.

Direct follow-up queries also produced evidence of resistance. In the words of one participant: "We did not accept everything they told us." This raises questions about the extent to which respondents answered the knowledge questions according to what they believed should be true or based on what they experienced to be true which, in the case of Afghan law, would quite often be contrary to the civil rights and protections provided by law but not yet practiced in society.

Regarding the positive knowledge gain in Afghan law for communities that received RLS-I outreach, one may suppose what advantages that outreach may have contributed. Examples include the use of durable materials that remain in the household and serve as an ongoing reference, simple and clear communications of legal rights and protections, and a communication medium that may include entertainment and/or novelty value improving attention to and retention of the legal content.

Hypothesis 2: The intervention will result in TDR decisions and *shura/jirga* members being perceived as more impartial

Measurement: In addition to the core measurement of impartiality, this hypothesis was considered to be the umbrella heading for broader access to justice issues. Accordingly a battery of experiential and attitudinal items was applied to citizens who had settled a dispute informally. The items were organized around four themes of procedural justice, subversion of the decision making process through corruption within the *jirga* or interference from local elites, freedom of forum, and equity of the outcome.

Finding: Mean values for procedural justice, freedom of forum, and equity of outcome fell in the treatment group, resulting in negative treatment effects that were statistically different from zero. The mean value for subversion of decision fell, but only slightly and the positive treatment effect was not statistically different from zero.

Conclusion: In the case of Hypothesis 2, there is more conclusive evidence that the data from baseline to endline are not comparable. The report details how the endline data collection applied the five-point strength of response scale more as a binary (yes/no) scale, ignoring the nuance afforded by the five-point response continuum, inflating index scores, and distorting the measurement such that baseline and endline were different measurements altogether, rather than the same measurement taken at different points in time. Other instances were documented in which the endline data collection firm

chose slightly different methodological approaches in respondent selection compared to baseline, raising the risk of producing different measurements. Finally, RLS-I monitoring documented differences in the enumerators themselves. It was reported that baseline enumerators tended to be older and more experienced, with several having made field survey interviewing a career choice through the data collection firm's 20-year history. Endline enumerators, on the other hand, were reported to be competent but also younger and not as seasoned.

If the data are taken at face value, theoretical explanations for the seemingly negative treatment effect are also available. One such explanation is that RLS-I programming may have raised awareness among citizens of their rights and protections under the law, regardless of what forum they chose to resolve a dispute. This could have had the effect of changing the respondent's assessment of the response scale itself from baseline to endline, thus making the negative treatment effect an actual program result. This was not included in the initial theory of change, and would be something of a reversal if true. The initial theory of change was that disputant perception was likely a lagged variable that may not be reflected in survey data until a longer time period had elapsed. If RLS-I activities and outreach had the effect of raising awareness of citizens such that they changed their conception of the response scale on the access to justice measures, the theory of change may need to be modified to reflect that the citizen demand for access to justice may precede an actual supply-side change in adjudication on the part of elders.

However interesting such lines of inquiry may be, the data remain too chaotic to support any such explanation. If RLS-I outreach and activities shifted respondents' own assessments of the response scale from baseline to endline, one would expect to see a fall in index scores in the treatment group across the board. In reality, some index scores in the treatment group fall while others rise. In the absence of a stronger theoretical explanation, then, incomparable data seems a more likely explanation of the negative treatment effects detected in the data. Subsequent discussion below will attempt to draw conclusions that do not depend on measurements from baseline to endline and across treatment and comparison groups.

Baseline and endline data collection on disputants proved very useful in terms of learning for development effectiveness, one of the primary goals of the impact evaluation exercise. For example, the endline enumerators more successfully engaged female disputants at endline data collection, allowing for more detailed measurement of the gender deficit in disputant perceptions and, through narrative responses, greater understanding of the contexts of women's disputes and participation in TDR. Internal dynamics of the dispute resolution process such as *jirgamar* selection, duration, costs, and the use of guarantees (*machalgha*) help learning about the implementation environment, while the determinants of disputant satisfaction, acceptance, and enforcement are of great interest in strengthening informal

justice as an extension of the formal justice sector and as a viable alternatives to Taliban justice in contested areas.

