



McKinsey & Company

MCKINSEY CENTER FOR GOVERNMENT

# GOVERNMENT PRODUCTIVITY

UNLOCKING THE \$3.5 TRILLION OPPORTUNITY  
*Technical appendix*

DISCUSSION PAPER  
APRIL 2017

**The McKinsey Center for Government (MCG)** is a global hub for research, collaboration, and innovation in government productivity and performance. MCG is part of McKinsey's Public and Social Sector Practice, whose mission is to help governments and non-governmental institutions improve the lives of citizens worldwide and to help solve the world's most pressing economic and social issues.

MCG provides government leaders with insights, new approaches, benchmarks, and connections to help them improve the lives of the citizens they serve, within the fiscal constraints they face. We focus on the critical and common challenges that governments around the world face and create opportunities for government leaders to learn from successful experience, evidence-based innovation, and different contexts. MCG aspires to play a global role in diffusing best practices across governments around the world. Our unique set of global perspectives and best practices is accessible through our research, knowledge, publications, and tools. We also offer a forum of experts around the world from various backgrounds, where exciting new ideas and innovative models for government can be debated, codified, and shared.

The work of McKinsey's Public and Social Sector Practice spans economic development, education, health care, public finance, infrastructure, operations and delivery, and defense and security. We work with the world's leading public-sector organizations to address major challenges such as macroeconomic vulnerability, population aging, urbanization, and digitization of government services. In the social sector, we support the world's leading institutions in developing and scaling up solutions to major social challenges. We forge cross-sector partnerships among government bodies, agencies, foundations, donors, and businesses, and we work on the ground alongside our clients to achieve real impact in the communities of greatest need. McKinsey has served more than 250 government agencies and hundreds of social sector institutions across 110 countries and has completed more than 4,500 engagements in the past five years.

The partners of McKinsey & Company fund MCG's research; it is not commissioned by any business, government, or other institution. For further information about MCG and to download our publications, please visit [www.mckinsey.com/mcg](http://www.mckinsey.com/mcg). We welcome comments: please email us at [mcg@mckinsey.com](mailto:mcg@mckinsey.com).

This publication is written by experts and practitioners in McKinsey & Company's Center for Government along with other McKinsey colleagues.

This publication is not intended to be used as the basis for trading in the shares of any company or for undertaking any other complex or significant financial transaction without consulting appropriate professional advisers.

No part of this publication may be copied or redistributed in any form without the prior written consent of McKinsey & Company.



© Monty Rakusen/Getty Images

# TECHNICAL APPENDIX

## KEY TERMS AND METRICS USED

In *Government productivity: Unlocking the \$3.5 trillion opportunity*, we repeatedly use specific terms to explain our results and observations. In addition, for each of the sectors we use a set of metrics to analyze and compare countries. In this section, we define what we mean by the most-relevant terms, as well as how and why we chose the metrics that appear in our study.

### Overall metrics

We use a common set of metrics across all the sectors studied in the Government Productivity Scope (GPS) analysis. Exhibit A1 provides a definition of basic terms we use to analyze and describe the results across sectors, while Exhibit A2 sets out the specific metrics we use for each of the seven sectors in our study. We chose these metrics based on relevance, data availability, and data quality across the countries covered.

#### Exhibit A1

##### Key elements of the McKinsey Center for Government GPS analysis

Metric	Definition
Input	Government expenditure in sector (opex and capex, national and local) plus private expenditure unless otherwise indicated
Output	Volume-based indicator of coverage (e.g., number of citizens, number of students, number of passengers per kilometer)
Cost per unit (efficiency)	Input divided by output, unadjusted for differences in the outcome (e.g., dollars spent per student, without any adjustment for differences in PISA scores); <sup>1</sup> unit cost is used to measure efficiency (i.e., a country with a lower cost per unit is described as more efficient)
Outcome (effectiveness)	Quality-based indicator (e.g., PISA scores, healthy life expectancy); outcomes; used to measure effectiveness (i.e., a country with a higher outcome is described as more effective)

<sup>1</sup> See technical notes for adjustments in cost per unit metrics.

