European Commission

JRC SCIENTIFIC AND POLICY REPORTS

Assessment of the Web Index

Survey questionnaire calibration and uncertainty analysis

Annoni P. , Weziak-Bialowolska D. and Nardo M.

2012

Report EUR 25476 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for Security and Protection of the Citizens

Contact information
Paola Annoni, Dorota Weziak-Bialowolska, Michela Nardo
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 361, 21027 Ispra (VA), Italy
E-mail: paola.annoni@jrc.ec.europa.eu; dorota.bialowolska@jrc.ec.europa.eu; michela.nardo@jrc.ec.europa.eu
Tel.: +39 0332 78 5968
Fax: +39 0332 78 5733

http://composite-indicators.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

This publication is a Reference Report by the Joint Research Centre of the European Commission.

JRC73977

EUR 25476 EN

ISBN 978-92-79-25988-3
ISSN 1831-9424

doi:10.2788 /43830

Luxembourg: Publications Office of the European Union, 2012

Printed in Luxembourg

ASSESSMENT OF THE WEB INDEX

Annoni P. , Weziak-Bialowolska D. and Nardo M.
European Commission, Joint Research Centre
Econometrics and Applied Statistics Unit

Survey questionnaire calibration and uncertainty analysis

Contents

Executive summary

The purpose of this analysis is a comprehensive assessment of the Web Index 2011 (WI), published by the World Wide Web Foundation in September 2012. The WI aims to measure the *state and value of the Web* focusing on the impact of the Web on people and nations. The Index covers 61 countries worldwide and consists of 85 underlying indicators across seven components and three sub-indexes. Primary data, coming from an ad hoc expert assessment survey, and secondary data coming from official datasets are combined in the WI.

The usage of primary data is one of the innovative aspects of the first release of the WI. They play a remarkable role in the construction of the composite indicator as they account for about 60% of the WI indicators. They are sourced via an expert assessment survey and reviewed by national peers. Given that the expert assessment survey has been specifically designed for the first release of the Index, the analysis of the survey outcomes is of particular importance. To this aim a statistical model designed for the analysis of survey data is employed. Based on the model outcomes we provide suggestions on how to improve data gathering in future surveys.

The second part of the analysis contains the robustness analysis of the WI. Every composite index is the result of a number of choices on the framework, the number and identity of indicators to include, their normalization, the weights to attach to each indicator and component, the aggregation method and many others. As with every composite index, some choices are openly normative and subjective, driven by developers' and experts' opinion, others can be justified on the basis of statistical analysis, mathematical simplicity or common practice. The uncertainty analysis presented in this study aims at assessing to what extent these choices might affect the country scores and ranks based on the composite indicator. To this purpose six alternative scenarios are simulated each challenging one particular assumption made in the WI. The assessment of different scenarios is always done taking the official WI index, version 2011, as the reference one. In uncertainty analysis of composite indicators country rank volatility is generally caused by the country scoring relatively high in some indicators/components and low in others. Our analysis shows no cases of remarkable volatility. There are some countries with relatively high volatility for some scenarios. They are likely to feature as a sort of unbalance of scores in the different WI indicators/components.

Analysis of survey data

Primary data are the backbone of the WI. The survey consists of a detailed questionnaire submitted to the experts/professionals from 61 countries worldwide and assessed by national and regional peer reviewers. Designing a questionnaire is generally a difficult task. The WI case is particularly challenging given the complex nature of the topic surveyed and the wide coverage required. Our analysis of primary data aims at providing survey designers with some insights into possible problematic questions and/or unexpectedly behaving countries. A specific model belonging to the family of the Rasch models is employed. Rasch analysis is a statistical measurement tool originally conceived as a psychometric method for the social sciences and designed for the treatment of survey data. The analysis of WI primary data allows us to check for a series of issues: category redundancy, questions' unexpected answers , questions' relative difficulty and the validity of the selected framework. Results show that the questionnaire is balanced and the response structure organised in a ten-category scale is always appropriate. Few questions stand out as problematic: Q10 (To what extent does the government impose restrictions on access to websites (censorship)?), Q25 (Does the government have a specific Open Government Data initiative?), Q2a (Do the main political parties have websites?) and Q12 (To what extent do you think that the Web is making it easier to undertake criminal activities in your country?). Some of these questions do not seem to be clear enough for the respondents, while others appear to be too technical or counter-oriented with respect to the concept under measurement. The general suggestion for all of these questions is a rephrasing to make them clearer. No country shows a notable unexpected pattern of answers, confirming that the questionnaire was always scored by experts with their best efforts. Question difficulty is almost always as expected with a clear indication that gender bias does matter. Finally, survey data describe an almost unique factor in each WI component, as supported by the Rasch dimensionality analysis. This means that the grouping of the different survey indicators into different WI components is statistically appropriate.

Uncertainty analysis.

Scenario 1. Weighting. Weights assigned to each component/sub-index of the WI are changed for checking the volatility of scores/ranks with respect to the reference WI. Very extreme configurations are also tested by choosing a wide range of variability for the simulation weights. Overall the WI is not highly affected by the change in weights confirming the robustness of the Index with respect to the reference weighting structure. Equal weighting either at the sub-index level or at both component and sub-index levels is also tested and shows a maximum shift of 5 positions in the ranking. Iceland, Argentina and Namibia would be the most favoured countries if equal weighting were used for the WI. With more extreme weighting scenarios, distant from the reference one, the most affected countries would be Switzerland, Ireland, Singapore, Colombia, Poland, China and Russia, with shifts in rank higher than 10% of the maximum possible shift.

Scenario 2. Different aggregation for three indicators. The Communications Infrastructure component is meant to capture if people can (easily) access the Web, not how it is accessed. In order to take into account different access modalities for different countries, we adopt an alternative way to aggregate some of the indicators describing web access in the WI and check the impact on country scores and ranks at component, sub-index and Index levels. The WI is almost not affected by the change in the way Web access is included in the Index. A modest volatility in ranks is observed for the sub-index Readiness and the component Communications Infrastructure. For the Readiness sub-index differences in ranks are at most of 2 positions for Uganda (downward in the WI scale) and 3 positions for Pakistan (upward). In the case of Communications

Infrastructure the maximum shift amounts to 5 for Tunisia and 4 for China, they would then gain some positions.

Scenario 3. Inclusion of four additional indicators. The Institutional Infrastructure component of WI contains a set of indicators designed to describe possible gender biases in the access and use of the Web (*gender indicators*). In particular two indicators describe implicit gender bias in computer training and in focusing on science and technology expressed as a "distance" between respective levels for girls and boys. In order to take into account also the level of these indicators, four additional indicators are added to the Institutional Infrastructure component which measure the level of computer training and focusing on science and technology among girls and boys respectively. The addition has almost no effect on the final results. The highest observed difference in the WI ranking is of 1 position only. As expected, the volatility increases when the sub-index and the components are concerned. The biggest observed differences in the sub-index Readiness are of 4 (Morocco) and 3 (Benin) positions, while in the component Institutional Infrastructure the highest shift is of 5 positions (Ecuador and China).

Scenario 4. Different treatment for survey data. In the Index computation primary and secondary data are treated in the same way: after a statistical preliminary transformation, they are normalised and then aggregated across components and sub-indexes. In this scenario a different method is used to derive *'numbers'* from survey data, i.e. the Rasch method employed also for the overall analysis of the survey data. The replacement in the WI of the original survey indicators with the new statistically quantified indicators turns out to be the biggest challenge to the WI structure as the structure of four out of seven components are partially altered due to technical reasons related to the use of the Rasch model. Still the comparison between the reference WI and our simulations shows a rather robust Index: the largest changes are those for Australia and Philippines with a modest improvement of 4 positions in the WI ranking and by Singapore, Iceland and Benin which decline by 4 positions. A much higher ranking volatility can been seen at the sub-index and component level especially for the Web Content component where Indonesia could drop 14 positions in the WI ranking while Bangladesh and Ecuador would climb by 16 positions and South Africa by 13.

Scenario 5. Compensability. Can high web use or a high social impact compensate for poor institutional or communications infrastructure? The aggregation used in the WI assumes it can, as poor performances in some sub-indexes (components) are linearly compensated by good scores in others. We test a different aggregation where *bads* are less easily compensated by *goods*. WI passes the test easily: no country scores relatively high in some components and low in others so compensability does not seem an issue with this dataset.

Scenario 6. The contribution of each component and sub-index. In this scenario the contribution of each component to the Index is assessed by excluding one component at a time and comparing scores/ranks to the reference ones. Our analysis highlights the Political, Social and Economic Impact components as the three most influential ones, while the least influencing one turns out to be the Communications Infrastructure component. This reflects the weighting scheme of the WI where 60% of the overall weight is assigned to the sub-index Impact.

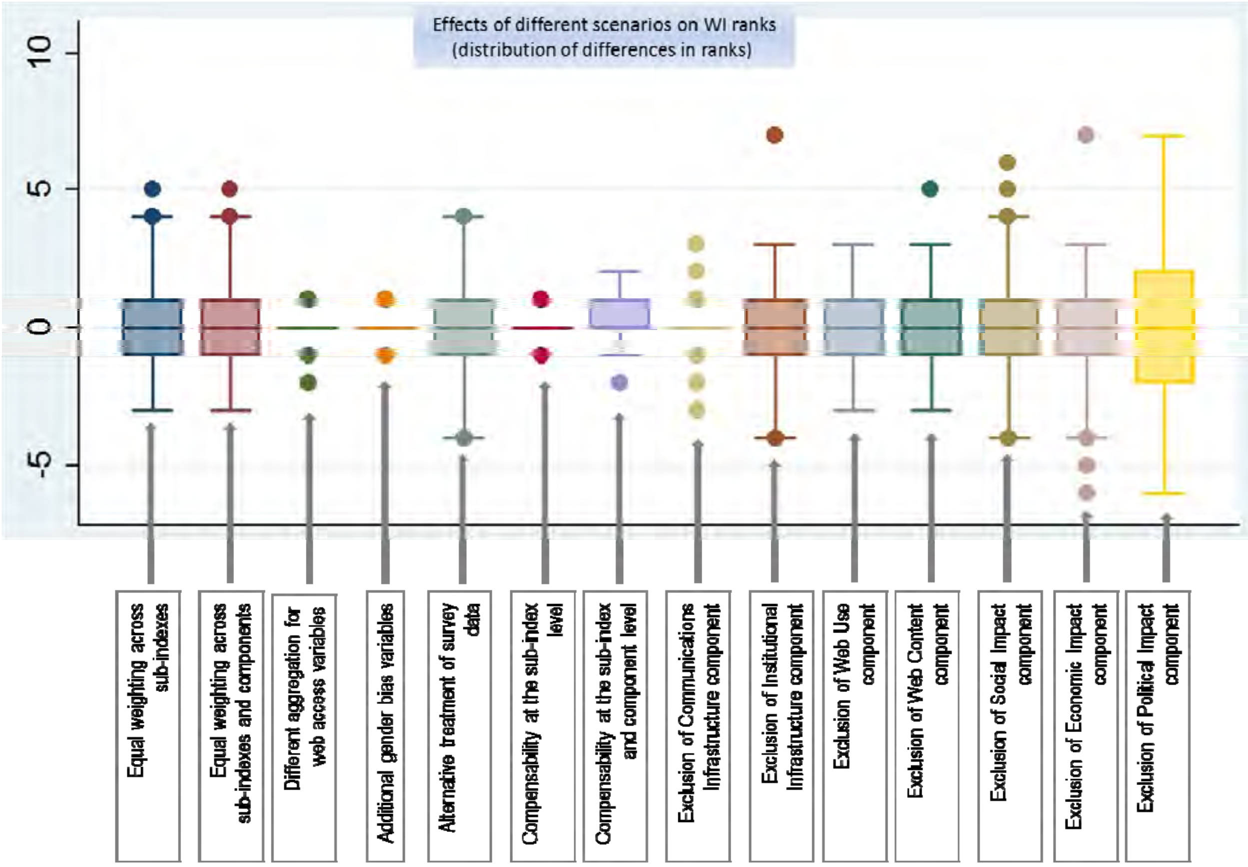
The correlation pattern of the WI is also tested. The weights assigned by developers to different sub-indexes and components, with the aim of attributing to these a pre-established scale of importance, are compared with the importance the same sub-indexes and components have as measured by a statistical measure. Our analysis finds the following:

Within Components: In the Communications Infrastructure indicator ITUG (% of population covered by a mobile cellular network) is much less important than what the weight assigned to it by the World Wide Web Foundation would suggest. The same happens in the Institutional infrastructure to the indicators WEFF (Burden of government regulation), Q9l (In your country, in tertiary education, what proportion of ICT graduates are women?), Q10 (To what extent does the government impose restrictions on access to websites?), Q16 (To what extent would you consider your country to be ranking amongst the World's best in training computer engineers?), Q25 (Does the government have a specific Open Government Data initiative?) and the cluster of Q9a-Q9d on gender bias.

We notice that the indicators WEFF, Q9l, and Q9a-Q9d are not significantly correlated with the WI components. They seem to follow a different behaviour as compared with all other indicators in the dataset. The same happens for Q12 (To what extent do you think that the Web is making it easier to undertake criminal activities in your country?) and to some extent also for WBC (ICT service exports as a share of GDP) in the component Economic Impact. These indicators count much less in the composite than the weight theoretically assigned to them.

Within sub-indexes. All the components and sub-indexes scores are highly correlated among themselves and with the WI. This means that whatever weights are assigned to the components or the sub-indexes the change in the WI is only marginal (as proved by our first scenario). Although to the sub-index Impact is assigned 3/5 of the overall weight, it actually weights much less being extremely correlated with the other two sub-indexes. In other words, the WI is not really "multi" dimensional as all components look pretty much the same from the statistical point of view. If the correlation structure is confirmed in other editions of the Index, there might be room for a reduction in the number of indicators included in the WI framework.

The overall picture of the effect of different tested scenarios on country ranks is shown in Box 1.



Box 1: Comparison of different scenarios on country ranks

1 Introduction

The Web Index (WI), developed by the World Wide Web Foundation, aims at measuring the *value and state of the Web* focusing on the impact of the Web on people and nations. The Index is computed for 61 countries worldwide and consists of 85 indicators across 7 components. The components describe available communication and institutional infrastructure (e.g., fixed and mobile Internet access, electrification rate, government regulation, government attitude and policies related to ICT usage and Web access, etc.); Web content and use (topics, languages, quantities, policies, practices, accessibility) and how use of the Web affects the economic, political and social fabric of each country (including the use of social networks, distance learning, impact of ICT on organizational models, etc.). All indicators have a positive polarity with respect to the *value and state of the Web*, that is the higher the better. Components are combined into 3 sub-indexes measuring the readiness of each country to take up the challenge offered by the Web; the extent of Web development and the impact observed so far. Finally the 3 sub-indexes are combined into the composite WI. The aggregation entails an equal weighting scheme within each component, but a differentiated weighting scheme for some components and the three sub-indexes[1]. Figure 1 shows the WI structure with the weights associated with the 7 components and the 3 sub-indexes.

The WI combines primary (survey) with secondary (hard) data and provides a score for each of the 61 selected countries worldwide. Primary data are obtained through country questionnaires originally including 63 questions (including main questions and components) each on a scale of 10 categories, with category “1” describing the lowest *value and state of the Web* and the “10” category describing the highest *value and state of the Web*. Secondary data are obtained from already existing sources such as ITU World Telecommunication Indicators Database, United Nations and the World Bank databases and other internationally recognized sources.

[1] For a detailed description see the World Wide Web Foundation website at http://www.webfoundation.org/projects/the-web-index

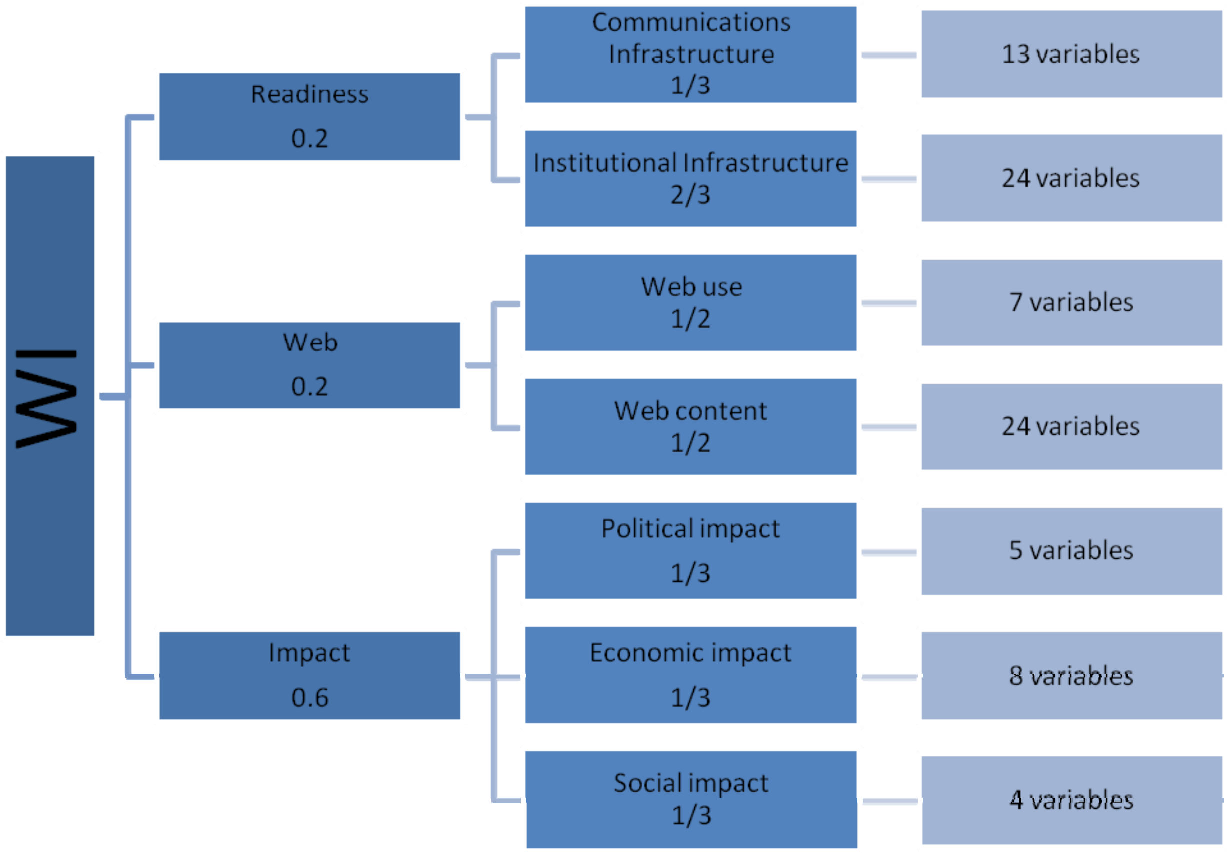


Figure 1. Web Index (WI) structure.

