
MEMORANDUM

To: Sophia Sahaf, MCC

From: Leonard Wantchekon, NYU and James Hollyer, NYU

Subject: Evaluation Design Report, Tanzania TCP

Date: January 19, 2010

In May 2006, the MCC and the Government of Tanzania signed a Threshold Country Agreement. This Agreement established the Tanzania Threshold Country Program (TCP), an \$11.15 million program designed to curb corruption in Tanzania. The program had four basic components: (1) Strengthening the rule of law, (2) Improving public procurement practices, (3) Increasing the ability of civil society to monitor government behavior, and (4) the creation of a Financial Intelligence Unit. The TCP closed on September 30, 2008.

In November of 2009, the MCC hired two external evaluators - Leonard Wantchekon and James Hollyer of New York University - to conduct an evaluation of the TCP's efficacy. The purpose of this evaluation is twofold: (1) to derive lessons from the successes and failures of the Tanzania TCP to inform both the Government of Tanzania and the MCC in their future activities; and (2) to fulfill the MCC's institutional commitment to measuring the results of its activities.

This report details the methods and measures to be employed in the evaluation of the TCP's impact. The report details evaluation design and measures for each element in the TCP. Section 1 discusses the evaluation of procurement reform. Section 2 details the strategies for evaluating rule of law interventions. These include treatments targeting both the Prevention and Combatting of Corruption Bureau (PCCB) and the Department of Public Prosecution (DPP). Section 3 deals with the TCP's interventions into civil society. These include (1) the training of Public Expenditure Tracking Survey (PETS) councils in local districts, (2) the provision of training to investigative journalists, and (3) support for the foundation of two Wildlife Management Areas (WMAs) in northern Tanzania. Finally, Section 4 discusses the creation of a Financial Intelligence Unit (FIU).

1 Public Procurement

Evaluation of the Public Procurement Regulatory Authority (PPRA)

Reforms to public procurement practices constituted one of the four pillars of the Tanzania TCP. \$3 million of the TCP budget was allocated towards this end. The PPRA was the Government of Tanzania partner agency involved in all efforts targeted at procurement

reform. The TCP involved four interventions that targeted procurement: (1) The (largely non-technical) training of PPRA officials; (2) Encouraging best practices, particularly by funding an information gathering trip to the US by PPRA Board members, who met with officials at US procurement agencies; (3) Funding a series of workshops with procuring entities to garner input as to challenges and problems facing current procurement policy; and (4) Funding a series of audits of procuring entities - conducted by Kilimanjaro International, in cooperation with the PPRA - which were made publicly available once complete.

The evaluation of these activities will proceed in two stages. The first will attempt to assess the effectiveness of activity 4, the audits funded by the TCP. The second stage will test for any cumulative impact of the four interventions described above. Unfortunately, it is not possible to disentangle the effects of interventions 1-3. Treatments take place at a national level and are not rolled out over time - implying that the identification strategy will have to rely on *before - after* comparisons, wherein measures taken prior to intervention are used as a control group and measures taken after intervention are used as a treatment group.¹ Data are only available at fiscal year intervals, making it impossible to distinguish between the various treatments in question.

To assess the efficacy of the audits conducted of procuring entities as part of the TCP, we will rely on data generated by the audits themselves and by follow-up audits conducted by the PPRA. Both the initial audits and the PPRA follow-up audits contain summary measures of overall rates of compliance with procurement regulations. The Kilimanjaro audits also contain disaggregated information regarding compliance with tender advertisement requirements, time for preparation of bids, bidder participation in tenders, time for processing bids, percentage of tenders involving bidder complaints, percentage of contracts requiring extensions and variations, percentage of contracts and tenders with incomplete information, and percentage of tenders wherein preference was applied. Presumably, the PPRA audits contain the same measures (though reports available on the PPRA website only report summary composite measures).

Simple difference-in-means tests can be used to assess the extent to which the target procuring entities improved on these measures between the initial audit and the PPRA's follow-up.² Unfortunately, the sample size is quite limited - only 20 audits were conducted - implying that standard errors are likely to be large and only very large effects will be significant and conventional levels of significance.³ However, large point-estimates indicating

¹This identification strategy relies on the strong assumption that any changes other than the interventions during the period between measurements are ignorable, i.e. that these changes are orthogonal to the measures described below.

²Improvement on any given measure is unlikely to be independent of improvements on other measures, so the test will likely rely on a Seemingly Unrelated Regression (SUR) specification including a post-treatment dummy.