SOURCE: McKinsey Center for Government GPS analysis

Exhibit A2

**Metrics and definitions by sector**

Sector	Type of metric	Metric	Definition
<b>Health care</b> 	Input	Total (private and public) expenditure on health care and public health, all functions	Final consumption of health-care goods and services, both government and privately financed
	Output	Total population	Midyear estimate of total number of residents, regardless of their legal status or citizenship
	Outcome	Healthy life expectancy	Number of healthy years an individual is expected to live at birth by subtracting the years of ill health—weighted according to severity—from life expectancy
<b>Primary education</b> 	Input	Total (private and public) expenditure on primary education	Expenditure on primary schools and other public and private educational institutions
	Output	Number of students enrolled	Total number of students enrolled in primary education
	Outcome	PISA score (average of reading, math, and science)	Programme for International Student Assessment result—math, reading, and science simple average
<b>Secondary education</b> 	Input	Total (private and public) expenditure on secondary education	Expenditure on secondary schools and other public and private educational institutions
	Output	Number of students enrolled	Total number of students enrolled in secondary education
	Outcome	PISA score (average of reading, math, and science)	Programme for International Student Assessment result—math, reading, and science simple average
<b>Tertiary education</b> 	Input	Total (private and public) expenditure on tertiary education	Expenditure on the highest level of education, covering private expenditure on schools, universities, and other private institutions, excluding R&D
	Output	Number of students enrolled	Total number of students enrolled in tertiary education
	Outcome	Composite metric based on graduation rate (33%), teaching quality (33%), income premium (17%), and employment premium (17%)	Composite metric using: <ul style="list-style-type: none"> <li>▪ Graduation rate (33% weight): number of graduates divided by number of enrolled students</li> <li>▪ Quality of teaching (33% weight): composite metric calculated by <i>Times Higher Education</i> that includes Academic Reputation Survey, staff-to-student ratio, doctorate-to-bachelor's ratio, doctorates-awarded-to-academic-staff ratio, and institutional income</li> <li>▪ Income premium of graduates aged 25–34 compared with those with only a secondary school qualification (17% weight): percentage difference between earnings of tertiary education graduates and those with only secondary education, including all earners</li> <li>▪ Employment premium of graduates aged 15–64 compared with those with only a secondary school qualification (17% weight): employment percentage difference between tertiary education graduates and those with only secondary education</li> </ul>

Sector	Type of metric	Metric	Definition
<b>Public safety</b> 	Input	Government expenditure on public safety (police, prisons, law courts, and fire safety)	Total public expenditure on police, prisons, law courts, and fire safety
	Output	Total population	Midyear estimate of total number of residents, regardless of their legal status or citizenship
	Outcome	Composite metric based on homicide rate (33% weight), confidence in police (17%), confidence in judiciary (17%), and feeling safe walking alone (33%)	Composite metric using: <ul style="list-style-type: none"> <li>▪ Homicide rate (33% weight): homicides per 100,000 inhabitants</li> <li>▪ Confidence in police (17%): survey measure. Percent of people who answered “yes” to “in the city or area where you live, do you have confidence in the local police force or not?”</li> <li>▪ Confidence in judiciary (17%): survey measure. Percent of people who answered “yes” to “in this country, do you have confidence in the judicial system and the courts?”</li> <li>▪ Feeling safe walking alone (33%): survey measure. Percent of people who answered “yes” to “do you feel safe walking alone at night in the city or area where you live?”</li> </ul>
<b>Road transport</b> 	Input	Government expenditure on roads (government agencies and state-owned enterprises)	Total public expenditure on road, including government, state, and local spending, as well as that by SOEs
	Output	Pkme (incorporates both freight and passenger movements)	Total number of yearly road passenger kilometers and freight-tonne kilometers traveled (converted to passenger kilometer equivalent, or pkme)
	Outcome	World Economic Forum quality of road metric	Survey measure. Answer (1–7 scale) to “how would you assess roads in your country?”
<b>Tax collection</b> 	Input	Overall expenditure for tax functions and related overhead	Total public expenditure on tax collection, including all public administrations and overhead costs
	Output	Total population (as a proxy for number of taxpayers)	Midyear estimate of total number of residents, regardless of their legal status or citizenship
	Outcome	Tax collection effectiveness (1 minus tax evasion over GDP)	Amount of tax evaded from legal activities over GDP, subtracted from 1