The aim of this analysis is to assess the quality of primary data, to possibly calibrate the questionnaire for future surveys, and to verify the robustness of the WI to the most important assumptions used for its construction. To this purpose the 2011 edition of the Index is analysed, which is the only one including both primary and secondary data.

Primary data are the backbone of the WI and come from an expert assessment survey designed by WI developers for the first release of the index. The survey consists of a detailed questionnaire submitted to experts/professionals from 61 countries worldwide and assessed by national and regional reviewers and by the Index developers themselves. Designing questionnaires is generally a difficult task and this is particularly valid in the WI case, given the complexity of the Index framework and its spatial coverage.

The first part of the analysis aims at providing the survey designers with some insights into possible pitfalls in the questions and/or strangely behaving countries. To this purpose a specific statistical model belonging to Rasch analysis, the Rating Scale model, is employed. Rasch analysis is a statistical measurement tool originally conceived as a psychometric tool for the social sciences and widely applied mostly in the medical and educational context. It consists of a family of statistical models designed for the treatment of qualitative data, like data collected in the surveys. Different models are available for different types of qualitative variables and the Rating Scale model is the most appropriate in the WI case. With Rasch analysis we were able to check for category redundancy in the questions, unexpected answers in some questions and/or by some countries, the relative difficulty of the questions and the validity of the framework, at least as far as survey data are concerned.

The second part of the study is focused on the robustness assessment of the Index. In every composite indicator the final index is the result of a number of choices on the framework, the number of available variables[2] to include, their normalization, the weights to attach to each variable, the aggregation algorithm, etc. Some of these choices are generally subjective, some others driven by statistical analysis, mathematical simplicity, experts' opinion or common practice. The aim of robustness analysis is to assess to what extent all these choices (or some of them) might affect the score/ranking of the composite indicator. In order to disentangle influential factors and fully understand the implication of their variability into WI, we chose to focus on some critical key assumptions of WI architecture, also in agreement with WF developers, and to check the overall influence of each of them on the Index. The check is usually done by calculating the differences in scores/ranks between the reference scenario (i.e. the reference scores and ranks provided by World Wide Web Foundation) and the "simulated scenario", i.e. the scenario including one of the alternative hypotheses. Ideally we would like to have an Index which is robust to the different assumptions made on the weighting structure, well balanced in terms of components and sub-indexes contribution to the overall score, and where survey and hard data are properly treated from the statistical point of view.

To stress the Index under different conditions, six alternative scenarios are simulated: the first scenario challenges the set of weights given to each component and each sub-index. To this purpose a Monte Carlo experiment is set-up to assess the impact of attaching different weights to the sub-indexes of the WI. Additionally equal weighting is adopted at the sub-index level and at both sub-index and component level. The second scenario tests a different way to cluster three indicators describing web access in the Communications Infrastructure component. The third scenario checks the inclusion of four additional indicators describing gender imbalance in the Institutional Infrastructure component. Scenario 4 evaluates the use of a different quantification procedure for

[2] The words "variable" and "indicator" are used as synonymous hereafter.

including survey data, based on Rasch analysis. Scenario 5 tests compensability issues by using a geometric aggregation instead of the linear one. Finally the last scenario provides an overall evaluation of WI by checking the role of each component with respect to the overall Index and the correlation structure of the Index.

The report is organised as follows. Section 2 describes the analysis of the survey questionnaire using Rasch analysis and major outcomes, while the robustness analysis with the detailed discussion of six different scenarios is presented in Section 3. Section 4 concludes.

2 Rasch analysis of primary data

Primary data are the backbone of the Web Index (WI) and were collected, for the first release of the Index, by means of an expert assessment survey. The survey consists of a detailed questionnaire answered by experts/professionals from 61 countries worldwide and reviewed by national peers. Each question is on a ten-category scale where category "1" is always the worst and category "10" always the best. For all questions no further specification is provided for intermediate categories. All the questions are designed by developers to have positive orientation with respect to the latent concept of the *state and value of the Web* (the higher, the better).

A specific statistical model, the Rating Scale model, belonging to the family of the Rasch models is employed to further investigate results from the expert survey with the purpose of both highlighting possible improvements of the questionnaire for the next WI releases and detecting specific behaviours both for countries and questions.

Rasch models allow for the construction of quantitative measures from qualitative data measured on a dichotomous or ordinal scales (Smith and Smith, 2004). Rasch scores are therefore a way to statistically quantify survey responses. Raw counts computed as the percentage of people scoring a certain category cannot be relied upon to serve as measures while the Rasch models have been designed to construct proper inference from observations of this kind (Bond, Fox 2007, pg.24).

In our analysis Rating Scale model is applied. For each of the following components a Rating Scale model is set up[3]: Institutional Infrastructure (14 questions); Web Content (22 questions grouped into three sub-components[4]); Web Use (6 questions); Economic Impact (4 questions) and Social Impact (3 questions). The components Physical Infrastructure and Political Impact are not included in the analysis due to the small number of questions included in the survey (two each). A total of seven Rasch models are run for a subset of 49 WI primary variables.

For each analysed component, Rasch country scores are eventually used as a scenario in the uncertainty analysis of the WI as alternative to the inclusion of raw primary data (see Section 3.4).

---

[3] Rasch models are run by means of Winsteps software package - ver. 3.74 (Linacre, 2012).

[4] Derived questions Q9a-b and Q9c-d could not be included in the Rasch model because of technical reasons described later in this Chapter. The list of survey questions is in the Appendix.

2.1 Rationale for Rasch analysis

Rasch analysis is a measurement tool originally conceived as a psychometric tool for the social sciences and widely applied mostly in the medical and educational context. It consists of a family of statistical models designed for the treatment of qualitative data, like data collected in surveys. Different models are available for different types of qualitative variables. The simplest Rasch model can handle dichotomous variables of the type "yes/no", "present/absent" or "right/wrong". Extensions of the simplest, binary model are available for variables with finer gradations of outcomes (categories) such as "Strongly Disagree/Disagree/Neutral/Agree/Strongly Agree", also called Likert items (Wright and Masters, 1982; Fischer and Molenaar, 1993, Smith and Smith, 2004). When the measurement scale is the same for all the variables the Rating Scale model is the most appropriate and it is used for the WI case (Smith and Smith, 2004).

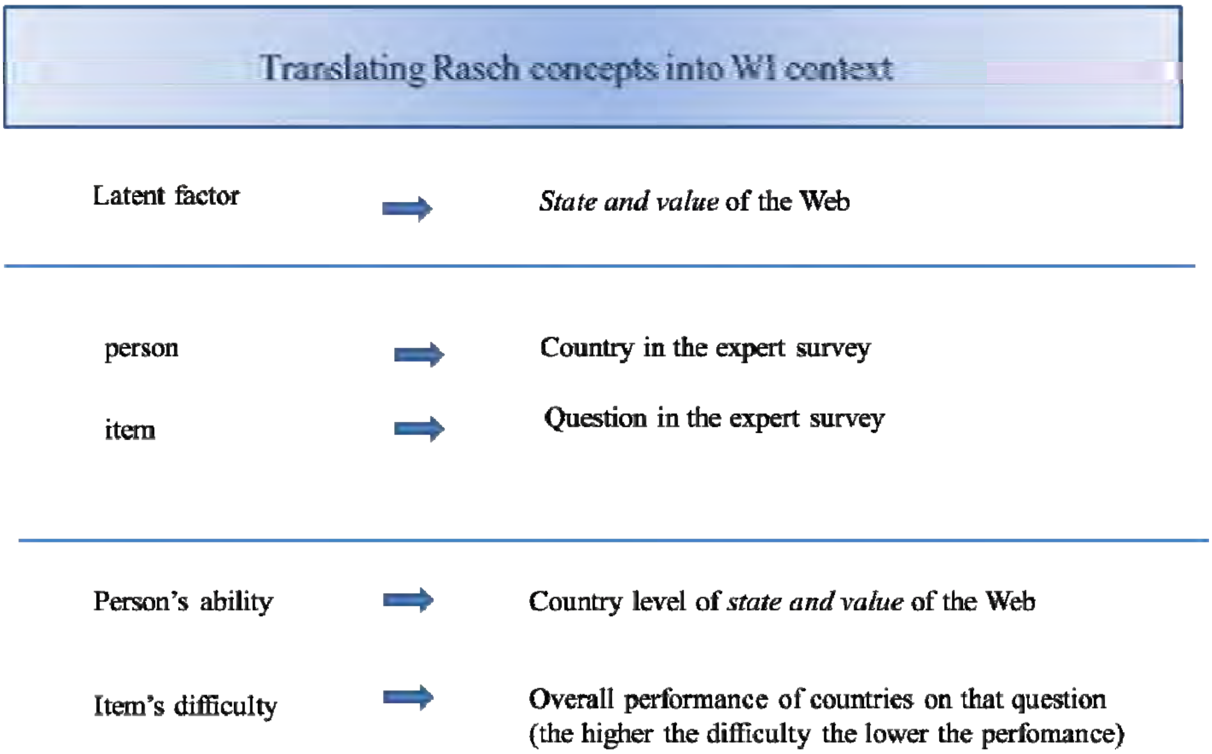
The Rasch approach is generally applied to the measurement of human performance, aptitude or perception when observed raw data are derived from commonly used surveys including Likert items representing a respondent's increasing inclination towards the concept questioned.

The starting point of the Rasch analysis is always a data matrix with ordinal scores obtained by a set of *persons* on a set of tasks (questions, physical or intellectual tests, etc…) called *items*. Persons and items are the two key elements in the Rasch terminology and the goal of the Rasch modelling is simultaneously estimating persons' *ability* and items' *difficulty* and their associated standard errors. To this aim, the Rasch approach provides statistical models to estimate the probability of a person with a certain ability succeeding on an item of a certain difficulty. The two key characteristics of persons and items are then ability θ and difficulty β respectively.

Since individuals are usually involved, ability and difficulty were originally conceived to refer to human perceptions. It is therefore necessary to fit these concepts in the context under investigation. In the WI case the latent concept, which is assumed to drive responses, is the *state and value of the Web*. Countries play the role of individuals and country's ability reflects the overall level of a certain aspect measured by the WI in that country. Questions included in the WI expert survey are the items. They are designed to capture different aspects of the WI in the analysed countries. The correspondence between the Rasch analysis terminology and the WI case is illustrated in Box 2.

In terms of Rasch parameters, questions which frequently score higher are associated with a lower difficulty – they are on average “easy” – while questions scoring on average lower are more difficult. In the Rasch models the performance of low ability countries is expected to be low in the most difficult questions, while high ability countries are expected to score high even on difficult questions. This basic assumption is linked to the concept of misfit of items (questions) and persons (countries) that will be shortly discussed.

Estimated countries ability θ_i and associated standard errors can be eventually used as country score. In the WI case even more interesting is the parameter on question difficulty β_j and its associated diagnostic which allows for the assessment and refinement of the survey questionnaire.



Box 2: Correspondence between the Rasch analysis concepts and the WI case

Box 3 provides a short technical description of the Rating Scale model, for further reading refer for example to Wright and Masters (1982) or Andrich (1988).

Basics of the Rating Scale model

The Rating Scale model is one of the extensions of the dichotomous model and handles cases with items on a multi-category scale equal for all the items, like the WI case.
In the RC model for WI data, the probability of observing category *h* for variable *j* on country *i* is modelled as:

$$P\{X_{ij} = h; (\theta_i, \beta_j, \delta_h)\} = \frac{\exp\{\theta_i - (\beta_j + \delta_h)\}}{\sum_{h=1}^{k_j} \exp\{\theta_i - (\beta_j + \delta_h)\}} \quad (1)$$

where θ_i is the *ability* parameter of country *i*; β_j is the *difficulty* parameter of question *j* and δ_h are category thresholds, equal for all the variables. Category thresholds are points at which the Rasch model assumes that the probability of opting for the next category is equal to that for the previous one (Linacre, 2012). The probability of passing or failing each threshold is described by the Rasch model.

By looking at eq. (1) it is evident that the model assumes that the probability for country *i* of scoring a certain value *h* on question *j* correlates positively with the country ability θ_i and negatively with the question difficulty β_j. It is then expected that the response reflects, at least to a certain extent, the level of the *state and value of the Web* in that country (ability) given the question difficulty. Thresholds δ_h stepwise increase the question difficulty: the overall difficulty of question *j* is indeed ($\beta_j + \delta_h$). Being based on statistical models, the Rasch analysis also provides estimated standard errors for both parameters θ_i and β_j .

It is interesting to note that the Rasch model is not designed to account for hierarchical dependencies among thresholds. Category thresholds are defined as difficulties of each successive step within the item, without any ordination among differential difficulties (Wright and Masters, 1982). Disordered thresholds do not necessarily imply disordered categories (Smith and Smith, 2004). Still, disordered thresholds reflect the existence of categories that do not have a higher probability of being observed as compared to other categories, at any value of the latent variable. It is up to the analyst to decide whether or not data are consistent with disordered thresholds. In our analysis of WI primary data a check on disordered thresholds is carried out for each Rasch model. Using Rasch modelling we focus on numerical information embedded in the response categories, setting aside their linguistic characteristics which have been set by experts. Thus, we assume that the response categories are clearly, properly and exhaustively defined. Rasch modelling provides diagnostics for potential flaws of the questionnaire.

Under the property of *local independence* row and column marginal frequencies of data matrix are sufficient statistics for the Maximum Likelihood estimate of parameters θ_i and β_j (Fischer and Molenaar, 1995). The local independence assumption requires that the success or failure on any item (question) should not depend on the success or failure on any other item (Bond and Fox, 2007). As will be shortly discussed (Section 2), the local independence requirement forced us to exclude a couple of variables (Q9a-b and Q9c-d) which are computed on the basis of other questions (Q9a, Q9b, Q9c and Q9d) in the Institutional Infrastructure component.

Box 3: Technical description of the Rating Scale model

2.2 Rasch model diagnostics and WI primary data assessment

The Rating Scale model is applied to WI primary data with particular focus on item analysis. The general rule in the Rasch modelling is that all items should share something in common but, at the same time, they should be as diverse as possible. For this purpose, a detailed diagnosis of model's diagnostics and item fit is discussed in this Section for the overall Rasch analysis of WI data. A separate discussion of each Rasch model for the different WI components is provided in Section 2.3.

Disordered thresholds and related redundancy in the number of categories

The first check on the quality of the Rasch model is on monotonicity in category measures and disordered thresholds. Lack of monotonicity can be due to low frequencies on certain categories for certain items. When monotonicity is violated, it is generally recommended either to collapse adjacent categories or not to include them in the analysis (Bond and Fox, 2007, Linacre 2002). Categories with low frequencies are problematic in the Rasch modelling as they do not provide sufficient observations for precise and stable estimation of thresholds δ_h. Such infrequently endorsed categories often indicate unnecessary or redundant categories. The same is also valid for items with category thresholds very close to each other or even reversed.

Despite the high number of categories allowed for in the expert assessment survey (from "1" to "10") no Rasch model on WI data shows disordered thresholds (apart from one single case in the Economic Impact model for category "6"). The distribution of frequencies across different categories is almost even for all the questions included in the survey. **This indicates that surveyed countries cover a wide range of variability in terms of *state and value of the Web* and all categories are well represented in all the questions. No recommendation is suggested with this respect for future surveys.**

Unexpected behaviours

Another criterion for assessing the quality of the Rasch model is the analysis of unexpected behaviours through misfit statistics: the outfit and infit mean squares (Bond and Fox, 2001; Smith and Smith, 2004). Outfit happens when unexpected observations occur on items that are relatively very easy or very hard for some persons (these persons are not expected by the model to perform in that way on those questions), while infit arises when unexpected patterns of scores (answers) are recorded by persons on items that are roughly targeted on them. By construction misfit statistics are expected to have a value close to 1.0: if this is the case countries and questions behave as the model expects them to do. Values greater than 1.0 indicate more variability in the observed data than what is predicted by the Rasch model (e.g. a fit statistic equals to 1.3 indicates 30% more variability in the observed data). This implies a consequent degradation of the quality of the model estimates and is interpreted as certain responses to the item being influenced by other features having no relation with the underlying phenomenon. These cases are considered as deserving further investigation.