³In such small samples, one must also make strong distributional assumptions to justify the use of standard

improved compliance across multiple measures may be taken as indicative of a treatment effect. These quantitative tests may be strengthened by additional qualitative assessments of the information contained both rounds of audits.

Two tests may be used to attempt to assess the cumulative effect of all interventions that targeted the PPRA. The first relies on data drawn from the initial round of 20 audits conducted by Kilimanjaro International (covering fiscal year 2006-7) and on data from an additional 30 audits covering the fiscal year 2007-8 (and released in July 2009). Under the admittedly quite strong assumption that the criteria by which procuring entities were selected to be audited remained constant across both rounds of auditing, differences in average compliance scores are a valid measure of effect of TCP treatments on procurement practices. The specification used in this evaluation would be essentially identical to that used in assessing the difference between initial and follow-up audits discussed above.

Additionally, the PPRA maintains records of awarded contracts for Goods, Consultancy Services, Non-Consultancy Services, and Works for Government Ministries, Local Governments, Parastatal Organizations, Independent Departments, and Government Agencies. The data run from fiscal year 2006\7 to 2008\9. It may be possible to examine changes in prices paid for analogous goods over time.⁴ One would expect improvements in procurement processes to lead to a decline in real prices paid for analogous goods between the 2007\8 period and the 2008\9 period. The presence of 2006\7 data allows for a placebo test - if reductions in prices between 2007\8 and 2008\9 are due to improvements in procurement stemming from TCP interventions, we should not expect to find a similar decline in prices between 2006\7 and 2007\8.

The quantitative tests discussed above unfortunately must rely on quite strong identifying assumptions. Because of the manner in which the TCP interventions were designed, we must rely on pre-intervention measures as a controls to assess any causal effect. Any results from the above should, therefore, be regarded as indicative rather than as definitive evidence of a causal effect. These results will be paired with qualitative results - relying largely on interviews with program participants and informed observers - which may provide further evidence regarding any results generated by quantitative tests.

2 Rule of Law

Evaluation of Aid to the Prevention and Combatting of Corruption Bureau (PCCB)

\$3 million of the TCP was devoted to programs meant to enhance the rule of law. The

t-tests.

⁴It seems less likely that services and works contracts are comparable over time.

PCCB was the largest of the two Government of Tanzania bodies affected by this element of the Program. (The other agency was the Department of Public Prosecution (DPP).) The PCCB benefited from (1) extensive training efforts funded through the TCP, (2) the provision of information technology and training in its use, and (3) the provision of 34 1998 vintage Toyota Rav4s. Training was provided by a variety of implementing partners (ICITAP, OPDAT, NYC DoI, and Crown Agents Consulting) and was rolled out gradually over the course of the TCP.

Our evaluation will attempt to exploit the gradual roll-out of training programs identify the causal effect of the training provided to the PCCB. The training programs generally targeted district bureau chiefs from across the country, with the intention that these chiefs would then convey the knowledge gained as a result of training to their subordinates. However, it was not possible to train all targeted officials at once - training was provided in waves to different groups of bureau chiefs. Our identification strategy will use these gaps between one wave of training and the next to construct a difference-in-differences design. We will compare changes in our measures of interest in bureaus that received training to those that did not in a given interval of time. This strategy assumes that the process by which bureau chiefs were assigned to a given wave of training is ignorable (e.g. either random or based on covariates orthogonal to changes in the measures of interest). This assumption will be discussed further below.

Our measures will be derived from data provided by the PCCB. The PCCB retains records of the number of allegations received, cases investigated, completed investigation files, administrative actions taken against Bureau employees, files transferred to other agencies, new cases brought to courts, cases prosecuted, convictions secured, number of acquittals, and the sums recovered. We are in possession of annual data on all such measures dating back as far as 1995. Data are collected from regional bureaus quarterly - and we hope to obtain data disaggregated at this level. Measures of particular interest include cases investigated as a percentage of allegations received, cases completed as a percentage of cases investigated, and cases prosecuted as a percentage of allegations received.⁵

We will thus assemble a time-series cross-sectional dataset containing these measures. Our dependent variables of interest are the the change in the number of cases investigated as a percentage of allegations received, the number of cases completed as the percentage of the number of cases investigated, and the number of convictions secured as a percentage of the number of allegations received. The main explanatory variable of interest is a dummy indicator that takes the value of 1 in the period during which a given bureau chief receives

⁵The number of allegations received likely reflects both the underlying level of corruption and citizens' willingness to come forward with information that they might have. These factors should be taken into account of in our final measures to be sure that any changes do not simply reflect a change in the willingness to report crimes or a change in the underlying rate of corruption.

training. The use of first differences in this specification ensures that all fixed bureau-specific characteristics are controlled for in our results. However, we do assume that - absent training - the rate of change in the various measures discussed above will be identical across bureaus. We can test this assumption by running placebo tests prior to the introduction of the TCP.

