Science, Technology & Innovation Indicators

Thematic Paper 1

Challenges in measuring the efficiency of national science, technology & innovation systems

SCIENCE TECHNOLOGY AND INNOVATION INDICATORS

A publication commissioned by the Ministry of Education,
Culture and Science.

Authors:
Pim den Hertog (Dialogic)
Cor-Jan Jager (Dialogic)
Arthur Vankan (Dialogic)
Robbin te Velde (Dialogic)
Jaap Veldkamp (Dialogic)
Dag W. Aksnes (NIFU)
Gunnar Sivertsen (NIFU)
Thed van Leeuwen (CWTS)
Erik van Wijk (CWTS)

Publication number:
2010.056-1435

Utrecht, the Netherlands
December 2014

For more information about STI² please visit
the website (www.sti2.nl)

This report is electronically available at www.sti2.nl

Graphic Design: Teatske sanne

SCIENCE TECHNOLOGY AND INNOVATION INDICATORS

innovatie • interactie

Nordic Institute for Studies in Innovation, Research and Education

Universiteit Leiden

Science, Technology & Innovation Indicators

Thematic Paper 1

Challenges in measuring the efficiency of national science, technology & innovation systems

2014

Table of contents

Thematic Paper 1

Challenges in measuring the efficiency of national science, technology & innovation systems

1. Introduction

Science, Technology and Innovation (STI) are seen as key determinants of economic competitiveness and growth as well as employment creation. The competitive advantage of knowledge-based economies is based on the capacity of a country to continuously generate new knowledge and introduce it to the market for economic and societal benefit (Metcalfe, S., 1995; Lundvall, B-Å, 1992).

Scientific literature often captures the full extent of innovation using the "National Innovation Systems" (NIS) approach. This approach takes into account research and development activities in the public and private domain as well as the determinants of innovation (Balzat & Hanusch, 2004). Developments in the field of NIS research have gradually evolved from a more country-specific analysis (Nelson, 1993) taking into account historical, political and cultural context, towards international performance comparison of national innovation systems (Furman, Porter and Stern, 2002).

These comparisons can and are being reflected increasingly in country rankings, input/output models or "innovation scores" in which a collection of science, technology and innovation indicators are analyzed on a macro level and compared with reference countries. A few examples are: The Innovation Union Scoreboard, The Global Competitiveness Report, and the Global Innovation Index. Policy-makers tend to be rather sensitive to these data-driven rankings and the rise or fall of a country in the rankings is exposed extensively in the media.

Most country comparisons mainly focus on input, output or a combination of both in which 'the usual suspects' of the richest economies perform well. A real policy challenge lies however in the delicate attribution of resources (input) to derive the maximum output. Despite well-established sets of indicators, few attempts combine both input and output indicators as they are often presented separately – some exceptions aside (e.g. May 1997 and 1998, King 2004, Leydesdorff & Wagner 2009). However, the attempts were scarce and the indicators have not been standardized or attained common acceptance as performance measures.

In general, we can say that indicator input-output models for NIS introduce several serious methodological issues that do not always receive the attention they deserve. With the increasing aim of country benchmarking comes the responsibility to address the merits as well as the pitfalls of multi-indicator analyses. In this paper, we address these issues in a structural manner and apply this knowledge to derive macro-based models for NIS efficiency.

We grouped the pitfalls of multi-indicator analyses into these three main categories:

- **Scope**: Indicators in rankings that are fundamentally different and/or unbalanced are combined to constitute an 'overall' score
- **Data quality**: There is a general disregard of adequate 'data quality' control
- **Comparability**: Current rankings take systemic heterogeneity between countries and NIS for granted.

The issues in the above categories are described more elaborately in the following chapter, after which we will structurally apply this knowledge in chapter 3 to critically select and reflect on indicators for inclusion in an efficiency model. The best-effort multi-indicator models will be introduced and assessed in chapters 4 and 5. Chapter 6 presents our concluding remarks and recommendations.

2. Current challenges

When describing the differences between national innovation systems (NIS) in terms of efficiency, we have to take into account many methodological issues. These issues are general in nature and affect any analysis dealing with international Science, Technology and Innovation statistics (Hall and Jaffe, 2012). This implies that the issues described here have a wider relevance beyond this STI[2] thematic paper. Most of these issues are described as sidesteps towards an efficiency model in literature, but not in a comprehensive and exhaustive manner.[1]
We argue that this systematic assessment is crucial to derive a useful NIS efficiency model and, more importantly, to be able to reflect on it in an enlightened and meaningful way. In the next paragraph (2.1), we will briefly introduce the three categories of challenges that arise from existing data collected at the national level and that are intended to provide cross-country comparisons. In the subsequent paragraphs (2.2 – 2.4), we will further extend these challenges to a typology of issues comprised of nine categories. Each of these nine categories will be introduced and illustrated with an example in the respective paragraphs, and, if possible, based on a topic relevant to our efficiency model.

2.1 Towards a typology of challenges in measuring innovation systems

The development in NIS research has gradually migrated from a focus on the structure of specific innovation systems towards performance based comparisons across the boundaries of these

1 See for example: Chen, C.P. & Yang, C. (2011); Liou, D.Y. (2009) or Nasierowski, W. & Arcelus F.J. (2003).

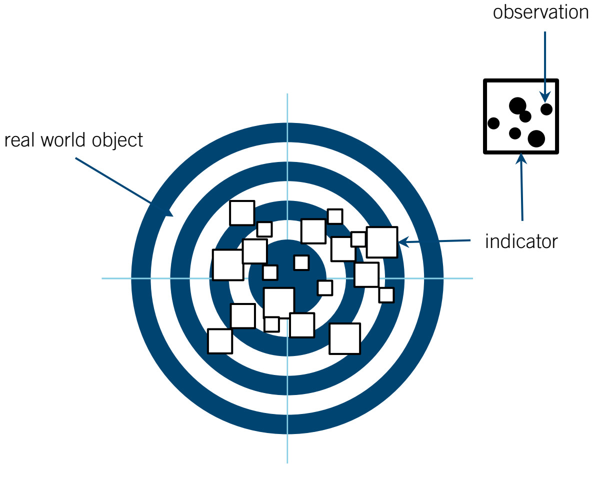
specific innovation systems (Balzat et al., 2004). These rather conflicting streams of research tend to go in opposite directions. At the one end, we observe an increasingly evolved and delicate distinction between actors or subsystems within a country, region or sector. At the other end, there is a growing desire for policy advice based on cross-country innovation system benchmarks, requiring an abstraction from this systemic heterogeneity.

The aims of the latter stream are apparent, as indicated in our introduction. It is therefore necessary to grasp this systemic heterogeneity in a way that validates cross-country analyses of innovation indicators. An important aim of OECD is to "improve the comparability of innovation indicators of its member countries" (OECD, 1999), so it comes as no surprise that most quantitative studies tend to use OECD data. The OECD has made extensive efforts to collect data using the same definitions as described for example in the Frascati Manual (OECD, 2002) for R&D statistics or Oslo Manual (2005) for innovation statistics. With these manuals, OECD is attempting to contribute to intergovernmental best practices for science and technology policies.

However, the indicators and metadata defined so adequately in these manuals merely provide the building blocks for innovation system (or efficiency) models. Combining and aggregating indicators into a more abstract model introduces other issues. What do the combined indicators tell us about the innovation system as a whole? Can the indicators be combined at all and can we even reflect on the respective model, without contextual background knowledge of the specific indicators (Hall et al., 2012). We often see that these finer points at the country level tend to be "lost in translation" when developing a broader model.

We introduce three main categories [A] Scope, [B] Data quality and [C] Comparability to systematically address these issues in the efficiency-models in Chapters 4 and 5. In general terms, scope issues can be found on the indicator level, data quality issues manifest themselves on the observation level, and comparability is all about the real world objects at stake here: national innovation systems that we want to measure. To visualize the sometimes rather theoretical issues, we have included a metaphorical picture throughout the remainder of this chapter.

Figure 1: Indicator (Scope), Observation (Data quality), Real World Objects (Comparability)



Scope

With scope issues, we refer to the extent to which we are able to [a] collect the appropriate indicators (content validity) in [b] a balanced and discriminate selection (instrument validity), and [c] at the right level (unit of analysis) in order to measure the innovation system. Carlsson, B., Jacobsson, S., Holmén, M., & Rickne, A. (2002) indicate the relevance of finding the appropriate unit of analyses in defining the innovation system and suggest that subsystems cannot always be 'lumped' together in an overall construct. This calls for a well-balanced and thoughtful selection of indicators.

Data quality

Although considerable efforts are being made by for example OECD and Eurostat (Oslo Manual, 2005; OECD, 2002; OECD, 2013) to come up with unified definitions and data preparation methods, data quality remains an important factor for consideration. With data quality we refer to issues regarding the [a] data collection process, [b] problems regarding the attribution of data to a specific system (data attribution bias), and [c] a potential indicator coverage bias in which a structural bias is present in an indicator. An example of the latter is the underrepresentation of humanities research in bibliometric indicators (Wendt, K., Aksnes, W., Sivertsen, G. & Karlsson, S., 2012).

Comparability

Having determined the adequate scope, selected the appropriate indicators, and controlled for the potential bias inherent in the indicators, we can still face severe comparability issues due to the specificity of an innovation system. The historically developed organizational and institutional structure of a country can play an important role in cross-country comparisons (Balzet et al., 2004). A very simple but convincing example is the language bias in science performance (Van Leeuwen et al., 2001; Wendt et al., 2012). Given the fact that the most widely used language in science is English, there is an implicit advantage for native English speaking countries in science output. To sum up: With comparability issues, we refer to [a] structural differences between countries or regions, [b] governance differences that impact indicator comparability, and [c] country specialization differences.

2.2 Issues with regard to scope

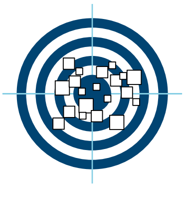
2.2.1 Unit of analyses

What is the appropriate unit of analysis? In other words, what or who is the major entity analyzed in a study.

Since we want to make generalizations at the level of countries, NIS seems to be the appropriate unit of analysis. This assumes that we can make meaningful observations about NIS as a whole.

The notion of a “national innovation system” is however quite broad. It encompasses at least three subsystems (Science, Technology and Innovation) and it remains to be seen whether these subsystems can be grouped together under one single heading. Although a recent IPTS study (Hardeman & van Roy, 2013) found a strong correlation between the output indicators for science and technology, a more detailed breakdown reveals that this correlation was only found for high quality scientific output and high quality technology output. When we look at science and technology output indicators in general, the correlation disappears. This is because composite indicators such as the one used for the Innovation Union Scoreboard (European Commission, 2014) average out the differences between the research and innovation subsystems and therefore mask the distinctive characteristics of NIS. Thus it is not worthwhile to make inferences about NIS as a whole. When comparing countries on the efficiency of their NIS, we should rather take their Science subsystem and Technology subsystem (and probably also their Innovation subsystem) as units of analysis. Although these subsystems are obviously nestled within an overall NIS, the functioning of the various subsystems is so distinctively different that their overall operation can only be understood at subsystem level. In practice, we observe that indexes presented for example in the Global Competitiveness Index, Global Innovation Index or Innovation Union Scoreboard strive to provide a comprehensive picture of the system as a whole, which can of course be a goal in itself. The biggest danger, however, is a one-size-fits-all arbitrary combination of indicators, merged in a score that is difficult to comprehend.

2.2.2 Content validity



Content validity is the extent to which the content of the measurement instrument matches a content domain associated with the construct. The issue here is that the devil is really in the detail. Slight changes in the definition of indicators sometimes lead to very different outcomes.

Whereas the selection of the unit of analysis refers to the appropriate delineation of the real world object that we want to describe, the general notion of validity refers to the match between this real world object and the measurement instrument (the theoretical concept) applied.[2] In other words, does the instrument measure what it is supposed to measure?

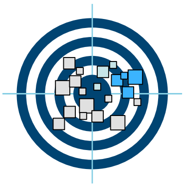
Imagine that we want to measure the efficiency of a system. To do so, we divide a (combination of) output indicator(s) by another relevant input indicator. Two main input indicators are money and people. These two types of indicators are closely related (e.g. R^2 of Business Expenditure on R&D [BERD] and ‘R&D personnel business’ is 0.85). However, since the correlation is <1, selecting either money or people as input indicator will still have an impact on the efficiency rate. These differences can be explained by a multitude of factors. One example is that BERD refers

2 In the literature on research methods, the overall notion of validity often refers to a specific type, namely construct validity. Ironically, given the aim of the literature, this causes a great deal of confusion.

to both capital and labor whereas the number of researchers refers to labor only.[3] According to classic economics, the amount of capital per labor is a determinant for output. The indicator 'number of researchers' does not take total investments into account. Moreover, funding statistics also capture researchers' wage levels. One can imagine that hiring a very productive researcher (potentially from abroad) will lead to a better output.

In short, the (construct) validity of the efficiency indicator that uses research funding is high as long as it is used to measure the dimensions capital and labor, but low when it is used to measure the specific dimension of labor.