See the conclusions of the "Extensions" section below for additional discussion establishing a link between attendance at RLS-I activities and improvement in disputant assessment.

Hypothesis 3: The intervention will result in a decrease in the number of TDR decisions that negatively impact women and children

From these data, no effect on *baad* or forced marriage can be determined. That being said, however, elders in the treatment group were asked about changes in their *jirgee* as a result of RLS-I interventions. Ten respondents specifically mentioned that their *jirgee* no longer used *baad* to resolve disputes but instead now use some compensation in the form of labor, land or money. Three respondents said that forced marriages were no longer acceptable in their communities as a result of the messages brought back from the RLS-I training.

The elder interview included a set of attitudinal questions about dispute resolution practices that are harmful to women and children. These practices include forced marriage, or marriage against the will of one or both parties, and the use of *baad* to resolve disputes. For questions on both *baad* and forced marriage, only a very small fraction of baseline respondents responded that they knew of any such cases. At endline, however, the frequency was nearer one-quarter of citizens and near one-tenth of elders. The degree of difference in these findings between baseline and endline may be related to the ways the two data collection firms asked questions of respondents. Endline enumerators may have been better able to establish rapport with respondents, or there may have been other factors at work. Survey fatigue or annoyance at being queried again after such a short period may have resulted in some respondents inflating responses or providing confounding responses. The exact circumstances of the data collection and the differences compared to baseline may be unknowable; however, some figures collected at endline are more concordant with other estimates of incidence.

Hypothesis 4: The intervention will result in an increased role for women in TDR processes as disputants, witnesses or decision-makers

For both elders and citizens, the interview included questions on women's roles in dispute resolution – their participation on decision-makers or witnesses and their ability to present their cases directly rather than through an intermediary. Elders were asked their opinions about various possible roles for women in TDR. For the majority of questions, little or no change was seen from baseline to endline in attitudes held by elders about women's participation in TDR. When elders were asked about specific cases of

women's involvement in TDR, endline respondents were more than twice as likely to report a case with women's involvement (7% at baseline compared to 15% at endline). This may be a result of the formation of *spinsary* groups as part of RLS-I activities in their districts, but comparison district respondents also showed an increase, though smaller, for this response. RLS-I may also have had some spillover effect with the advent of the *spinsary* groups or the result could represent interference from the differences between the two survey firms carrying out data collection at baseline and endline.

For citizen respondents who received outreach materials, the D-in-D changes were generally also small. The exception was for a question on whether women's testimony would be accepted in a jirga, with a statistically significant 9% difference for treatment districts.

Secondary research questions

Across the sample, the quantitative data show that documentation and registration of cases had improved substantially more for elders in treatment districts than those in comparison districts. These data were self-reported. RLS-I data on documentation and registration support the general pattern of self-reports, though with less incidence of documentation or registration. Given the overall pattern of gain for both treatment and comparison groups on these measures, along with anecdotal evidence, some spillover into comparison districts likely occurred.

Respondents were asked about long-standing disputes, if any, with which they might be familiar. As with questions about cases of forced marriage and *baad*, few elders knew of such long-standing disputes in their villages and communities, but many more respondents at endline (5 at baseline compared to 33 at endline) provided information on such disputes. While this may be related to the difference in data collection methods employed at the two points in time, it may also reflect greater willingness to discuss such disputes in light of program effects and spillover.

Extensions to the core analyses

The RLS-I development hypothesis is that skills- and knowledge-building of informal justice providers increases access to justice and citizen confidence in TDR mechanisms. While longitudinal measurements were clouded by issues of incomparability from baseline to endline, examination of relationships between elder knowledge, disputant perception, and various program metrics within only the treatment group suggested that knowledge was in fact important for improving disputant assessment of informal dispute resolution, most notably the index values of subversion of decision and equity of outcome.