SOURCE: McKinsey Center for Government GPS analysis

## SCOPE OF GOVERNMENT PRODUCTIVITY ASSESSMENT

### Sectors included in the analysis

We divide sectors into addressable and non-addressable. We define as addressable those sectors in which countries—and more specifically, governments—can make meaningful changes in the relatively short term. Non-addressable sectors, on the other hand, are either based on long-term commitments (for example, pensions or debt service) or used as levers for redistribution of income in pursuit of complex social goals such as fairness (for example, social spending). Additionally, we define addressable sectors as those sectors in which we can designate a meaningful outcome variable that is arguably related to expenditure. For example, health-care or education expenditure is primarily aimed at improving health or education outcomes, respectively. In contrast, citizens use expenditure on pensions

and unemployment benefits for many different purposes, so it is difficult to establish a connection between these types of expenditure and any specific outcomes.

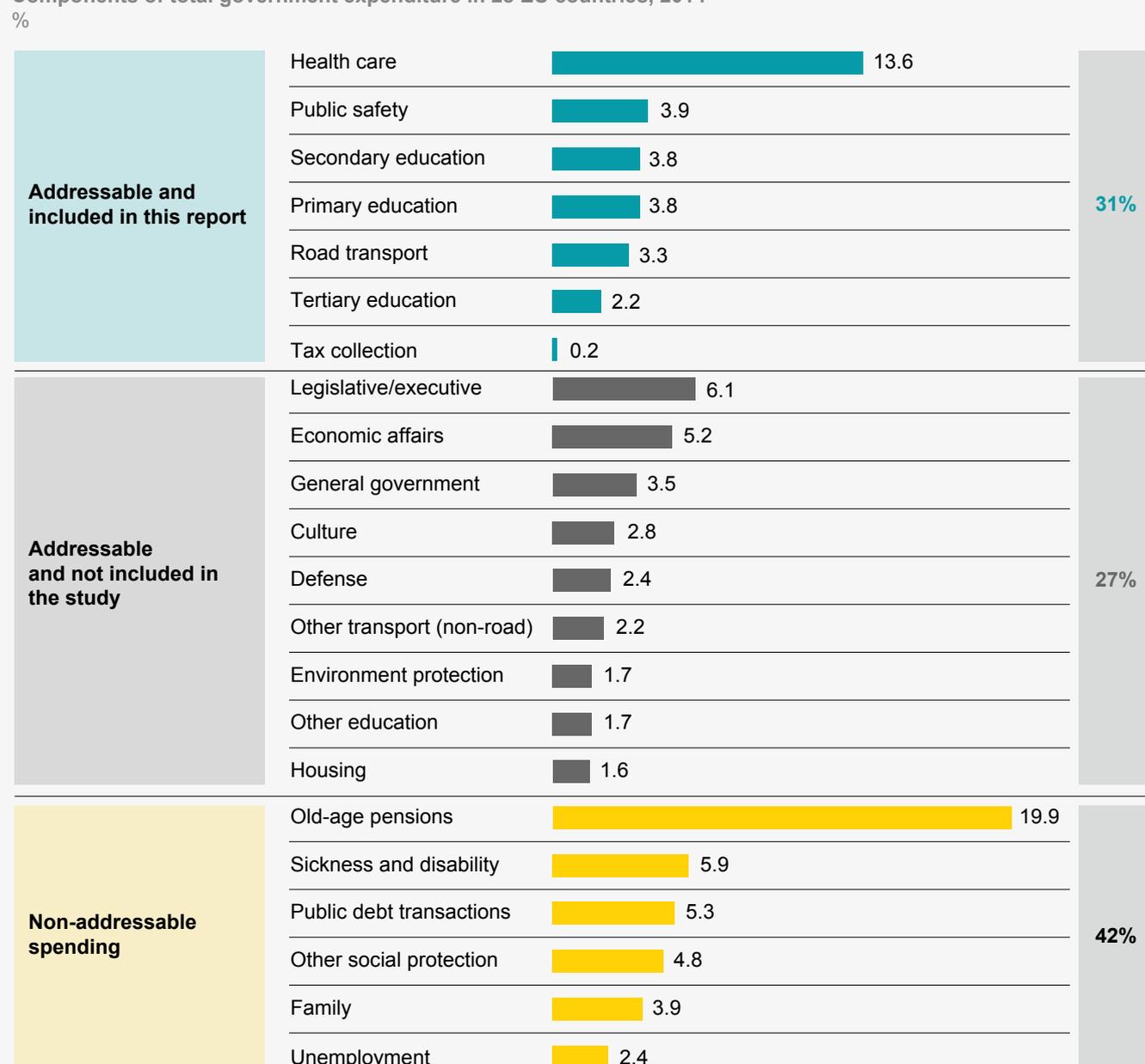
Among addressable sectors, we include only some sectors—based on the size of the sector, the existence of meaningful output and outcome variables, and data availability. We aim to increase the sector coverage of our analysis in future iterations of this work.

