MATHEMATICA
Policy Research

Evaluation of the Vocational Training Grant Fund in Namibia: Baseline Data User's Manual

November 4, 2015

Evan Borkum
Arif Mamun
Malik Khan Mubeen
Linus Marco

Submitted to:
Millennium Challenge Corporation
875 15th Street, NW
Washington, DC 20005
Project Officers: Algerlynn Gill and Emily Travis
Contract Number: MCC-10-0114-CON-20 (MCC-13-TO-0001)

Submitted by:
Mathematica Policy Research
1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Arif Mamun
Reference Number: 40233.332

This page has been left blank for double-sided copying.

CONTENTS

TABLES

I. INTRODUCTION

To promote economic growth and reduce poverty in Namibia, the Millennium Challenge Corporation (MCC) signed a $304.5 million compact with the Government of the Republic of Namibia in 2009. The compact, which was formally completed in September 2014, included three projects: tourism, agriculture, and education. The education project sought to address the shortage of skilled workers in Namibia and limitations in the education system's capacity to create a skilled workforce. It included a vocational training activity, which focused on expanding the availability, quality, and relevance of vocational education and skills training in Namibia. MCC contracted with Mathematica Policy Research to conduct an evaluation of the vocational training activity.

The Vocational Training Grant Fund (VTGF) subactivity was one of the components of the vocational training activity. It involved awarding grants to training providers for high-priority vocational skills programs; training providers that received these grants used them to award scholarships to eligible applicants to participate in these programs. Mathematica is conducting an impact evaluation of the VTGF subactivity that relies on a random assignment design, in which eligible applicants to each VTGF-funded training were randomly assigned to a group that received the offer of VTGF funding (treatment group) and a one that did not (control group). To inform this evaluation, a baseline survey of eligible applicants was conducted between late-2011 and mid-2014.

This manual provides information about the sample, data collection, and data cleaning for the VTGF baseline survey (these data were analyzed in our baseline report, Borkum et al 2015). It also describes the content and format of the baseline data files that Mathematica submitted to MCC.

This page has been left blank for double-sided copying.

II. SAMPLE

The targeted sample for the VTGF evaluation consists of the 1,892 applicants who applied to the 28 VTGF-funded trainings listed in Table II.1. Eleven different providers conducted these trainings, some of which provided multiple trainings. The list of trainings in Table II.1 does not cover the full set of trainings funded by the VTGF subactivity. Specifically, it excludes 26 trainings for which there was no control group (typically because there were sufficient slots to accommodate all applicants), 21 trainings for which the follow-up survey date (one year after the end of training) would fall outside of the evaluation period, and 9 trainings for which there were severe violations of random assignment (the first three intakes of COSDEC Benguela). Although baseline data were collected for some of these 56 excluded trainings because the evaluation design had not been finalized at the time of data collection, we only cleaned and analyzed the baseline data for applicants to the 28 trainings in Table II.1. Therefore, only applicants to these trainings are included in the data files provided to MCC.

Table II.1. Targeted sample for the VTGF evaluation

Training provider	Course	Intake	Training start date	Number of treatment applicants	Number of control applicants
NATH	Tour Guiding	1	4-Oct-10	50	33
Wolwedans	Hospitality & Tourism	1	11-Jan-11	31	3
Wolwedans	Hospitality & Tourism	2	11-Jul-11	35	25
Wolwedans	Hospitality & Tourism	3	7-Feb-12	39	11
ABTCC	Food & Beverage/ Housekeeping	1	4-Sep-12	15	16
ILSA	Reception Management & Office Administration	1	1-Oct-12	118	27
IUM[a]	Hospitality & Tourism	1	5-Jan-13	16	59
IUM[a]	Hospitality & Tourism	1	5-Jan-13	16	243
VVTC	Front Office	1	3-Jun-13	12	6
VVTC	Food Production	1	3-Jun-13	10	21
VVTC	Housekeeping & Food Preparation	1	3-Jun-13	13	7
VVTC	Food & Beverage Services	1	3-Jun-13	12	6
OVTC	Hospitality & Tourism	1	4-Mar-13	35	22
ZVTC	Plumbing	1	8-Jul-13	20	68
ZVTC	Hospitality & Tourism	1	8-Jul-13	20	168
ZVTC	Office Administration & Computing	1	8-Jul-13	16	212
ZVTC	Bricklaying	1	8-Jul-13	20	24
KAYEC	Carpentry	1	1-Oct-13	15	18
KAYEC	Shuttering	1	1-Oct-13	15	4