2.2.3 Instrument validity



In the previous category we addressed individual variables to describe trends in specific content domains. Using one single variable to measure a content domain is often too narrow. Hence in practice, composite indicators are generally used to capture the full richness of such a domain. This raises two questions concerning the construction of the measurement instrument:

- Which items to include – and which not
- How to group the various items

In the ideal scenario, each individual item measures one particular dimension of the content domain, and the items are not correlated. However, the very fact that the items are not correlated raises the question whether they could be conceptually lumped together at all (see 2.2.1). If, on the other hand, slight variations of the same item are used, there is a real risk that the instrument becomes biased towards the items that are more or less duplicated.

The benefit of composite indicators is that it is easier to interpret them than identify common trends across many separate individual variables. This is one of the reasons they are increasingly being used to compare countries (Bandura, 2006). The construction of composite indicators is however a difficult art. One should always have a clear prior conceptual understanding of the data as different techniques may identify dimensions that do not necessarily help to reveal the clustering structure in the data and may actually mask a taxonomy (e.g., a hierarchy) already present in the data.

3 What is at stake here, are the differences in researchers' costs. These vary widely across countries. Parts of these cost differences are adjusted by using purchasing power parity (PPP) numbers. China, with its relatively low labor costs, performs indeed much better if productivity is measured in terms of economic resources. However, the results for some other countries are rather cross-intuitive. The UK has a high number of R&D personnel compared to the volume of economic expenditure (on par with China) and this also applies to Australia and Finland. Therefore these countries fare less well when using numbers of R&D personnel. Sweden and Austria have high ratios, in other words, high costs per work-years R&D, and these countries are improving. In fact, the number of PPP$ per R&D personnel is almost twice as high in Sweden as in Finland.

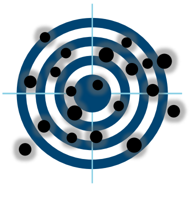
A severe drawback of multivariate analysis techniques is that the results are not valid if the sample is small compared to the number of indicators. A rule of thumb is that you need around 12 cases per independent variable, in order to have sufficient observations for different combinations of independent variables. It is typically the case in cross-country comparisons that only 15 to 50 observations are available.

Another challenge is to find the appropriate balance within the instrument. There might be good reasons to deviate between indicator weights in developing an overall construct. The World Economic Forum's well-known Global Competitive Index (GCI) does contain many variables. In its 2008 Global Competitiveness Report, the Forum revised the GCI with a two-stage principal component to aggregate 16 individual indicators into two sub indicators (Porter et al., 2008), but the individual indicators were not weighed. A potential focus bias in these 16 input indicators might lead to an unbalanced composite indicator.

2.3 Issues with regard to data quality

A measurement instrument is valid if there is a match between the theoretical construct (the measurement instrument) and the real world object that is supposed to be measured. Validity per se does not say anything about the reliability of the measurement results. This is mainly due to the quality of the underlying data. In the particular case of analyzing cross-country differences, one major issue is that each country uses its own specific data sets, measurement methods, and definitions. Although supranational organizations such as the OECD and EU have made great progress in unifying international data collection, substantial differences still exist between countries. The basic problem is that it is sometimes difficult to tell whether (or to what extent) the differences in a model between countries are real or rather constructs due to differences in measurements.

2.3.1 Data collection reliability



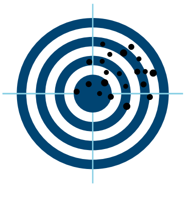
Are there any reliability issues in the data collection method for the selected indicators? Especially if one indicator is measured in different ways, the results may be incomparable.

The number of data sources for internationally comparable science, technology and innovation statistics is rather limited. The most widely used sources are OECD's Main Science and Technology Indicators and the S&T data series from Eurostat (including the Community Innovation Surveys, CIS). There are several issues regarding the quality of the underlying national data. A first significant difference between countries is the way data is collected. Even if a uniform measurement instrument is used (such as CIS), the widely varying results suggest that the instrument is not being applied in a uniform manner. For example, in several countries the scores on the narrower CIS item "innovations new to the market" exceed the scores for the broader CIS item "innovations new to the firm". From a conceptual point of view, this is a non-valid result. It raises questions about the overall reliability of the measurements. Another common error is the use of different definitions. Although these differences are often seemingly small, we have already concluded that small differences can lead to large differences in outcomes. For instance, countries have adopted different criteria for defining a 'researcher'.[4]

Countries also sometimes change definitions unilaterally over time. If sudden shifts occur in trends, one should always be mindful of such changes in definition. One example is the remarkable 18% rise in BERD (as a percentage of GDP) in the Netherlands of the 2011 data. In hindsight, more than three quarters of the increase can be attributed to a less stringent definition of 'R&D' by the Dutch national statistics office.[5] Another reason for this increase is the notion that since 2011 the R&D data collection has been expanded with companies with 1-10 employees, which also explains this break in series.

Long-range time series are dotted with such changes in definitions. Thus one should pay attention to the small print below time series – if such disclaimers are provided at all. Conspicuous changes in trends can sometimes be traced to disclaimers in the original national data sources if they are not mentioned in the consolidated international data source. In some cases, time series can be corrected retrospectively. Often though, the best one can do is to faithfully adopt the disclaimer that describes the changes in the underlying data collection process. One could also shorten the time series (removing all values prior to the change) or drop the entire time series for the country concerned. Obviously, in both cases the number of observations decreases. As for structural differences in the use of definitions and/or applications of an instrument (such as in the case of CIS), a radical option would be to omit the entire set of measurement results.

2.3.2 Indicator coverage bias



Even if a measuring instrument is valid and the reliability of the underlying data as such is satisfactory, there can still be serious issues with regard to data quality. This is because there is a structural bias in the structure of the data set (thus not in the measurement instrument itself). The measurement then inherits this structural bias from the underlying data set.

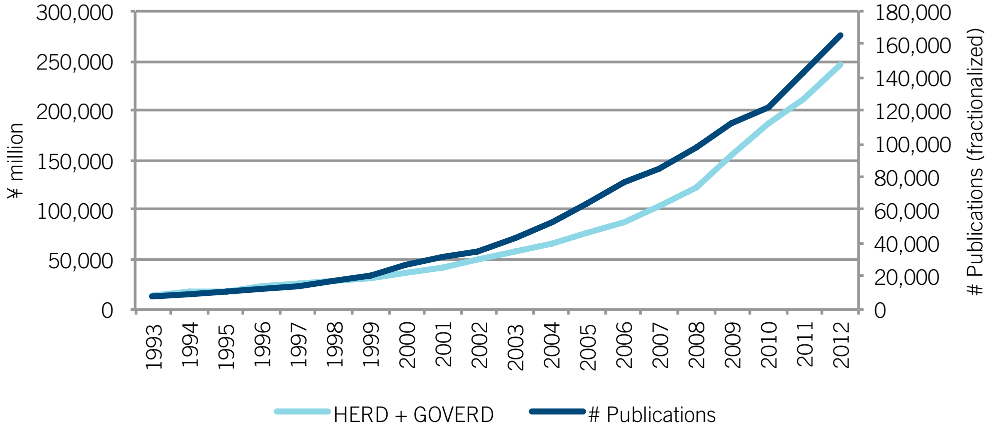
4 See the following quote from the Frascati Manual (OECD, 2002: p.33): "Two systems are now used by OECD member countries to classify persons engaged in R&D. Chapter 5, Section 5.2, contains definitions both for a classification by occupation, linked as far as possible to the International Standard Classification of Occupation – ISCO (ILO, 1990), and for a classification by level of formal qualification based entirely on the International Standard Classification of Education – ISCED (UNESCO, 1997). While it would be desirable to have data based on both classifications, most member countries use only one. As data are available by occupation for most OECD countries, the fact that a few still collect only qualification data for some or all sectors means that serious problems of international comparability remain."

5 http://www.rathenau.nl/nc/web-specials/de-nederlandse-wetenschap/nieuws/2012/11/sterke-groei-van-de-nederlandse-rd-uitgaven-van-bedrijven-in-2011-bijgesteld-dd-04-12-12.html

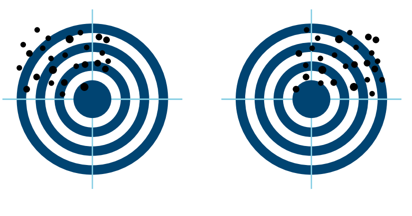
There are only two worldwide publication and citation databases: Thomson Reuter's Web of Science (WoS) and Elsevier's Scopus. Although the two databases differ in terms of general coverage, this coverage is suboptimal for both sources (Falagas et al., 2008). Alas, both databases also share a number of biases. Most importantly, social sciences and humanities are weakly covered. This introduces a bias in favor of countries (such as China) that have a relatively weak position in social sciences and humanities.

Another issue is the fact that an increasing amount of journals is being added to the two databases. An interesting example is China, since more and more journals published in Chinese are being included in the databases (especially Scopus). Although the number of China's publications is increasing proportionally to expenditure (as shown by the light blue line in Figure 2), this increase can (besides extra investments) also be attributed to the autonomous growth (coverage) in the underlying data sets and the focus of Chinese scientists to write papers in journals covered in WoS and Scopus, not in the country's scientific performance per se.

Figure 2: The number of fractionalized publications and R&D expenditures in China's higher education and government sectors (national currency), from 1993 to 2012



2.3.3 Data attribution bias



The assignment of an observation to a particular class is not always obvious. Sometimes, observations are attributed to classes that are known to be wrong (because the most adequate class is unknown). In other cases, observations can be attributed to multiple classes at the same time.

Research on science, technology, and innovation mostly makes inferences at the level of bigger aggregates such as countries, industry, sectors, and so on. A practical problem is that it is not always evident how to group individual data points (observations) under these aggregates. The issue at stake is the attribution of data points to specific classes. This is especially a challenge when we are dealing with cross-border activities and obviously an important factor when comparing countries. The proper assignment of financial data from a multinational company to countries of residence is a well-known thorny problem for two main reasons. Firstly, although multinationals usually operate in a multitude of countries, most internal operations are not (made) visible to the outside world. Financial data is usually aggregated in one location, mostly the headquarters. Thus we can no longer split overall turnover and so forth between the individual countries of origin. Secondly, the headquarters' location is often chosen for legal reasons, not based on where the head office is actually located (e.g. the headquarters of the French-German multinational EADS is in the Netherlands).

For the geographical assignment of patents, we run into similar problems. A patent file can actually have three addresses but often none of these addresses corresponds with the location where the invention has actually been made.[6] Ideally, the address of the "inventor" should refer to this location but many (larger) firms use the address of the 'applicant' for this field. Moreover, similar to the aforementioned headquarter issue, some firms tend to use the legal location of their headquarters (while others use the location of their R&D plant – which would be more valid than the address of their headquarters). Thus, in the case of the Dutch high-tech multinational Philips, all patents are assigned to one specific address in Eindhoven, the Netherlands. This sole fact greatly boosts the output of patents for the Netherlands (often used as an indicator for the performance of a Technology sub-system). Similarly, in international S&T statistics it has turned the Eindhoven region into one of the most innovative regions in the world. Obviously, especially for smaller countries, the presence (or absence) of even one large high-tech multinational has a big impact on country totals (Nokia in Finland being a well-known case).

With specific reference to output indicators for the Science sub-system, we encountered a serious assignment issue concerning the geographic origin of a scientific paper. Many publications are internationally co-authored, and are the result of collaborative efforts involving more than one country. Different principles and counting methods are applied in bibliometric studies. The simplest method would be to only take the address of the first author. However, this would not do justice to the efforts of all the other authors. Hence the most common is "whole" counting, in which every author (or country) gets full credit – thus articles are double counted and the total number of articles is greatly inflated. To correct for the latter problem, a system of "fractionalized counting" is used. With this method, credits are divided proportionally between the participating countries. For example, if an article has contributions from three departments in different countries, each country is credited 1/3 article (0.33).[7]

6 The 'Authority' (where the patent is being filed – can be any country in the world), the 'Applicant' (the owner of the patent – usually the firm – can be located anywhere in the world), and the 'Inventor' (is usually an employee of the 'Applicant'. Ideally this address corresponds with the location where the invention was actually made.)

7 One can argue that these counting methods are complementary: The whole count gives the number of papers in which the country has "participated". A fractional count gives the number of papers "creditable" to the country (Moed, Glänzel, & Schmoch, 2005).

The choice of a particular counting method obviously has a large impact on the output variable as the proportion of internationally co-authored publications varies significantly across countries (e.g. 24% for China and 66% for Switzerland, 2011). Using fractionalized counting reduces the publication numbers by 11% for China compared to whole counting and 44% for Switzerland (2012). These are the extremes, whereas the reduction for the Netherlands is 35% and the US 17%.

Table 1: Reduction in total number of publications due to fractionalization, selected countries (2012)

Country	Reduction in total number of publications due to fractionalization
China	11%
Denmark	37%
Netherlands	35%
South Korea	14%
Sweden	38%
Switzerland	44%
United Kingdom	30%
United States	17%

Both “whole counting” and “fractional counting” have their pros and cons. In the context of measuring productivity, the latter method has a higher validity than the former. This is because under “whole counting” countries with extensive foreign collaboration would be credited much of the research output done by scientists in other countries. Still we could argue that the use of fractionalized counting “punishes” international collaborative papers, which in other contexts are seen as particularly valuable types of research (e.g. the extent of internationalization is generally regarded as a quality indicator for the Science sub-system). In addition, we could say that while collaborating with other authors, an individual author is aware of the full content of the article (and therefore whole counting would be more valid).