On the other hand, values of misfit statistic lower than 1.0 indicate that the model predicts the data "too well" and observed data show less variability than modeled. For example a misfit value of 0.60 indicates 40% less variation in the observed response pattern than predicted by the model. This is an interesting case as, if this occurs, the associated questions are considered redundant: they add too little information as compared to the information already provided by other questions.

Critical values of fit statistics[5] are shown in Table 1. It is important to remark that high values are generally more critical than low ones in terms of model quality.

Table 1: Misfit statistics critical values (Linacre, 2012, p. 328)

Misfit value	Interpretation
> 2.0	Distorts or degrades the measurement system.
1.5 – 2.0	Unproductive for construction of measurement, but not degrading.
0.5 – 1.5	Productive for measurement.
< 0.5	Less productive for measurement, but not degrading.

In the WI case, only 4 out of 49 questions are spotted as misfitting, with values higher than 1.5. These are questions Q10 and Q25 in the Institutional Infrastructure component, Q2a in the Web Content component and Q12 in the Economic Impact component. These results are thoroughly discussed in the following Section (2.3) where the Rasch outcomes are described separately for each component. A possible reason for this behaviour is also suggested.

[5] Critical values are derived by approximate *t* distribution, after proper transformations of infit and outfit statistics.

Overall the analysis of WI question misfit is quite satisfactory. Only 4 out of 49 questions present some critical issues. The recommendation is to further refine these questions - Q10, Q25, Q2a andQ12 - for future surveys.

As far as countries are concerned, in all the models both low-fitting and high-fitting countries are present. This is expected from surveys of this type but we nonetheless checked for the constant presence of particular countries misfitting in all the Rasch models. This would mean that experts in these countries showed a systematic behaviour in filling-in the questionnaire.

Figure 2 allows for a quick inspection of results. It shows the percentage of cases each country is classified as misfitting, either high or low, across all the Rasch models on the WI components. We remind the reader that the most critical misfit type is that with high values.

At a first glance, Figure 2 shows that all the countries do misfit at least once. This is a clear indication that a certain level of misfit is intrinsic and structural in these types of surveys and there is not a group of countries that clearly comes out as problematic ones. Tanzania and Turkey are the most "low" misfitting ones, being misfitting in almost all the models (7 times Tanzania and 6 times Turkey) but the misfit type is the least serious one: they show less variability than expected. In the case of Tanzania this is due to unexpected low values on quite a number of questions (Figure 3), while in the case of Turkey this is due to unexpected middle scale values on quite a number of questions (Figure 4). Other misfitting countries are Egypt, France, Finland, Senegal, Namibia, Yemen, Uganda and South Africa each misfitting more than 70% of the times with a mixed type of misfit (apart from Yemen that is always misfitting "low").

Overall we can say that no systematic pattern is detected by the analysis of country misfit. **This supports the belief that country experts filled in the questionnaire to their best capabilities.**

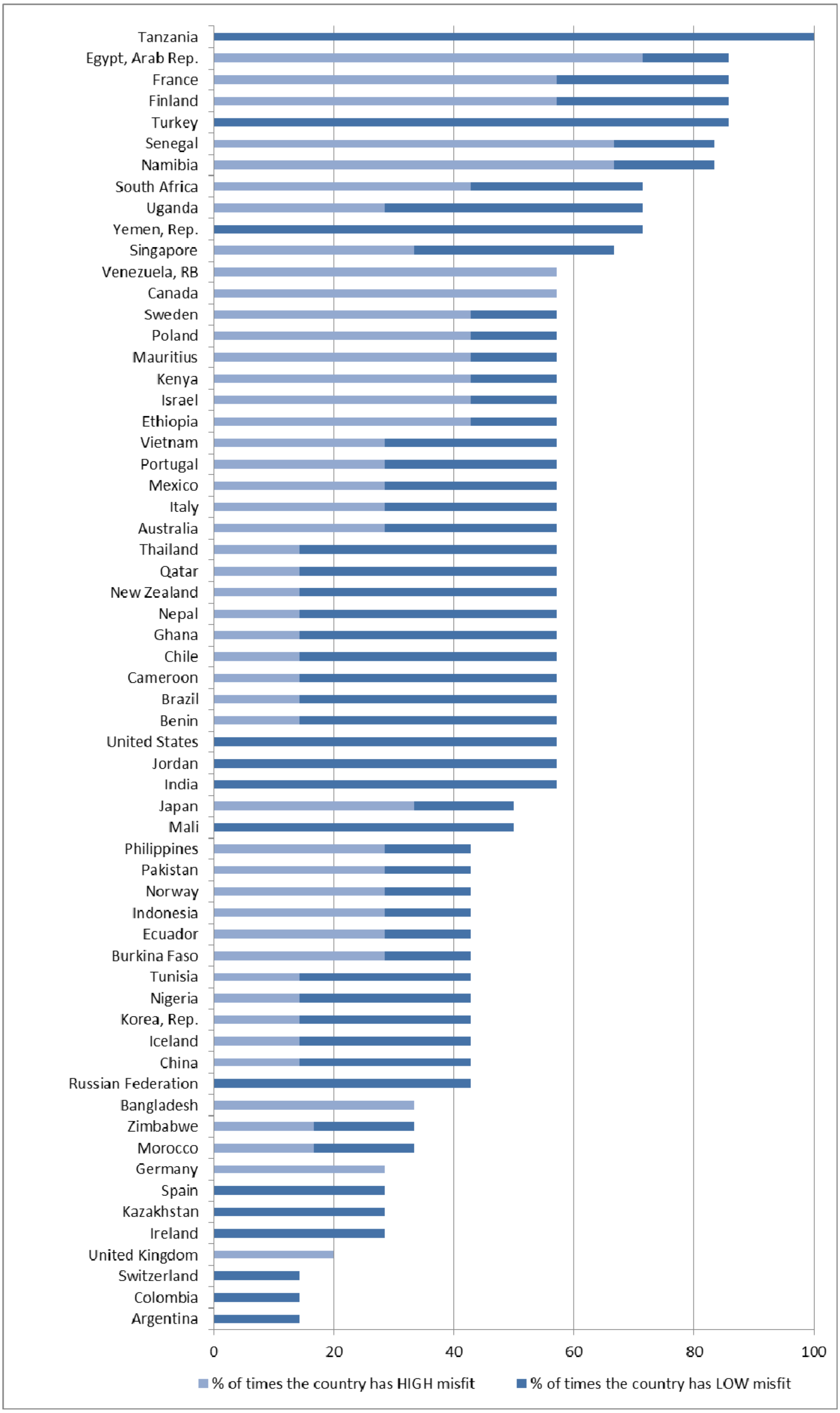


Figure 2: Overall country misfit analysis

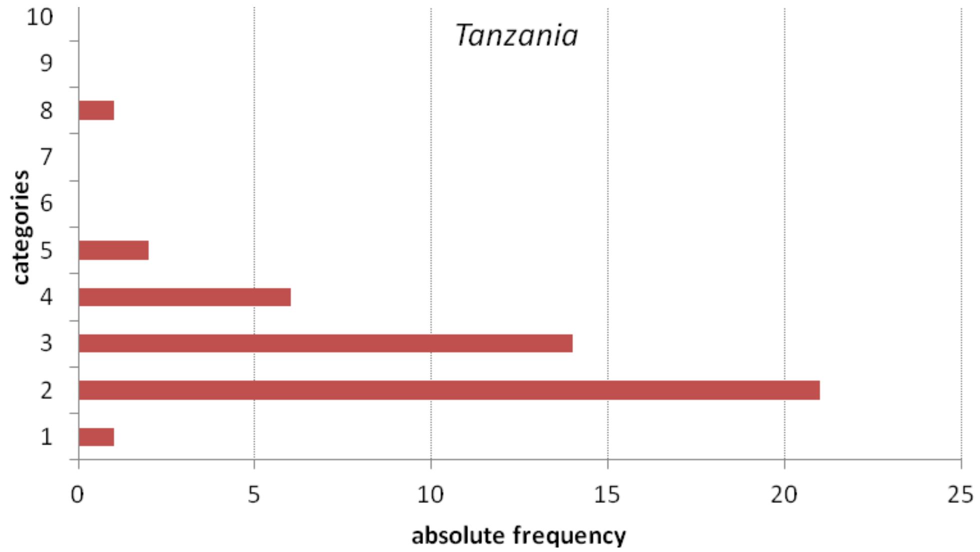


Figure 3: Tanzania – Category frequency distribution

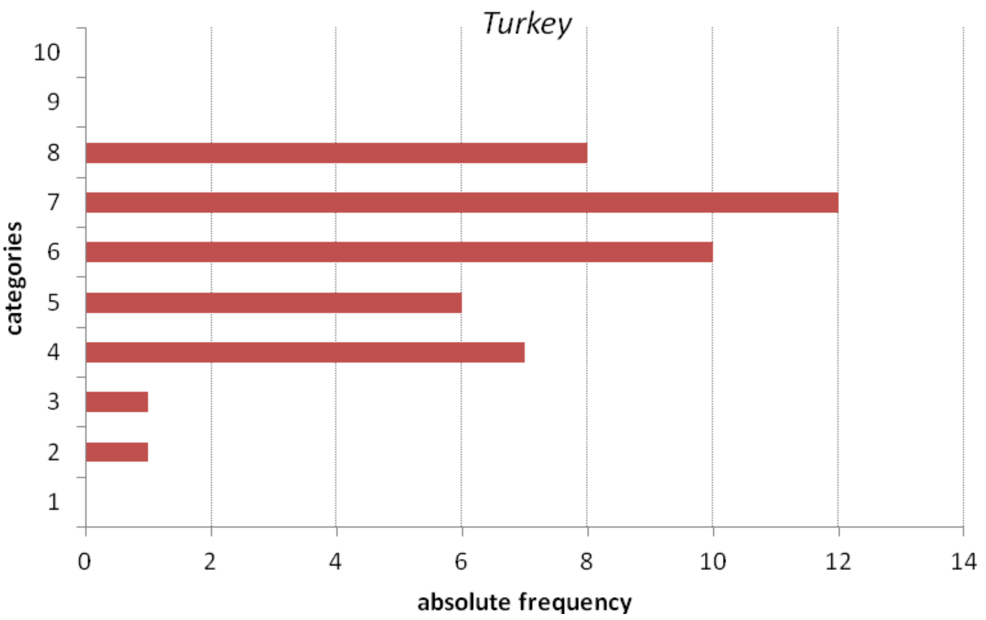


Figure 4: Turkey – Category frequency distribution

Single latent factor assumption

Rasch analysis relies on the assumption that item responses are driven by a common, unique factor which is not directly measurable but is indirectly measured by the item set. A related assumption is that all information contained in the data is explained by this latent, driving factor. The rule is that all items must be related to this latent factor while, at the same time, being as different as possible among them in order to carry complementary pieces of information for portraying the latent factor. It is important to remark that the uni-dimensionality assumption of the Rasch analysis does not imply that the performance on items is due to a single process. Actually, a multiple set of processes are usually involved in responding to a set of questions. As long as they function in unison, uni-dimensionality will hold (Bond and Fox, 2001, pg. 103). This is indeed the general assumption when measuring multi-facets phenomena like in the case of composite indexes which are usually computed as aggregated measures of a set of components. In the WI case, this assumption is tested by the Rasch dimensionality analysis for each component separately.

Rasch model diagnostics allow for testing the single latent dimension assumption by means of the analysis of the correlation pattern of standardized residuals, by means of principal component analysis (Morrison, 2005). If residual variability, that is the amount of variance in the data which is not explained by the model, is sufficiently low then no other latent dimension is shared by the data other than the Rasch dimension (Smith and Smith, 2004; Linacre, 2012). Real data always differ to some degree from theoretical assumptions, the question is whether the distance from the ideal situation is tolerable enough. Principal component analysis of standardized residuals is of help in identifying features shared in common by residuals which are often indications of secondary dimensions in the data. If this is the case, proper actions should be undertaken.

The analysis of standardized residuals for the WI shows that in all the components the single latent factor assumption is satisfied.

Table 2 summarizes results: the percentage of variance accounted for by the model is higher than 60%, with a minimum of 62.4% for the Institutional Infrastructure component and a maximum of about 86% for the Q5 sub-component of the Web Content component. **This confirms that all the questions fit well the WI framework.**

Table 2: Dimensionality analysis

Model	Percentage of variance explained by the model
Institutional Infrastructure	62.4%
Web Content: Q5 questions	86.3%
Web Content: Q23 questions	72.7%
Web Content: remaining questions	65.3%
Web Use	80.4%
Economic Impact	81.7%
Social Impact	72.7%

Questions and country measures

The two key parameters of the Rasch models, person's ability and item's difficulty, are the final outcomes. In the WI they represent an aggregated, quantitative measure of the *state and value of the Web* for each component in each country (country's ability) and of the overall difficulty of each question. Both are simultaneously estimated by the model together with an associated standard error (SE). In the WI case, we provide a ranking of questions and countries according to their Rasch measure together with the interval of uncertainty on the basis of the estimated SE. Questions can be reordered according to their difficulty level: the higher the difficulty, the lower, on average, the level of *state and value of the Web* captured by that question. Similarly, countries are reordered based upon their Rasch measure separately for each WI component. The ranking is not unique as it is associated to standard errors. Therefore a better approach would be identifying a "range of ranks" for each country or groups of statistically equivalent countries.

The following Section discusses results for each component.

2.3 Component by component analysis

2.3.1 Institutional Infrastructure

The Institutional Infrastructure component originally includes 16 survey questions. However the first run of the Rasch model is carried out on 14 questions only. Two questions (Q9ab and Q9cd) are discarded as they do not comply with the assumption of local independence (see Box 3).

The first run of the Rasch model on the 14 survey questions revealed two misfitting questions: Q10 and Q25 have both infit and outfit statistics higher than 2. These questions thus add noise to the measurement process, showing more variability than expected and distorting the Rasch measurement. To investigate possible reasons for such result we analyze both the content and the wording of these problematic questions:

> Question Q10: *To what extent does the government impose restrictions on access to Websites (censorship)?*
>
> Question Q25: *Does the government have a specific Open Government Data initiative?*

For question Q10 the worst score "1" means "*Extensive restrictions are imposed & many Websites are blocked*" while the best score "10" means "*No restrictions are imposed & no Websites are blocked*". The question appears to be worded in a misleading way as its orientation with respect to the latent concept – *state and value of the Web* – seems to be reversed. It indeed refers to "restrictions" likely inducing respondents to assign high scores when they feel that restrictions are high. This question is counter-oriented with respect to the WI latent concept (the higher the worse) which is detected by the Rasch model by misfit statistics. **The recommendation is then to carefully reword either the question Q10 for future surveys by changing its orientation with respect to the WI underlying concept or the labels of the response categories.**

In question Q25 the worst score "1" means "*No full Open Government data initiative*" and the best score "10" means "*Full Open Government Data initiative*". The question has two potential pitfalls: *i*) it is a very technical question; *ii*) the question itself is worded in a way that a dichotomous answer may be expected. However, this second issue is not supported by the analysis of the distribution of category frequencies, as shown in Figure 5 where it is clear that respondents considered all possible categories for question Q25. **We recommend WI developers to further investigate on question Q25.**

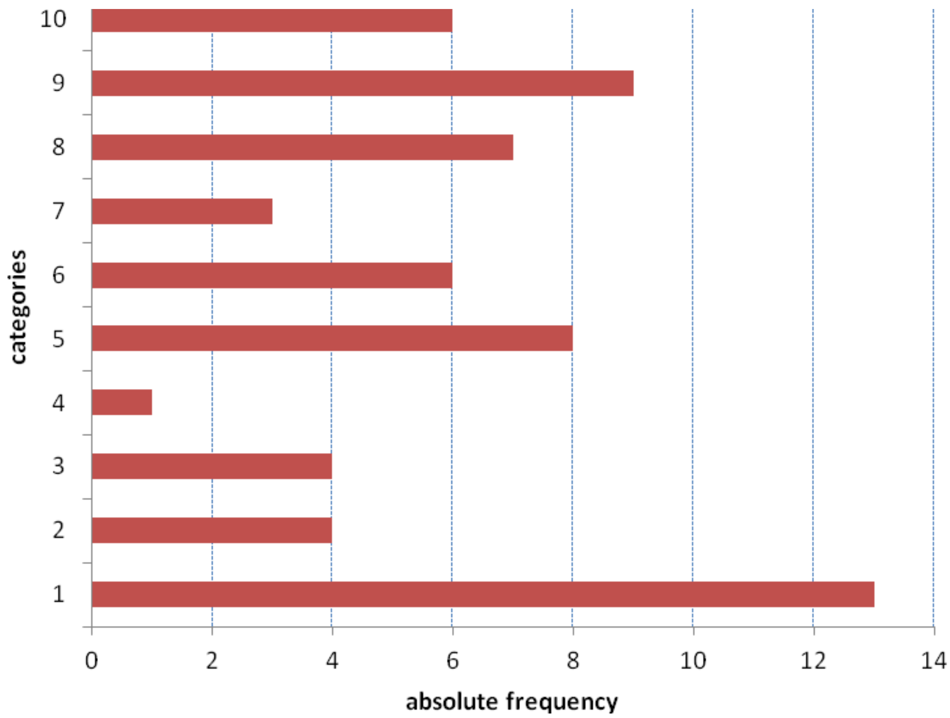


Figure 5: Variable Q25 – Category frequency distribution

The two misfitting questions are discarded from the analysis of this component and the Rasch model run again with the remaining variables. For all the questions fit statistics turn to be within the acceptable range of values [0.5; 1.5]. The Rasch model explains a good portion of data variability (62.4%) and is clearly uni-dimensional implying that the proposed set of questions is driven by a common, unique factor.