Table II.1. *(continued)*

Training provider	Course	Intake	Training start date	Number of treatment applicants	Number of control applicants
KAYEC	Concrete Work	1	1-Oct-13	15	16
KAYEC	Concrete Work	2	17-Mar-14	9	1
COSDEC Benguela	Office Administration & Computing	4	14-Apr-14	30	16
NamWater	Grader	2	19-May-14	10	4
NamWater	Bulldozer	2	19-May-14	10	2
NamWater	Forklift	2	19-May-14	20	5
KAYEC	Shuttering	3	25-Jun-14	30	25
KAYEC	Carpentry	3	25-Jun-14	30	22
KAYEC	Concrete Work	3	25-Jun-14	23	4
Total	**--**	**--**	**--**	**955**	**937**

Notes: Table excludes 26 trainings with no control group (2 NATH trainings, 2 ZVTC trainings, 5 KAYEC trainings, 10 RVTC trainings, 1 NamWater, 4 NAMCOL, and 2 COSDEC Benguela trainings); 21 trainings not covered by the evaluation period (4 NAMCOL trainings, 14 NIMT trainings, and 3 NamWater trainings); and 9 trainings with severe violations of random assignment (9 COSDEC Benguela trainings).

Number of treatment and control applicants corrects for multiple applications; applicants are linked to the first included training to which they applied.

[a] IUM hospitality and tourism trainings were conducted at two separate campuses with separate random assignment; these are treated as separate trainings for evaluation purposes.

III. DATA COLLECTION

A. Timing

The baseline data were collected between December 2011 and July 2014. The long fielding period for the baseline survey reflects the fact that VTGF grants were awarded (and training providers conducted random assignment) at several points throughout the compact period, and the baseline survey for applicants to each training was expected to be conducted soon after random assignment. (In practice, for reasons described in Borkum et al 2015, the baseline survey almost always was often conducted several months after training had started). MCA-Namibia collected baseline data for the initial cohorts of applicants between December 2011 and August 2012. NORC (in partnership with Survey Warehouse, a local data collection firm) took over the data collection for subsequent cohorts in February 2013, and continued to collect baseline data after Mathematica joined the evaluation in mid-2013.

B. Questionnaire

The VTGF baseline questionnaire was initially developed by MCA-Namibia, and contained several sections (Table 2). It collected data on basic demographic characteristics of the applicants, together with a range of outcome measures relevant to the evaluation research questions. These outcomes focused on the applicants' vocational training history, employment status, and earnings and income. It also gathered extensive contact information for applicants to facilitate their being contacted for the follow-up survey.

Some changes were made to the questionnaire over time. NORC made a handful of changes when it took over the data collection from MCA-Namibia in February 2013, and Mathematica made a small number of further changes when we joined the evaluation in mid-2013. The changes made to the questionnaire over time were relatively minor, and involved adjusting the wording of some questions, adding or removing some questions, and making some changes in question order and skip patterns. Despite these changes, the basic questionnaire and methodology remained similar over time, enabling us to combine data from different periods for the analysis. The questionnaire is included in the data package and indicates changes that were made over time when NORC took over the data collection (highlighted in yellow) and when Mathematica joined the evaluation (in blue). The version of the survey used for each observation in the public use file (MCA-Namibia, NORC before Mathematica joined, or NORC after Mathematica joined) is specified in the variable t0_survey_version.

Table III.1. VTGF baseline survey sections

Section	Key topics covered
Identifying and contact information	Name; identification number; date of birth; region of residence; town and region of origin; telephone contact numbers; email address; postal address; contact information of friend or relative.
Demographic information	Age; gender; marital status; nationality.
Education (excluding vocational training)	Highest level of education; whether moved for education; desire for further education; challenges to further education
Vocational training	Enrollment in vocational training (in previous five years and as of survey date); total months of vocational training; sectors and skill areas of vocational training; job attachments; perceived quality of vocational training; dropout from vocational training; completion of vocational training (including sectors, skill areas, and institutions); accessibility of vocational training (*).
Employment and earnings	*Employment status:* whether currently employed; availability for employment; whether actively seeking employment. *Among those employed:* number of jobs currently held; hours and days worked; type of employment (part-time, full-time, or self-employed); help received in finding employment; relevance of employment to training; whether employment is paid; job tenure; job satisfaction; size and sector (formal or informal) of workplace; source of information about job; occupation and sector of employment; monthly income from employment; number of dependents on earnings. *Among those unemployed:* duration of unemployment; reasons for unemployment; whether previously employed (including satisfaction and reason for leaving); willingness to consider vocational training in the future (*).
Household demographics and income	Household size; ownership status of dwelling (*); monthly household income; main sources of household income; relationship of respondent to head of household; parental education; orphan status.