The issue is further aggravated when analyzing long-term periods. This is due to the fact that in recent decades, the proportion of internationally co-authored publications has increased significantly. Retrospective fractionalized counting may thus punish ‘old policies’, e.g. in the case of EU FPs, which boosted cooperation and consequently output across Europe.

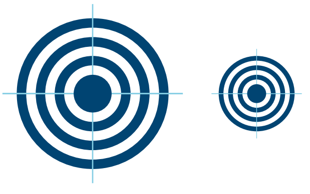
2.4 Issues with regard to comparability

In the previous paragraphs we discussed the validity of measuring instruments and the quality of data. Our final paragraph on methodological issues looks at the comparability of real world objects being studied. In terms of an experimental setting, the issue at stake is whether the differences found can solely be attributed to the effect of the experiment or are they simply inherent in the basic characteristics of the objects being compared; in other words, were these differences already present anyway, before and regardless of the effect of the experiment. One should not compare apples and oranges.

In the particular case of comparing self-conscious objects (such as people, or larger social aggregates such as organizations or countries), one should also consider the strategic behavior of the objects. That is, people (or firms, or countries) tend to develop conscious strategies to deal with their structural differences, and those of their peers. In the long run, such strategies might coagulate into semi-structural differences. This is the cultural dimension of evolution. The results are semi structural because they are not directly based on the structural traits of the object, but since they are based on a long term process (and often intertwined with other long term processes), they are still hard to adapt. Thus in the short term, these differences are a given.

In general, when comparing social entities, we should bear in mind all three types of differences (structural, governance, specialization). Although these are distinctive types, they are also conceptually linked.

2.4.1 Structural differences



Structural differences exist between research subjects (such as countries). When comparing subjects, these (background) differences should be taken into account.

One obvious structural difference between countries is size. We can measure country size in many different dimensions, such as area, number of inhabitants, and GDP, which can all be used as baseline to correct for country size. From an economic point of view, it makes little sense to correct for area. Both the number of inhabitants (indicators expressed in per capita figures) and GDP (indicators expressed in $) are widely used as denominators. Using GDP introduces a bias in favor of low income countries. Hence one should be careful when comparing countries that are in a (very) different economic development stage.[8] Likewise, using the number of inhabitants sec introduces a bias against countries with (very) high populations such as China and India. In this case, it would be better to use a more precise and relevant basis for correction such as the number of researchers or R&D personnel. Moreover, one should take care when using a simple linear correction method where the numerator is very small. For example, a small change in the absolute total number of patents (or publications) could cause huge shifts for very small countries such as Luxembourg.[9]

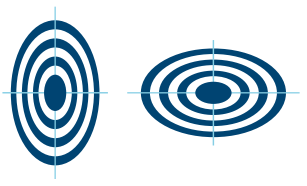
Another structural characteristic of a country is the language that is spoken (more precisely, whether it is English or not). While the process of internationalization results in a diminishing effect of language bias in the natural sciences (King 2004), the combined use of English and

8 For this reason, the World Bank uses four classes to compare countries based on GDP (income): low-income economies, lower-middle-income countries, upper-middle-income countries, and high-income countries (including a subclass of high-income OECD members). Source: http://data.worldbank.org/about/country-classifications/country-and-lending-groups

9 One way to deal with this size effect is to consider the values for very small countries by definition as outliers. Consequently, they are removed from the set of reference countries.

the national language continues to be important in e.g. profession oriented scientific journals in clinical medicine and in the general publication patterns of the social sciences and humanities (Van Leeuwen, 2013). Both the exclusion of journals in national languages and the inclusion of them in some fields may affect bibliometric indicators (Archambault et al. 2006, van Raan et al. 2011). There are also large variations with regard to the use of national languages among disciplines in the social sciences and the humanities (Sivertsen, 2009).

2.4.2 Governance differences



Differences in governance refer to the design of the unit of analysis. These are semi-structural differences in terms of the structure and functioning of countries, and the Science, Technology and Innovation subsystems within a country.

The distribution of political control and budgets over the various levels of government (national, regional, local level) varies widely between countries. Distinctive differences can also exist within countries in terms of governance structures and arrangements between the Science, Technology, and Innovation subsystems. Obviously, in a federal state, more power (and money) is controlled at the state (regional) level than in unitary states. Ideally, for a fair comparison of government funding for research, we should examine a country's total budget. However, most international statistical indicators (such as HERD and GOVERD) are only available at the national/federal level. This introduces a large bias against federal states since the often substantial government budgets at the state level are not included. GBAORD does include government research expenditure on all government levels, but a drawback is that it is even broader: it also includes all research funding that flows to foreign entities (e.g., transnational research organizations), whereas HERD and GOVERD are neatly delineated within national boundaries. More importantly, GBAORD covers the provisional and final budgets for government R&D expenditure.[10] The budget that is eventually actually spent on R&D in a country might deviate from GBAORD. Thus it remains to be seen whether GBAORD is a valid indicator of R&D expenditure in a particular country's public sector.

There are also structural differences in how a public research system is organized. While in some countries the majority of applied research takes place in universities (e.g. Sweden), others have a larger independent public research institute sector specializing in applied research.[11] A calculation method based on a combination of the two sectors is, therefore, well justified. For example, since a government can choose to allocate its applied research funding to either universities or public research institutes, HERD and GOVERD are more or less communication vessels and should be lumped together. At the same time, with regard to output, we should acknowledge that while scientific journal publishing is a main output channel at universities, other types of

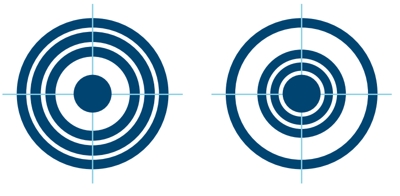
10 The final figures for the Netherlands represent budgets spent.

11 Public research institutes such as Fraunhofer and Helmholtz in Germany, INRIA in France, TNO in the Netherlands, CNR in Italy, and VTT in Finland.

research dissemination may be equally or more important in the governmental research sector. Thus, although 75% of public sector R&D expenditure (HERD+GOVERD) in the Netherlands is allocated to the higher education sector (and hence 25% to public research organizations), it accounts for 91% of the total national publication output.

A government can also decide to integrate public research institutes into universities, as Denmark did in 2007. Hence governance arrangements are a typical example of semi-structural differences. A more volatile trait of governance systems is the prevailing strategy leaning towards science, technology, and/or innovation. For example, over the past two decades, China's science policy has been strongly focused on science and engineering and its proportion of social sciences is negligible. Consequently, this enormous increase in scientific production from China has had a significant worldwide impact on the scientometric statistics for natural sciences and engineering (because a world average is used for field normalization) but much less so on social sciences.

2.4.3 Specialization differences



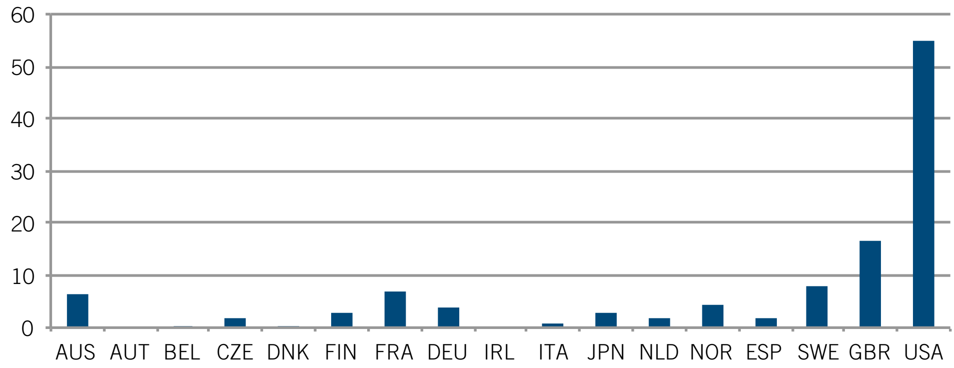
Besides the structural and governance differences, the units of analyses can also differ due to e.g. path dependency or governance differences. As a result, the units of analyses focus/specialize on different aspects.

The current science, technology, and industry base of a country is the result of a process that is highly path-dependent. History is an important determinant. Scientific specialization patterns – just like technological specialization patterns – tend to co-evolve with broader R&D structures. That is, they depend on investment patterns as well as on industrial structure (Peter & Bruno, 2010).

These structures are remarkably persistent. A country's export specialization can remain surprisingly stable over a very long period of time. For example, the relative strengths of the Netherlands (measured in revealed comparative advantages) are still in low-tech products such as dairy, foodstuff, and colonial products, despite the fact that the Dutch economy is highly developed and very open (Dialogic et al., 2012).

Similarly, for various strategic reasons, countries can decide to maintain a large army and make substantial investments in defense-related R&D. In some countries such as the US, the majority of public spending on R&D is in defense, but the proportion of defense-related research is also substantial in some EU Member states as illustrated in Figure 3.

Figure 3: Proportion of public spending on defence-related R&D (2011)



Source: Defense Budget R&D as a percentage of Total GBAORD (OECD MSTI).

High scores in the percentage of public research spending on defense also affect the international comparability of the input and output S&T statistics (Wendt et al., 2012). For example, research results are obviously not published in public channels such as scientific journals or patents, even when large parts of the research are performed by public research institutes or universities (as in the US).

3. Model development

In this chapter, we will initiate the process of appropriate indicator selection for building an efficiency model in the following chapters 4 and 5. A reflection on the concept of efficiency is also given in paragraph 3.3.

3.1 Data selection and country selection

3.1.1 Data selection

Efficiency is defined by output per unit of input. To select valid and legitimate indicators that describe NIS, we started by creating a long-list of available and suitable science, technology, and innovation indicators. In order to create the long-list, we consulted a number of international indicator sources (among others, OECD, Eurostat, IUC-database). The long-list consists of approximately 200 indicators (see 'Appendix A | long-list of indicators').

The long-list was analyzed with respect to nine issues, distributed over the three categories as addressed in chapter 2:

a. Scope
b. Data quality
c. Comparability

Where scope, data quality or comparability is deemed insufficient, the indicator is left out of the analysis. Note that, although presenting it as a linear process, we followed an iterative process. For example, indicators with a perfect scope but poor data quality could be substituted for other indicators with a less perfect scope but good data quality. Our long list was consequently narrowed down to a smaller set of indicators that will serve as basis for the reference model. Examples of discarded indicators are:

- United States Patent & Trademark Office (USPTO) grants by technological field were considered for Technology output, but discarded due to indicator coverage bias (2.3.2). The USPTO would be biased towards countries with a strong presence / market in countries like Canada, the UK and of course the US itself. The alternative (PCT and Triadic patent families) is a more balanced alternative.
- CIS (Community Innovation Survey) data was considered for innovation output. However, data collection reliability was deemed insufficient for reasons explained in paragraph 2.3.1.
- The number of doctoral graduates could pose as candidate indicator for the input science sub-system (this indicator is considered an 'enabler' in the Innovation Union Scoreboard). However, since the retention rate of doctoral graduates within the academic world differs greatly per country, we do not consider this a valid input construct for the science system. Better alternatives are included in Table 2.

The resulting list of indicators is shown in Table 2 and Table 3.

Table 2: Science indicators for inclusion in efficiency model

	Indicator	Description	Source
[1]	HERD	Expenditure on R&D in the Higher Education Sector (million current PPP)	OECD: MSTI
[2]	GOVERD	Government Intramural Expenditure on R&D (million current PPP)	OECD: MSTI
[3]	Publication output	Total number of publications per country per year in Web of Science - fractional count[12]	Web of Science
[4]	Citation impact	Field-normalized citation impact – fractional count[12]	Web of Science
[5]	Publications top 10%	Number of publications in top 10% of most cited publications compared to statistical expected count in top 10% most cited publications per country (in %)[12]	Web of Science
[6]	Higher Education R&D personnel	Higher Education total R&D personnel (fte)	OECD: MSTI
[7]	Government R&D personnel	Government total R&D personnel (fte)	OECD: MSTI

12 In our model we decided to use a fractional counting scheme for the calculations instead of whole counting. (See paragraph 2.3.3 for a further elaboration on the two options). We believe this is the fairest way of comparing the research output of different countries. Countries with extensive foreign collaboration would otherwise be credited with much of the research output that is basically done by scientists in other countries (Aksnes, Schneider & Gunnarsson, 2012).

The science indicators can basically be brought down to input indicators (HERD and GOVERD) and output indicators on quantity (publication output) and impact (citation impact, top 1%, top 10%) publications. The indicators relating to technology input and output are somewhat larger, as shown in Table 3 below.