Questions can be reordered on the basis of their estimated Rasch measure to get an "*easy-difficult scale*" (Figure 6). This scale allows for a quick glance at the average scores received by each question: the higher the position of a question on that scale, the lower, on average, the score it got (and vice-versa).

Comparing the "twin questions" Q9a with Q9b and Q9d with Q9c we notice that the ones referring to males (Q9a and Q9d) receive on average higher scores, being situated more towards the bottom of the "*easy-difficult scale*", than their female counterparts. It is remarkable that the four top questions, all referring to female implication in ICT, receive generally lower scores than the others, rising a gender bias issue.

Furthermore, questions Q9f, Q9h, Q9i and Q9l, while explicitly referring to the situation of female population, do not have the male counterpart. **It may be worth including the male counterpart of these questions in future surveys**.

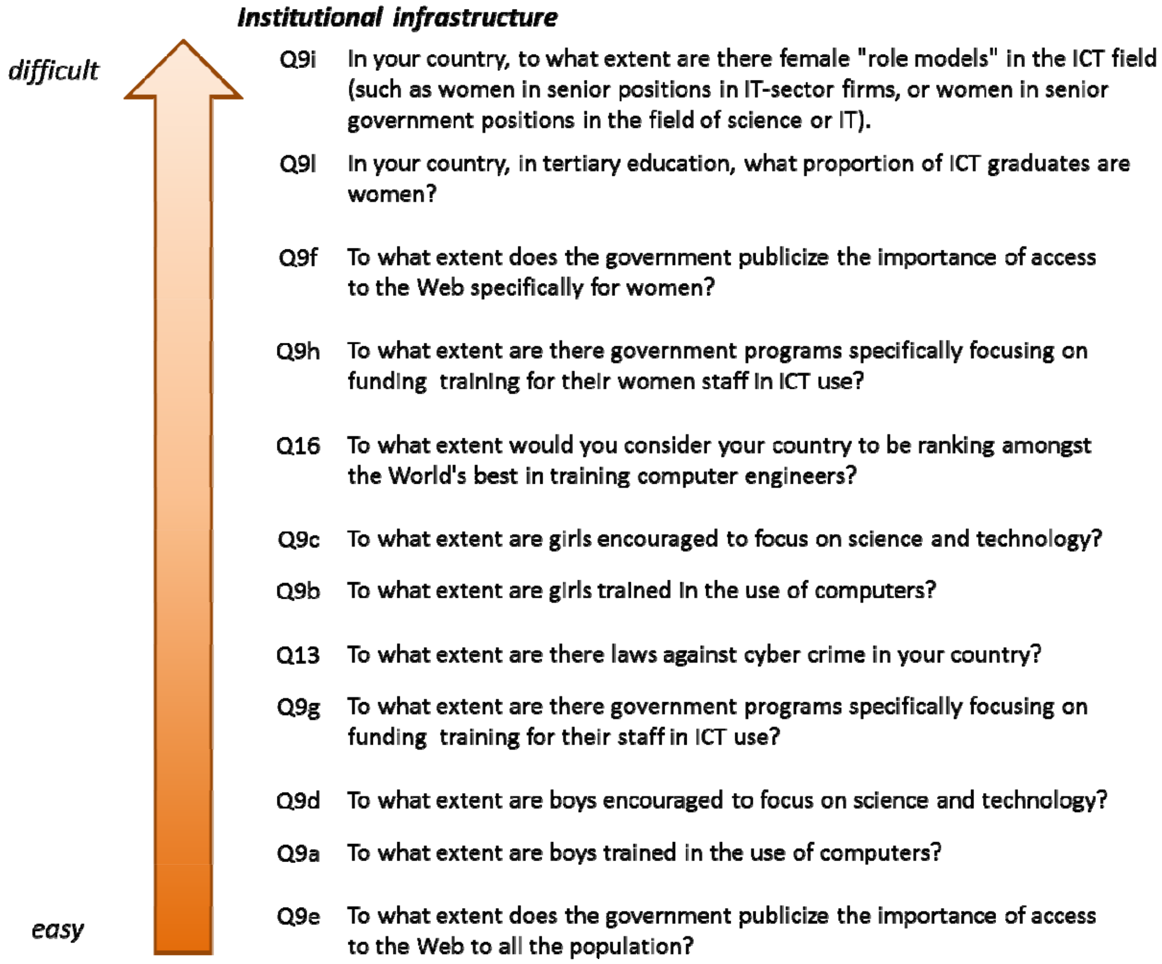


Figure 6. Order of questions in Institutional Infrastructure component.

Having established the measurement scale of Institutional Infrastructure we proceed with the assessment of country performance in the 12 survey questions. The ranking of countries according to their country measure and standard error is presented in Table 3 and Figure 7.

Table 3. Institutional Infrastructure component - 12 questions: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
United States	1.14	0.21	1		Chile	0.1	0.17	31
Singapore	1.1	0.21	2		Philippines	0.07	0.17	32
Spain	1.02	0.2	3		Mexico	0.04	0.17	33
Sweden	0.98	0.2	4		Argentina	0.01	0.17	34
Australia	0.87	0.19	5		Russian Fed.	0.01	0.17	34
Canada	0.83	0.19	6		France	-0.02	0.17	36
Germany	0.83	0.19	6		Jordan	-0.05	0.18	37
Iceland	0.76	0.19	8		Tunisia	-0.08	0.18	38
Ireland	0.76	0.19	8		Thailand	-0.17	0.18	39
New Zealand	0.76	0.19	8		Cameroon	-0.21	0.18	40
Norway	0.76	0.19	8		South Africa	-0.27	0.18	41
Finland	0.73	0.18	12		Pakistan	-0.3	0.18	42
Kazakhstan	0.56	0.18	13		Senegal	-0.3	0.18	42
Poland	0.56	0.18	13		Ghana	-0.34	0.18	44
UK	0.53	0.18	15		Kenya	-0.63	0.2	45
Italy	0.44	0.18	16		Benin	-0.72	0.21	46
Switzerland	0.41	0.17	17		Egypt	-0.72	0.21	46
China	0.38	0.17	18		Uganda	-0.72	0.21	46
Indonesia	0.34	0.17	19		Vietnam	-0.72	0.21	46
Qatar	0.34	0.17	19		Ecuador	-0.81	0.21	50
Korea, Rep.	0.31	0.17	21		Ethiopia	-0.85	0.22	51
Mauritius	0.31	0.17	21		Nigeria	-0.85	0.22	51
Turkey	0.31	0.17	21		Bangladesh	-1	0.23	53
Venezuela	0.28	0.17	24		Burkina Faso	-1.3	0.26	54
India	0.25	0.17	25		Yemen, Rep.	-1.3	0.26	54
Brazil	0.22	0.17	26		Nepal	-1.37	0.27	56
Portugal	0.22	0.17	26		Tanzania	-1.37	0.27	56
Colombia	0.19	0.17	28		Mali	-1.52	0.28	58
Israel	0.19	0.17	28		Namibia	-1.6	0.29	59
Japan	0.13	0.17	30		Zimbabwe	-1.89	0.33	60
					Morocco	-2.13	0.36	61

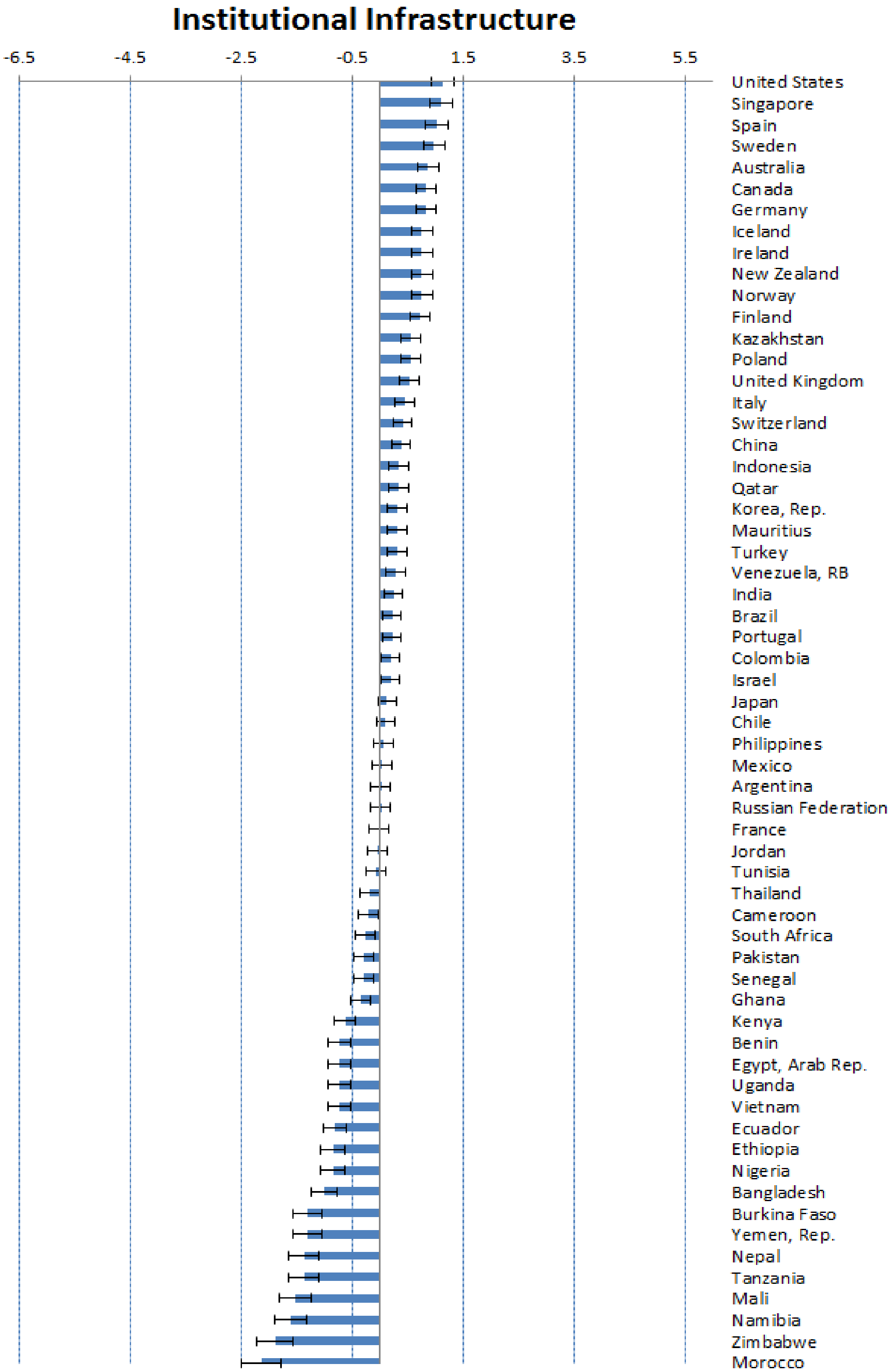


Figure 7. Countries performance in Institutional Infrastructure component, analysis based on 12 survey questions (error bars indicate ± 1SE).

2.3.2 Web Use

The Web Use component of WI includes six survey questions describing the access to the web by people with different disabilities. All of them fit the model with fit statistics within the tolerance range. As much as 80.4% of variance is explained by the model. As can be seen by Figure 8, people with hearing disability are those who have the most effective access to the Web while illiterate people have the least effective access.

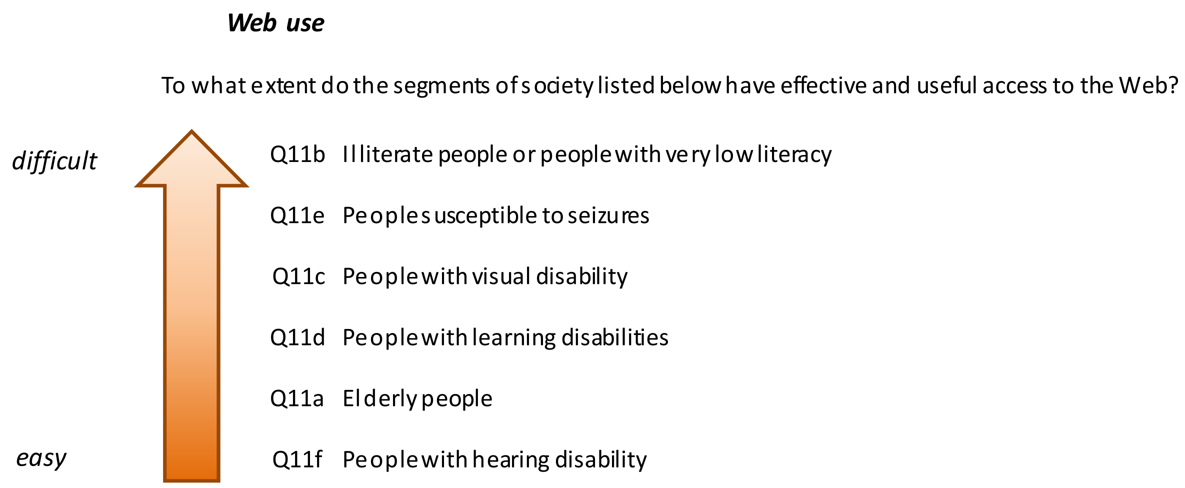


Figure 8. Order of questions in Web Use component.

As for the other components, we provide country ranking, measures and standard errors (Table 4 and Figure 9).

Table 4. Web Use component: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
Switzerland	4.16	0.74	1		Venezuela, RB	-1.52	0.39	27
Iceland	3.12	0.69	2		China	-1.67	0.4	32
United States	1.99	0.52	3		Nepal	-1.67	0.4	32
Portugal	1.74	0.48	4		Thailand	-1.67	0.4	32
Finland	1.34	0.42	5		Brazil	-1.84	0.41	35
Canada	1.03	0.38	6		Japan	-2.01	0.43	36
Germany	0.76	0.35	7		Tunisia	-2.01	0.43	36
Singapore	0.76	0.35	7		Argentina	-2.2	0.44	38
Italy	0.64	0.35	9		Jordan	-2.2	0.44	38
UK	0.64	0.35	9		Mauritius	-2.4	0.46	40
Ireland	0.52	0.34	11		Indonesia	-2.63	0.49	41
Australia	0.41	0.33	12		South Africa	-2.63	0.49	41
Kazakhstan	0.19	0.33	13		Uganda	-2.63	0.49	41
New Zealand	0.19	0.33	13		Cameroon	-2.88	0.52	44
Norway	0.08	0.33	15		Qatar	-2.88	0.52	44
Sweden	0.08	0.33	15		Vietnam	-2.88	0.52	44
Israel	-0.13	0.33	17		Benin	-3.17	0.55	47
Poland	-0.13	0.33	17		Nigeria	-3.17	0.55	47
Spain	-0.13	0.33	17		Tanzania	-3.17	0.55	47
Mexico	-0.71	0.35	20		Egypt	-3.5	0.6	50
France	-0.83	0.36	21		Kenya	-3.91	0.67	51
Philippines	-0.96	0.36	22		Morocco	-3.91	0.67	51
Russian Fed.	-0.96	0.36	22		Yemen	-3.91	0.67	51
Colombia	-1.23	0.37	24		Burkina Faso	-4.43	0.78	54
Turkey	-1.23	0.37	24		Ecuador	-4.43	0.78	54
Ghana	-1.37	0.38	26		Ethiopia	-5.24	1.05	56
Chile	-1.52	0.39	27		Bangladesh	-6.53	1.85	57
India	-1.52	0.39	27		Mali	-6.53	1.85	57
Korea, Rep.	-1.52	0.39	27		Namibia	-6.53	1.85	57
Pakistan	-1.52	0.39	27		Senegal	-6.53	1.85	57
					Zimbabwe	-6.53	1.85	57

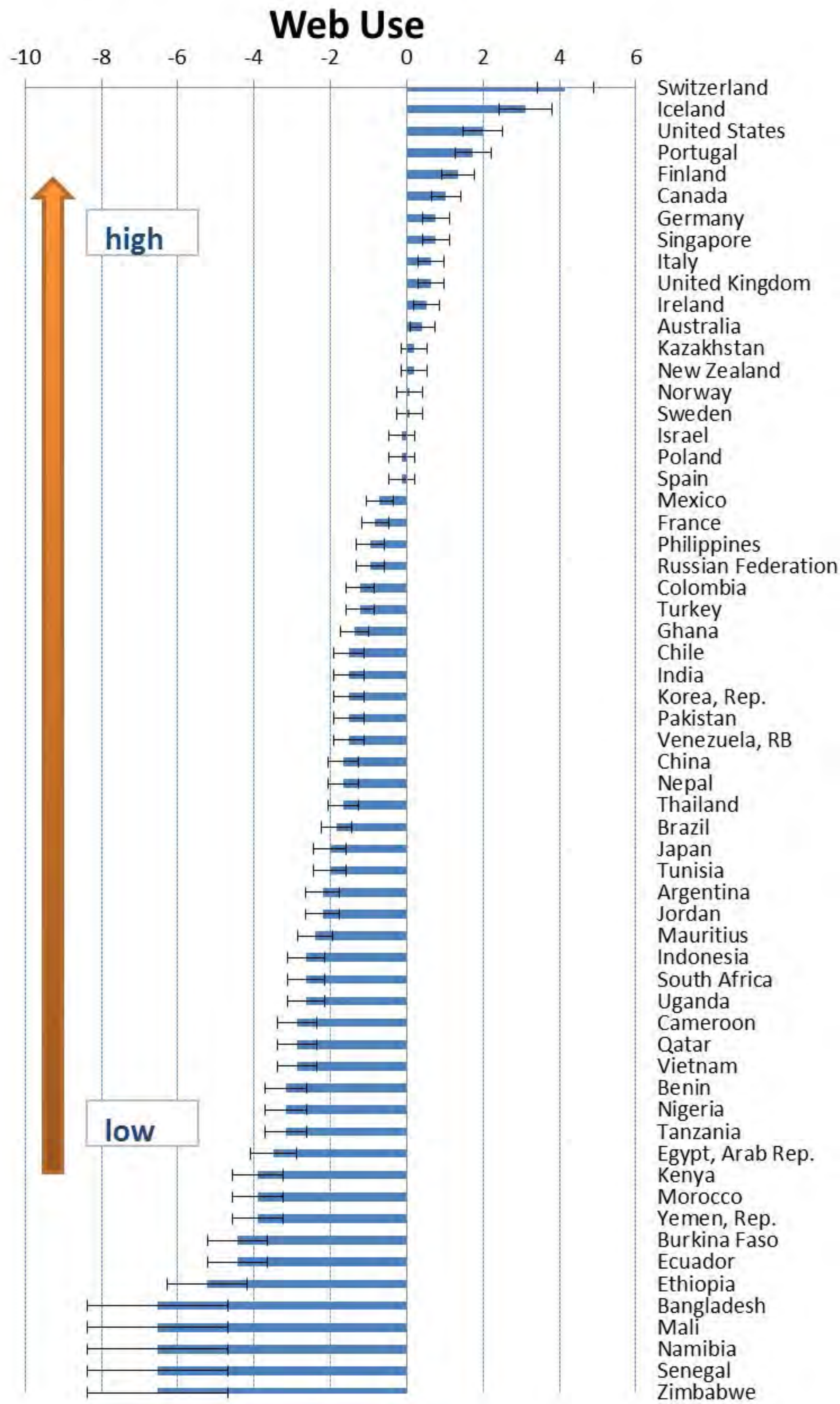


Figure 9. Countries performance in Web Use survey questions (error bars indicate ± 1SE).