The data also suggested that the benefit of RLS-I was not transmitted through the mechanism of increased elder knowledge; rather, it was attendance at RLS-I activities among local elders that

positively affected disputants' assessments. Furthermore, the absolute size of the district cohort passing through the RLS-I core curriculum was predictive of both learning Afghan law and improving disputant perception. This defines a definite role for network and peer effects in program success, exactly as is supposed by the development hypothesis and solicitation design.

The size of the district cohort had a strong relationship with knowledge of Afghan law (and a possible negative relationship with knowledge of non-Afghan law). The correlations between size of the district cohort and disputant perceptions are strong, though only procedural justice stands out when the data is modeled in regression format. It is interesting to note that in the previous section knowledge of Afghan law was linked to the disputant index values of subversion of decision and equity of outcome, while procedural justice seems to be affected more by the size of the district cohort.

Wherever the cohort variable is associated with elder knowledge and disputant perception there is the potential to establish benchmarks for graduation.

The initial motivating question behind these additional findings was whether knowledge actually mattered as an intervening process leading to change in disputant perception. The answer seems to be yes in general, and yet the current data cannot answer the question of whether RLS-I is affecting disputant perception through the mechanism of improved elder knowledge.

On the other hand, the data provide evidence that RLS-I has positively affected disputant perception irrespective of knowledge. This is seen through the program metrics of activity attendance and size of the district cohort, neither of which depend on elder knowledge. One interpretation could be that RLS-I activities are filling a crucial governance gap in contested areas that helps establish a normative framework of both Afghan law and *Shari'ah* that elders and citizens may follow in the absence of a functioning formal justice sector and as an alternative to Taliban justice.

VI. RECOMMENDATIONS

Recommendations emerging from the endline research focus on two areas: programming and impact evaluation research.

Programming

In order better to serve future program participants and ensure their acquisition of the new knowledge and skill sets imparted in RLS-I training, RLS-I recommends the following:

1. *Improve training and reinforce learning*

TDR workshops should put significant emphasis on comprehension and retention by low-literate adult audiences. Specific suggestions include:

- Ensure that workshops are of sufficient duration to cover content and permit active learning. For some subjects, workshop timelines will need to be two to three days rather than one day to permit additional contact time with key themes, both in theory-based lectures and hands-on activities.
- Emphasize training of trainers, for both the form and content they are responsible for imparting. Reinforce key content knowledge in their repertoire, strengthening their ability to impart the training content to the specific audiences, taking into account specialized learning needs, and monitoring their performance and coaching for improvements.
- Enlist the support of the scholar-trainers to bolster training in areas where test scores are lowest. This would include finding ways to strengthen training in Afghan law and the Constitution. Some participants and even some trainers appear to see these themes as opposing *Shari'ah*; this dynamic needs to be collaboratively and explicitly addressed.
- Utilize short-term technical advisors to strengthen the trainers' use of adult learning principles, particularly for low-literate audiences, and create appropriate materials for participants such as simplified illustrated handouts and audio-visual lecture aids. Adult learning also requires that workshop leaders are able to differentiate instruction based on participants' knowledge levels and progress; scholars will need to be supported in assessing participants *in situ*.
- Include participatory or active learning techniques in each workshop, such as role-playing, case studies and examples for discussion that are drawn from participant experience such as specific instances of resisting outside influence, being impartial, dealing with long-standing disputes, etc., rather than lectures or abstract cases. These techniques are particularly useful for low-literate audiences and those who must use new skills on their own once they return to their villages.

2. *Reach remote program participants*

Capacity building programming for rural or hard-to-reach populations is particularly challenging in that participants are more isolated from sources to scaffold their new learning. Participants need greater support in their home environments to reinforce and extend training content to fit into these environments. RLS-I developed village-based training for women in Phase 2, for example, which could be used as well for future program participants or for other village elders

who are non-participants. Local-level engagement provides essential support between and after training events, so that capacity building can take root through application of the new knowledge and skills.