(*) = Removed from the survey when Mathematica joined the evaluation in mid-2013.

C. Data collection procedures

The baseline survey was conducted by telephone, using contact information collected by training providers as part of the application and random assignment process. For the initial cohorts covered by MCA-Namibia, interviews were conducted by MCA-Namibia staff, who entered the data into a spreadsheet. Interviews for subsequent cohorts managed by NORC used a proprietary computer-assisted telephone interview system (Liberty). This portion of the data collection was conducted by interviewers from Survey Warehouse from their call center in Windhoek, Namibia; NORC staff conducted trainings for Survey Warehouse interviewers and also provided continued oversight of the data collection effort. Data entered into the Liberty system during the interviews were uploaded to NORC's server in the US in real time.

D. Response rate

A total of 1,406 applicants in the treatment and control groups from the 28 targeted VTGF trainings in Table II.1 responded to the baseline survey (741 in the treatment group and 665 in the control group). This sample reflects an overall baseline survey response rate of 74 percent (78 percent in the treatment group and 71 percent in the control group).

IV. FILE PROCESSING AND CONTENT

The full set of pubic use files provided to MCC, which we describe in this section, is listed in Table IV.1. These files include the raw data file, baseline public use data file, the cleaning and analysis programs, a restricted use file, and a codebook (consisting of a summary file and a full codebook).

Table IV.1. Baseline files provided to MCC

File type	File name
Raw data file*	Nam_VTGF_bline_Raw data
Public use data file	Nam_VTGF_bline_Public use file.dta
Cleaning do file	Nam_VTGF_bline_cleaning.do
Construct and analysis do file	Nam_VTGF_bline_analysis.do
Restricted use data file*	Nam_VTGF_bline_Restricted use file.dta
Codebook summary	Nam_VTGF_bline_Public use for codebook.dta_contents.xlsx
Codebook	Nam_VTGF_bline_Public use for codebook.dta_summary statistics.txt

* Contains personally identifiable information. Raw data file also includes individual identifiers such as name and ID numbers.

The baseline public use data file "Nam_VTGF_bline_Public use file.dta" includes the following types of variables:

1. Cleaned survey variables, denoted with prefix t0_.
2. Constructed variables used in the analysis, denoted with prefix t0_x_.
3. Variables related to random assignment obtained from training providers, denoted with prefix MPR_.

To obtain the baseline public use data file, the raw survey data file provided by NORC (which included interviews completed by MCA-Namibia), was processed and cleaned as follows:

- The raw data provided by NORC included applicants to some trainings that were not among the 28 trainings included in the evaluation (Table II.1). Mathematica merged the raw data file with information on random assignment obtained from training providers. The random assignment information included several instances of the same applicant applying to multiple trainings. We consolidated the training information for each unique applicant before merging with the raw data file; any applicant that did not apply to at least one of the 28 included trainings was then dropped from the data file.[1]

[1] The raw data file provided to MCC also only includes applicants from the 28 trainings included in the evaluation.

- Based on the random assignment information for each respondent from training providers, we created a binary treatment assignment variable, MPR_treat (1 for treatment and 0 for control), and a unique code for each training, MPR_unique_traincode1. For applicants who applied to multiple included trainings, this information is based on the first training to which they applied—consistent with the analysis plan for the impact evaluation.
- Mathematica removed more detailed information about the trainings, including the random assignment date, provider name, course name, intake, and training start and end dates, to reduce the risk that respondents could be identified (as described in further detail below). These variables (denoted with a prefix MPR_) are available in a separate restricted use file, "Nam_VTGF_bline_Restricted use file.dta"; users who are given access to this file can merge this information with the public use file using the variable t0_rid.
- Cases identified as survey incompletes in the raw data files from NORC were dropped from the public use data file. The numbers of observations of complete cases for relevant trainings in the baseline data file is 1,406.[2]
- Certain variables in the raw data file have been excluded. Some have been dropped because they contain direct identifying information, like names, telephone numbers, and national ID numbers. Additional changes and exclusions to protect anonymity of the respondents are described in the anonymization section below. Other variables were excluded because they are survey administration variables, like internal NORC ID numbers and records of respondents' demeanor during the interviews.
- Mathematica cleaned the raw survey variables. The cleaning process included checking the validity of variable values and ranges; verifying skip patterns; cleaning and back-coding common "other-specify" responses; creating binaries of categorical variables; checking and correcting for duplicate observations (applicants who applied to multiple trainings and were surveyed twice); and recoding skips, missing data, and other non-response values to standardized lettered indicators (Table IV.2).
- "Nam_VTGF_bline_cleaning.do" documents cleaning for the baseline data file; cleaned survey variable names are denoted with a prefix t0_ and correspond to the sections and the numbering of the questions in the survey instruments.
- "Nam_VTGF_bline_analysis.do" creates constructs created based on the survey items, denoted with a prefix t0_x_, and conducts the baseline analysis reported in Borkum et al (2015).