2.3.3 Web Content

The Web Content component contains 22 survey questions, with two sets of clustered questions[6]. Following the structure of the component, we distinguish three sub-components (see the list of questions in Appendix):

1. Web Content – questions Q5a, Q5b and Q5c),
2. Web Content – 10 questions from Q23a to Q23j,
3. Web Content – 9 remaining survey questions.

Each of them is analyzed separately.

Web Content – questions Q5a, Q5b and Q5c

The Rasch model is run for the set of 3 questions describing the availability of on-line curricula for primary, secondary and tertiary education. All of them fit the model perfectly, with fit statistics in the range of [0.5; 1.5]. The Rasch measurement scale is confirmed to be uni-dimensional and explains 86.3% of data variability. . The order of the questions on the "*easy-difficult scale*" (Figure 10) is as expected, leading to the conclusion that they provide unique pieces of information to the measurement process. The question on the availability of curricula in tertiary education (Q5c) receives on average higher scores as tertiary education is the one for which on-line services are more developed and used; while lower scores are given to the availability of curricula of primary education (Q5a). The two most difficult questions (referring to on-line curriculum of primary and secondary education) are more similar to each other than the remaining one (on-line curriculum of tertiary education).

[6] The clustering is carried out by computing arithmetic average of the scores after z-score standardization.

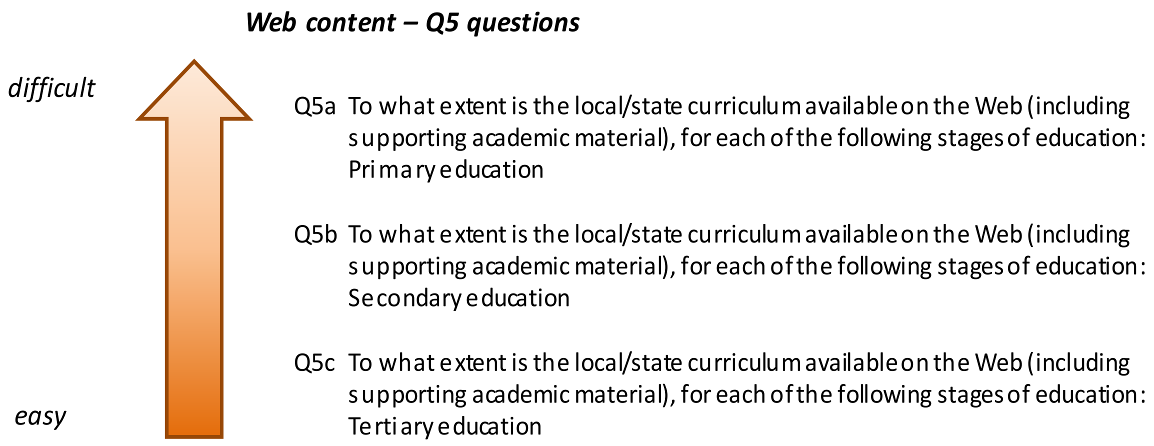


Figure 10. Order of questions in Web Content – questions Q5a, Q5b, Q5c.

The ranking of countries according to their Rasch measure and associated standard errors are presented in Table 5 and Figure 11.

Table 5: Web Content – questions Q5a, Q5b, Q5c: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
Japan	5.98	1.76	1		Burkina Faso	0.03	0.55	30
Singapore	5.98	1.76	1		India	-0.24	0.51	32
UK	5.98	1.76	1		Jordan	-0.24	0.51	32
France	4.86	0.97	4		Colombia	-0.7	0.45	34
Portugal	4.86	0.97	4		Ghana	-0.7	0.45	34
Norway	4.17	0.73	6		Argentina	-0.89	0.44	36
Poland	4.17	0.73	6		Italy	-0.89	0.44	36
South Africa	4.17	0.73	6		Kazakhstan	-0.89	0.44	36
Sweden	4.17	0.73	6		Philippines	-0.89	0.44	36
Ireland	3.69	0.66	9		Qatar	-0.89	0.44	36
Mexico	3.69	0.66	9		China	-1.08	0.42	41
United States	3.69	0.66	9		Egypt	-1.08	0.42	41
Canada	3.26	0.65	13		Thailand	-1.25	0.42	43
Israel	3.26	0.65	13		Yemen, Rep.	-1.25	0.42	43
New Zealand	3.26	0.65	13		Vietnam	-1.43	0.42	45
Finland	2.83	0.68	16		Mauritius	-1.6	0.42	46
Brazil	2.34	0.72	17		Kenya	-1.79	0.43	47
Iceland	2.34	0.72	17		Uganda	-1.79	0.43	47
Korea, Rep.	2.34	0.72	17		Cameroon	-1.98	0.45	49
Switzerland	2.34	0.72	17		Namibia	-1.98	0.45	49
Turkey	2.34	0.72	17		Tanzania	-1.98	0.45	49
Australia	1.8	0.75	22		Venezuela	-1.98	0.45	49
Chile	1.26	0.72	23		Benin	-2.19	0.47	53
Ecuador	1.26	0.72	23		Nepal	-2.19	0.47	53
Tunisia	1.26	0.72	23		Ethiopia	-2.73	0.58	55
Germany	0.77	0.67	26		Indonesia	-2.73	0.58	55
Pakistan	0.77	0.67	26		Mali	-2.73	0.58	55
Spain	0.77	0.67	26		Zimbabwe	-3.15	0.71	58
Russian Fed.	0.36	0.61	29		Nigeria	-3.85	1.01	59
Bangladesh	0.03	0.55	30		Senegal	-3.85	1.01	59
					Morocco	-5.09	1.84	60

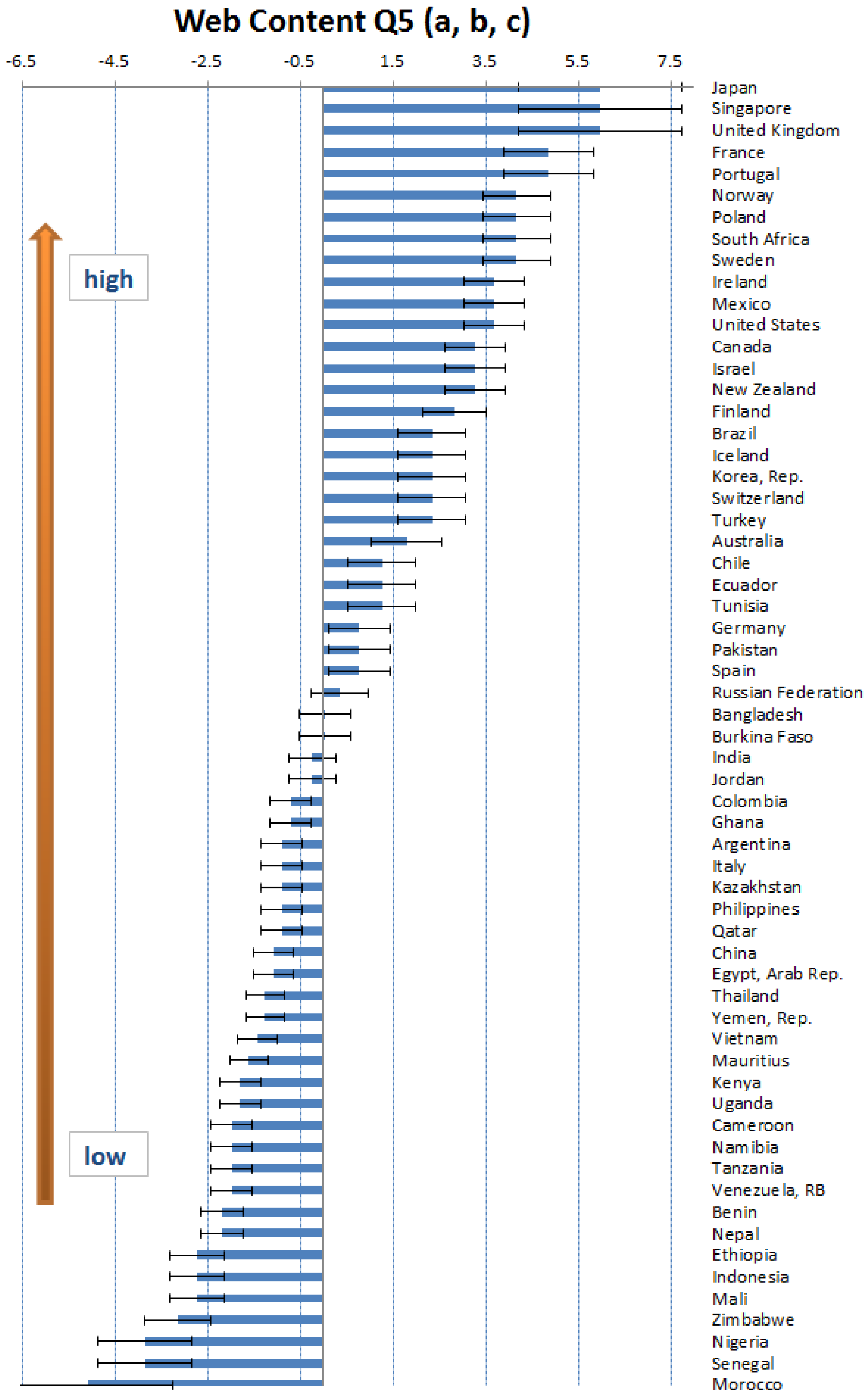


Figure 11. Countries performance in Web Content – questions Q5a, Q5b, Q5c (error bars indicate ± 1SE).

Web Content – questions Q23a to Q23j

In this sub-component the Rasch model is run for a set of ten questions. All of them fit the model perfectly with fit statistics values in the range of [0.5; 1.5]. Data share one common factor with 72.9% of data variance explained by the model. The order of the questions, based on difficulty measure, is presented on the "*easy-difficult scale*" in Figure 12).

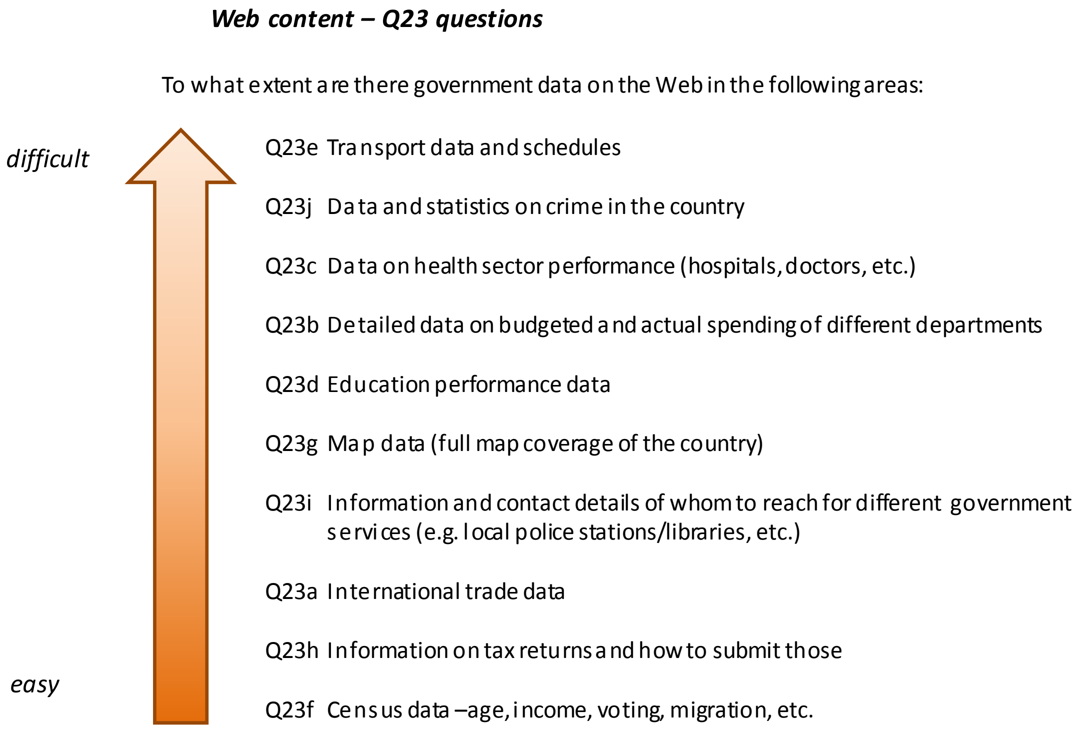


Figure 12. Order of questions in Web Content – Q23a to Q23j.

The ranking of the country performance in this group of questions is presented in Table 6 and Figure 13.