3. ***Track specific applications of knowledge in events***

   Use ongoing project M&E to uncover outcome-level results from training. Workshops and other events should include an event monitoring function to track the extent to which knowledge is applied in workshops, discussion sessions and network meetings. When asking participants to evaluate their experiences in training, the M&E team should also ask about application of new learning, both inside and outside the event. These types of measures provide evidence of the level of exposure participants receive and can be combined with impact evaluation data on outcomes.

4. ***Test assumptions on critical mass and saturation***

   Design M&E and impact evaluation tools collaboratively with the programming team, in order to devise ways of testing the appropriate numbers and types of project interventions necessary for the desired results. The peer effects described in this report form part of the theory of change, including the level of saturation in a community ("critical mass") necessary to support elders in their work in their home villages. Enhancements to training to meet capacity building goals should be tested and documented to assist programming in deciding which are both effective and cost-effective. These might include appropriate materials, increased workshop exposure, wider invitations to networking and discussion events, workshop content on how to share new knowledge, or enlisting leaders from among the workshop participants – such as from a district *shura* – to support cascading the workshop content to others in villages.

   It is important to recognize that knowledge gain, in the context in which RLS-I has operated, is both a more complex measurement and more difficult to achieve for numerous reasons, compared to more its measurement in more traditional training-based programming. Success and failure of RLS-I workshop messaging are highly contingent on the external environment.

5. ***Develop and adapt theoretical models according to the data***

   A theory of change is only as valuable as its explanatory power with practical, real-world situations. Data and their interpretation need to be incorporated into the theory of change for that theory to be useful in programming and evaluation. The RLS-I theory of change has proposed that changes in elder knowledge would improve dispute resolution in communities, which in turn will cause positive perception changes among disputants. However, the exact relationship between program inputs and changes in disputant perceptions is not theorized.

With initial research possibly showing that it is not only knowledge gain among elders that changes perceptions, it is important to understand the process within communities that results in this apparent effect. Further qualitative analysis of the disputant narrative data would provide information on this process, as would targeted research with disputants in treated villages.

Impact evaluation research

1. ***Build upon the research initiated by this impact evaluation***

   Impact evaluation research is still infrequent in conflict-affected environments, but under USAID's 2011 Evaluation Policy and increased awareness of the uses of impact data, more studies have been funded in the last five years. These data provide a wealth of important learning for development effectiveness, across sectors and geographic areas. It is important for program designers, managers and implementers to follow these findings and build upon the evidence base regarding development effectiveness. The impact evaluation model adopted for RLS-I has become stronger through Phase 2 with the M&E system that has been constructed in concert with the impact evaluation research. Planning for further impact evaluation work should capitalize on these gains and utilize a longer time horizon to test assumptions more robustly, take advantage of opportunities for trend line capture with more than one cohort of treatment districts, and work closely with programming to refine instrumentation (including knowledge questions) in line with curricular changes. Retain the use of comparison groups as well, to estimate the counterfactual and prepare for eventual treatment (see the section on "Adopt a pipeline evaluation approach to program expansion", below).

2. ***Use secondary data to strengthen and understand the findings***

   Data from research by national statistics bureaus, non-governmental organizations and donors can be used to contextualize and interpret findings on more limited evaluation questions in a given impact evaluation design. RLS-I analyses have integrated village-level data to measure socio-economic development and distance from district centers as predictors of knowledge and disputant perception. Similarly, secondary data from the formal justice sector analyzed with the impact evaluation data allow RLS-I to understand the contextual conditions that may be conducive to success. Using contextual data and uncovering correlations opens lines of inquiry that can strengthen programming.

3. ***Take opportunities to randomize where possible***

   Randomization in conflict-affected environments is a major challenge for robust evaluation designs. Often, the imperatives of intervention in particularly sensitive areas are paramount, and selection of treatment subjects is therefore done before evaluators are engaged. However, there may be opportunities to randomize within populations that can add inferential strength to evaluation design. One potential opportunity involves collaboration between project programming and project M&E to jointly and simultaneously mobilize and select participants in all future program districts. Randomization can then be carried out within treatment districts for the impact evaluation sample.