Table 3: Technology indicators for inclusion in efficiency model

	Indicator	Description	Source
[8]	Triadic patent families	Number of triadic patent families (priority year)[13]	OECD: MSTI
[9]	Patent application PCT	Number of patent applications to the PCT (priority year)[13]	OECD: MSTI
[10]	High tech export	Export flow High-technology industries – Total trade in goods – thousand USD – Partner country: World	OECD: STAN[14]
[11]	High tech import	Import flow High-technology industries – Total trade in goods – thousand USD – Partner country: World	OECD: STAN[14]
[12]	Medium tech export	Export flow Medium-high technology industries – Total trade in goods – thousand USD – Partner country: World	OECD: STAN[14]
[13]	Medium tech import	Import flow Medium-high technology industries – Total trade in goods – thousand USD – Partner country: World	OECD: STAN[14]
[14]	ICT export	Export flow Information Communication Technology manufactures (ICT) – Total trade in goods – thousand USD – Partner country: World	OECD: STAN[14]
[15]	ICT import	Import flow Information Communication Technology manufactures (ICT) – Total trade in goods – thousand USD – Partner country: World	OECD: STAN[14]
[16]	BERD	Expenditure on R&D in the Business Enterprise Sector (million current PPP)	OECD: MSTI
[17]	BE R&D personnel	Total Business Enterprise R&D personnel	OECD: MSTI

3.1.2 Country selection

The set of reference countries in this paper is confined to the set of WTI[2] reference countries that have been of central interest in previous years (see Table 4). Data was collected on the above indicators for all these countries for most recent years.

13 As discussed in paragraph 2.3.3, a patent can have three addresses but often none of these corresponds precisely with the location where the invention was actually created. In our model, patents are presented according to the inventor's country of residence. Furthermore, we used the priority date as reference, since this is closest to the invention date.

14 STAN Bilateral Trade in Goods by Industry and End-use.

Table 4: Reference countries in efficiency models

Reference countries		
Australia	France	Austria
Belgium	Ireland	Germany
Canada	Japan	United Kingdom
China	Korea	United States
Denmark	The Netherlands	Sweden
Finland	Norway	Switzerland

3.1.3 Controlling the source data

To control for country size (see paragraph 2.4.1), we decided to normalize all country-size proportional absolute values for the selected indicators (e.g. the number of publications) by dividing the respective values by Gross Domestic Product – million current PPP $. Purchasing Power Parities (PPPs) are currency conversion rates that both convert to a common currency and equalize the purchasing power of various currencies. In other words, they eliminate the differences in price levels between countries in the process of conversion. In doing so, we mitigate the risk of comparing countries against different price levels and/or deviating inflation developments.

In order to develop composite indicators that combined more than one indicator (e.g. HERD + GOVERD), we had to devise a method that would enable us to compare them according to the weights assigned to the individual indicators. There are multiple ways to normalize, of which the most important ones are based on (i) average & standard deviation and (ii) minimal & maximal values. Since the data of the different indicators are not distributed normally, we chose the second method. By doing so, we migrated the individual indicators to the proportion of the maximum value on that indicator (of any given country). So for example: if Switzerland has the highest relative publication output of all reference countries (after controlling for country size), its respective value is set at "1". All other countries are proportionally migrated on this indicator (so all with a value of 0<1).

3.2 Division into Science, Technology and Innovation subsystems

Finding the appropriate unit of analysis is an important step in creating a model. As discussed in paragraph 2.2.1, the notion of a "national innovation system" is quite broad and can be divided into at least two (Science, Technology) or three (Science, Technology and Innovation) subsystems.

To decide on the appropriate unit of analysis, we collected dozens of indicators of each subsystem in our long-list, as described in the previous paragraph. A critical assessment on the topics of scope, data quality, and comparability, resulted in two fundamental choices for scoping this paper.

a. Excluding the innovation subsystem from the analysis.
b. Making a clear-cut division between the Science and Technology subsystems.

3.2.1 Exclusion of innovation subsystem

The reason for excluding a separate model dedicated to the Innovation subsystem, is that the quality of most (if not all) innovation indicators is insufficient to derive a meaningful model and interpretation. CIS indicators would be a logical option to include, but as discussed in paragraph 2.3.1, the reliability of the data is questionable. Furthermore, the most straightforward input indicators (R&D-expenditures or researchers) have a strong bias against the soft side of innovation (e.g. service innovation).[15]

3.2.2 Division between Science and Technology model

Two distinct efficiency models (one for Science and one for Technology) turned out to be the most meaningful. The rationale behind this choice is that we assumed the scientific subsystem would be substantially different from the technology (eco)system (e.g. writing publications versus protecting new technology by means of a patent). A recent study performed by Hardeman & van Roy (2013, forthcoming) for the Joint Research Centre on efficiency in the production of excellent research, reveals a high correlation between output indicators for science and technology. However, this relationship is based on high quality science output and high quality technology output. When we extend this to the more general output indicators such as the number of citations and publications, this correlation falls apart. In fact, our own (factor) analysis (see Table 5) reveals that output indicators for science and technology for the EU27 countries lead to one specific Factor for science output (51% variance) and one Factor for technology output (35% variance).

Table 5: Rotated Component Matrix (Extraction Method: Principal Component Analysis, Rotation Method: Varimax with Kaiser Normalization and Rotation converged in 3 iterations)

	Component	
	1	2
Publications	.864	
Citation impact	.967	
Top 1%	.969	
Top 10%	.941	
Triadic patent families		.935
Patent applications PCT		.974
Export minus Import		.685

15 "[..] the desire for better information on R&D in service activities has been expressed. The basic definitions in this Manual were originally developed for the manufacturing industry and research in the natural sciences and engineering. Specific problems therefore arise for applying them in the social sciences" (OECD, 2002: p 20).

The table shows the factor loadings for indicators, illustrating a clear distinction between "Technology output" indicators and Science output indicators. This hampers the rationale to combine the two and justifies the approach to reflect separately on the science and technology subsystems of Research and Innovation.

3.3 Theoretical complications of an efficiency model

In general terms, efficiency is the ratio between output and input. Calculating efficiency is relatively simple for e.g. 'energy conversion'. After all, to ascertain the energy efficiency of a wind turbine, you can measure input and output both in terms of energy. However, the efficiency model of NIS is considerably more complex. In the following section we address two significant theoretical complications when looking at the efficiency models of complex systems like NIS.

Path dependency

In a simple model, more input will automatically result in more output. To revert to the energy example, after an adjustment in the power plant, the 'new' (energy conversion) efficiency can be measured instantly. However, 'changing the parameters' of NIS (e.g. by restructuring the higher education sector, reshaping industrial policies, adjusting specialization patterns) requires a lot of time and resources. But changing the system will not lead to another efficiency level immediately. Extra input, e.g. in the educational system, is not directly transformed into more and/or better output. For example, it would take many years of continual additional investments for the Netherlands to achieve its relative number of S&E graduates on a par with leading countries such as Finland or Sweden (a well-known weakness of the Dutch system).

In other words, the current composition of a country's economic activities is very much determined by its historical development – something we have to bear in mind when analyzing NIS efficiency.

Economics of marginal returns

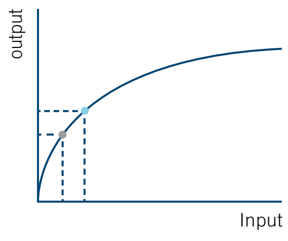
When improving a production process in terms of efficiency, the first steps are easy (e.g. cut overheads). But the next improvements will be more costly (e.g. switching from hand-made to machine-made products). And at some point, further improvements will not be beneficial because the costs no longer outweigh the benefits. In other words, there are diminishing returns. This phenomenon (or rather: economic law) affects the efficiency performance of NIS in two ways.

1. Modest innovators versus innovation leaders

The Innovation Union Scoreboard ranks countries as (i) modest innovators, (ii) moderate innovators, (iii) innovation followers, and (iv) innovation leaders. Innovation leaders are countries with, among other indicators, a great deal of STI input (e.g. R&D) and output (e.g. patents). The modest innovators, on the other hand, are countries with fewer STI activities. In spite of the presumably normative labelling, *based on the theory of marginal returns, it is likely that the less innovative countries are on average more efficient*. This is depicted in Figure 4. The curved line shows the return, i.e. the target's additional yield caused by one additional unit of input (e.g.

euros). The grey and the blue dots plot two countries on the return curve. As shown, the 'grey' country (in this example a moderate innovator) is on average far more efficient, but the 'blue' country (=innovation leader) has more innovative activities. In other words, measuring *average* efficiency favors countries with fewer STI activities.

Figure 4: Efficiency of two countries



2. Efficiency versus effectiveness

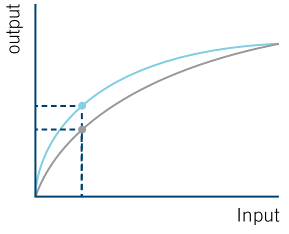
The countries' marginal return curve can differ and thus also the size of the optimal innovation system.[16] For instance, if a country is large, the shape is steeper at the start, which implies a large innovation system in equilibrium. If a country specializes in traditional sectors, the technological opportunities are low, which appears as a rather flat yield curve (and thus a small optimal innovation system). In contrast, if a country specializes in high tech, the slope of the yield curve keeps rising steeply, leading to a large innovation system. This is depicted in Figure 5 by showing two different return curves.

Since the return curve can differ between countries, it is necessary to critically consider the notion of efficiency versus effectiveness. Only in the theoretical situation when a country is 100% efficient in translating input to output, can it be plotted on the optimal marginal efficiency curve (the 'grey' country in Figure 5). In practice, a country will always be less efficient in translating STI input to STI output, that is, it performs below the theoretical optimum. Consequently, even with the same input, more output could be generated (e.g. by adjusting the higher education system within the current budget). This is depicted by the 'blue' country in Figure 5.

In the absence of details on the *individual* return curves, it is difficult to draw correct conclusions about the actual efficiency of an NIS. Although the 'blue' country is more effective, in terms of its maximum output level, it is less efficient given its input level, in translating input to output. Without putting extra resources in the innovation system, the 'blue' country can already generate more STI output. The grey country on the other hand cannot perform better without adding extra input to the system; considering the idiosyncratic context, it is more efficient than the 'blue' country. Especially the notion of the *potential* (rather than theoretical) efficiency gains (given a country's idiosyncratic context) is interesting for policy makers.

16 In economic terms, the optimal size of an innovation system is when the marginal efficiency is 1 (because it is the additional yield divided by the additional costs). Investing more in the innovation system will yield less than the investments.

Figure 5: Efficiency versus effectiveness



4. Science model

In paragraph 4.1, we present a reference model for science efficiency based on the decisions in the previous chapter. In subsequent paragraphs, we will reflect on this model in two steps. We distinguish between:

a. Issues in the model that cannot be solved completely, but do provide the tools (indicators) to show the impact of these issues. We will introduce some of these deviations by including alternative indicators or removing existing ones to potentially reveal differences in deemed science efficiency. This 'robustness check' of the reference model is presented in paragraph 4.2.
b. Another category of issues, described in paragraph 4.3, are the more fundamental ones, which cannot be solved, despite the fact they actually should. These comprise the list of disclaimers for the science model and their impact is discussed more thoroughly in the concluding chapter.

4.1 The reference model for science

Having discussed the indicators for the science subsystem in paragraph 3.1, we started creating a reference model for the science subsystem in which we develop composite indicators for both input and output.

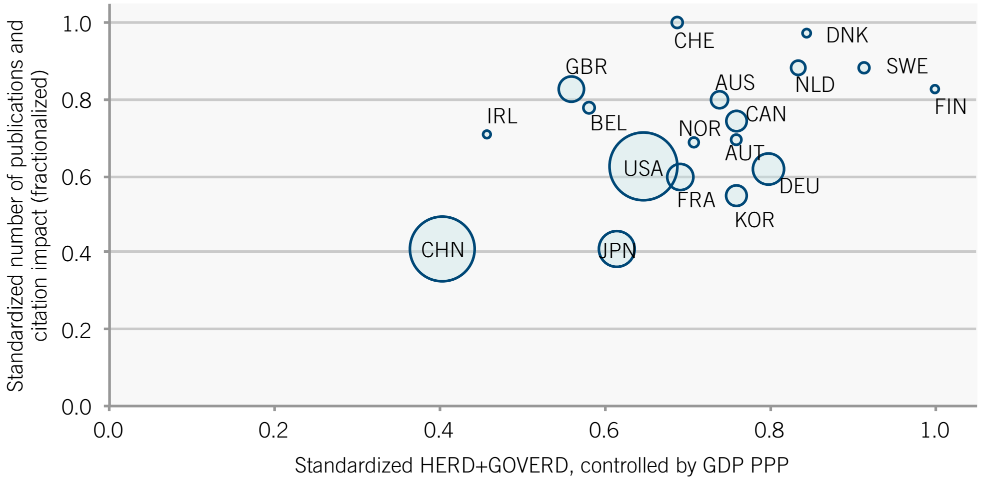
As science input, we used the sum of HERD + GOVERD; indicators 1 and 2 in Table 2, controlled for country size (GDP in PPP) and normalized. Two fundamental choices were made here. First, we decided to primarily focus on 'money' as means of input, because this comprises a more complete overview of both human and physical resources to 'run' the respective science systems. Moreover there are fewer problematic definition issues than with the sole focus on human resources. Second, we decided to incorporate GOVERD in the model for reasons fully described in Chapter 2. The main argument is that the relative size of higher education and government sectors varies significantly across countries due to the fact that the public research system is organized differently. While in some countries the majority of the applied research takes place in universities, others have a larger publicly funded sector specializing in applied research. A calculation method based on a combination of the two sectors is, therefore, well justified. This methodology was also chosen in a productivity study by Leydesdorff & Wagner (2009).