Table 6. Web Content – questions Q23a to Q23j: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
UK	5.47	1.82	1		Russian Fed.	0.35	0.21	29
Portugal	3.23	0.57	2		Brazil	0.3	0.21	32
Germany	2.95	0.5	3		Colombia	0.3	0.21	32
Singapore	2	0.34	4		India	0.3	0.21	32
Switzerland	2	0.34	4		Indonesia	0.3	0.21	32
Israel	1.88	0.33	6		Norway	0.3	0.21	32
Korea	1.88	0.33	6		Tunisia	0.22	0.2	37
United States	1.78	0.32	8		Ghana	0.14	0.2	38
Japan	1.68	0.31	9		Philippines	-0.02	0.19	39
South Africa	1.68	0.31	9		Uganda	-0.02	0.19	39
Finland	1.59	0.3	11		Namibia	-0.13	0.19	41
Spain	1.59	0.3	11		Thailand	-0.13	0.19	41
Australia	1.49	0.3	13		Jordan	-0.16	0.19	43
Ireland	1.49	0.3	13		Ethiopia	-0.2	0.19	44
Mexico	1.49	0.3	13		Nigeria	-0.27	0.19	45
New Zealand	1.32	0.29	16		Kenya	-0.31	0.19	46
Canada	1.24	0.28	17		Nepal	-0.38	0.19	47
France	1.02	0.26	18		Pakistan	-0.41	0.19	48
Iceland	1.02	0.26	18		Vietnam	-0.48	0.19	49
Sweden	1.02	0.26	18		Venezuela	-0.59	0.19	50
Poland	0.82	0.25	21		Argentina	-0.66	0.19	51
China	0.71	0.24	22		Benin	-0.9	0.21	52
Turkey	0.71	0.24	22		Mali	-0.94	0.21	53
Kazakhstan	0.65	0.23	24		Bangladesh	-1.03	0.22	54
Chile	0.54	0.23	25		Burkina Faso	-1.08	0.22	55
Italy	0.54	0.23	25		Tanzania	-1.08	0.22	55
Qatar	0.54	0.23	25		Yemen	-1.13	0.22	57
Ecuador	0.49	0.22	28		Zimbabwe	-1.65	0.29	58
Egypt	0.35	0.21	29		Cameroon	-2.08	0.37	59
Mauritius	0.35	0.21	29		Senegal	-2.08	0.37	59
					Morocco	-2.41	0.45	61

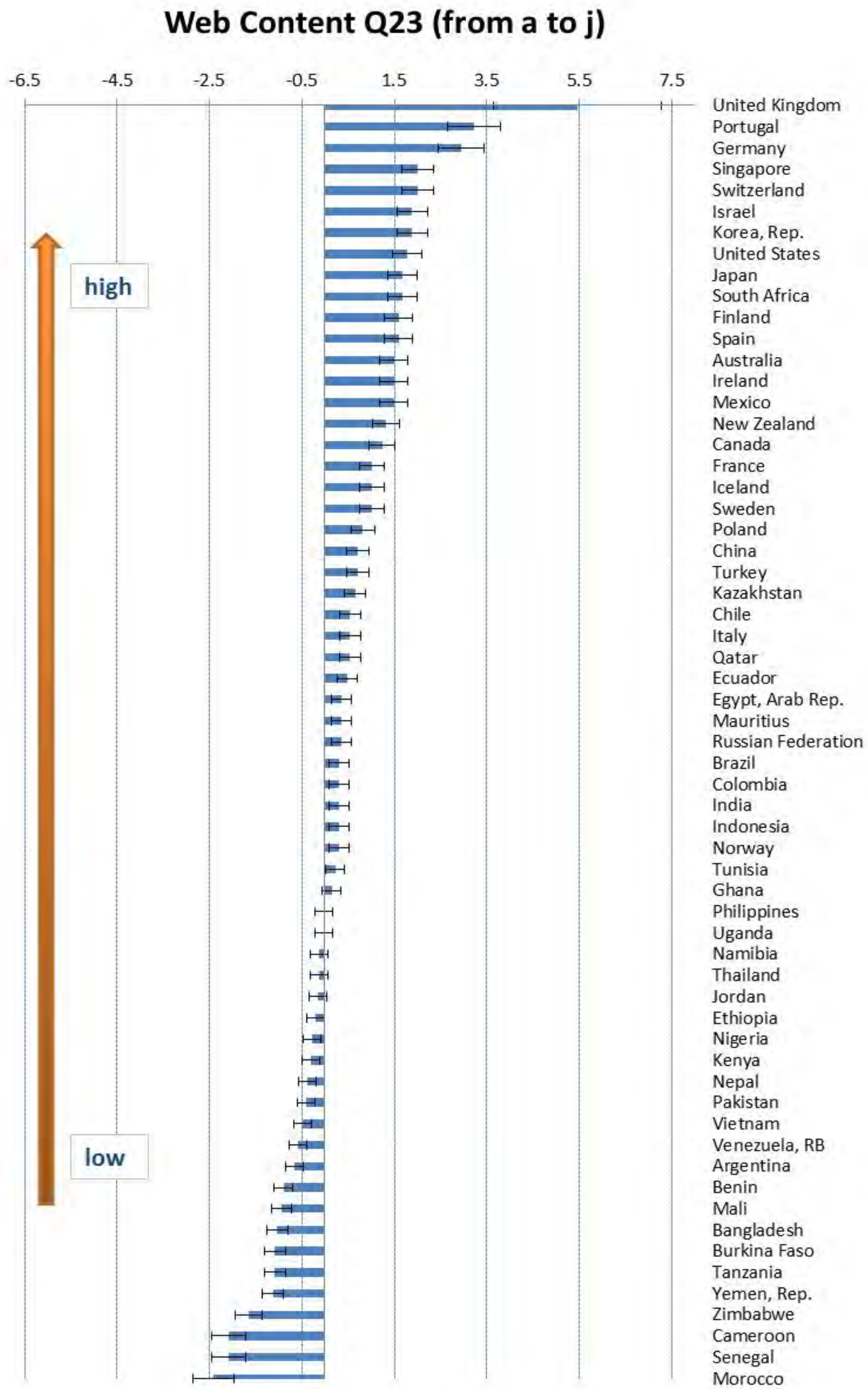


Figure 13. Countries performance in Web Content – questions Q23a to Q23j (error bars indicate ± 1SE).

Web Content – remaining questions

The remaining nine survey questions are analyzed together. Question Q2a turns out to be misfitting, showing high values of misfit statistics.

Question Q2a: *Do the main political parties have web-sites?*

For this question the worst category "1" means "*None have Websites*" while category "10" means "*All have Websites*". As for all the questions, no other specifications are provided for intermediate categories which, in this specific case, is likely to be troublesome. Indeed the way the question is worded may lead the respondent to answer either yes or no . From answer distribution (Figure 14) the question seems to be dichotomous despite its ten-category scale with a clear predominance of category "10".

The recommendation for future surveys is to reword question Q2a.

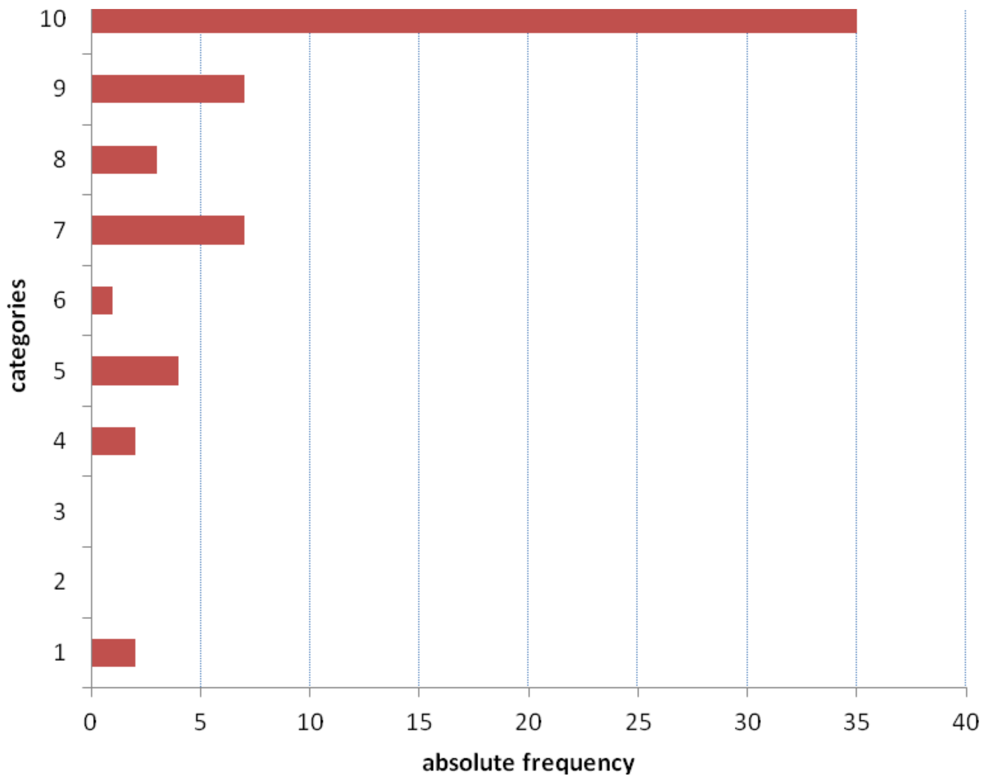


Figure 14: Variable Q2a – Category frequency distribution

The Rasch analysis on this group of questions is thus carried out excluding question Q2a. The order of difficulty of the questions is shown in Figure 15. None of them is here misfitting and the model explains 65.3% of total variance. The country ranking, measures and associated standard errors are shown in Table 7 and Figure 16.

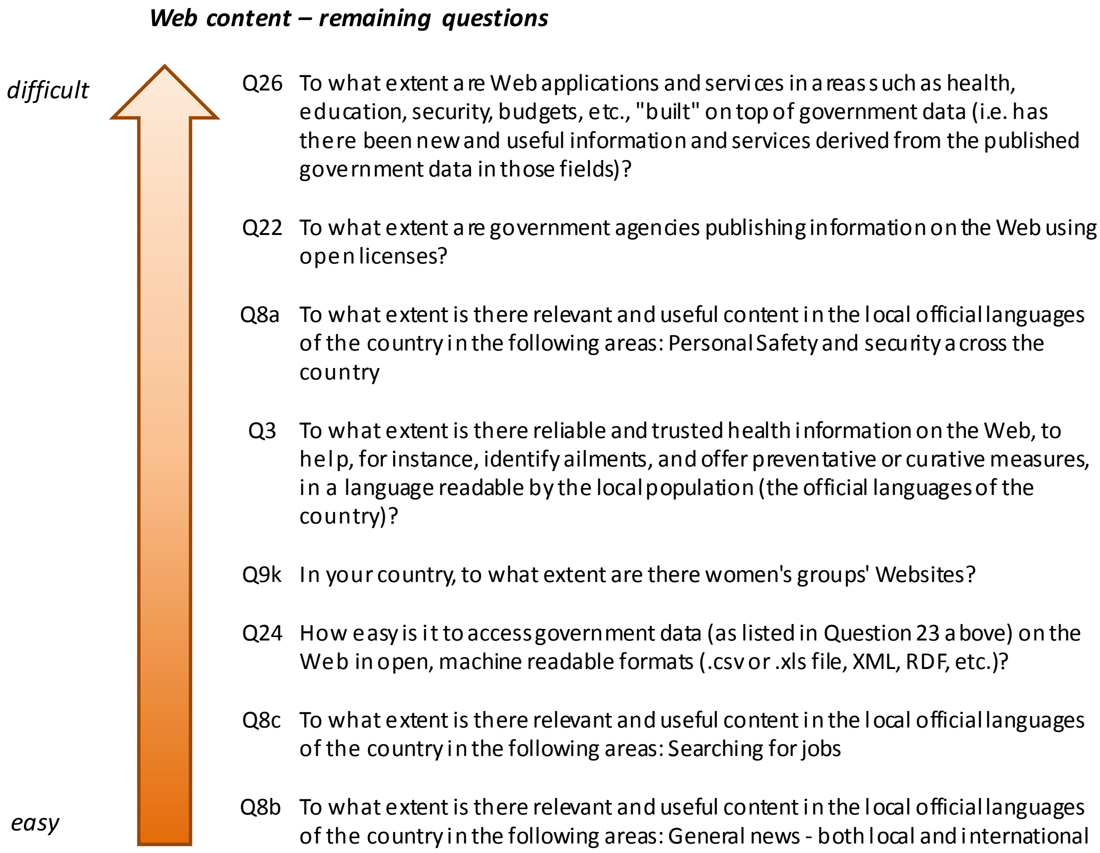


Figure 15. Order of remaining questions in Web Content component.

Table 7. Web Content – remaining questions: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
United States	1.99	0.41	1		Brazil	0.05	0.2	31
Singapore	1.83	0.38	2		Venezuela	0.05	0.2	31
UK	1.47	0.32	3		Colombia	0.01	0.2	33
New Zealand	1.28	0.29	4		India	0.01	0.2	33
Canada	1.2	0.28	5		Thailand	0.01	0.2	33
Australia	1.12	0.27	6		Pakistan	-0.02	0.2	36
Finland	0.85	0.25	7		Argentina	-0.1	0.2	37
Ireland	0.85	0.25	7		Qatar	-0.1	0.2	37
Sweden	0.85	0.25	7		Nepal	-0.14	0.2	39
Germany	0.79	0.24	10		Kenya	-0.18	0.2	40
Portugal	0.79	0.24	10		South Africa	-0.18	0.2	40
Spain	0.79	0.24	10		Uganda	-0.18	0.2	40
Switzerland	0.73	0.24	13		Jordan	-0.22	0.2	43
Japan	0.68	0.23	14		Ghana	-0.26	0.2	44
Kazakhstan	0.68	0.23	14		Ecuador	-0.3	0.2	45
Poland	0.68	0.23	14		Egypt	-0.3	0.2	45
Norway	0.63	0.23	17		Benin	-0.55	0.21	47
Iceland	0.58	0.22	18		Mauritius	-0.55	0.21	47
Israel	0.58	0.22	18		Vietnam	-0.55	0.21	47
Russian Fed.	0.58	0.22	18		Nigeria	-0.64	0.22	50
Italy	0.53	0.22	21		Tanzania	-0.64	0.22	50
Mexico	0.53	0.22	21		Morocco	-0.74	0.22	52
France	0.48	0.22	23		Burkina Faso	-0.79	0.23	53
Chile	0.39	0.21	24		Mali	-0.79	0.23	53
Korea, Rep.	0.3	0.21	25		Yemen, Rep.	-0.79	0.23	53
China	0.22	0.2	26		Cameroon	-0.85	0.23	56
Indonesia	0.22	0.2	26		Senegal	-0.96	0.25	57
Philippines	0.09	0.2	28		Namibia	-1.02	0.25	58
Tunisia	0.09	0.2	28		Bangladesh	-1.09	0.26	59
Turkey	0.09	0.2	28		Zimbabwe	-1.32	0.29	60
					Ethiopia	-1.41	0.31	61

Figure 16. Countries performance in Web Content – remaining questions (error bars indicate ± 1SE).

2.3.4 Economic Impact

Four survey questions are originally included in this component. However question Q12 shows more variability than expected (high misfit). This question is:

> Question Q12: *To what extent do you think that the Web is making it easier to undertake criminal activity in your country?*

The worst category "1" for this question means "*The Web is making it much easier to undertake criminal activities in the country*" and the best category "10" means "*Not at all: the Web has not made it easier to undertake criminal activities in the country*". Low scores are meant to be related to a facilitated criminal activity by web use, while high scores describe the opposite. In our opinion the wording is not straightforward. The question contains two words "easier" and "criminal" which, put together, are likely to cause misunderstanding. The problem is the difficulty of correctly understanding the question. At a first glance, question Q12 seems to be counter-oriented. **The recommendation is to reword the question making it clearly oriented with respect to the WI latent concept.**

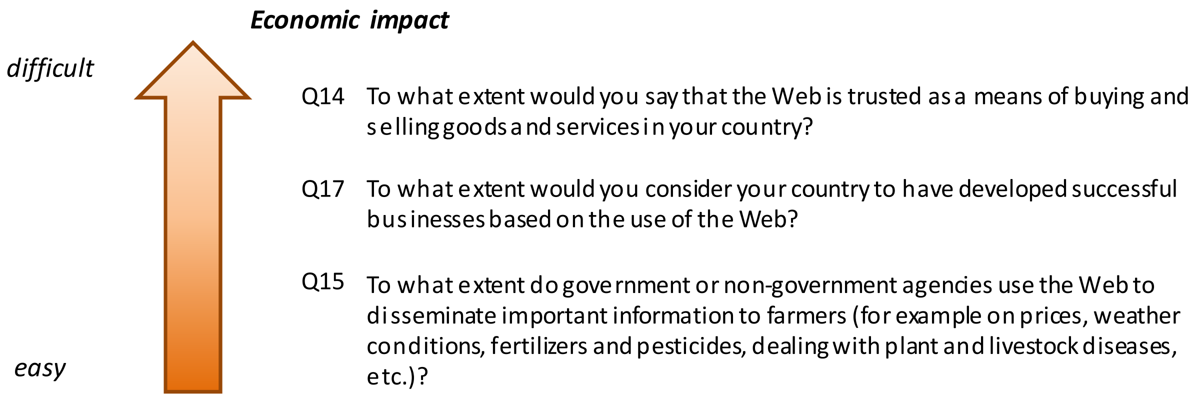


Figure 17. Order of questions in Economic Impact component.

The Rasch model without question Q12 shows no misfitting questions with the model explaining 81.7% of total variance. Table 8 and Figure 18 display Rasch measures, standard errors and ranking for all countries.