4. ***Adopt a pipeline evaluation approach to program expansion***

   Impact evaluation designs often raise resistance because of the use of control and comparison groups. These respondents are asked to participate in the study design but are not provided treatment or incentives. Those assigned to, or representing, these groups may call the process unfair or even unethical. By adopting a pipeline approach to program expansion, also called delayed control or delayed comparison design, some of these concerns can be assuaged. Strictly speaking, this means that the control or comparison populations would participate in a follow-on treatment cohort, after having provided the point of comparison for analysis. How this is handled will vary by project but it is generally best not to communicate this to comparison district respondents at the time of data collection.

5. ***Use the same data collection methods for baseline and endline***

   To account accurately for differences between baseline and endline, the process of collecting data from the field needs to be substantially parallel, even for relatively simple questionnaires. With more complex, perception-based, and qualitative data collection, this parity is more difficult and more essential. For RLS-I, switching data collection partners for the endline may have corrected for sources of bias in baseline data collection, but also may have invalidated the

evaluation measurements. Engaging one firm or organization for both data collection efforts can also make multiple solicitations unnecessary, minimize contracting delays and help ensure that field techniques will be the same in all phases of data collection.

6. ***Ensure sufficient time between baseline and endline***

Narrow programming timelines often present significant challenges for rigorous evaluation data collection. Donors and evaluators must understand the need for sufficient implementation time to effect the changes presumed in a given theory of change. Intervention durations that are too short can result in lackluster evaluation results. They can also cause survey fatigue among respondents or even distress or harm. When the subject matter is sensitive, this threat is more pronounced. The endline RLS-I data collection caused significant distress to some Afghan communities by asking disputants to recount highly negative, charged experiences very soon after they occurred. One of the benefits of TDR is the process of reconciliation and return to equilibrium when disputes are resolved, but the limited time frame from which to draw cases for endline data collection in a sense forced individuals to re-live these disputes in nearly real time, while the *jirga* dispute resolution process is supposed to help disputants reconcile and move past their disputes and restore community harmony.

7. ***Integrate the research into an M&E system capable of robust inference***

Impact evaluation is not the only means to uncovering outcome- and impact-level results, but it does provide important lessons for ongoing M&E. Hallmarks of industry-leading M&E include engaged and committed local leadership; effective and durable capacity building; grounded instrumentation and data capture; and straightforward off-the-shelf technology that is adapted for project use. These characteristics make RLS-I's M&E a fertile proving ground for learning for development effectiveness in Afghanistan's conflict-affected environments. M&E field work can support impact evaluations in capturing an array of contextual and secondary data for cross-fertilization of findings. Impact evaluation also often proves useful for the development of refined M&E processes and tools. The two functions (monitoring and evaluation) should be considered part of a paired and iterative process.

8. ***Choose evaluation questions that can be meaningfully and reliably measured***

Complex theories of change require deft impact measurement, and not all outcomes are equally evaluable within the logical chains they propose. Variables that are likely to be significantly affected by intervening or mediating conditions outside the project's control or whose effects are likely to be lagged result in data that are more subject to interpretation and to unobserved factors. While statistical methods such as regression are useful in controlling for some exogenous effects, it is important to keep in mind the complex relationships involved in many

development environments. For example, the RLS-I impact measurements of disputant perceptions can be problematic for these reasons. However, disputants' experiences with the TDR system, including both women's and men's experiences, are critical for seeing change over time in perceptions of impartiality and justice. Attention should be shifted from the large disputant sample size needed for making inference, to better connecting the contextual qualitative narratives with the numeric assessments. The selection of evaluation questions for impact evaluation needs to consider the degree to which an outcome can be readily measured within a given context. More complex measures may need to be dealt with using observational or other designs.

VII. ANNEXES

Evaluation Measurements: Annex A

Annex A: Impact evaluation indicators by data collection tool

The following table presents the complete list of indicators first identified by the impact evaluation. Not all indicators could be measured, while other indicators were not viably measured.

