README for the Reproducibility Package for Business Ready (B-READY) 2024

Data Availability Statement/Data Sources

In this reproducibility package, we used two major data sources for the Business Ready (B-READY) Project:

- 21 Expert Surveys for 10 Topics: Currently these datasets are private. This reproducibility package includes a deidentified, clean version of the data in 01_Raw/03_Pub_Raw. This deidentified data is the starting input of the reproducibility package. The cleaning and deidentification codes are included, but they will not possibly run and are not sourced by the main do-file.
- Enterprise Surveys at Economy Level: Data was obtained from a custom query from the Enterprise Surveys site selecting all indicators for the economies included in the B-READY report. Data was accessed on May 31, 2024.

On specific topic scoring, we used additional data sources for standardizing units of measurement, etc. Specifically:

- List of currencies ISO codes [https://www.iso.org/iso-4217-currency-codes.html]
- Exchange rate data [https://www.xe.com/currencytables/?from=USD&date=2023-05-01#table-section]. We used exchange rate data as of May 1, 2023.
- WDI Database of GNI Per Capita [https://data.worldbank.org/indicator/NY.GNP.PCAP.CD] used to standardize the LCU based inputs.

Additionally, we used data provided by relevant Topic Teams and Front Office to make data corrections. Issues are explained and documented in the Issue Log.

For the data analytics submodule (all files included in the folder "B-READY_Analytics_Reproducibility_Package) we used these additional data sources:

- WBL2024-2-0.xlsx: WBL 2024 region and income grouping https://wbl.worldbank.org/content/dam/sites/wbl/documents/2024/WBL2024-2-0.xlsx
- Historical-data---COMPLETE-dataset-with-scores.xlsx: Doing Business Historical Data downloaded from the official archived website here: https://archive.doingbusiness.org/en/doingbusiness
- GNI_Working_Pop.dta: Selected indicators from the World Development Indicators including WBG lending status, GNI per capita, and total population. Indicators wee manually selected from the data URL: https://datatopics.worldbank.org/world-development-indicators/
- WDIEXCEL.xlsx: Bulk download from the World Development Indicators downloaded from https://datatopics.worldbank.org/world-development-indicators/ in Excel format on May 20, 2024.
- API_BX.KLT.DINV.CD.WD_DS2_en_excel_v2_1585271.xls: Foreign direct investment, net inflows data from the World Bank Website. https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD

Folder Structure

- **00_Documentation**: includes two documents
  - B-READY_Issue_Log: documenting reported errors in relevant topic datasets and used as basis for changes in data cleaning, with details including topic and sub-topic, error type, unique identifier when relevant, required actions and actions taken in 03_cleaning.do.
  - B-READY_Topic_Scoring_Rules: documenting scoring rules for each of indicators, especially the bundling rules.

- **01_Raw**: includes the raw inputs for the B-READY Project:
  - **01_Suso_Raw**: 21 expert surveys in zip files downloaded directly from the data collection platform Survey Solutions and their unzipped version in folders (not included in the reproducibility package).
  - **02_Temp_Raw**: temporary folder for storing the 21 expert survey datasets after processing the roster (if any) and correcting the documented errors (not included in the reproducibility package).
  - **03_Pub_Raw**: ready for publishing datasets after further removing personal identifiable information at expert level.
  - **04_Topic_Data**: Additional datasets either listed as external sources or produced as part of the scoring process, including:
    - [Utility Services] US_Pillar3_Fix.dta: used for correcting the pillar 3 score calculation to exclude the missing values in ES data to be aligned with the scoring design.
    - [Labor] Labor_Unit_of_measurement.dta: used to correct the unit of measurement for pillar 1 variables
    - [Labor] wdi_percapita.dta: used in labor scoring to calculate scores for two of the pillar 1 variables
    - [International Trade] List of currencies.xlsx: used in scoring one of the pillar 1 variables for the international trade topic (not included in the reproducibility package).
    - [International Trade] USD_exchange May 1 – XE site. Xlsx: used in scoring one of the pillar 1 variables for the international trade topic (not included in the reproducibility package).
  - **05_ES**: Enterprises Survey data shared by the ES team to calculate Pillar 3 Efficiency score for relevant topics.
  - **06_Note**: Included field notes shared by the topic team for additional cleaning.
    - country_names_id.dta: internal master dataset including country names and the associated individual standard code for reshaping during scoring.
    - [Business Entry, Business Location, Financial Services] GNI_Working_Pop.dta: Foreign Exchange data used to convert variables expressed in local currency units into a uniform USD measure. Furthermore, the working age population number is used in FS_CI in filtering out unrepresentative responses from credit bureaus/registries.
    - [Business Location] envp_efficiency_ed_FO.xlsx, pt_efficiency_ed_FO.xlsx: team lead made and signed off on respondent response edits to facilitate ease of scoring text entries.