Regarding output, we chose both a quantity as well as a quality indicator for our reference model. The most basic (and least disputed) way to measure scientific output is in terms of the total number of scientific publications produced by a country (read: an author who is affiliated to an institute established in a particular country). Limiting the indicators to merely fractionalized publication output alone (indicator 3 in Table 2), does however introduce an obvious bias towards quantity over quality. For this reason, we also included an indicator that captures the field-normalized citation impact of a country (indicator 4 in Table 2). Both indicators were given equal weight.

There is a certain time lag from investments in the research system until published articles (Rousseau & Rousseau, 1998), and an additional lag once the published results start to receive citations. This fact should be taken into account in research productivity indicators. A two year lag has been considered as appropriate at this high level of aggregation (Leydesdorff & Wagner, 2009). This means that the 2012 publication data should be compared with the 2010 R&D expenditures and a corresponding time lag used in the temporal analyses.

Figure 6 summarizes the resulting reference model for science efficiency. The size of the blue dots represents country size (in GDP PPP).

Figure 6: Basic efficiency model for Science

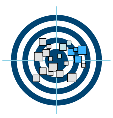
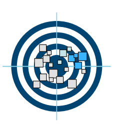


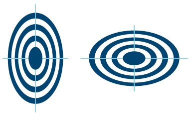
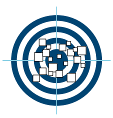
Note that the results do not tell us how well the countries are performing in absolute terms, but rather in relative terms: how well are they doing compared to each other. We observed a pattern in which Ireland, the United Kingdom and Switzerland have high outputs compared to the resources they put in. On the other side we see Germany, Korea, and Japan have relatively low efficiency in their science subsystems.

4.2 Variations on the science based model

This paragraph introduces a series of modifications to the reference model using the 'vocabulary' in Chapter 2. Putting aside all fundamental reliability points in the reference model discussed in the next paragraph, we can switch a number of 'buttons' *within* the reference model or *use alternative* indicators. This would introduce some slight 'deviations' in efficiency rankings, but does provide a better grip on the 'robustness' of the reference model. Table 6 below addresses the variations made to the reference model.

Table 6: Variations on science reference model

Model variation	**Description**
Instrument validity  **Alternative assessment of 'quality' in science output**	The field-normalized citation impact per country is one way of operationalizing quality, but not necessarily the best. As an alternative, we switched this indicator in the composite output construct with the number of publications in top 10% of most cited publications (see indicator 5 in Table 2). This indicator might approximate the concept of 'quality' better than citation impact. **Model adjustments:** We substituted field-normalized citation impact with an index of presence in top 10% cited papers.
Instrument validity:  **Exclusion of 'quality' in science output, merely focusing on quantity**	Since citation impact is normalized (worldwide average = 1), this indicator is not linear scalable. Since the citation impact score depends on other countries' impact scores, it would require relatively immense resources to double the impact score. Given that linear scalability does apply to our input indicators, it is up for debate whether we could conceptually use citation impact as output indicator. Although including some means of output quality in the model seems hardly debatable, an overall additional insight would be obtained if science output is restricted to hard publication counts only. **Model adjustments:** We excluded citation impact from science output, thereby only leaving the number of publications as output construct.

Governance differences  **Solely using HERD as input indicator**	Since our model focusses on capital productivity (see paragraph 2.2.2), we took R&D expenditures as input. Therefore we combined the expenditures of higher education and government sectors. At the same time, we should acknowledge that while scientific journal publishing is a main output channel at universities, other types of research dissemination may be equally or more important in the governmental research sector. **Model adjustments:** Both the inclusion and exclusion of GOVERD in the model causes distortion. As a robustness check, we explored excluding GOVERD.
Content validity:  **Changing time lag between input and output indicators**	As indicated, we used a two-year time lag between the composite input and output indicators for the science subsystem. Reality however, does not imply nor justify such a confined time lag for all publications. To assess the overall impact of alternating time lags, we calculated their impact on the reference model. **Model adjustments:** Instead of using a two-year time lag between input and output scores on the selected indicators, we calculated the impact of a one-year time lag.

Content validity: **Labor indicators as alternative for capital indicators in the Science model**	As discussed in paragraph 2.3.1, human resources can be an alternative input measurement. The reason for using R&D expenditure in the reference model rather than HR indicators is because the latter strand brings additional methodological challenges. In the OECD's R&D statistics, there are two relevant categories, one containing data on the number of R&D personnel and one on the number of researchers. The first category contains all personnel involved in R&D activities while the latter is limited to researchers. The majority of publications will obviously be produced by the population of researchers. Thus this category would be the most relevant for comparing the publication output. Unfortunately, countries have adopted different criteria for defining a researcher. It is therefore difficult to make cross-national comparisons using the number of researchers. R&D personnel are defined more broadly, cause fewer methodological problems and are more suitable as alternative input for the science system. **Model adjustments:** Instead of HERD + GOVERD as input for the Science system, we used R&D personnel working in Government and/or Higher Education as input indicator.

All these variations were calculated and their impact on the science reference model is reflected in Table 7. The countries are "ranked" based on the reference model, with the most efficient countries on top. For the sake of comparison, we divided the (composite) output construct by the (composite) input construct and normalized the values column-wise to an average of 1 as represented by the grey bars. The red and green arrows represent a respective rise or fall in relative position of at least four places. For example: China drops more than four places compared to the reference model if we only consider the number of publications as scientific output.

Table 7: Impact of variations on science reference model for efficiency

	Science Reference model	Top 10 publications instead of citation impact	Only publications as output	Solely HERD as input	Time lag = 1 year	Science R&D personnel as input	Science and government R&D personnel as input
Ireland	1,49	1,48	1,44	1,18	1,66	1,29	1,46
United Kingdom	1,42	1,44	1,49	1,28	1,31	0,75 ↓	0,89 ↓
Switzerland	1,39	1,44	1,60	0,96 ↓	1,42	1,10 ↓	1,40
Belgium	1,28	1,31	1,34	1,17	1,28	1,05 ↓	1,14
Denmark	1,10	1,13	1,31	0,79 ↓	1,14	0,86 ↓	1,04
Australia	1,03	1,02	1,15	1,01	1,02	0,77 ↓	0,81 ↓
Netherlands	1,02	1,04	1,07	0,87 ↓	1,06	1,25 ↑	1,24
China	0,98	0,92	0,56 ↓	2,05 ↑	0,95	1,24 ↑	0,69 ↓
Canada	0,94	0,94	0,98	0,82 ↓	0,92	1,08	1,09
Norway	0,93	0,92	0,86	0,94	0,98	1,17 ↑	1,00
Sweden	0,92	0,92	1,10 ↑	0,74 ↓	0,93	1,21 ↑	1,37 ↑
United States	0,92	0,97	0,64 ↓	1,15 ↑	0,86		
Austria	0,88	0,87	0,84	0,71 ↓	0,89	1,08 ↑	1,20 ↑
France	0,83	0,82	0,74	0,92 ↑	0,85	0,91	0,81
Finland	0,80	0,77	0,95 ↑	0,78	0,84	0,65	0,62
Germany	0,75	0,75	0,65	0,91 ↑	0,72	1,16 ↑	0,87 ↑
Korea	0,69	0,65	0,77 ↑	1,00 ↑	0,60	0,75	0,72
Japan	0,63	0,58	0,51	0,73	0,57	0,67	0,66
Correlation with basic model	-	0,99	0,88	0,32	0,98	0,3	0,61

Reflecting briefly on the table, we conclude that the reference model is rather stable regarding the output side. The efficiency is only mildly affected by a holistic focus on the number of publications as output and even less so by introducing an alternative ‘quality’ indicator. Also the time lag between input and output does not seem to have a big impact.

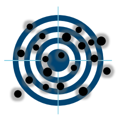
In contrast, however, on the input side, we do observe some large differences in alternative indicators. The sole use of HERD as input leads to a striking ‘efficiency increase’ for China – stressing the relevance of including GOVERD as input indicator. It also reveals the fragility of these rankings. This could be even more true, if we considered using labor not capital as alternative input construct. Quite a few deviations seem to emerge from the reference model. There are numerous explanations – one of which might be explained by the relative wage differences not captured in GDP in PPP between countries. For example we see that China with relatively ‘cheap’ R&D workers drops in efficiency, whereas Sweden, Norway, Germany, and Austria do better in the more holistic efficiency model. This might justify going one step further than a PPP correction and control for more granular subsystem differences. In the meantime, it does encourage the notion that these models should be assessed, analyzed, and interpreted against a deeper understanding of their (composite) constructs – a topic of further debate in the next paragraph.

4.3 Remaining disclaimers for the science model

To conclude this chapter, some issues that cannot be solved in the model and thus should be taken into account when interpreting the model(s).

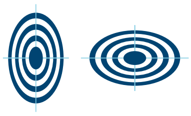
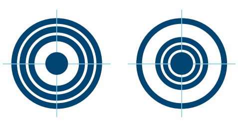
Table 8: Remaining issues with regard to the science model

Issue	Description

Data collection reliability

Comparability of macro-based R&D statistics
	A factor that may affect the validity of measurements for research efficiency is potential problems with the comparability of national R&D statistics. The following disclaimers apply to the reference countries in this study.
	- HERD:
	    - The data for Austria and Sweden are 'estimates/projections'.
	    - OECD figures for the Netherlands are overestimated compared to Statistics Netherlands (on average 10%).
	    - The R&D data for the USA exclude capital expenditures ("expenditures creating future benefit").
	- GOVERD:
	    - The data for Australia, Austria, and Sweden are 'estimates/projections'.
	    - USA and Switzerland cover only federal/central government activities. State and local government establishments are excluded.
	    - Germany and the Netherlands 'include other classes' in their totals.
	- Gross Domestic Product (Million current PPP$): Three countries (Australia, Canada and USA) capitalize R&D in their national accounts. Since we normalize all variables based on GDP (PPP), the normalized values are an underestimate compared to the other countries. After all, the denominator for these three countries is relatively high

Indicator coverage bias **Field coverage problems in Web of Science**	Data on scientific publishing are retrieved from the CWTS WoS database,[17] which is generally regarded as a satisfactory representation of international mainstream research. However, the limited coverage of the social sciences and humanities in publication databases may affect countries in different ways. Data from OECD MSTI show that the proportion of humanities/social science research in government and higher education sectors differs considerably among countries. Some use less than 10% of their research expenditure in these fields (e.g. around 5% in the UK and China), while others may spend several times more (e.g. 25-30% in Norway, Finland, Denmark, Austria, Ireland). In contrast, the proportion of social sciences and humanities publications in WoS varies from 2% (China, Japan) to 14-16% (Australia, USA, UK). Thus there is a gap, and part of the research within these fields is not visible in WoS. A similar issue is present in the engineering sciences, which are only partially covered in the ISI WoS (around 50% of the publications according to Moed et al. (2005). Countries specializing in engineering sciences (e.g. many Asian countries) therefore face a comparative disadvantage when productivity indicators are calculated. Consequently, the limited coverage affects the quality of the number of publications and the citation impact.
Indicator coverage bias **Language bias towards native English speaking countries**	As discussed in paragraph 2.4.1, language bias has often been an issue when interpreting country representation based on the ISI WoS database.
Indicator coverage bias **Average research productivity differs per scientific discipline**	In some scientific disciplines, the number of publications per researcher per year will be higher than in other disciplines. This affects the publication output. There will be a bias towards countries specializing in disciplines (such as medicine) with a relatively high output. In theory, one could control for this by using the average number of publications per researcher in a particular field. However, the underlying data to perform this particular control is not available (i.e. it is impossible to determine the average number of publications per person per field). In practice, this is not a major issue because the two disciplines that dominate the database (medicine and the natural sciences) have similar publication patterns (with some exceptions on a more detailed level).

17 This version of WoS includes the Science Citation Index Expanded, the Social Sciences Citation Index, and Arts and Humanities Citation Index; the database does not include the Conference Proceedings Index.

Governance differences  **Expenditure statistics are not all included in a country's macro-figures**	In the reference model we used expenditure (monetary resources) as input indicator. However, MSTI indicators only provide data based on national expenditures. If federal states also finance scientific research, these numbers are not included in the MSTI data. This poses a serious problem at the input level. Although in some federal states (such as Switzerland) relatively little research is financed by the state, in other countries such as Germany and the United States, this introduces a major distortion in the financial totals. In theory, the issue could be solved by collecting primary data at state level within each country, however this very labor-intensive exercise still delivers results that cannot be used directly for cross-country comparisons (due to e.g. differences in definitions).
Specialization differences  **Defense related R&D**	Another important difference between countries that is somewhat between governance and specialization, is the extent to which research is being conducted on defense-related matters. Obviously, one could correct for these structural differences between countries by excluding defense-related research, thus focusing solely on civil research. Funding figures for civil only research (e.g., using GBAORD, excluding defense-related research) are indeed available – and output figures are by definition limited to public, civil research. We would thus simply exclude defense-related research at both the input and output side. This is the fairest way to compare countries that differ greatly in terms of specialization pro or contra defense. However, an important disclaimer is that in the USA and Israel, a great deal of basic and blue sky research is actually being done under the label of defense-related research ('dual use research'). Focusing on civil research would thus exclude large chunks of highly relevant and often state-of-the-art research actually being done in the public sector.