Evaluation of Aid to the Department of Public Prosecution (DPP)

Approximately \$610,000 of the \$3 million granted to Rule of Law programs under the TCP was targeted towards the DPP. This aid was not allocated in the initial TCP plan. Rather, it was issued in response to the GoT's decision - in 2007 - to civilianize its prosecutorial force. As a result, the DPP needed to hire a large number (approximately 160) of new civilian prosecutors. These new recruits, it was felt, would require training before entering the DPP. Thus, USAID devoted some of the funds from the Rule of Law program under the TCP to training the entire class of new DPP recruits. Training was conducted by Kilimanjaro International and largely focused on the implications of the new anti-corruption act (passed in 2007), which vastly expanded the number of charges that prosecutors could bring against defendants.

Unfortunately, developing rigorous quantitative assessment of the effect of this training is likely to prove highly problematic. There does not exist a clear control group with which to compare officials affected by this intervention. *All* new recruits received training. Moreover, changes to the training program coincided with the civilianization of the DPP, implying that new recruits affected by this intervention are unlikely to be comparable to those who entered in previous years. And comparisons of outcomes before and after this intervention took place will reflect *both* the influence of training and the influence of civilianization itself.

Therefore, we will rely on qualitative assessments of the efficacy of this training. We have and will continue to conduct interviews with participants in the design of this program and informed outside observers. And we will likely attempt to talk to participants in the training process itself to garner their impressions of its efficacy. It is worth noting in this regard that the DPP has continued this training procedure after TCP support has been withdrawn, which may be taken as a positive assessment of its efficacy.

3 Strengthening Civil Society

Evaluation of Public Expenditure Tracking Surveys (PETS)

As part of the program to strengthen Civil Society Programs, the TCP called for training to be provided for local PETS councils in a subset of Tanzania's 123 districts. PETS councils would examine the budgetary records of district level governments to detect signs

of graft or of spending on priorities other than those set by village councils. They would post the results of their investigations on local signboards, to inform other locals of any discrepancies discovered. PETS training was provided by Pact-Tanzania and the Campaign for Good Governance and implemented through civil society organizations (CSOs) active in participating districts. CSOs chose whether or not to take part in PETS training and determined the districts that would receive training.

Unfortunately, it will not be possible to assess the degree to which district level spending corresponded to the local demand as expressed through village councils in a rigorous quantitative manner. However, it is possible to assess changes in various measures of budgeting transparency and indicators of corruption in response to PETS training. We rely on measures derived from the Council Auditors General (CAG) reports. These measures include questioned revenue as a percentage of total revenue, questioned expenditures as a percentage of total expenditures, personnel charges as a percentage of total expenditures,⁶ and the auditor's opinion of district accounts (a trichotomous variable taking the values 'adverse', 'qualified', and 'clean').⁷ High levels of questioned expenditures\revenue as a percentage of total expenditures\revenue or high levels of personnel expenditures as a percentage of total expenditures may be taken as indicative of poor accounting and are likely to be correlated with corruption. Much the same is true of 'adverse' auditor opinions. To the extent that PETS councils are effective in their intended aims, district officials - fearing that they will be publicly revealed as incompetent or corrupt - will devote effort towards improving these measures.

The greatest difficulty in assessing the effect of the PETS program lies in the non-random assignment of districts to treatment. To the extent that CSOs that received intervention do not resemble those that did not receive the intervention, simple comparisons between these two groups may be biased. We attempt to address this difficulty first by employing a difference-in-difference design that controls for all district level factors that do not change over time. The changes in (roughly) continuous measures detailed above - the percentage of questioned revenues\expenditures and the percentage of expenditures devoted to personnel - will be regressed on a dummy indicator that takes the value one when a district receives treatment. For the auditor's opinion measure, we will construct an ordinal index $\{-1, 0, 1\}$ indicating, respectively, a worsening, no change, and an improvement on the auditor's index. An ordered probit regression can then be used to assess the probability of any change in this index.