Hypothesis 1: The intervention will result in TDR decisions that better reflect and/or are based in Afghan law, *Shari'ah*, and human rights norms	**Elders**	**Disputants**	**Citizens**
% responding change in adjudication compared to six months ago; coding of qualitative response on what has changed	x		
Aspects of case adjudication that may be attempted from RLS-I programming, but not succeed	x		
Knowledge increase: Afghan law	x		x
Knowledge increase: Family law	x		
Knowledge increase: Inheritance	x		
Knowledge increase: Property/Deeds	x		
#, % of cases where parties could exercise veto right on decision- makers		x	
#, % of case resolutions accepted by parties	x	x	
#, % of respondents perceiving (Afghan, *Shari'ah*, customary) law as source of adjudication	x	x	
Freedom of venue		x	
Hypothesis 2: The intervention will result in TDR decisions and shura/jirga members being perceived as more impartial	**Elders**	**Disputants**	**Citizens**
Extent of external influence over process, outcome	x	x	
Subversion	x	x	x
Source of law	x	x	x
Equity of outcome	x	X	x
Quality of process	x	x	x
#, % of cases where bond was collected, cross-referenced against decision accepted, satisfaction, fairness, justice, etc (Use of machalga as coercive/corrupt mechanism)	x	x	
#, % of respondents disagreeing with some aspects of decision, regardless of whether they accepted		x	

Hypothesis 2: The intervention will result in TDR decisions and shura/jirga members being perceived as more impartial	Elders	Disputants	Citizens
#, % respondents voicing satisfaction with process and outcome of dispute resolution		x	
#, % respondents voicing satisfaction resolution (outcome) of dispute		x	
#, % respondents perceiving justice in dispute resolution (outcome)		x	
Hypothesis 3: The intervention will result in a decrease in the number of TDR decisions that negatively impact women and children	**Elders**	**Disputants**	**Citizens**
Marriage against will	x		x
% of forced marriage leading to disputes	x		x
Incidence of baad	x		x
Attitudinal items - forced marriage	x		x
Attitudinal items - baad	x		x
Disputant case types, outcomes with accompanying assessment data		x	
Hypothesis 4: The intervention will result in an increased role for women as disputants, witnesses or decision-makers	**Elders**	**Disputants**	**Citizens**
% responding is it possible for women to sit as jirga members	x		x
% present case directly vs. via intermediary	x	x	x
% responding that women would be called to testify before jirga	x		x
% responding that women's testimony before a jirga would generally be accepted	x		x
Incidence of women playing role as disputant, witness, jirga member	x		x
Attitudinal items - Women as disputants, witnesses, and jirga members	x		x
Disputant case types, outcomes with accompanying assessment data		x	

Additional Findings: Annex B

The following tables are generated through linear regression modeling and are used to generate predicted values of knowledge or disputant scores. Each table is referenced in the text.

Annex Table 1: Elder knowledge and disputant perception

In the following table, elder knowledge scores are juxtaposed against the two knowledge topics as Afghan law (*non-Afghan law*) and jointly cross-referenced against the four index values of disputant perception. The control variables for region are similarly ordered.

Elder knowledge and perceptions from parties to disputes the elder helped mediate	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
Constant	4.77 *(4.79)*	1.49 *(1.13)*	4.88 *(4.90)*	4.58 *(4.73)*
Afghan law (*non-Afghan law*[17])	0.037 *(-0.011)*	-0.464 *(0.089)*	-0.062 *(-0.088)*	0.336** *(0.118)*
South (*South*)	-0.380 *(-0.380)*	1.45 *(1.50)*	-0.402 *(0-.416)*	-0.233 *(-0.230)*
North (*North*)	-0.725 *(-0.726)*	1.65 *(1.68)*	-0.664 *(-0.670)*	-0.662 *(0-.667)*
Proportion of variance accounted for	0.264 *(0.264)*	0.259 *(0.255)*	0.141 *(0.141)*	0.220 *(0.210)*

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

[17] Coefficients values are from single variable regressions for each topic. Including both topics tended to reduce the value of the coefficients, but the change is thought to be driven largely due to a high degree of overlap (co-linearity) between each variable.