- ▪ [Financial Services] FS_ST_Text_Variables_ed_1.xlsx, FS_ST_Text_Variables_edited.xlsx: team lead made and signed off on respondent response edits to facilitate ease of scoring text entries.

- **02_Template**: includes scoring template for each topic, and subfolder "reweight" included the reweighting table provided by Topic Teams to assign additional weights to the calculated indicators and scores.
    - o Note: Subfolder "_processed" is currently not in use but included the processed version when we tried to harmonize the scoring process.
    - o Subfolder "reweight" currently only includes 3 topics, the reweighting tables for other topics are included in the scoring templates as additional tabs. We can see further harmonization can be incorporated and would love to hear the advice from DIME to improve the overall efficiency of this process.

- **03_Code**: Includes all the Stata dofiles used in the data process and reflects each key step in the project.
    - o [01_unzip.do] Unzip the 21 SuSo Raw zip files and unzip all into proper folders
    - o [02_reshape.do] Reshape for topics with rosters, merge back to main surveys
    - o [03_cleaning.do] Clean variable naming, mistake values, etc. (used the issue log for documentation)
    - o [04_deidentify.do] Drop the PIIs and save to publishable datasets folder
    - o [05_aggregate.do] Loop all topics and create aggregated datasets, including the topic-specific before-aggregation processing
        - ▪ 05_aggregate_FS_CI is a special case to not aggregate FS CI data since the scoring is at individual/expert level, not economy level.
    - o [06a_ES.do] processed the ES data
    - o [06b_merge.do] merge for multi_questionnaire topics and overall with the ES dataset. Due to the data availability in Enterprise Surveys, Angola (4), Israel (81), Jamaica (83) and Jordan (85) are dropped from the results in Pilot 1.
    - o [07_Topic_Scoring] run the topic specific scoring process, commonly each topic folder would include dofiles:
        - ▪ organize scoring template and reweight tables
        - ▪ create indicators/bundles with non-median rules for Pillar 1 and 2
        - ▪ calculate Pillar 3 indicators if not using ES data
        - ▪ match with scoring template
        - ▪ calculate the scores
        - ▪ organize and save final scores
    - o [08_BReady.do] produces the overall B-READY scores.
    - o [88_AND_OR_Assessor.do] is a program for processing the logic operators used in topic scoring

- **04_Output**: includes multiple outputs from the data process. This folder populates after the code runs.
    - o **01_Aggregated**: Datasets aggregated from the expert level to economy level
        - ▪ **Topic_Aggregated**: folder with 20 sub-topic aggregated datasets
        - ▪ BREADY_Overall.dta: the main dataset at the economy level including aggregated variables from all 20 expert surveys and the ES data. This is the starting point for topic-specific scoring.

- ▪ FS_CI_aggregate.dta: special sub-topic that is scored at expert/individual level due to the scoring rules agreed upon by B-Ready leadership, therefore there is no aggregation being done at this stage.
  - o **02_Template**: 10 processed and standardized scoring templates for scoring
  - o **03_ES**: Processed ES dataset with Pillar 3 variables created.
  - o **04_Score:** Calculated economy level scores for 10 topics in Excel.
    - ▪ **Financial Services – Credit Information**: This is a special case of scoring where instead of assessing the aggregated responses, we assess the individual responses first and then aggregate by taking the higher scoring responses by country and using that as the country score.

- **05_Questionnaire**: 21 questionnaires in PDF format for each expert survey used for collecting the raw data with Survey Solutions.

- **06_Codebook**: 21 standardized codebooks using the DIME iecodebook command for each expert survey dataset.

- **B-READY_Analytics_Reproducibility_Package:** folder with the data analytics submodule of the reproducibility package. It contains the code and data to reproduce the analytical graphs and tables in the B-READY 2024 report.
  - o **data:** includes datasets:
    - ▪ 01_final_scores: scores of 10 topics in Excel. This data is the result of the code in the upper directory of the reproducibility package.
    - ▪ 02_other_data: Additional datasets used with details in previous section.
  - o **code:** Includes all the Stata dofiles used for producing multiple tables and graphs in the report.
  - o **output:** includes sub-folders for tables and graphs in the report with datasets used for analysis.
  - o Kindly note that the manuscript has gone through typesetting process, which changed the exact aesthetics of the figures and tables, but those are the same tables and figures as produced by the code.