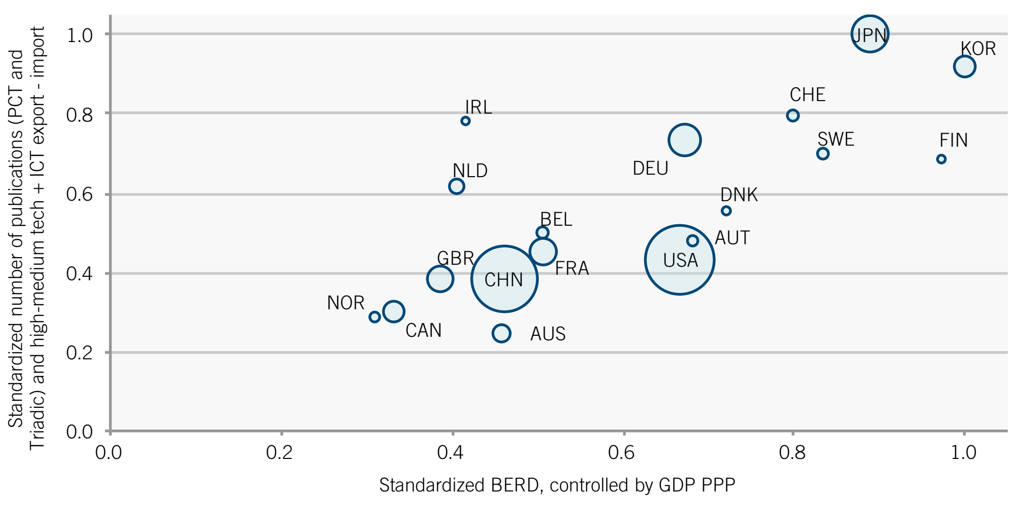
5. Technology model

To create a model that provides insight in the efficiency of a national technology system, we used the structure detailed in chapter 4. First we discuss all the issues that can be solved and present our 'reference model' to measure technology efficiency (paragraph 5.1). Paragraph 5.2 describes our 'robustness check' for variations on the reference model and in paragraph 5.3 we list the remaining disclaimers with regard to the model.

5.1 The reference technology model

Figure 7 summarizes the model for technology efficiency. The size of the blue dots represents country size (in GDP PPP). Just like with the science model, we decided to primarily focus on 'money' as means of input (BERD). Output is a combination of patent statistics and net export, both of equal importance. Since we have two patent variables and only one net export variable, we gave the three output indicators the following weights: (i) number of patent applications PCT [weight=0.25], (ii) number of triadic patent families [weight=0.25], and (iii) high tech, medium high-tech and ICT manufacturers net exports [weight=0.5].[18] As discussed in paragraph 3.1.3, all the indicators are first corrected for GDP (PPP, except export/import) and after that normalized based on maximum value.

Figure 7: Basic efficiency model for Technology



The results not only tell us how well the countries are performing in relative terms. Based on this specific model, we can say that especially Ireland is rather efficient in turning technology input into output, followed by the Netherlands, Japan, and Germany. The relation between the input and output of each individual country is depicted in Figure 7.

5.2 Variations on the technology base model

In order to develop the technology model, methodological choices had to be made, some of which are debatable. To assess the impact of our decisions on the model, we can adjust it slightly and compare this new model with our reference model. In Table 9 we present these choices and discuss how we can adjust the model to gain insight in the sensitivity of the reference model.

18 Since net exports can be <0, we had to recalculate every value in order to obtain only positive values. After all, net export should conceptually add to the technological output, but not 'punish' patent output.

Table 9: Choices made in the reference model for technology

Choices	Description
Content validity: **Selection of output indicator**	Although we started out with dozens of output indicators, by analyzing each indicator based on the nine different challenges, we ended up with just 'patents' and 'export minus import' as the best indicators for technological output. However, these two types of indicators are rather different (correlation < 0.2) thus raising the question whether we should combine them into one output-value. Their rather 'weak' relationship shows they are distinct concepts, and thus have different types of technological output (see also 2.2.3). **Model adjustments:** To gain insight in the impact of our decision to combine two types of output indicators, we also made two models focusing only on (i) patents and (ii) net export.
Content validity: **Most appropriate time lag between input and output indicators**	Just like the science model, time dimension is also an issue in the technology model. Similar to the science model, a two year lag was considered appropriate. This means that the 2012 patent and export data should be compared with the 2010 R&D expenditure. **Model adjustments:** We made a model with only a one year time lag.
Instrument validity: **What weights to use for the output indicator?**	In our reference model we combined three variables into one output indicator (two relating to patents and one relating to import/export). Since there were no reasons to attribute different weights, we decided to attribute equal weights to the two different types of technology outputs. Thus the two patent indicators have weight 0.25, and the net high and medium high tech export have weight 0.5. However, this choice is unfounded and other weights could be attributed. **Model adjustments:** We can show the impact of our decision using the two models we already made to demonstrate our 'selection of output indicator'. After all, the extreme opposite of equal weights is focusing the model on only one type of technology output.

Content validity:	
Selection of input indicator	In the basic model we used expenditure (monetary resources) as an input indicator, rather than human resources (see 2.2.2), which can also be a reasonable indicator (bearing in mind data quality issues). **Model adjustments:** To check the impact of selecting monetary resources, we also ran the model with R&D personnel as input indicator input.

Based on the four new models, we can analyze the sensitivity of the reference model. The efficiency is measured by dividing output by input. Just like in Table 7, all columns are normalized to an average of 1. Red and green arrows show when the ranking position of a country rises or falls at least four places.

Table 10: Sensitivity analysis of the reference model

	Basic model Technology	Only patents (PCT) as output			Only netto export as output			Time lag = 1 year		Business R&D personnel as input		
Ireland	1,96	0,70	↓	14,00	2,90		0,00	2,02	0,00	2,19		0,00
Netherlands	1,59	1,81		-1,00	1,41		0,00	1,61	0,00	1,44		0,00
Japan	1,17	1,61		-1,00	0,66	↓	11,00	1,14	0,00	1,28		0,00
Germany	1,13	1,12		3,00	1,06		2,00	1,11	0,00	1,25		1,00
Switzerland	1,04	1,19		-1,00	0,86	↓	6,00	1,08	0,00	1,26		-1,00
Belgium	1,03	0,82	↓	7,00	1,16		-2,00	0,98	2,00	1,16		0,00
United Kingdom	1,03	0,92	↓	5,00	1,06		0,00	0,99	0,00	0,97		2,00
Norway	0,98	0,99		0,00	1,08		-3,00	0,99	-2,00	0,84		2,00
Korea	0,96	1,13		-3,00	0,90		1,00	0,88	3,00	1,02		-2,00
Canada	0,94	0,95		0,00	0,99		-2,00	0,98	-2,00	0,53	↓	6,00
France	0,94	0,91		2,00	0,92		-1,00	0,94	0,00	0,80		2,00
Sweden	0,87	1,25	↑	-9,00	0,59	↓	5,00	0,89	-1,00	0,87		-3,00
China	0,86	0,42	↓	5,00	1,28	↑	-10,00	0,81	1,00	0,45	↓	4,00
Denmark	0,80	0,96	↑	-5,00	0,70		-2,00	0,83	-1,00	0,63		1,00
Austria	0,74	0,80		-1,00	0,70		-3,00	0,75	1,00	0,74		-2,00
Finland	0,74	1,18	↑	-10,00	0,47		3,00	0,76	0,00	0,81	↑	-4,00
United States	0,68	0,70		-2,00	0,63		-1,00	0,68	0,00			
Australia	0,55	0,53		-1,00	0,64		-3,00	0,58	0,00	0,73	↑	-4,00
Correlation with basic model	-	0,66			0,77			0,99		0,72		

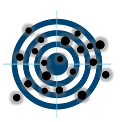
The four 'new' models are all highly correlated to the reference model. Especially adjusting the time lag to just one year does not affect the efficiency rate. This is not surprising, since the correlation between both models is >.99. More interesting is the model in which we used human resources as input instead of expenditure. This adjustment also does not have a significant effect on a country's efficiency ranking. Only Canada and China drop respectively six and four places, while Finland and Australia rise four places. The overall picture stays, to a certain extent, the same. This also applies to the adjustments on the output side. The variation in most countries' ranking is small. On the other hand, the adjustment has a major impact on five countries. Without export, Ireland falls 14 places, Finland rises 10 places, Sweden rises 9 places, and Belgium falls 7 places. Without patents, Japan drops 11 places and China rises 10 places. In other words, these countries perform very well in only one of the two types of indicators. Without one of them, their ranking changes radically.[19]

5.3 Remaining disclaimers for the technology model

Table 11 shows the remaining issues to be taken into account when drawing conclusions.

Table 11: Remaining issues with regard to the technology model

Issue	Description

Data collection reliability

Comparability of macro-based R&D statistics
	- A factor that may affect the validity of measurements for research efficiency is the potential problem with the comparability of national R&D statistics. For the countries ultimately selected, the following disclaimers apply:
	- BERD:
	- The values for three countries (Austria, Ireland, and Sweden) are ‘estimates/projections’.
	  - As discussed in paragraph 2.3.1, the Netherlands showed a remarkable rise in BERD in 2011. Since this was mainly due to a new definition of R&D[5], we recalculated the 2010 values based on the increase in 2011 that can be attributed to the less stringent definition of R&D [0.23*(14/18))+1 = 1.05].
	  - The USA excludes capital expenditures (“for the Business Enterprise sector, depreciation is reported in place of gross capital expenditures.”).
	  - Switzerland is interpolated.
	- Import/export of high tech / medium high tech:
	  - The extent to which a technology is indicated as medium/high tech is defined by sector, not the technology per se.
	  - It should be borne in mind that mirror flows often do not match between two countries. In other words, the export values from country A to country B (reported by country A) may not correspond with the import values to country B from country A (reported by country B). Although asymmetries exist for almost all trade flows, the differences observed may be relatively small.
	- Gross Domestic Product (Million current PPP$): see paragraph 4.3.

19 Performing well in export: Ireland, Belgium, and China. Performing well in patents: Japan, Sweden, and Finland.

Structural differences **Bias towards countries with high international output**	Large countries have a larger internal market. As a result, the need to export or have worldwide patent protection is less prominent. Since we used indicators with an international focus (PCT, triadic patents, export), this model has a bias towards smaller to medium sized countries with an open economy.
Specialization differences **Bias towards industries with a focus on patents**	Strategies with regard to intellectual property rights differ widely across industries, and so does the propensity to patent (Mansfield, 1986; Arora et al., 2008). This is due to the fact that the effectiveness of patents is linked to the specific characteristics of the technology and R&D process as well as the nature of the market and the patterns of competition (Orsenigo and Sterzi, 2010). Hence if industries appear relatively less in patent databases, this is not necessarily an indication that they are less knowledge-intensive but rather that they have other ways to protect their intellectual property. Having said this, as Mansfield already noted, even in industries where most of the inventions would be introduced without patent protection, at least half of the patentable inventions were still patented. Furthermore, even industries hitherto thought to attach relatively less importance to patents (such as software), value patent terms much more than previously assumed (Sukhatme & Cramer, 2014). However, the reference model still favors countries (with sectors) that prefer patents to protect their intellectual property.

6. Conclusions

Literature on National Innovation Systems (NIS) has gradually evolved from country-specific analyses, focusing on their idiosyncratic characteristics, to more general cross-country performance comparisons using composite constructs derived from a multitude of indicators. Although this aim makes perfect sense from a policy perspective, there might be a fine line between having bold ambitions and jumping to conclusions too soon. There tends to be a general disregard for the overall construct of the pitfalls inherent in the individual indicators, the scope they represent or broader comparability beyond the country context.

In this paper, we analyze the Science, Technology and Innovation systems of reference countries from a macro perspective with due care devoted to the selection, control, and deeper understanding of the indicators used in this model. As point of departure, we decided to move away from the holistic view of focusing on the input and output of the respective systems in isolation, but rather reflect on them collectively by looking at efficiency. In this chapter we sum up our main conclusions and add some recommendations.

[1] All nine categories of identified methodological issues apply with different degrees of impact and need to be consciously balanced to develop the most appropriate model. Any analysis that deals with international Science, Technology, and Innovation statistics should take many methodological issues into account as indicated in Chapter 2 of this theme paper. Nine categories of issues were identified, and by structurally assessing a broad variety of (200) indicators, we conclude that, for our model, the majority of indicators has to be discarded because of their unsuitability for inclusion. Those that remain need to undergo a conscious selection for inclusion. In this process, it is necessary to continuously weigh the scope, reliability, and comparability issues implicitly inherent in the (combination of) these indicators.

[2] The merits of the model will always be bounded by its pitfalls, requiring deeper understanding of its underlying constructs. The two reference models for science and technology were presented here as 'best-effort' models. Although we were able to control for factors like country size, price level, and carefully select the most reliable indicators, their expression power was still bounded by flaws like data collection differences between countries that exist in supranational data sources. Including or excluding GOVERD in the Science model is one telling example from paragraph 4.2 in which both alternatives cause distortion in the model (e.g. looking at the deviation in China's relative position).

Including more indicators in composite constructs, which is often the case in international comparative studies, does not automatically lead to a more reliable model; rather it can even exacerbate distortion due to stacked biases in underlying data, making it difficult to interpret the model as a whole.

Some pitfalls can be circumvented, some can be controlled, and others merely provide a proxy to measure input or output. In the end – there is no one-size-fits-all model (yet) that can capture the concept of efficiency in a uniform and undisputed way. Consequently, all models should be construed in the context of the (country specific) limitations of their underlying constructs.