The use of the difference-in-differences estimator requires the assumption that - absent the PETS intervention - changes in the variables of interest would be identical in observa-

⁶This measure is often used as an indicator of corruption in the existing literature

⁷Additional measures may be used once the content of the CAG reports has been reviewed more thoroughly.

tionally similar treated and non-treated districts. We can test this assumption by running placebo tests for years prior to the beginning of the TCP using the CAG data.

As a robustness check, we can control for selection of districts based on observable characteristics through the use of a propensity score matching algorithm.⁸ We can then compare our variables of interest in treated and untreated districts using the balanced sample to gain a measure of the causal effect of PETS training.

Evaluation of Investigative Journalist Training

As part of its program to Strengthen Civil Society to fight against corruption, the TCP created a program of investigative journalist training. This training aimed to equip journalists with the skills necessary to investigate and disseminate evidence of government corruption. Training was conducted by the implementing partner Pact-Tanzania. The content of training changed over time. Initially, journalists selected by local media industry organizations were given a week's worth of lectures on topics such as how to locate sources. Roughly 220 journalists took part in such classes. In late 2007 or early 2008, the training program was revised to include travel grants meant to help reporters develop a story and a forum in which reporters submitted stories for critique by trainers and other participants. Roughly 80 journalists received this more involved training.

To measure the effect of this training, we will rely on newspaper clippings of corruption related reports collected by Pact-Tanzania from newspapers circulated in Dar es Salaam. We will also make use of the names of participants and the date of their training, also provided by Pact. Pact has already demonstrated that the total number of corruption related articles is rising over time. However, many potential confounds - from increasing newspaper publication overall to an increasing awareness of corruption unrelated to the TCP - may account for this increase. We intend to instead measure the individual propensity to publish corruption-related articles and the 'quality' of these articles and to assess how these are associated with training.

To measure the individual propensity to publish corruption related articles, we will construct a time-series cross-sectional database containing count information on the number of corruption-related articles published by a given author in a given period of time (e.g. number of corruption-related articles per month). We will also construct dummy variables that take the value of 1 after a journalist received either the initial week or subsequent 3-stage training. The specification would involve a count (e.g. negative binomial) model

⁸Using data on district level characteristics such as poverty levels, agricultural productivity (both drawn from the World Bank), population levels, age distributions (from the Tanzanian Socio-Economic Database), and representation by government or opposition parties we can predict the probability that given district is selected for intervention. This probability (propensity score) can be used to construct a sample that is balanced on observed covariates across treatment and control groups.

regressing the number of corruption articles published in a given period on journalist-specific fixed-effects, training dummies, and a time trend measure (e.g. a cubic time trend). Such a specification would assess whether the average number of corruption articles published by a given journalist increased following training - controlling for time trends affecting all journalists.

These measures may be biased downwards due to measurement error. Pact reports that many journalists choose to publish their articles anonymously. Journalist anonymity will be a particular problem if authors grow more likely to hide their identity after receiving training. However, even if the probability with which a journalist seeks anonymity is unaffected by training, the non-classical nature of the measurement error will tend to bias our results downward since our measures can only understate the number of articles published by a given author.⁹ Therefore, results from this specification should be treated as a lower bound on the magnitude of any effect of training captured.

Journalist training was no doubt intended not only to increase the quantity of corruption-related articles produced by journalists, but also to increase their quality. Since quality is an inherently subjective measure, it would be best to assess this change by constructing a clear measurement model. We intend to have two coders read each clipping from the Pact archive that can be attributed to a named author (both readers will be asked to read each article, to minimize on reader specific effects and enhance measurement validity). Each reader will be asked to answer a set of binary questions regarding each article. For instance, ‘Does the article present information in a clear manner?’, ‘Does the article include numerous grammatical errors or typos?’, ‘Does the article attempt to relate specific instances of corruption to broader economic or political issues, or is does it concentrate only on a specific instance of corruption?’, ‘Does the article cite any sources? If so, does it cite more than one?’, etc. We can treat article quality as a latent variable explaining the outcomes to these binary questions, and we can thus extract a measure of quality from either a factor analysis of these responses or from an item response theory (IRT) model using this data. The results will constitute a dependent variable for subsequent analysis.

Quality can be modeled as a function of journalist specific fixed effects, time trends, and training indicator variables (much as in the model above, however employing a linear rather than a negative binomial specification). These results should give an indication as to whether the average quality of an author’s articles increases following training, controlling for time effects. This measure may be taken as an indication of the effect of training.

Evaluation of the Support of Two Wildlife Management Areas (WMAs)

⁹Assume for instance that a given author publishes one corruption-related article per month before treatment, and two such articles after treatment. Now assume that said author seeks anonymity with probability 0.5. In expectation, the observed difference in publishing before and after treatment is 0.5 articles per month, even though the true causal effect of training is an increase of one article per month.