[2] The raw data file provided to MCC also only includes completed surveys.

Table IV.2. Missing data codes in the VTGF baseline data

Type	Explanation	Missing data code
Legitimate skip	No value recorded due to a skip pattern in the survey	.a
Missing (illegitimate skip)	No value recorded but should have been recorded	.e
Don't know	Option "Don't Know" selected by respondent	.d
Not applicable	Option "Not Applicable" selected by respondent	.n
Refused	Option "Refused" selected by respondent	.r
Illegitimately non-missing	Question should have been skipped, but respondent selected a response.	.m
Missing in MCA-Namibia data	Missing code in MCA-Namibia portion of the dataset (was supposed to denote legitimate skip, but is not always the case)	.c
Construct missing	Constructed variable is missing because components are missing or constructed variable is not defined	.v

All cleaned survey, construct, and random assignment variables can be found in the accompanying codebook. Entries for each variable include the variable name, question text, universe, variable type, mean, minimum and maximum value for numeric variable, frequency of binary and categorical variables, and number of nonmissing responses.

This page has been left blank for double-sided copying.

V. DATA ANONYMIZATION

We reviewed the risks to the rights and privacy of individual respondents to the baseline survey, and determined the appropriate data anonymization and data access strategies that balance the need to minimize such risks with MCC's need to be transparent and provide adequate access to the data for policy research and the public good. The draft guidelines for data documentation and anonymization from MCC guided our review and decisions about potential risks and the level of anonymization for the baseline survey. In the process, we took into account issues such as the likelihood of compromise of the data and the value and potential uses of the data if compromised. In this section, we discuss these risks and risk mitigation strategies.

A. Removing individual identifiers

The baseline survey was administered to applicants to VTGF-funded trainings who were randomly assigned. The survey data included geographic identifiers of the trainee such as region (of residence, origin, and postal address) and town of origin, as well as direct individual identifiers such as name, national (or other) ID number, telephone number, email address, postal address, and the name and telephone number of a family member or friend to help contact the respondent for the follow-up survey. The data also included the name and address of the respondent's employer (for those who were employed at the time of the survey)

As part of the standard procedure to limit risks to privacy of the survey respondents, we removed all direct individual identifiers from the public and restricted use data files submitted to MCC.[3] We also made the following adjustments for the public use file:

- As a further precaution, we also removed the date of birth of the respondent, which was unique in many cases (age in years is available in the survey, so removing date of birth will not compromise the utility of the data).
- Because the town of origin variable is unique in many cases—and has only a handful of observations in most other cases—we removed this variable.
- We made a minor adjustment to the region of origin variable, by combining two regions with few observations (less than 10 each) into a single "other" category. With this adjustment, we retained region information, because these high-level geographic identifiers are not specific enough to identify individual respondents.
- We also removed the name and address of the respondent's employer, which was unique in many cases and could be used to identify the respondent.

The date of birth of the respondent, town of origin variable, unadjusted region of origin variable, and employer information are available in the restricted use file, which can be merged to the public use file using the t0_rid variable.

[3] These direct individual identifiers are included in the raw data file provided to MCC, which reflects the full data as received from NORC.

B. Truncating outlier values of continuous variables

We also examined continuous variables to determine whether outlier values could potentially be used to identify individuals. The only variables that we identified as posing a potential risk were the age of the respondent, the age at which the respondent's mother or father died (if relevant), and the number of household members. We made the following adjustments to the public use file:

- The age of the respondent poses a risk because there are relatively few respondents in the sample who are in their teens, as well as relatively few in their forties or fifties. We therefore top- and bottom-coded respondent age at the 95th and 5th percentile of the distribution, respectively. This truncated variable is denoted with a suffix _t in the data file.
- The age at which the respondent's mother or father died poses a risk because there are several values that only appear for a unique respondent, or a handful of respondents. We therefore recoded the cleaned version of these variables into categories (0-4 years, 5-9 years, 10-14 years, 15-19 years, and 20 years or older), and removed the raw variables.
- The number of household members poses a risk because there are relatively few households with a large number of members. Therefore, we collapsed this variable into two categories for households with more than 10 members: one category for 11-15 members, and another for 16 or more members. This recoded variable is denoted with a suffix _r in the data file.