Annex Table 2: Change in elder knowledge and disputant perception

The following table repeats the analysis of Annex Table 2, but uses change in knowledge rather than level of knowledge at endline.

Elder knowledge and perceptions from parties to disputes the elder helped mediate	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
Constant	4.77 *(4.77)*	1.43 *(1.14)*	4.87 *(4.88)*	4.75 *(4.77)*
ΔAfghan law (Δ*non-Afghan law*)	0.026 *(-0.027)*	-1.41***(-0.484)	-0.029 *(0.071)*	0.183 *(0-.023)*
South (*South*)	-0.379 *(-.0383)*	1.71 *(1.75)*	-0.417 *(-0.411)*	-0.231 *(-0.240)*
North (*North*)	-0.853 *(-0.857)*	0.749 *(0.947)*	-0.893 *(-0.888)*	-0572 *(-0.598)*
Proportion of variance accounted for	0.284 *(0.284)*	0.428 *(0.368)*	0.239 *(0.239)*	0.149 *(0.140)*

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

Annex Table 3: Change in knowledge of Afghan and non-Afghan law

The following table uses the change score in non-Afghan law to predict the change score in Afghan law.

Elder gain in knowledge of Afghan law as a function of gain/loss in non-Afghan law	Δnon-Afghan law
Constant	0.087
Δ Afghan law (Δ *non-Afghan law*)	-0.243
South (*South*)	-0.062
North (*North*)	-0.107
Proportion of variance accounted for	0.057

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

The mean change score in non-Afghan law (-22%) predicts a change score for Afghan law of 5.3% – half that of the actual mean value of 10.2%. The change score in non-Afghan law at the 75th percentile (-7%) predicts a change score in Afghan law of 1.7%, while a change score in non-Afghan law at the 25th percentile (-38%) predicts a change score in Afghan law of 9.1%.

Annex Table 4: Change in elder knowledge and change in disputant perception

Elder knowledge and perceptions from parties to disputes the elder helped mediate	Δ Access Rights	Δ Decision Subverted	Δ Freedom of Forum	Δ Outcome Just
Constant	0.508*(.450)*	0.116 *(-0.037)*	0.380 *(0.283)*	0.090*(0.092)*
Δ Afghan law (Δ *non-Afghan law*)	-0.265 *(-0.201)*	-1.28** *(-0.248)*	-0.386 *(-0.376)*	0.515 *(-0.243)*
South (*South*)	-0.208 *(-0.232)*	1.57 *(1.54)*	-0.183 *(-0.225)*	0.157 *(-0.147)*
North (*North*)	-1.15 *(-1.11)*	0.727 *(0.879)*	-1.15 *(-1.09)*	-0.668*(-0.715)*
Proportion of variance accounted for	0.241 *(0.238)*	0.317 *(0.261)*	0.194 *(0.191)*	0.146 *(0.125)*

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

Annex Table 5: Elder knowledge and exposure to RLS-I activities

The following table shows regression coefficients for the marginal effect of an RLS-I activity on the dependent variables of knowledge of Afghan law, non-Afghan law (consisting of Family and Property topics, possibly serving as a crude proxy for *Shari'ah*), and the combined overall score. Control variables are included for north and south regions. The coefficient multiplied by a given number of trainings attended yields the predicted knowledge score for a given topic.

Relationship between RLS-I activities attended and knowledge levels at endline	Afghan law (8 items)	non-Afghan law (8 items)	Overall (16 items)
	Coefficient	Coefficient	Coefficient
Constant	0.740	0.634	0.683
# activities attended	-0.008	-0.003	-0.005*
South	-0.058	-0.324	-.0225
North	-0.131	-0.132	-0.128
Proportion of variance accounted for	*0.049*	*0.355*	*0.491*

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

To generate a predicted value, multiply an illustrative value of elder knowledge at endline with the coefficient listed with # activities attended. Mean and quartile values will provide a range of predictions well-situated within the sample data, while mean values disaggregated by district (See Extensions section) may also be helpful.