While not a part of the original TCP; some \$200,000 was allocated to the support of two WMAs (Burunge WMA and Enduimet WMA) in northern Tanzania. These activities received funding in response to an RFP issued in October of 2007, which was issued under USAID's initiative as a result renewed progress by the GoT on its WMA program.¹⁰ The WMAs were intended to (1) improve wildlife management in villages bordering national parks by constructing local representative institutions charged the oversight of wildlife and ecotourism, (2) to give locals a stake in the preservation of wildlife by increasing their share in the profits from ecotourism, and (3) to eliminate petty corruption involving tourist operators bribing local officials to secure permission to build tourism facilities at low cost. Aid under the TCP was provided through the African Wildlife Foundation (AWF), which helped the Burunge and Enduimet WMAs draft business plans, provided information about the going rate for paid by tourism providers, and constructed a best practice toolkit for WMA participants.

Unfortunately, given the small number of participating villages involved in the WMAs and the low level of aid involved, a rigorous quantitative investigation of the effect of this aid is not feasible for this study. Rather, we will rely on a qualitative evaluation of these activities. While any evidence produced by such a study will not be definitive; it can at least provide some sense of whether or not one should expect to find an effect of this intervention.

To conduct our qualitative evaluation, we will rely on interviews with participants in the TCP - notably AWF management and USAID officials. We will also seek to interview members of the Wildlife Division to gain their impressions of the WMA program as a whole and to gain access to any data regarding the Burunge and Enduimet WMAs that might be indicative of the effectiveness of interventions in these Areas. We will also attempt to gain access to the records of any meetings of WMA boards and information regarding the selection of WMA representatives. These materials should help to give some indication as to whether the WMAs functioned in the intended manner.

4 Creation of a Financial Intelligence Unit

Evaluation of the Creation of a Financial Intelligence Unit (FIU)

The creation of a Tanzanian Financial Intelligence Unit (FIU) was one of the four central pillars of the TCP, and \$1.5 million was allocated to this task. The FIU is an agency operating within the Tanzanian Ministry of Finance and Economic Affairs that is charged with combatting money laundering and the financing of terrorism. To this end, it is meant (1) to collect reports of irregular financial dealings from private sector financial

¹⁰Progress on this front followed the removal of the incumbent Director of the Wildlife Division.

institutions, (2) to disseminate information to private sector financial institutions regarding what constitutes an irregular transaction and how such a transaction should be reported, and (3) launch its own investigations into suspected money laundering or terrorist financing. These activities were intended to enforce the provisions of the Anti-Money Laundering (AML) Act of 2006, which expanded the list of predicate offenses for money laundering and was also a requirement of the TCP.

Unfortunately, a number of factors conspire to make the rigorous quantitative evaluation of the effect of the creation of the FIU problematic. First, since the creation of the FIU affected the entire country, the construction of an adequate control group is difficult. Second, adequately measuring the level of money laundering and/or the financing of terrorism is an extremely difficult endeavor, and well beyond the scope of this study. Third, if rational money launderers believed that the FIU would have an effect, they would doubtless alter their activity even before the Unit began operations - biasing any measure of causality that attempted to compare measures before and after the FIU began operations. Fourth, the FIU is only beginning investigative and data collection activities now (in 2010). To date most efforts have been devoted to the formation of the Unit and informing participants in the financial sector of the requirements of the AML Act and of the methods for reporting suspicious transactions.

For these reasons, we will instead conduct a qualitative evaluation of this element of the TCP. While such evidence cannot be taken as definitive for the presence or absence of any treatment effect; it can at least provide some information as to whether or not we would expect to find such an effect.

To conduct this evaluation we will rely on information from several sources. First, we have and will continue to interview key actors in the TCP and informed observers of the FIU. These include the FIU chairman, USAID contacts, several of the implementing partners, and officials at the Bank of Tanzania and the Revenue Bureau. We will also rely on third party reports - particularly the *Mutual Evaluation Report* of the FIU published in December of 2009 by the Eastern and Southern Africa Anti-Money Laundering Group (ESAAMLG). A particular focus of this evaluation - given the information we have collected so far - will be the staffing of the FIU.¹¹

¹¹The appointment of a chair of the FIU ran behind the schedule set out in the TCP. At present, in addition to the Chairman, the FIU has only two non-secretarial staff, both of whom are seconded from the Bank of Tanzania.