Table 8. Economic Impact component: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
Switzerland	4.52	1.1	1		Kazakhstan	-0.17	0.44	27
UK	3.61	0.85	2		Ecuador	-0.17	0.44	27
Canada	2.96	0.76	3		Jordan	-0.36	0.45	33
Sweden	2.96	0.76	3		Thailand	-0.36	0.45	33
United States	2.42	0.7	5		Indonesia	-0.36	0.45	33
Finland	2.42	0.7	5		Philippines	-0.36	0.45	33
Norway	2.42	0.7	5		Turkey	-0.57	0.45	37
Australia	1.97	0.65	8		Pakistan	-0.57	0.45	37
Ireland	1.97	0.65	8		Vietnam	-0.77	0.45	39
France	1.59	0.58	10		Kenya	-0.77	0.45	39
New Zealand	1.59	0.58	10		Nepal	-0.77	0.45	39
Germany	1.59	0.58	10		Tunisia	-1.2	0.47	42
Singapore	1.03	0.49	13		Colombia	-1.2	0.47	42
Korea, Rep.	1.03	0.49	13		Egypt	-1.2	0.47	42
Spain	0.8	0.46	15		Venezuela	-1.2	0.47	42
Japan	0.8	0.46	15		Mauritius	-1.43	0.49	46
Chile	0.6	0.44	17		Ghana	-1.68	0.52	47
Iceland	0.41	0.44	18		Yemen, Rep.	-1.98	0.56	48
Israel	0.41	0.44	18		Namibia	-1.98	0.56	48
Qatar	0.22	0.44	20		Senegal	-1.98	0.56	48
Poland	0.22	0.44	20		Benin	-2.32	0.62	51
China	0.03	0.44	22		Cameroon	-2.32	0.62	51
Brazil	0.03	0.44	22		Burkina Faso	-2.32	0.62	51
South Africa	0.03	0.44	22		Tanzania	-2.32	0.62	51
Russian Fed.	0.03	0.44	22		Mali	-2.32	0.62	51
Mexico	0.03	0.44	22		Uganda	-2.75	0.69	56
Italy	-0.17	0.44	27		Nigeria	-2.75	0.69	56
Portugal	-0.17	0.44	27		Bangladesh	-3.3	0.8	58
India	-0.17	0.44	27		Morocco	-3.3	0.8	58
Argentina	-0.17	0.44	27		Ethiopia	-4.14	1.07	60
					Zimbabwe	-4.14	1.07	60

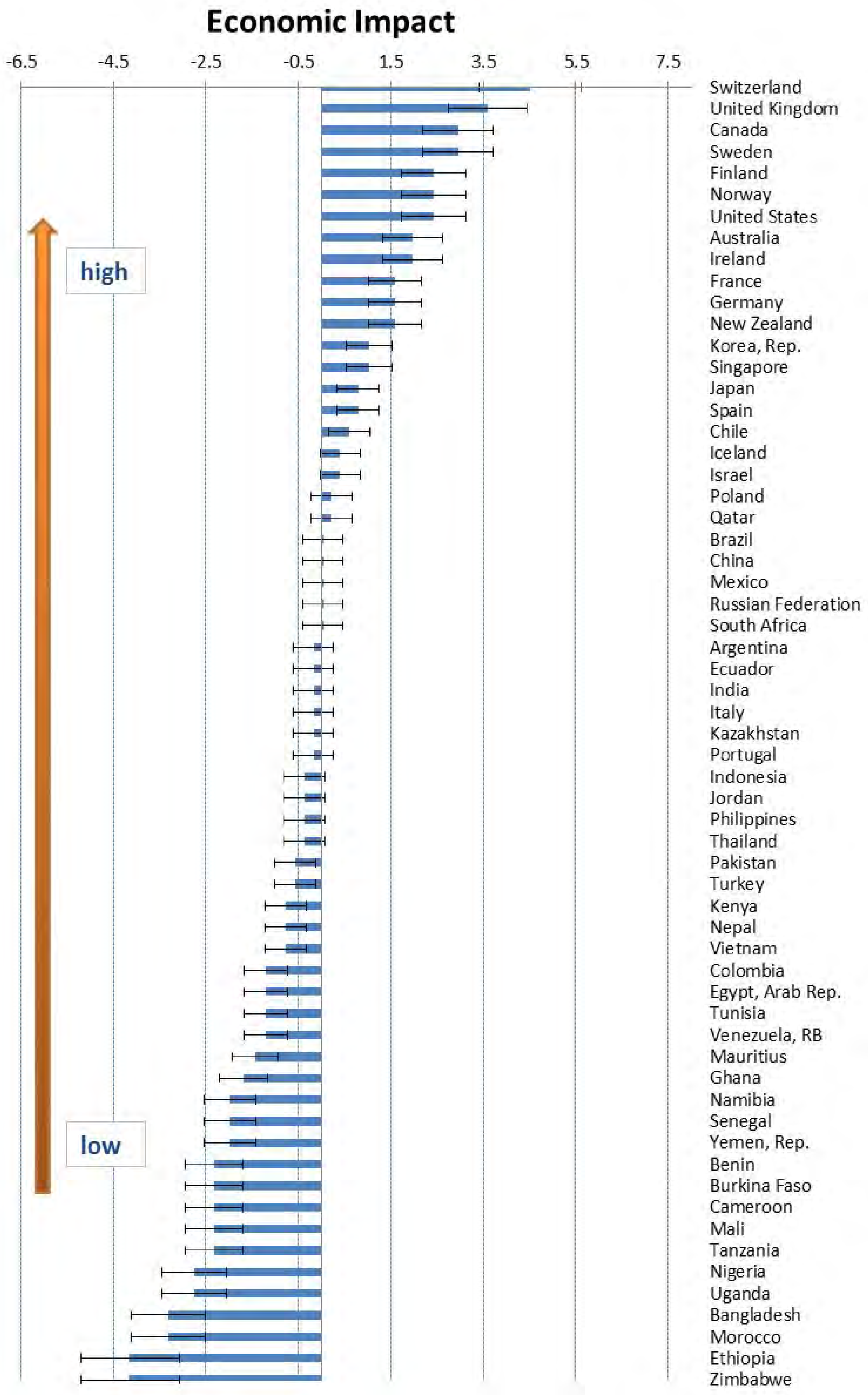


Figure 18. Countries performance in Economic Impact survey questions (error bars indicate ± 1SE).

2.3.5 Social Impact

The Social Impact component includes three survey questions all fitting well the model. Among the three, the one scoring higher on average is that related to usage of social networking sites. The lowest scores are associated to distance learning for teachers (see Figure 19).

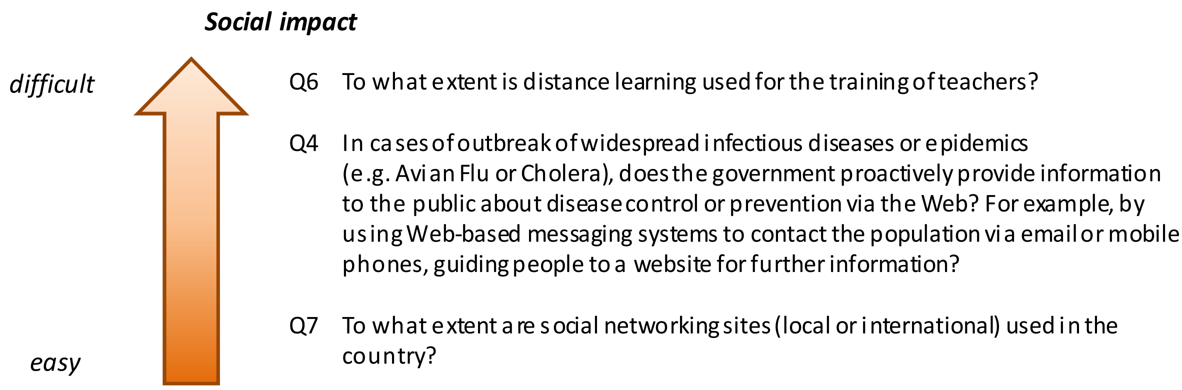


Figure 19. Order of questions in Social Impact component.

The model explains 72.7% of total variance. Rasch measures, standard errors and ranking are provided in Table 9 and Figure 20.

Table 9. Social Impact component: country scores and ranking

Country	Measure	S.E.	Ranking		Country	Measure	S.E.	Ranking
Canada	3.64	1.05	1		Venezuela	0.04	0.42	27
Australia	1.97	0.57	2		Argentina	-0.13	0.41	32
New Zealand	1.97	0.57	2		Indonesia	-0.13	0.41	32
United States	1.97	0.57	2		Tunisia	-0.13	0.41	32
Ireland	1.67	0.52	5		India	-0.29	0.41	35
Sweden	1.67	0.52	5		Jordan	-0.29	0.41	35
Norway	1.42	0.49	7		Singapore	-0.29	0.41	35
UK	1.42	0.49	7		Egypt	-0.46	0.4	38
China	1.19	0.47	9		Thailand	-0.46	0.4	38
Finland	1.19	0.47	9		Bangladesh	-0.62	0.4	40
Iceland	1.19	0.47	9		Russian Fed.	-0.62	0.4	40
Japan	0.98	0.45	12		Ghana	-0.78	0.41	42
Korea, Rep.	0.78	0.44	13		Mauritius	-0.78	0.41	42
Brazil	0.59	0.44	14		Nepal	-0.78	0.41	42
France	0.59	0.44	14		Ecuador	-0.96	0.42	45
Chile	0.4	0.43	16		Pakistan	-0.96	0.42	45
Germany	0.4	0.43	16		Tanzania	-0.96	0.42	45
Philippines	0.4	0.43	16		Cameroon	-1.14	0.44	48
South Africa	0.4	0.43	16		Morocco	-1.14	0.44	48
Colombia	0.22	0.42	20		Nigeria	-1.14	0.44	48
Israel	0.22	0.42	20		Burkina Faso	-1.35	0.47	51
Mexico	0.22	0.42	20		Ethiopia	-1.35	0.47	51
Portugal	0.22	0.42	20		Uganda	-1.35	0.47	51
Spain	0.22	0.42	20		Kenya	-1.58	0.51	54
Switzerland	0.22	0.42	20		Namibia	-1.58	0.51	54
Turkey	0.22	0.42	20		Vietnam	-1.58	0.51	54
Italy	0.04	0.42	27		Senegal	-1.87	0.57	57
Kazakhstan	0.04	0.42	27		Yemen, Rep.	-1.87	0.57	57
Poland	0.04	0.42	27		Benin	-2.25	0.67	59
Qatar	0.04	0.42	27		Mali	-2.25	0.67	59
					Zimbabwe	-2.81	0.84	61

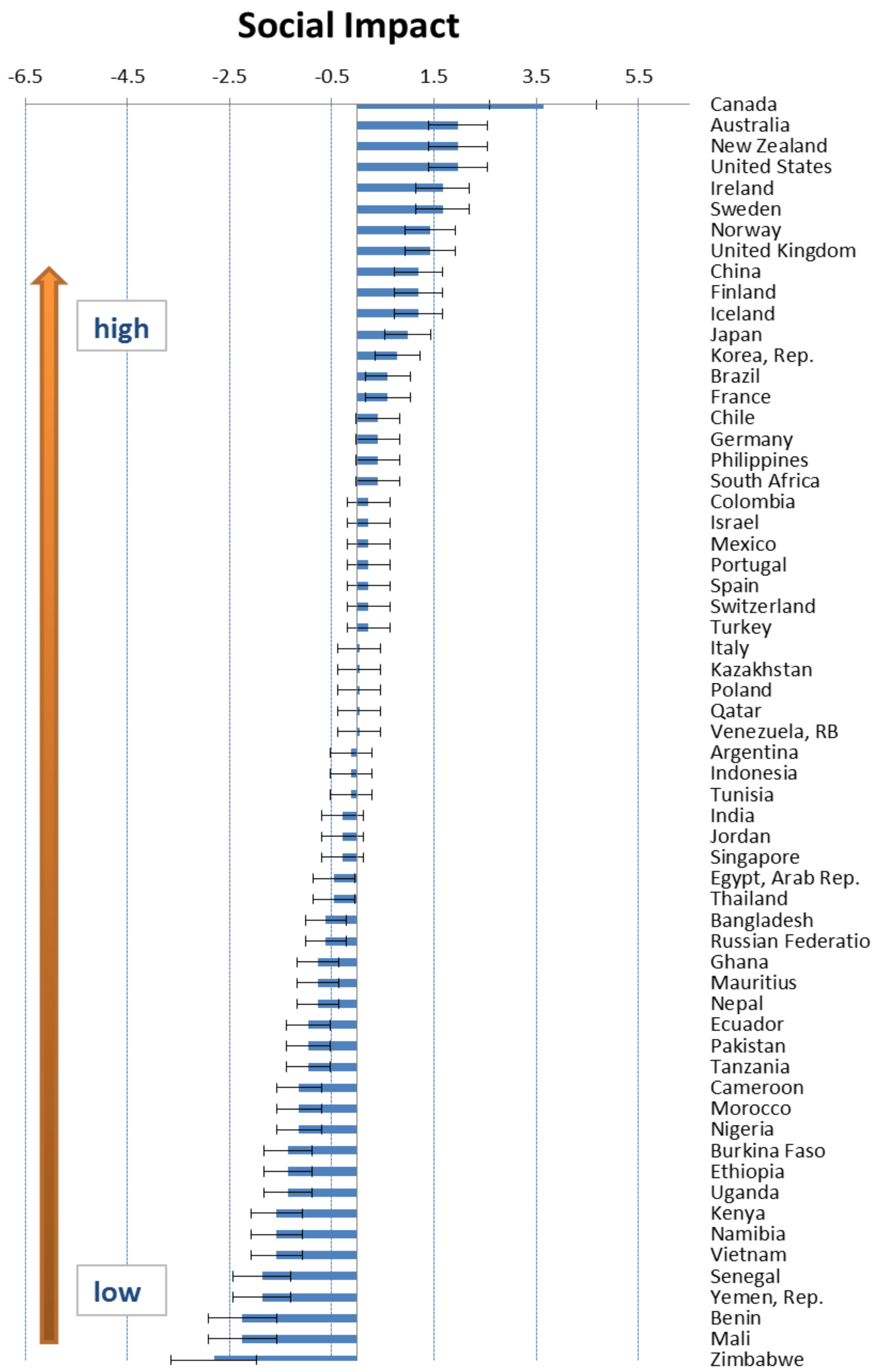


Figure 20: Countries performance in Social Impact survey questions (error bars indicate ± 1SE).

2.4 Do survey questions function in the same way?

The WI is computed for a wide set of countries world-wide with intrinsically different characteristics. It may occur that in some countries the understanding of certain questions is different than in others due, for example, to cultural and religious differences among the countries. All Rasch models assume that the questions in the survey function in the same way across all the countries. If not, the survey is likely to favor some countries while penalizing others and Rasch estimates will be biased. But more than model correctness, we are here interested in detecting questions which possibly behave differently for different groups of countries thus implying that scores on these questions are also driven by factors exogenous with respect to the *state and value of the Web*.

Rasch models provide the analyst with a statistical tool for detecting this source of bias that is called Differential Item Functioning (DIF). DIF analysis first requires to split the countries into different groups according to an external criterion. In the WI case the choice is to cluster the countries into five groups according to their geographical location, as also done by WI developers for regional analysis of the index (Table 10).

Table 10: Group of countries

Africa	Americas	Asia Pacific	Europe	Middle East & Central Asia
Benin	Argentina	Bangladesh	Finland	Israel
Burkina Faso	Mexico	India	France	Jordan
Cameroon	Colombia	Indonesia	Germany	Qatar
Egypt	Ecuador	Korea	Italy	Yemen
Ethiopia	Brazil	Nepal	Iceland	Kazakhstan
Ghana	Canada	New Zealand	Turkey	
Kenya	Chile	Pakistan	Poland	
Mali	United States	Philippines	Portugal	
Mauritius	Venezuela	Singapore	Ireland	
Morocco		China	Norway	
Namibia		Japan	Russia	
Nigeria		Thailand	Spain	
Senegal		Australia	Sweden	
South Africa		Viet Nam	Switzerland	
Tanzania			UK	
Tunisia				
Zimbabwe				
Uganda				

DIF analysis spots out possible interaction effects between these groups and the scores obtained on questions. If different groups of countries all behave almost the same for all the questions (null hypothesis, H_0), then it is reasonable to assume that estimated question difficulties do not differ if computed separately for each group or for the overall set of countries. The occurrence of statistically significant differences in question difficulties is instead a signal that in some countries the understanding of a certain question, and of the problem, is different than in the others. As aforementioned, it may be connected for example to the cultural differences (cultural bias) among countries. That is why it is particularly interesting to screen the WI survey questions with respect to DIF[7]. DIF is computed for all the questions, only the problematic ones are displayed in the following. The null hypothesis tested of no DIF is (Linacre 2012):

H_0: the question has the same difficulty as its average difficulty for all groups (Table 11);

Table 11: Summary of hypothesis H_0 verification

Component	Group	Question	Question difficulty	Scores
Institutional Infrastructure	Africa	Q9d	lower than expected	higher than expected
	Europe	Q9d	higher than expected	lower than expected
	Asia Pacific	Q9l	higher than expected	lower than expected
Web Content Q5 questions				
Web Content Q23 questions	Africa	Q23g	higher than excpected	lower than expected
	Europe	Q23g	lower than expected	higher than expected
Web Content remaining questions	Africa	Q22	lower than expected	higher than expected
	Europe	Q22	higher than expected	lower than expected
	Africa	Q24	lower than expected	higher than expected
Web Use				
Economic Impact				
Social Impact	Africa	Q6	lower than expected	higher than expected
	Europe	Q6	higher than expected	lower than expected

Three WI components do not show any problematic questions: Web Content, Web Use and Economic Impact. In the remaining components only few questions show a significant DIF. In the Institutional Infrastructure component question Q9d (*To what extend are boys encouraged to focus on science and technology*) performs differentially for countries in Europe and Africa. For European countries the scores are lower than expected while the opposite occurs in African countries. In the case of African countries it means that this question scores higher than modelled by the Rasch model. It implies that males are encouraged to focus on science and technology to a higher extent than the model predicts taking into account the scores on the other Institutional Infrastructure

[7] Since WI analysis is based on the Rasch, model uniform DIF is applied here, that is DIF is simply computed as the difference between question difficulty across groups of countries (Wilson 2005, pg.165).

questions. The opposite occurs for European countries where males are encouraged to focus on science and technology to a lower extent than the model predicts, given all the variables included in the component. This suggests that there is a gender issue in particular in the field of science and technology with a wide gap between African countries and European ones.

In this component question Q9l (*In your country, in tertiary education, what proportion of ICT graduates are women?*) also appears problematic. It shows DIF for the countries from the Asia Pacific region. Observed scores are lower than expected meaning that proportion of ICT female graduates is lower than predicted by the Rasch model.

In the Web Content component only one question is detected as functioning in a different way, Q23g (*To what extent are there government data on the Web in the following areas: Map data*). Countries from Africa score on this question too low and countries from Europe too high. Again a difference is detected between African and European groups of countries.

In the Web Content – remaining questions components – two questions show a significant DIF: Q22 (*To what extent are government agencies publishing information on the Web using open licenses*) and Q24 (*How easy is it to access government data on the Web in open, machine readable formats*). For question Q22 the analysis is the same as for the case of the Institutional Infrastructure, with African countries scoring higher than expected and European countries lower than expected. The interpretation is difficult in this case, still a significant difference between Africa and Europe is highlighted. For question Q24, African countries score higher than expected by the model.