Exhibit A3 shows the sectors currently included in the GPS analysis and their respective shares of government expenditure.

### Exhibit A3

#### Scope of sectors covered and not covered in the GPS analysis

##### Components of total government expenditure in 28 EU countries, 2014



SOURCE: Eurostat

## Time lags, country coverage by sector, and data sources

Exhibit A4 provides an overview of metrics, time lags, and key data sources for metrics used for benchmarking across each sector. For a discussion of specific metrics, refer to the “Sector-specific notes” section in the technical notes.

### Exhibit A4

#### Time lags, country coverage, and data sources for metrics by sector

Sector	Metric	Description	Lag from input	Number of countries	Data source <sup>1</sup>
<b>Health care</b> 	Input	Total (private and public) expenditure on health care and public health, all functions		42	Organisation for Economic Co-operation and Development (OECD)
	Output	Total population	0 years		World Bank
	Outcome	Healthy life expectancy	1 year		Euromonitor, based on the World Health Organization
<b>Primary education</b> 	Input	Total (private and public) expenditure on primary education		35	OECD
	Output	Number of students enrolled	0 years		OECD and World Bank
	Outcome	PISA score (average of reading, math, and science)	5 years		OECD
<b>Secondary education</b> 	Input	Total (private and public) expenditure on secondary education		33	OECD
	Output	Number of students enrolled	0 years		OECD and World Bank
	Outcome	PISA score (average of reading, math, and science)	1 year		OECD
<b>Tertiary education</b> 	Input	Total (private and public) expenditure on tertiary education		28	OECD
	Output	Number of students enrolled	0 years		UNESCO
	Outcome	Composite metric based on graduation rate (33%), teaching quality (33%), employment premium (17%), and income premium (17%)	5 years		OECD, <i>Times Higher Education</i> , UNESCO, World Bank
<b>Public safety</b> 	Input	Government expenditure on public and safety (police, prisons, law courts, and fire safety)		36	International Monetary Fund (IMF)
	Output	Total population	0 years		World Bank
	Outcome	Composite metric based on homicide rate (33% weight), confidence in police (17%), confidence in judiciary (17%), and feeling safe walking alone (33%)	3 years		Gallup and United Nations Office on Drugs and Crime
<b>Road transport</b> 	Input	Government expenditure on road (government agencies and state-owned enterprises)		26	IMF and OECD
	Output	Passenger km equivalent (incorporates both freight and passenger movements)	3 years		OECD
	Outcome	World Economic Forum quality of road metric	3 years		World Economic Forum
<b>Tax collection</b> 	Input	Overall expenditure for tax functions and related overhead		28	OECD
	Output	Total population (as a proxy for number of taxpayers)	0 years		World Bank
	Outcome	Tax collection effectiveness (1 minus tax evasion over GDP)	1 year		Buehn and Schneider (2012) and World Bank

<sup>1</sup> The listed source is the main data source. We use complementary sources (e.g., Eurostat, local sources) to transform existing data (e.g., split capex-opex) and to supplement the primary data sets with missing data.