Annex Table 6: Disputant perception and exposure to RLS-I activities

The following table shows regression coefficients for the marginal effect of an RLS-I activity on the disputant index values of Access Rights, Decision Subverted, Freedom of Forum, and Outcome Just. Control variables are included for north and south regions. The coefficient multiplied by a given number of activities attended yields the predicted value of the disputant index score.

Relationship between RLS-I activities attended and knowledge levels at endline	Access Rights (7 items)	Decision Subverted (4 items)	Freedom of Forum (3 items)	Outcome Just (4 items)
Constant	4.81	0.811	5.00	4.67
# activities attended	0.020*	0.035	-0.004	0.021
South	-0.610	2.24	-0.634	-0.309
North	-0.876	1.58	-0.995	-0.643
Proportion of variance accounted for	*0.552*	*0.489*	*0.369*	*0.277*

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

To generate a predicted value, multiply an illustrative value of elder knowledge at endline with the coefficient listed with # activities attended. Mean and quartile values will provide a range of predictions well-situated within the sample data, while mean values disaggregated by district (See Extensions) may also be helpful.

Annex Table 7: RLS-I activity attendance and change in disputant perception

Relationship between RLS-I activities attended and knowledge levels at endline	Δ Access Rights	Δ Decision Subverted	Δ Freedom of Forum	Δ Outcome Just
Constant	-0.700	0.389	-0.780	-0.884
# activities attended	0.108**	-0.043	0.093*	0.110**
Effect size	0.414	0.056	0.385	0.499
Proportion of variance accounted for	*0.132*	*-0.019*	*0.108*	*0.215*

** Significant at 10%* ***Significant at 5%* **** Significant at 1%*

Annex Table 8: Size of cohort and elder knowledge

Relationship between RLS-I activities attended and knowledge levels at endline	Afghan law	Non-Afghan law
Constant	0.484	0.907
cohort	0.002	-0.003***
Effect size	0.432	0.530
South	0.017	-0.488
North	0.093	-0.311
Proportion of variance accounted for	*0.086*	*0.432*

** Significant at 10% **Significant at 5% *** Significant at 1%*

Annex Table 9: Size of cohort and change in elder knowledge

Relationship between RLS-I activities attended and knowledge levels at endline	Δ Afghan law	Δ non-Afghan law
Constant	-0.186	0.121
Cohort	0.003***	-0.003***
Effect size	0.532	0.512
South	0.065	-0.342
North	0.073	-0.304
Proportion of variance accounted for	*0.160*	*0.241*

** Significant at 10% **Significant at 5% *** Significant at 1%*

Annex Table 10: Size of district cohort and disputant perception

Relationship between RLS-I activities attended and knowledge levels at endline	Access Rights	Decision Subverted	Freedom of Forum	Outcome Just
Constant	4.23	0.877	4.45	4.72
Cohort	0.004***	0.002	0.003	0.001
Effect size	0.355	0.089	0.209	0.050
South	-0.158	1.08	-0262	-0.312
North	-0.187	1.60	-0.260	-0.378
Proportion of variance accounted for	*0.237*	*0.235*	*0.132*	*0.109*

** Significant at 10% **Significant at 5% *** Significant at 1%*

Annex Table 11: Size of district cohort and change in disputant perception

Relationship between RLS-I activities attended and knowledge levels at endline	Δ Access Rights	Δ Decision Subverted	Δ Freedom of Forum	Δ Outcome Just
Constant	-0.866	1.06	-1.29	-0.693
Cohort	0.011	-0.010*	0.013	0.006*
Effect size	0.544	-0.516	0.670	0.357
Proportion of variance accounted for	*0.264*	*0.233*	*0.424*	*0.089*

** Significant at 10% **Significant at 5% *** Significant at 1%*