Reproduction instructions

The materials included in the reproducibility package allow users to reproduce all indicators and exhibits, except for the indicators for the International Trade topic scores. Users can reproduce the results by following these steps:
1. Add the global path in line 86 of "03_Code/00_Master.do"
2. Add the global path in line 26 of "B-READY_Analytics_Reproducibility_Package/code/00_Master.do"
3. Run both do-files sequentially. The first main do-file reproduces the B-READY topic scores indicators. The second reproduces the analytical graphs and tables of the B-READY 2024 report.

Users who obtain access to the restricted data sources "List of Currencies ISO Codes" and "Exchange Rate Data" can additionally reproduce the indicators for the International Trade topic scores by activating the lines 234-241 of "03_Code/00_Master.do" and following the previous instructions.

The complete linkage between the code outputs (figures and tables) and the exhibits of the B-READY 2024 report is available in the reproducibility verification report of this reproducibility package. Some tables and graphs in the Executive Summary are re-used (duplicated) with the tables and graphs in the main report, the linkage is as below:

Table/Graph in Executive Summary	Duplicate Table/Graph in Main Report
Table ES.1	Table 2.1/Table 3.1
Table ES.2	Table 2.2
Figure ES.4	Figure 2.1
Figure ES.5	Figure 2.2
Figure ES.6	Figure 2.3
Figure ES.8	Figure 3.2
Figure ES.9	Figure 3.3
Figure ES.10	Figure 3.5
Figure ES.11 a/b/c	Figure 3.6 a/b/c
Figure ES.12	Figure 3.7
Figure ES.13	Figure 3.8
Figure ES.14	Figure 4.2
Figure ES.15	Figure 4.1
Figure ES.16	Figure 4.3

Match corresponding question in PDF questionnaires and scoring template

To read the question associated with a given Survey Solutions variable presented in Excel, please search the variable name in the accompanying PDF questionnaire. The variable name appears on the top right corner of each question in the PDF.

1. Does the legislation require entrepreneurs to verify and to have the company name approved in order to register a new business?

SINGLE-SELECT
01 ◯ Yes
02 ◯ No

Variable Name BE_1_1_1

For questions that are designated as multi-select in the PDF questionnaire, note that the variable names in Excel are the combination of the name on the top right corner and the corresponding option number, generated by Survey Solutions as "var_name__option" with a **double underscore "__"** in between.

For example, for the question below, you will find 5 variables in the corresponding Excel representing each of the multi-select choices: BE_1_2_2__1, BE_1_2_2__2, BE_1_2_2__3, BE_1_2_2__4, BE_1_2_2__88.

8. Does the legislation require companies to register the following type of information about [beneficial owners](BE_1_2_2 "The natural person who ultimately owns or controls a company, even if the title to the property is under another name.")?

Please select all that apply
BE_1_2_1==1

MULTI-SELECT: YES/NO   BE_1_2_2

01   ☐/☐ Full name

02   ☐/☐ Identification number
           and information

03   ☐/☐ Date of birth

04   ☐/☐ Address

88   ☐/☐ Other

Variable names and expert-level responses updates

For some variables, their Survey Solutions variable names were updated in subsequent data processing to rectify naming errors in questionnaire prefixes or suffixes to ensure naming consistency. Those cases are listed below by topic.

Utility Services: the topic contains 3 questionnaires with corresponding Excels. The naming updates made are as follows.

- **Electricity**
    o Removed additional underscore: all variable names starting with "U_E" are renamed as "UE" for consistency. For example, U_E_2_3_2__1 from the PDF questionnaire will be found as UE_2_3_2__1 in corresponding Excel.
    o Variable UW_5_1_2c was corrected as UE_5_1_2c.

- **Water**
    o Removed additional underscore: All variable names starting with "U_W" are renamed as "UW" for consistency. For example, U_W_3_1_8__1 from the PDF questionnaire will be found as UW_3_1_8__1 in corresponding Excel.

- **Internet**
    o Removed additional underscore: All variable names starting with "U_I" are renamed as "UI" for consistency. For example, U_I_5_1_7 from the PDF questionnaire will be found as UI_5_1_7 in corresponding Excel.
    o Corrected variables generated for multi-select question from UE_5_2_2 to the correct UI_5_2_2. For example, variable named UE_5_2_2__1 in the Internet Questionnaire will be found as UI_5_2_2__1 in the corresponding Excel.

For Labor, variables with names starting from L1_4, L1_38, L1_41, L1_44, L1_48, L2_15 and L2_17, and ending in a lower-case letter a/b/c/d/e were corrected with an additional single underscore "_" before the last letter. For example, L1_4a from the PDF questionnaire will be found as L1_4_a in the corresponding Excel.