SOURCE: McKinsey Center for Government GPS analysis

Beyond the main sector metrics, we use other metrics and variables throughout the report. The following is a list of these metrics, in order of source:

- *Edelman Trust Barometer*: trust in business, trust in government
- *Global Terrorism Database*: terrorist attacks
- *IHS Markit*: gross domestic product (GDP)
- *International Monetary Fund (IMF)*: government revenue, government expenditure, government fiscal balance, government debt, wages, government expenditure on social benefits
- *International Labor Organization*: public employment (absolute)
- *Organisation for Economic Co-operation and Development (OECD)*: proportion of public vs. private expenditure in different sectors, public employment (percent)
- *United Nations*: demographic structure, population forecasts
- *World Bank*: population size, purchasing power parity (PPP) conversion factors, GDP deflator
- *World Health Organization (WHO)*: public expenditure on health

### Countries included in the analysis

Our full sample contains 42 countries—including all OECD and G20 countries except for Argentina and Saudi Arabia, due to data-availability constraints. The only included country that does not belong to either of these two groups is Singapore, which is included thanks to its reliable data availability. Altogether, these 42 countries account for roughly 80 percent of the world's GDP. As shown in Exhibit A4 above, our analysis does not cover all 42 of these countries in every sector, due to data-availability constraints.



## TECHNICAL NOTES

### Methodological choices

#### Time periods and lags

Our snapshot and productivity-improvement analyses are based on the most-recent periods possible given data-availability constraints. In our productivity-improvement analysis, we consider changes in efficiency and effectiveness over five-year periods in all sectors except for primary and secondary education, for which we examine a six-year period because students take Programme for International Student Assessment (PISA) tests every three years. Exhibit A5 shows the date ranges used and explains the rationale behind date and lag choices in every sector.

Exhibit A5

#### Time periods for analysis

Sector	Snapshot analysis		Productivity-improvement score range <sup>2</sup>	Rationale
	Input and output year	Outcome year		
 <b>Health care</b>	2014	2015	2009 to 2014	The one-year lag between input and outcomes reflects the way healthy life expectancy (HLE) is estimated, taking into account current rates of mortality and morbidity in the population <sup>3</sup>
 <b>Primary education</b>	2010	2015	2004 to 2010	The five-year lag reflects the fact that PISA scores are taken around five years after the last year of primary education; the latest PISA scores are from 2015—therefore our period of analysis in spending stops at 2010
 <b>Secondary education</b>	2014	2015	2008 to 2014	In most countries, students are still in secondary education when they take the PISA, so we apply a one-year lag to account for the fact that expenditure may not happen at the beginning of a particular year and to correct for some potential statistical biases
 <b>Tertiary education</b>	2010	2015	2005 to 2010	Since the composite outcome metric contains income and employment premiums, which take time to react to spending, we use a five-year lag between spending and outcomes
 <b>Public safety</b>	2012	2015	2007 to 2012	Based on expert interviews and literature reviews, changes in police systems and prisons take, on average, three years to significantly affect outcomes
 <b>Road transport<sup>1</sup></b>	2010 <sup>1</sup>	2013	2005 to 2010	The three-year lag between spending and output is the average time it takes for major road projects to be completed; the latest year with good country coverage of outputs (pkme) is 2013
 <b>Tax authority</b>	2009	2010	2004 to 2009	The latest year available for tax collection effectiveness data is 2020; we include a one-year lag between spending and outcomes to account for the fact that expenditure may not happen at the beginning of a particular year and to correct for some potential statistical biases

<sup>1</sup> In transport, the lag is applied between input and output instead of output and outcome. So 2013 is the year of output as well as outcome.

<sup>2</sup> Defined by input years.

<sup>3</sup> The literature often uses no lags, and there is no clear consensus on the appropriate number. In any case, the results are not very sensitive to changes in this parameter.

SOURCE: McKinsey Center for Government GPS analysis

## Public vs. private expenditure

As a general rule, we include both public and private expenditure in our analysis because outcomes are a result of both types of spending and because it is not possible to distinguish between outcomes that come from public spending and those that are a result of private spending. Moreover, expenditure is mostly public—for example, within our sample, public expenditure represents on average more than 90 percent of the expenditure in primary and secondary education and 70 percent of that in tertiary education and health care.

In public safety, road transport, and tax collection, we exclude private expenditure for the following reasons:

- *Public safety:* There is no standard international definition of “private security,” nor is there a centralized database with comparable data across countries and years. Moreover, private security often includes types of expenditure such as corporate security guards, security clearance systems, and alarm systems, each of which generally have little to do with the outcomes we measure (homicide rate, confidence in police, confidence in judiciary, and feeling safe walking alone).
- *Road transport:* While part of the expenditure in this sector may be private, it is very complex to account for private spending in a consistent way across countries, since all systems are different and account for private expenditure in different ways. For example, in some countries all revenue collection and expenditure is carried out by private companies; in others, private companies receive public subsidies.
- *Tax collection:* There is no private expenditure on tax collection.