[3] Efficiency is a disputable concept, but does provide a refreshed view on innovation performance. Even in a world where data could perfectly reflect the input and output of an innovation system, the concept of efficiency itself would still create problems in operationalization. The economics of marginal returns show that measuring efficiency is rather complex, for two reasons. Firstly, the relative 'position' of a country should be taken into consideration. Generally speaking, countries that are underperforming could probably increase their efficiency with relatively fewer resources than countries that are reaching their ceiling. Secondly, the ideal combination of output per unit of input might differ between countries (in cases where an increased output is impossible without an increased input as well). Countries that are perfectly able to achieve their potential, but are active in traditional sectors in which they have reached an optimum output / input ratio, given the profile of their country, cannot be directly compared with countries heavily involved in high-tech markets. On the other hand, we wish to point out that the concept of efficiency does bring a refreshed notion of productivity to bear. Most studies focus on output in which "the usual suspects", namely the thriving EU economies, outperform the smaller or less prosperous economies. From that point of view, a fresh look at performance provides a new perspective on the performance of innovation systems.

[4] The apparent absence of a clear correlation between the Science and Technology subsystems, paves the way for considering them separately. Considering the outputs in science and technology together (controlled for country size and price level), reveals the striking pattern that the two domains are divergent rather than convergent, as illustrated in chapter 3. With that notion in mind, we could decide whether to lump the two subsystems together or consider them separately. We chose the latter because the two systems appeared to differ considerably in their output; and we did not want to even out those differences by combining them. The price to pay for this division is two models with a more limited scope than one overall model for science and technology efficiency. Unfortunately, the innovation system lacks reliable indicators to enable us to consider them all together in an efficiency model.

[5] Dutch science and technology efficiency is above average. Based on our reference model, we can state that the Netherlands' science system has an above average efficiency compared to reference countries, and comes in 7th place. When considering a narrower scope of input, namely R&D personnel, its efficiency is even higher. Ireland, the United Kingdom and Switzerland are on average more efficient. The efficiency in the Technology system is less prone to deviations and the Netherlands is ranked here first or second compared to reference countries, after Ireland, but ahead of Japan and Germany.

[6] The science model is robust in its output constructs, but fragile on the input side. The science reference model is rather stable with regard to output. Its efficiency is only mildly affected by a holistic focus on the number of publications as output and even less so by introducing an alternative 'quality' indicator. A possible explanation is the fact that our reference countries are all positively biased on the selected output indicators. On the positive side: consequently, the model is rather robust with regard to output. The input side however does pose problems with the use of alternative indicators. Some methodological explanations are given in paragraph 4.2 and need to be kept in mind when interpreting the results of the presented alternative models. To improve the reliability of indicators, more emphasis should be placed on the actual implementation (in practice) of the data collection method of these indicators per country.

[7] Technology output is difficult to operationalize in a uniform way. In contrast with the science model, we observed a large difference between the efficiency model based on patents only or on net export of high and medium tech, ICT products, and services. A reflection on both individually showed that the two outputs are rather diverse, leading to different efficiency rankings. Unlike the division between science and technology, we decided not to split the technology model in two, because there is no rationale to dedicate BERD resources to either of them, so two models would be difficult to interpret for all countries. A closer look at the two models individually does reveal interesting differences between countries. Finland, Denmark, Sweden, Japan, and Korea benefit from a 'patents only' output model, whereas China and Ireland benefit most from the 'net export only' model. These deviations can be explained quite rationally by looking at the country profiles, but pose the problem of what adequate 'weights' should be given to the individual indicators in the reference model.

[8] Data quality and data collection reliability should be continually improved. In all our input and output variables, we identified indicator coverage biases (e.g. limited coverage of social sciences in WoS and Scopus) and data distribution biases (e.g. geographical assignment of patents). Although these biases cause shifts in the countries' rankings, both types are less severe than the problems arising from data collection reliability. This is because the first two issues introduce a structural bias: all countries are more or less affected in a similar manner – and if breaks occur, we can (at least partially) control for biases. A lack of data collection reliability on the other hand introduces more or less random distortions between countries. Bibliometric data is less affected by this problem because data is collected in a uniform manner and the definition of the core elements (e.g. 'publication', 'author', 'citation') is the same for all countries. MSTI statistics on the other hand are subject to data collection problems. For example, more than half the countries we used as cases in our models have at least one 'disclaimer' regarding STI expenditure data.

When MSTI data is used independently, the meta data supplied by OECD provides enough insight to interpret the differences between countries. That is also the reason why we decided to resort to basic techniques and keep the number of indicators in the model within manageable proportions. Unless the data quality is improved, it makes little sense to develop complicated models comprised of a multitude of indicators that are prone to distortion.

[9] Use quantitative models that explain NIS as a starting point for further analyses. Although quantitative models are useful for cross-country comparisons, they also come at a price: most country-specific details get lost along the way. Differences in NIS performance are explained in terms of rather sterile variables, whereas many of the discrepancies between countries could be attributed to their idiosyncratic characteristics and/or NIS. Systemic heterogeneity between countries should not be taken for granted. We are fully aware of the fact that we subsequently need to grasp which factors make a system 'tick'. We have to complement the insights gained from this type of macro-analysis (where we used the straight-jacket of a few macro-indicators), with more fine-grained policy analyses in our search for the variables that influence the correlation between input and output.

7. References

- Aksnes, D.W., Schneider, J.W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalized counting methods. *Journal of Informetrics*, 6(1), 36-43.
- Archambault, É., Vignola-Gagne, E., Cõté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68 (3), 329-342.
- Arora, A., Ceccagnoli, M., & Cohen, W.M. (2008). R&D and the patent premium. *International Journal of Industrial Organization, 26*, 1153-1179.
- Balzat, M. & Hanusch, H. (2004). Recent trends in the research on national innovation systems. *Journal of Evolutionary Economics, 14,* 197-210.
- Bandura, R. (2006), A Survey of Composite Indices Measuring Country Performance. Quoted in: OECD (2008), Handbook on Constructing Composite Indicators. Methodology and user guide. Paris: OECD.

- Carlsson, B., Jacobsson, S., Holmén, M. & Rickne, A. (2002). Innovation systems: analytical and methodological issues. *Research Policy, 31,* 233-245.
- Chen, C.P. & Yang, C. (2011) Chen, C. P., Hu, J. L., & Yang, C. H. (2011). An international comparison of R&D efficiency of multiple innovative outputs: The role of the national innovation system. *Innovation: Management, Policy & Practice, 13*(3), 341-360
- Dialogic, NIFU & CWTS (2012). Science, Technology & Innovation Indicators. Internationalization and specialization of the Dutch STI system. http://www.dialogic.nl/documents/2010.056-1235.pdf
- European Commission (2014). Innovation Union scoreboard report: 2014 edition.
- Falagas M.E., Pitsouni E.I., Malietzis G.A., and Pappas G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 2008;22(2):338-42. Epub 2007. PubMed PMID: 17884971; HLWIKI Canada. Scopus vs Web of Science 2013.
- Furman, J. L., Porter, M. E. and Stern, S. (2002) The determinants
- of national innovative capacity, Research Policy 31(6): 899–933.
- Hall, B.H. & Jaffe, A.B. (2012). Measuring Science, Technology, and Innovation: A Review. *Report prepared for the Panel on Developing Science, Technology, and Innovation Indicators for the Future,* National Academies of Science.
- Hardeman & van Roy, 2013. Hardeman, S., & Van Roy, V. (2013). An analysis of national research systems (II): Efficiency in the production of research excellence. Publications Office of the European Union.
- King, D.A. (2004). The scientific impact of nations: what different countries get for their research spending. *Nature, 430,* 311-316.
- Leydesdorff, L. & Wagner, C. (2009). Macro-level indicators of the relations between research funding and research output. *Journal of Informetrics, 3,* 353-362.
- Liou, D. (2009). Two-stage R&D Efficiency Evaluation from the Perspective of National Innovation System. *Proceedings of the 2009 PICMET conference in Portland, Oregon USA.*
- Lundvall, B-Å. (ed.) (1992). National Innovation Systems: *Towards a Theory of Innovation and Interactive Learning*, Pinter, London.
- Mansfield, E. (1986). Patent and Innovation: an empirical study. *Management Science, 32*, 173-181.
- May, R. M. (1997). The Scientific Wealth of Nations. Science, 275, 793-796.
- May, R. M. (1998). The Scientific investments of Nations. Science, 281, 49–51.
- Metcalfe, S. (1995), "The Economic Foundations of Technology Policy: Equilibrium and Evolutionary Perspectives", in P. Stoneman (ed.), *Handbook of the Economics of Innovation and Technological Change*, Blackwell Publishers, Oxford (UK)/Cambridge (US).
- Moed, H., Glänzel, W., & Schmoch, U. (2005). Handbook of Quantitative Science and Technology Research – The use of publication and patent statistics in Studies of S&T systems. Kluwer Academic publishers Dordrecht.
- Nasierowski, W. & Arcelus, F.J. (2003). On the efficiency of national innovation systems. *Socio-Economic Planning Sciences, 37,* 215-234.
- Nelson, R. R. (1993). National innovation systems: a comparative analysis.University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.
- OECD (1999) Managing national innovation systems. OECD, Paris.

- OECD (2002). *Frascati Manual: Proposed standard practice for surveys on research and experimental development*. Paris, France: OECD Publications Service.
- OECD (2013). Main Science and Technology Indicators http://stats.oecd.org/Index.aspx?DataSetCode=MSTI_PUB
- Orsenigo, L. and Sterzi, V. (2010). Comparative Study of Use of Patents in Different Industries. Bocconi University (KITes working paper 33/2010).
- OECD Oslo Manual (2005). Guidelines for collecting and interpreting innovation data (3rd edition)
- Peter, V. and Bruno, N. (2010). International Science & Technology Specialisation: Where does Europe stand? Paper prepared for the EC, DG Research (RFK2009).
- Porter, M.E., Delgado, M., Ketels, C., Stern, S. (2008) Moving to a New Global Competitiveness Index. In: World Economic Forum. The Global Competitiveness Report 2008-2009 (Chapter 1.2, pp.43-63).
- van Raan, A. F. J., van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics*, 88(2), 495-498.
- Rousseau, S. and R. Rousseau (1998), “The Scientific Wealth of European Nations: Taking Effectiveness into Account,” Scientometrics, vol. 42, no. 1, pp. 75-87.
- Sivertsen, G. (2009). Publication patterns in all fields. In Åström, F., Danell, R., Larsen, B., & Schneider, J. W. (eds.), Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th birthday. Leuven, Belgium: International Society for Scientometrics and Informetrics.
- Sukhatme, N.U. and Cramer, J.N.L. (2014). Who Cares About Patent Term? Cross-Industry Differences in Term Sensitivity. Princeton Dept. of Economics (SSRN2293245).
- van Leeuwen, T.N., Moed, H.F., Tijssen, R.J.W., Visser, M.S. and Van Raan (2001) A.F.J., Language biases in the coverage of the science Citation Index and its consequences for international comparisons of national research performance, Scientometrics, 51, 1, pp. 335-346.
- van Leeuwen, T.N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship ? Bibliometrie - Praxis und Forschung, 2013, 1-18
- Wendt, K., Aksnes, D. W., Sivertsen, G., Karlsson, S. (2012): Challenges in cross-national comparisons of R&D expenditure and publication output, paper for International Conference on Science and Technology Indicators (STI), Montreal, Canada, 2012.