The original versions of all these variables are available in the restricted use file, which can be merged to the public use file using the t0_rid variable. Variables related to earnings and household income, for which outlier values typically pose a risk to identification, were recorded in categories rather than as continuous variables; these variables are therefore not a source of risk for the respondents to the baseline survey.

C. Addressing unique and rare observations

We examined all categorical variables to determine any that might pose a risk to subject identification because of a limited number of observations with a particular value. To make this exercise as objective as possible, we simply identified all variables for which a given value was reported by fewer than 10 respondents. Then, on a case by case basis, we determined whether and how these variables should be adjusted. Based on our review, we made adjustments to the following variables for the public use file:

- Marital status: rare categories (for example, "divorced") were collapsed into the "other" category.
- Level of education: categories below grade 8 completion (no education, less than grade 7, completed grade 7) were rare, so we collapsed all of these into a single category.
- Sectors/areas of vocational training received and completed: sector/areas with fewer than 10 observations were collapsed into a single "other" category.
- Institution of vocational training: institutions with fewer than 10 observations were collapsed into a single "other" category.

- Relationship to household head: there were only 3 observations in the "niece" category, which were recoded into the broader "other" category.
- "Other, specify" variables: the responses recorded for variables were highly variable—most responses are reported by only a handful of respondents (usually by one respondent). We therefore removed these variables from the restricted use file.

All adjusted variables are denoted with a suffix _r in the public use data file. Again, all the original variables are available in the restricted use file. Although some of the remaining variables in the public use file have fewer than 10 observations per category, we determined that these variables do not pose a risk to identification because they are: (1) unlikely to be public knowledge (for example, the opinion of the respondent), and/or (2) applied to the experiences of the respondent during a specific period that occurred several years ago (for example, hours worked).[4]

D. Removing training information

The analysis data file also included information about the (first) specific VTGF training to which each respondent applied, including training provider name, intake number, course name, and course date. This was obtained by matching survey data to random assignment information and is used in the baseline analysis. Because some of these trainings only had a small number of applicants, including this information would potentially allow respondents to be identified in combination with information on their individual characteristics and outcomes. Therefore, we removed the training information from the public use file. Instead, we included an arbitrary identifier for each training (defined by a specific training provider, intake, and course), MPR_unique_traincode1, to enable users to include training fixed effects in the analysis. Although we also removed the course dates because they could be used to identify the specific training to which an individual applied, we retained constructed variables for the difference in timing between the training start date and survey date, grouped into categories (these variables were used in the baseline analysis). As mentioned above, the restricted use file includes the full training provider information, and can be merged to the public use file using the t0_rid variable. In this way, that users who spent time processing the public use file can easily link that to the training provider information if/when they obtain access to the restricted use file.

E. Low remaining risk to identification of subjects

The steps implemented and described above to anonymize the baseline data file reduce the risks to the rights and privacy of individuals but do not eliminate them. It may still be possible for a persistent intruder to use the remaining information in the data files to individually identify a respondent. In particular, in our effort to anonymize the data sets, we did not thoroughly investigate whether unique and rare *combinations* of responses to some categorical items would permit identification of individuals. As a result, there may exist some situations where

[4] These variables include the following: months of vocational education in past 5 years; reason for dropping out of vocational training; variables related to employment at the time of the survey (such as the number of jobs and hours worked); number of dependents on earnings at the time of the survey; income from sources other than job at the time of the survey; variables related to unemployment at the time of the survey (such as the duration of unemployment); and additional skills needed to find a job.

individuals could be identified using a combination of responses to several items. However, the chance of this happening is low, both because the potential benefits of doing such identification are unclear, and because this would require considerable effort on the part of an intruder. Therefore, while the risk of deductive disclosure of identity of individual subjects is not completely removed, it is reasonably low. We therefore believe the data reduction strategies implemented for the baseline data file is adequate to allow MCC to make the public use version of the survey data file available to the public.

REFERENCES

Borkum, Evan, Arif Mamun, Malik Mubeen, and Linus Marco. "Evaluation of the Vocational Training Grant Fund in Namibia: Baseline Report." Final report submitted to the Millennium Challenge Corporation. Washington, DC: Mathematica Policy Research, September 2015.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research