## Capital expenditure vs. operating expenditure

We include both capital expenditure (capex) and operating expenditure (opex) in all sectors except tax collection, where capex is small and data on the ratio of capex to opex is very scarce.

For all sectors, we split capex and opex and treat them differently. For opex, we apply it all to the year in which the cash flow took place—that is, if the amount was paid in 2005, it will count as a 2005 expenditure. Regarding capex, we spread it across 20 years, which we take as an average period over which capital is depreciated and has an impact on outcomes. For example, if there was a fixed capital investment in 2005, we put only one-twentieth of the amount in the 2005 expenditure, and we spread the rest equally over the next 19 years.

## Extrapolation of missing data

Where data is missing at the beginning of a time series, we take the earliest year for which we have data and extrapolate backward by a maximum of five years at a constant level. Similarly, where data is missing at the end of a time series, we extrapolate forward by up to five years at a constant level. Where data is missing in intermediate years, we linearly interpolate.

## Cross-section and time-series comparability

To make expenditure data comparable across countries (cross-section) and time (time series), we express all data in Part I of the report in 2010 dollars at PPP. This means that we have converted all currencies using PPP exchange rates, and changes over time are not affected by inflation.

## Methodological limitations

As mentioned in the discussion paper, this analysis is a first step in measuring government productivity and does not attempt to provide final answers or definitive judgments on productivity levels or changes in them. Rather, the analysis aims to identify broad trends and raise relevant questions with the objective of helping countries improve, mostly through learning from best-practice peers. Hence, we acknowledge a number of limitations in our methodology.

First, our choice of metrics is based on a combination of relevance and the availability of cross-country and time-series data. Therefore, the metrics we use cover some of the most-relevant aspects within each sector but necessarily leave out some other aspects—especially outcomes—which may also be valuable to citizens.

Second, we have not attempted to put an absolute value on efficiency or effectiveness measures, meaning that we cannot assert a country is *efficient* or *effective* per se but simply *more or less efficient or effective* than comparable countries. All statements about productivity, therefore, are relative, and we make no absolute judgments.

Third, we do not correct for structural factors that may affect inputs, outputs, or outcomes, for example, cultural aspects, diet, habits, political systems, genetic factors, population density, and country size. Thus, we make direct comparisons with care, especially those direct comparisons that are static. To partially correct for this bias, we compare similar countries and add GPS improvement scores, which rely on changes over time.

Finally, we measure cost per unit and outcomes, but we do not formally test nor do we imply direct causality between these two sets of metrics. We acknowledge that other factors beyond expenditure—such as the ones outlined above—may influence outcomes.

## Sector-specific notes

In several of the sectors we analyze, we adopt specific elements of the methodology, which we note below.

### Primary and secondary education

To reflect the fact that students spend several years in both primary and secondary education, we use rolling averages of spending over a five-year period.

### Tertiary education

To reflect the fact that students spend around three years in tertiary education, we use rolling averages of spending over a three-year period to obtain average expenditure per student over the whole period they were in tertiary education.

We calculate the quality of teaching using a metric (“teaching score”) from *Times Higher Education*, where all tertiary education institutions receive a teaching score between one and 100 depending on their teaching quality. To obtain a quality-of-teaching number for each country, we take the simple average of the best institutions in the country and choose the number of institutions based on the country’s population. We use the following process:

1. We establish the number of institutions that we will include in each country’s average—expected number (E)—allotting one institution per 250,000 students. For instance, the quality-of-teaching score of a country with 500,000 enrolled students will be the average score of its two best institutions, while that of a country with 10,000,000 students will be the average of its top 40 institutions.
2. We check the number of actual institutions (A) per country included in the *Times Higher Education* database.
3. In most cases,  $A > E$ , meaning that the number of institutions covered by the database is enough to calculate the quality-of-teaching score per country. In these cases, a country’s quality of teaching score will be the simple average of the scores achieved by its E top universities. In the few cases in which  $A < E$ , we assume that the remainder of institutions ( $E - A$ ) in a particular country has the lowest score that a university of that country has ever had. We then calculate a weighted average of the two components, A and ( $E - A$ ). For example, if a given country has 2,500,000 enrolled students,  $E = 10$ . Then, let’s imagine that  $A = 8$ , the average score of the 8 included is 30, and the country’s lowest-scoring university ever has a score of 10. This country will have a final score of 26. The calculation is the following:  $(0.8 \times 30) + (0.2 \times 10) = 26$ .
4. We only apply this estimation methodology if A is at least 20 percent of E. For example, if a country’s average should be made of 10 universities ( $E = 10$ ) given the number of enrolled students it has, but it only has one university included in the ranking ( $A = 1$ ), the country is not included in our analysis since  $1 \div 10 = 10$  percent, which is less than 20 percent.

## Road transport

On the output side, we look at passenger kilometers<sup>1</sup> and freight-tonne kilometers<sup>2</sup> traveled over the course of a year in each country. This approach takes into account not only the size of transport networks but also how extensively they’re used—thus we evaluate a country’s transport system as more efficient and productive if it has higher load factors. To compare countries with different mixes of freight and passenger traffic, we combine the two outputs into a single metric: passenger kilometer equivalent (pkme). We use energy consumption (in megajoules) per passenger kilometers and tonne kilometers to convert freight-tonne kilometers into pkme, obtaining a 1:5 equivalence rate. For example, a road system is considered to produce the same output by moving one passenger one kilometer or five tonnes of freight one kilometer.<sup>3</sup>

---

<sup>1</sup> A passenger kilometer represents the movement of a passenger across the distance of one kilometer. Two passenger kilometers could be the result of one passenger traveling two kilometers or two passengers traveling one kilometer each.

<sup>2</sup> A tonne kilometer represents the movement of one tonne of freight across a one kilometer distance.

<sup>3</sup> Based on OECD estimates of energy intensity for road systems.

We use the World Economic Forum (WEF)'s "quality of road infrastructure" survey as our outcome indicator. It is calculated as part of their *Global Competitiveness Report*. The WEF constructs this measure by asking residents "how would you assess roads in your country?" Respondents answer on a scale of one to seven, with one being "extremely underdeveloped" and seven being "extensive and efficient by international standards." In order to have a more stable metric, we use a three-year rolling average.

### Tax collection

We use tax-collection effectiveness as our outcome (that is, an indicator of the quality of the tax system). We define tax-collection effectiveness as (1 minus tax evasion), and we define tax evasion as the percentage of tax not collected by governments over GDP, including economically legal but illegally hidden activities and excluding illegal activities. To calculate tax-collection effectiveness, we use tax evasion over GDP time series data from Buehn and Schneider (2012). We then calculate (1 minus tax evasion) with the rest of sectors for consistency purposes (so a higher outcome means greater efficiency).

### The GPS improvement score

The objective of the GPS improvement score is to compare countries' trajectory in productivity (both in terms of effectiveness and efficiency) across time. We give each country a number, or score, which we use to compare it to other countries—especially those in each country's peer group.

The calculation process consists of several steps, as follows:

- For each country, we look at the change in both cost per unit (efficiency) and outcomes (effectiveness) between two particular years. For example, for health care, we look at the increase in spending per person from 2009 to 2014 and in healthy life expectancy (HLE) from 2010 to 2015 since we use a one-year lag between inputs (expenditure) and outcomes (HLE). See Exhibit A5 for the time ranges used for each sector.
- We calculate the median and the standard deviation of the change in cost per unit across all countries (in health care we cover 42 countries, so there are 42 changes). We do the same for outcomes.
- For each country, we calculate how many standard deviations away from the median the country is—again, both for cost per unit and for outcomes. For example, consider outcomes. Let us say that the median increase in HLE from 2010 to 2015 was one year, and the standard deviation was 0.3. If Country A increased its HLE in this period by 1.6 years, then its HLE increased by 2 standard deviations above the median. In our scoring system, this gives Country A 2 points for its productivity-improvement score. Now let us say that Country A increased spending per person by an amount that equates to 0.4 more standard deviations compared with the median. For this, Country A receives a negative 0.4 points in our scoring system. (If it had reduced spending per person by 0.4 standard deviations more than the median, since spending less means higher efficiency, it would get a positive 0.4 points added to its score.)
- We add the two scores, effectiveness (2) and efficiency (−0.4) to arrive at Country A's overall productivity-improvement "raw" score: 1.6.