For **Financial Services – Operation of Credit Bureau and Registry**, depending on the reporting entity (i.e., credit bureau or credit registry), questions prefixed with FS5_* (for bureaus) or FS6_* (for registries) are presented in the questionnaire. These two sets of questions are identical, thus, economy-level answers from all credit registries are merged under FS5_* variables in the Excel file.

For **International Trade**, to distinguish variable names across the three questionnaires (Regulations in Goods; Regulation in Services; Public Services), prefixes for all variables in Regulations in Goods are systematically updated to TRG_*; prefixes for all variables in Regulations in Services are systematically updated to TRS_*; prefixes for all variables in Public Services remain as IT_*. To locate the corresponding question in the PDF questionnaires, please revert the prefixes to IT_* for all variables.

Additionally in International Trade, for variable TRG_1_3_2_1 (i.e., IT_1_3_2_1 in the Excel) of the Regulations in Goods questionnaire and variables IT_2_1_1_2a; IT_2_1_1_3a; IT_2_1_1_4a; IT_2_1_1_5a of the Public Services questionnaire, expert-level responses were updated to binary yes/no depending on the existence of certain features.

For **Dispute Resolution**, for variables DR_B_1_4; DR_B_1_5; DR_B_1_7; DR_B_2_1; DR_B_2_2; DR_B_2_3; DR_B_2_4; DR_B_2_5; DR_B_3_2; DR_B_4_1; DR_B_4_2; DR_B_5_5; DR_D_2_2; DR_E_1_1; DR_E_1_4; DR_F_1_1; DR_F_1_2; DR_F_1_3; DR_F_1_4; DR_F_1_5; DR_F_1_6; DR_F_1_7; DR_F_1_8; DR_F_1_9; DR_F_2_1; DR_F_2_2; DR_F_2_3; DR_F_2_4; DR_F_2_5; DR_F_2_6; DR_F_2_7; DR_F_2_8, expert-level responses were updated to binary yes/no depending on the existence of certain features.

For Market Competition – Procurement, depending on whether an e-procurement platform is centralized or non-centralized, two sets of identical questions (rosters MCp_7_r and MCp_10_r) are presented in the questionnaire. These questions are first appended at expert level before being merged back to generate the economy-level answer under a new prefix of MCp_7_10_r.

Find Scored Variables and their Scoring Rules

All the scored Survey Solution variables are documented in the corresponding Scoring Template column Suso_Var. For example, question BE_1_1_1 in Business Entry is used to construct 01BE_Indic1 in the Scoring Template together with BE_1_1_2. They are linked by the "AND" logic, which means both variables must be both "Yes" after aggregation to be scored as 1 point for Social Benefit as design.

Special labels for economy-level aggregated answers

Economy-level answers, in standard cases, can be yes, no or a numeric value depending on the question type. Additionally, economy-level answers can be of special value in some scenarios, and these special values are explained below.

For binary and multi-select questions, special cases may include:

- **0.5**: economy-level answer returns "**0.5**" when responses at expert-level have equal number of "yes" and "no".
- **Yes; 50% skip logic (-8)**: economy-level answer returns "**Yes; 50% skip logic**" when half of the experts from a given economy skipped the question, and the other half responded with a "yes".
- **No; 50% skip logic (-7)**: economy-level answer returns "**No; 50% skip logic**" when half of the experts from a given economy skipped the question, and the other half responded with a "no".
- **Even split; 50% skip logic (-6)**: economy-level answer returns "**Even split; 50% skip logic**" when half of the experts from a given economy skipped the question, and the other half responded with equal number of "yes" and "no".

- **Skip logic (-9)**: economy-level answer returns "Skip logic" when more than half of the experts from a given economy skipped the question.

For categorical single-select or continuous (numeric) questions, special cases may include:

- **Single mode; 50% skip logic (-6)**: economy-level answer returns "**Single mode; 50% skip logic**" when half of the experts from a given economy skipped the question, and the other half responded with a single answer option. The option value is retained in an additional variable starting with "v_", followed by the variable.
- **Multiple modes (-8)**: economy-level answer returns **"Multiple modes"** when more than half of the experts from a given economy responded with more than one overlapping answer option OR non-overlapping answer options. The multiple option values are retained in an additional variable starting with a "v_", followed by the variable name.
- **Multiple modes; 50% skip logic (-7)**: economy-level answer returns "**Multiple modes; 50% skip logic**" when half of the experts from a given economy skipped the question, and the other responded with non-overlapping answer options. The multiple option values are retained in an additional variable starting with a "v_", followed by the variable name.
- **Skip logic (-9)**: economy-level answer returns **"Skip logic"** when more than half of the experts from a given economy skipped the question.