Appendix A | Long-list of indicators

STI indicators
Finance
Total intramural R&D expenditure (GERD) (by source of funds)
BERD (by source of funds)
HERD (by source of funds)
GOVERD (by source of funds)
PNPERD (by source of funds)
GBAORD
Gross domestic expenditure on R&D (GERD) by performance sector and field of science
Gross domestic expenditure on R&D (GERD) by performance sector and socio-economic objective
Total intramural R&D expenditure (GERD) by type of R&D activity
Business enterprise expenditure on R&D (BERD) by type of R&D activity
Governmental expenditure on R&D (GOVERD) by type of R&D activity
Higher Education R&D expenditure (HERD) by type of R&D activity
Private non-profit sector on R&D (PNPERD) by type of R&D activity
Business enterprise expenditure on R&D (BERD) by economic activity (NACE Rev. 2)
Business enterprise expenditure on R&D (BERD) by economic activity (NACE Rev. 1.1)
BERD by economic activity (aggregated sectors, NACE Rev. 1.1)
BERD by economic activity (OECD ANBERD; includes estimates for unavailable data; does not cover all EU member states)
BERD by economic activity (aggregated, according to their technological intensity, NACE Rev. 2)
BERD by economic activity (aggregated, according to their technological intensity, NACE Rev. 1.1)
R&D personnel at national and regional level
R&D expenditure by foreign affiliates
Basic research expenditure as a percentage of GDP
Public research funded through institutional block funding
Public support to business R&D and innovation
Direct funding of business R&D and innovation
Government budget appropriations or outlays on R&D
Government budget appropriations or outlays on R&D - GBAORD according to NABS 2007
Government budget appropriations or outlays on R&D - GBAORD according to NABS 1992

GBAORD as a percentage of total general government expenditure
Innovation statistics (CIS2010)
Total number of employees in 2008/2010 - Enterprises with innovation activity
Total number of employees in 2008/2010 - Total enterprises
Enterprises with % of employees with university education - Enterprises with innovation activity
Total turnover in 2008/2010 - Enterprises with innovation activity
Total turnover in 2008/2010 - Total enterprises
General information about enterprises (market)
Enterprises by type of innovation
Product and process innovation
Innovation activities and expenditures in 2010
Public funding for innovation activities
Highly important sources of information for product and process innovation
Types of co-operation partner for product and process innovation
Highly important objectives for product and process innovation
Hampered innovation activities
Organizational and marketing innovation
Implementation type of a new organizational method
Highly important objectives for marketing innovation
In-house and external skills available in the enterprises
Methods for stimulating new ideas or creativity
Patents
Patents (Full counting)
EPO filings by sector of economic activity (Full counting)
EPO filings by technological field (Full counting)
EPO filings by IPC35 (Full counting)
USPTO grants by technological field (Full counting)
USPTO grants by IPC35(Full counting)
University Patents
indicators of international cooperation
Patents filed by universities and public labs (per GDP)
International co-patenting (PCT patent applications [%])
Patenting firms less than 5 years old (per GDP)

Share of domestic inventions that are foreign owned in the total number of patents
Share of foreign inventions that are domestically owned in the total number of patents
Patent citations
Share of patents filed by PRIs
Trademarks statistics
Trademark applications by time
Trademark applications to the Office of Harmonization for the Internal Market (OHIM) (by class, 1999-2012)
Bibliometric data
Number of publications by FP7
Number of publications by main field
Number of publications by NACE
Co-publications by FP7 thematic priority and partner country
Co-publications by main field and partner country
International co-publications by FP7 thematic priority
International co-publications by main field
Number of co-publications by partner's country
Single author publications by FP7 thematic priority
Single author publications by main field
Single country co-publications by FP7 thematic priority
Single country co-publications by main field
Unclassified co-publications by FP7 thematic priority
Unclassified co-publications by main field
Number of citations by FP7
Number of citations by main field
Number of citations by NACE
Number of top-10% most highly cited publications by FP7
Number of top-10% most highly cited publications by main field
Number of top-10% most highly cited publications by NACE
Average of relative citations (ARC) by FP7
Average of relative citations (ARC) by main field
Average of relative citations (ARC) by NACE
Average of relative citations (ARC) of all publications
Average of relative citations (ARC) of co-pubs

Average of relative citations (ARC) of international co-publications
Average of relative citations (ARC) of single author publications
Average of relative citations (ARC) of single country co-publications
Average of relative citations (ARC) of unclassified co-publications
Average of relative impact factors (ARIF) by FP7
Average of relative impact factors (ARIF) by main field
Average of relative impact factors (ARIF) by NACE
Specialization index (SI) by FP7
Specialization index (SI) by main field
Specialization index (SI) by NACE
Publications in the top-quartile journals (per GDP)
Science Citation Index (web-of-science)
Grant programmes
Incoming Marie Curie fellows per thousand researchers
FP6 & FP7 collaborative links per researcher
Higher education and human capital
Active population by age groups and highest level of education attained
Graduates/new entrants in ISCED 5-6 programmes
Number of students by level of education and region
Education expenditures by level of education
HRST and education variables
HR by age, sex and occupation
HR by sector of economic activity
Number of graduates by broad scientific field
Share of tertiary enrolment by age groups
R&D personnel by sector of employment and occupation
R&D personnel by sector of employment and qualification
R&D personnel by sector of employment and field of science
Business enterprise R&D personnel by industry
15-year old top performers in science (%)
Doctoral graduation rate in science and engineering
S&T occupations in total employment (%)
Total researchers (FTE), by sectors of performance

Share of non-citizen students that attend advanced research programs as a share of the total number of students that attend advanced research programs
Overall mobility flows
Mobility flows between different institutional sectors of the economy
Mobility flows that cross geographical (i.e. national) boundaries
Knowledge-intensive services
Employment NACE (1.1/2)
Localization quotient NACE (1.1/2)
High-tech statistics
Employment in high tech industries and knowledge-intensive services
High-tech industries
High-tech exports
High-tech imports
Venture capital investment by aggregated stage of development (1989-2006)
Venture capital investment by detailed stage of development (from 2007)
Co-operation, external control and knowledge sourcing
Indicators of co-operation
Indicators of external control and knowledge sourcing
EU Framework Programme indicators
FP data (FP6 and FP7 (up to late 2009)
EU Structural Funds indicators
Structural Funds (2000-2006/2007-2013)
Economic variables
Economic variables
Direct investment flows in high tech manufacturing
Direct investment positions in high tech manufacturing
Gross value added by industry
Number of local units
Sectoral gross fixed capital formation
Technological balance of payments
Volume of external trade: export (+share of total)
Volume of external trade: import (+share of total)
Innovation Union (IU) Scoreboard - enablers - human resources
Percentage youth aged 20-24 having attained at least upper secondary level education

Percentage population aged 30-34 having completed tertiary education
New PhD graduates (ISCED 6) per 1000 of the population aged 25-34
Innovation Union (IU) Scoreboard - enablers - Open, excellent and attractive research systems
International scientific co-publications per million population
Non-EU doctorate students per million population
Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country
Innovation Union (IU) Scoreboard - enablers - Finance and support
Venture capital (early stage, expansion and replacement) as percentage of GDP
Public R&D expenditures as percentage of GDP
Innovation Union (IU) Scoreboard - firm activities - firm investments
Non-R&D innovation expenditures as percentage of turnover
Innovation Union (IU) Scoreboard - firm activities - linkage & entrepreneurship
SMEs innovating in-house as % of SMEs
Public-private co-publications per million population
Innovative SMEs collaborating with others as % of SMEs
Innovation Union (IU) Scoreboard - firm activities - Intellectual Assets
PCT patents applications per billion GDP
Community designs per billion GDP
Community trademarks per billion GDP
PCT patent applications in societal challenges per billion GDP
Innovation Union (IU) Scoreboard - outputs - innovators
SMEs introducing marketing or organizational innovations as % of SMEs
SMEs introducing product or process innovations as % of SMEs
Fast-growing innovative firms
Innovation Union (IU) Scoreboard - outputs - economic effects
Knowledge-Intensive Services exports as percentage of total service exports
Sales of new to market and new to firm innovations as percentage of turnover
Contribution of high- and medium-tech products to the trade balance as percentage of the trade balance
License and patent revenues from abroad as percentage of GDP
Employment in Knowledge-Intensive Activities (manufacturing and business services) as % of total employment
European Research Area (ERA) performance indicators

New academic oriented tertiary education degree (ISCED 6 and 5A) graduates per 1000 population aged 25-34 - Total [per 1000]
New PhD graduates (ISCED 6) per 1000 of the population aged 25-34 - Total [per 1000]
National citizens with a doctorate having lived/stayed abroad in the past ten years [Percentage]
Number of researchers by citizenship [Absolute value]
Amount of public (government+EC) funding of R&D as percentage of GDP
National public funding allocated as project-based funding as percentage of GBAORD
Share of national public funding to trans-nationality coordinated research as percentage of GBAORD
SF allocations on core RTDI as a percentage of all SF allocations
R&D fiscal incentives as percentage of GDP
Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country
Employment in Knowledge-Intensive Activities (manufacturing and business services) as % of total employment
PCT patent applications in societal challenges per billion GDP (in PPS€) (climate change mitigation; health)
Comparative performance of science and innovation systems
Top 500 universities (per GDP)
Top 500 corporate R&D investors (per GDP)
Ease of entrepreneurship index
internet use for innovation
Fixed broadband subscribers (by population)
Wireless broadband subscribers (by population)
Networks (autonomous systems) (by population)
E-government readiness index
Key figures
Labor productivity
Environmental productivity
Overview of national innovation policy mix
Public research university-centered
Public research basic research oriented
Public research civil oriented
Public research generic research
Revealed technology advantage (RTA) in selected fields

RTA
IPP.Stat
Fiscal environment and tax incentives
Legal and regulatory business environment

Dutch Summary

Onderzoeken op het gebied van Nationale Innovatie Systemen (NIS) zijn geleidelijk geëvolueerd van specifieke kwalitatieve analyses op nationaal niveau naar bredere kwantitatieve internationale vergelijkingen. Het vergelijken van NIS (waaronder het opstellen van landenranglijsten) kent een aantal methodologische uitdagingen. Deze uitdagingen cq. valkuilen krijgen meestal weinig aandacht. In deze studie behandelen we voor het eerst op een gestructureerde wijze de belangrijkste issues die spelen. Deze zijn verdeeld over drie hoofdcategorieën:

- **Scope**: Hoe worden indicatoren die fundamenteel van elkaar verschillen gecombineerd in één model?
- **Kwaliteit van de gegevens**: Hoe wordt omgegaan met data waarvan de kwaliteit niet optimaal is?
- **Vergelijkbaarheid**: Hoe gaan modellen om met de diversiteit in de populatie? Dat wil zeggen in hoeverre houden ze rekening met specifieke eigenschappen van landen?

Veel ranglijsten richten zich op input-variabelen en/of output-variabelen. Hierdoor scoren de 'usual suspects' (de landen met de hoogste BBP's) veelal goed. Het is echter niet zo moeilijk om met veel input veel output te genereren. Vanuit het oogpunt van een doeltreffende besteding van publieke middelen is het zaak om met zo min mogelijk input zoveel mogelijk output te behalen. In deze studie maken we daarom een vergelijking tussen landen op basis van rendement. We laten daarbij zien (i) welke methodologische tekortkomingen er veelal zijn, (ii) hoe met deze tekortkomingen omgegaan kan worden en (iii) hoe de uiteindelijke uitkomsten geïnterpreteerd moet worden (rekening houdend met de tekortkomingen).

Het vertrekpunt is een longlist van circa 200 wetenschap-, technologie- en innovatie-indicatoren. Elk van deze indicatoren is beoordeeld op basis van de checklist van (negen) issues. Alleen de indicatoren die voldoende scoren op de drie hoofddimensies scope, datakwaliteit of vergelijkbaarheid zijn gebruikt in de verdere ontwikkeling van het model. Deze exercitie zorgt niet alleen voor een fikse indikking van de lijst met indictoren maar brengt ook de noodzaak naar voren om twee fundamentele keuzes te maken:

1. Innovatie-indicatoren zijn weggelaten uit het model.
2. Er is een duidelijke onderscheid gemaakt tussen de subsystemen Wetenschap en Technologie.

Uiteindelijk zijn er twee basismodellen opgesteld. Die beschrijven hoe efficiënt de Nederlandse wetenschaps- en technologiesystemen presteren ten opzichte van die van vergelijkbare landen. Vervolgens is de robuustheid getest en zijn de modellen verbeterd.

Op basis van de algemene analyse naar mogelijk tekortkomingen in NIS-ranglijsten en de ontwikkeling van de twee eigen efficiëntiemodellen komen we tot de volgende conclusies:

1. Het opstellen van een ranglijst van NIS-ranglijsten en van NIS- efficiëntiemodellen is methodologisch gezien bepaald geen sinecure. Als daarbij niet of onvoldoende rekening wordt gehouden met de issues die wij in deze studie behandelen, zullen de uitkomsten meestal minder betrouwbaar zijn.
2. Elk NIS-model bevat mede daarom bijna altijd disclaimers. Deze voetnoten worden vaak niet gelezen maar zijn van essentieel belang bij de juiste verwerking en interpretatie van de uitkomsten.
3. Het onderling vergelijken van de efficiëntie van NIS moet ook zorgvuldig gebeuren. Zowel de uitgangspositie als de optimale verhouding tussen output en input verschilt per land. Toch geeft de notie van efficiëntie nuttige aanvullende informatie over de prestaties van een NIS.
4. In termen van efficiëntie vertonen wetenschaps- en technologiesystemen een heel ander gedrag. Ze moeten daarom ook als aparte subsystemen geanalyseerd en beschreven worden.
5. Op basis van de uitkomst van onze modellen kan worden geconcludeerd dat zowel het Nederlandse wetenschapssysteem als het technologiesysteem relatief efficiënt zijn ten opzichte van de systemen van vergelijkbare landen.
6. De keuze voor andere outputvariabelen heeft weinig invloed op de uitkomst van het wetenschapsmodel (dat wil zeggen het model is robuust). De keuze van andere inputvariabelen (zoals alternatieve definities van financiering) heeft wel significante invloed op de uitkomst.
7. Voor het technologiemodel geldt dat het, vanwege de diversiteit tussen de subsystemen van landen, moeilijk is om een uniforme set van outputvariabelen te definiëren die recht doet aan de specifieke kenmerken van een land (met name sterke gerichtheid op export of op patenten).
8. De huidige kwaliteit van internationaal vergelijkbare data over wetenschaps- en technologiesystemen laat nog veel te wensen over. Er zijn aanzienlijke verbeteringen nodig, bijvoorbeeld in de uniforme verzameling en verwerking van statistieken. Gegeven de problemen met de datakwaliteit heeft het op dit moment nog weinig zin om complexe modellen met samengestelde indicatoren te gaan ontwikkelen.
9. De meerwaarde van kwantitatieve internationale vergelijkingen is dat ze een goede eerste indruk kunnen geven de bredere trends die er spelen. Ze zijn een bruikbaar startpunt voor verder (diepgaander) specifieke kwalitatieve analyses op nationaal niveau.