- Since this methodology necessarily leads to many negative scores, we add +2.5 to the effectiveness score and +2.5 to the efficiency score of all countries. This addition has no effect on the relative results but results in scores that are always positive and easier to compare.
- To finish with the example, the final score for Country A in health care would be:  $2.5 + 2 + 2.5 - 0.4 = 6.6$ .

### Sizing the prize

To estimate the potential benefits of countries learning from best practice and achieving similar results to the best improvers in their peer group, we follow a systematic approach to “size the prize” based on GPS improvement scores. The final number we estimate answers the question “how much could governments save per year by 2021 if they adopted best practices and thus were able to improve their scores as much as the best improvers in their peer groups?”

The calculation process consists of the following nine steps:

1. For each sector, we classify all countries in quartiles by outcome, creating peer groups. This approach means that a country can be in different quartiles and compared to different countries for different sectors (for example, a country can be in the first quartile in health care and in the third quartile in public safety).
2. In each sector’s quartile, we find the best improver, namely the country with the highest GPS improvement score.
3. We then take each country and calculate the difference between its GPS improvement score and the best improver’s score. This number is the starting point for estimating the potential improvement for each country. For example, if Country A has a score of 2 and the best in its quartile has a score of 5, we conclude that Country A can, at best, improve by 3 points (that is, standard deviations).
4. Since each point in the score is a standard deviation and we know the percentage of expenditure each standard deviation represents, we can then calculate the country’s potential savings if it were to improve at the same rate as the best country in its peer group. For instance, following the previous example, if one standard deviation represents a 5 percent difference in cost per unit, it means that Country A could reduce its final cost per unit by 15 percent. Note that in this calculation, we assume that countries use all the improvement potential to reduce expenditure rather than to drive improved outcomes.
5. However, we do not allow for any level of savings: we cap potential savings by the greatest reduction in savings in the whole sample, excluding outliers. For instance, if the country with the greatest spending reduction exhibited a savings of 10 percent and Country A has an initial cost per unit (input/output) of 10, our methodology does not allow Country A to achieve a cost per unit lower than 9—that is, it will not be allowed to reduce its cost per unit by more than 10 percent from its initial point. When selecting the country with a greatest reduction—namely the country that will establish the

percentage cap—we exclude countries that reduced spending due to extraordinary circumstances (for example, a deep recession that caused a default) or that saw a significant deterioration of outcomes following spending cuts.

6. We cap not only the change in cost per unit but also its level—that is, we cap the potential cost per unit after savings of each country to the lowest level of cost per unit in its quartile. Following our previous example, the cap on the change in cost per unit did not allow Country A to go lower than 9. Once this new constraint is added, if the country with the lowest final level of spending within Country A's quartile has a cost per unit of 9.5, Country A will be capped at that level, achieving lower savings than the original 10 percent. Conceptually, what this cap does is make sure that no country can reduce spending to a lower level than the most efficient country among its peers.
7. Once we get a result for the seven sectors and all the countries included in our sample, we scale it up to all countries worldwide using government expenditure. For example, if the countries for which we have data in a particular sector cover 60 percent of the world total expenditure, and we estimate that their savings in that particular sector could amount to \$500 billion, we will assume that the countries that are not included will be able to save the same average proportion, amounting to a total of \$833 billion in savings worldwide.
8. We also scale up the result to what we define as addressable but excluded sectors (legislative/executive, economic affairs, general government, culture, defense, other transport, environment protection, other education, and housing), assuming that the saving percentage achieved in included sectors can be replicated in non-included sectors. We assume no savings in pensions, sickness and disability payments, interest payments, social protection, family transfers, or unemployment benefits, given that we have classified these items as non-addressable for the purposes of this study.
9. Finally, once we determine which countries could have saved money in the analyzed period, we estimate how much countries could save from 2016 to 2021 and the impact this would have on government budgets and the fiscal balance by 2021. To do so, we take IMF future estimations of GDP, government expenditure, and fiscal balances, and we assume that in the next five years countries should be able to replicate the proportion of savings they could have achieved in the past five years had they saved at the rate of the best improver in their peer group. ■