In the Social Impact component only question Q6 (*To what extent is distance learning used for the training of teachers*) is differently functioning. The differences are again between European and African countries with the same pattern as for the Institutional Infrastructure and the Web Content – remaining questions – components.

Overall, the DIF analysis spots out differences between Africa and Europe for four questions (out of 49). It is important to note that this may be due to the fact that only these two groups of countries include similar countries with respect to cultural background and level of economic development. Choosing different grouping criteria would result in different DIF outcomes. In this case study we adopt the WI classification in order to provide the index developers with further insights into possible differences among the five groups of countries.

2.5 Highlights

The main recommendations derived from the Rasch analysis on WI primary data are summarized in Box 4.

Possible issue	Outcomes from Rasch analysis		Recommendation
Categories redundancy	Never detected	→	No action: the country sample is rather heterogeneous
Questions unexpected behaviour	**Q10**: higher variability than expected	→	Rewording: counter orientation with respect to the WI
	Q25: higher variability than expected		Rewording: too technical for respondents not experts in the field
	Q2a: higher variability than expected		Rewording: a dichotomous answer seems to be expected
	Q12: higher variability than expected		Rewording: misleading orientation with respect to the WI
Countries unexpected behaviour	All the countries are misfitting at least in one model	→	No action as this indicates that no specific pattern of responses is detected for a specific country
Single, underlying latent factor	The explanatory power of all the models is always satisfactory	→	No action: the framework of WI is confirmed

Box 4: Summary of Rasch analysis results

3 Uncertainty analysis of the WI

In every composite indicator the final index is the result of a number of choices on the framework (driven by socio-economic aspects and experts' opinion), the number of available variables to include, their normalization, the weights to attach to each variable, the aggregation algorithm, and so on. Some choices are openly normative and subjective, driven by developers' and experts' opinion, others can be justified on the basis of statistical analysis, mathematical simplicity or common practice. The aim of Uncertainty Analysis of composite indicators – UA - is to assess to what extent these choices affect the composite indicator scores and country ranking.

Due the structure of the WI we decided to avoid plugging all "uncertainties" in a big experiment with the aim of checking their simultaneous, joint influence on the final score/rank. This approach would have made it difficult to disentangle influential factors and to fully understand the implication of their variability. Instead, we chose to focus on some critical key points, defined and discussed together with the Index developers, and to check their influence on the WI scores - at the component and sub-index level - and on the country ranking. The check is done by calculating the differences in scores/ranks between the reference scenario (i.e. the reference scores and ranks provided by World Wide Web Foundation) and the "simulated scenario", that is the scenario including one of the alternative hypotheses.

The robustness assessment of the WI consists of the following steps. Six different scenarios are set up each containing a different assumption with respect to the reference WI:

1. Scenario #1: Weighting. The reference scenario for WI assigns unequal weights on the basis of experts decision, as discussed in Section 1. In this scenario we set up a Monte Carlo experiment to assess the impact of different weighting schemes used both at the component level and at the sub-indexes level. In this scenario we also compare the WI with the alternative hypothesis of equal weighting to all components and sub-indexes.
2. Scenario #2: Different aggregation method for 3 indicators describing web access modality in the Communications Infrastructure component. To each country the maximum value of the three indicators is assigned.
3. Scenario #3: Inclusion of 4 additional indicators Q9a, Q9b, Q9c, Q9d, describing the level of possible gender bias in the Institutional Infrastructure component;

4. Scenario #4: Different quantification of survey data. Rasch scores are used to replace raw survey data in the WI computation.
5. Scenario #5: Compensability. Geometric aggregation is employed instead of linear aggregation for combining components and sub-indexes.
6. Scenario #6: The contribution of each component and sub-index to the WI is assessed by evaluating the role of each component and the correlation structure of the dataset.

All simulations are based on ad-hoc Matlab scripts (The MathWorks^{TM}, ver. R2009a). For technical reasons countries are referred to with labels displayed in Table 12.

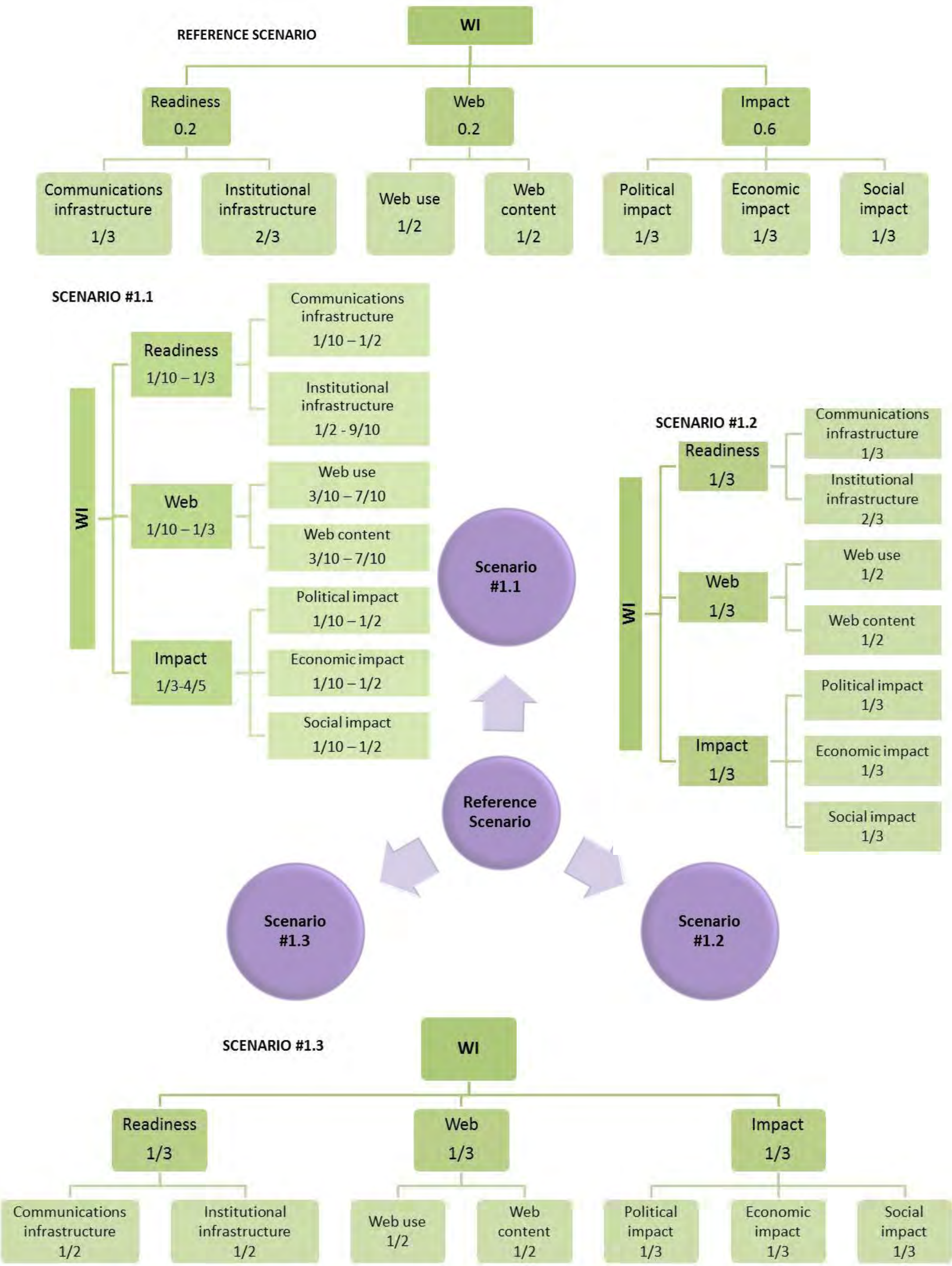


Figure 21: Uncertainty analysis – scenario#1.1, scenario #1.2, scenario #1.3

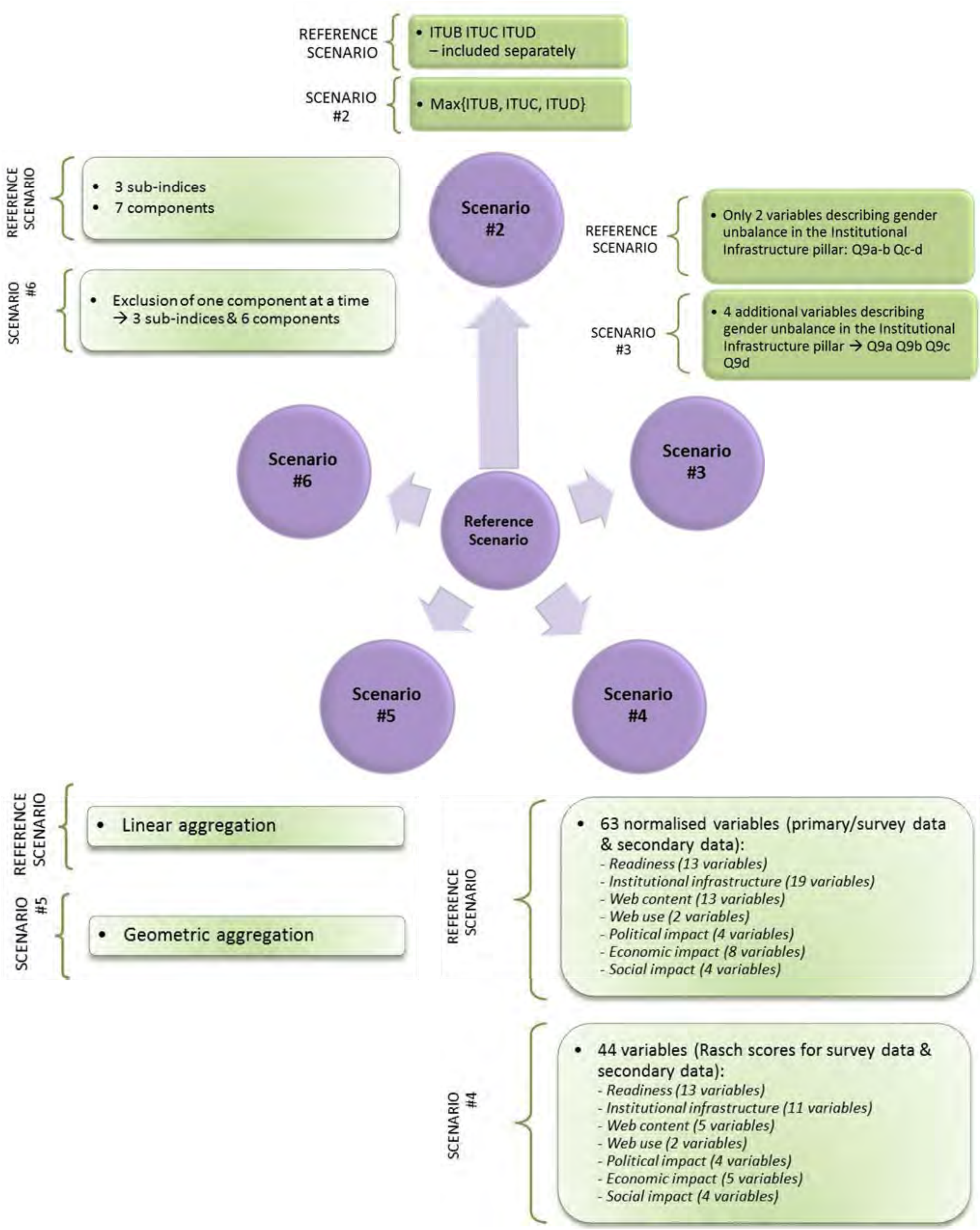

REFERENCE SCENARIO
• ITUB ITUC ITUD – included separately
SCENARIO #2
• Max{ITUB, ITUC, ITUD}
REFERENCE SCENARIO
• 3 sub-indices
• 7 components
SCENARIO #6
• Exclusion of one component at a time → 3 sub-indices & 6 components
REFERENCE SCENARIO
• Only 2 variables describing gender unbalance in the Institutional Infrastructure pillar: Q9a-b Qc-d
SCENARIO #3
• 4 additional variables describing gender unbalance in the Institutional Infrastructure pillar → Q9a Q9b Q9c Q9d
Scenario #2
Scenario #6
Reference Scenario
Scenario #3
Scenario #5
Scenario #4
REFERENCE SCENARIO
• Linear aggregation
SCENARIO #5
• Geometric aggregation
REFERENCE SCENARIO
• 63 normalised variables (primary/survey data & secondary data):
- Readiness (13 variables)
- Institutional infrastructure (19 variables)
- Web content (13 variables)
- Web use (2 variables)
- Political impact (4 variables)
- Economic impact (8 variables)
- Social impact (4 variables)
SCENARIO #4
• 44 variables (Rasch scores for survey data & secondary data):
- Readiness (13 variables)
- Institutional infrastructure (11 variables)
- Web content (5 variables)
- Web use (2 variables)
- Political impact (4 variables)
- Economic impact (5 variables)
- Social impact (4 variables)

Figure 22: Uncertainty analysis – scenario#2 to scenario#6

Table 12: Country labels used in the analysis

Country name	Country label	Country name	Country label
Argentina	ARG	Mauritius	MUS
Australia	AUS	Mexico	MEX
Bangladesh	BGD	Morocco	MAR
Benin	BEN	Namibia	NAM
Brazil	BRA	Nepal	NPL
Burkina Faso	BFA	New Zealand	NZL
Cameroon	CMR	Nigeria	NGA
Canada	CAN	Norway	NOR
Chile	CHL	Pakistan	PAK
China	CHN	Philippines	PHL
Colombia	COL	Poland	POL
Ecuador	ECU	Portugal	PRT
Egypt	EGY	Qatar	QAT
Ethiopia	ETH	Russia	RUS
Finland	FIN	Senegal	SEN
France	FRA	Singapore	SGP
Germany	DEU	South Africa	ZAF
Ghana	GHA	Spain	ESP
Iceland	ISL	Sweden	SWE
India	IND	Switzerland	CHE
Indonesia	IDN	Tanzania	TZA
Ireland	IRL	Thailand	THA
Israel	ISR	Tunisia	TUN
Italy	ITA	Turkey	TUR
Japan	JPN	Uganda	UGA
Jordan	JOR	United Kingdom	GBR
Kazakhstan	KAZ	United States	USA
Kenya	KEN	Venezuela	VEN
Korea (Rep. of)	KOR	Viet Nam	VNM
Mali	MLI	Yemen	YEM
		Zimbabwe	ZWE

3.1 Scenario #1: weighting

3.1.1 Scenario 1.1: Country volatility as function of weights

The WI is computed by using a differential weighting scheme (Section 1) which aims to capture the different relevance of the components/sub-indexes in describing the *state and value of the Web*. This predefined set of weights is modified allowing weights to vary within predefined intervals decided together with WI developers. The WI sensitivity to different weighting schemes is then checked calculating, for each country, the difference between the reference score (rank) and the modified score (rank). Component and sub-index weights play the role of uncertain parameters of the UA. The WI case includes 10 uncertain parameters: the seven component weights plus the three sub-index weights (labels used for the WI weights in Table 13).

Table 13: Labels used for weights at the component and sub-index level

Component level						
Comm. Infrastructure	Inst. Infrastructure	Web Use	Web Content	Social Impact	Economic Impact	Political Impact
weight1	weight2	weight3	weight4	weight5	weight6	weight7
Sub-index level						
Readiness		The Web		Impact		
w_readiness		w_web		w_impact		

The distribution of each parameter is assumed to be a continuous uniform centred in the corresponding WI reference value (Table 14). The choice of the range of uncertainty is driven by two opposite needs: on the one hand, there is the need to ensure a 'wide enough' interval to have a meaningful check on scores volatility. On the other hand, it is recommended to avoid spoiling the rationale of the WI weighting scheme because it is designed to take into account the different relative relevance of the different aspects measured by the Index.

A total number of 1200 different simulations are run by randomly sampling across the different weight distributions. Each scenario corresponds to a different set of values for the weights of the seven components and the three sub-indexes. The sampling procedure is not completely at random as the following constraints on weights must hold:

$$\begin{aligned} &\text{weight1} + \text{weight2} = 1 \\ &\text{weight3} + \text{weight4} = 1 \\ &\text{weight5} + \text{weight6} + \text{weight7} = 1 \\ &\text{w\_readiness} + \text{w\_web} + \text{w\_impact} = 1 \end{aligned} \tag{1}$$

The sampling strategy for weights is adapted to take into account constraints on weights. In other words weight values cannot be sampled from these distributions in a fully independent way. In order to balance the sampling strategy, sampling is carried out by permuting the order selection of the weights. For this reason the final distributions of the simulated weights are no longer perfectly uniform as the presence of tails can be noted. However this is not affecting the validity of the uncertainty analysis. Figure 23 to Figure 26 show the actual weight distributions (absolute frequencies).

Table 14: Uncertainty intervals assigned to weights

Parameter	WI Reference value	Range of variability
weight1	1/3	U[1/10,1/2]
weight2	2/3	U[1/2,9/10]
weight3	1/2	U[3/10,7/10]
weight4	1/2	U[3/10,7/10]
weight5	1/3	U[1/10,1/2]
weight6	1/3	U[1/10,1/2]
weight7	1/3	U[1/10,1/2]
w_readiness	1/5	U[1/10,1/3]
w_web	1/5	U[1/10,1/3]
w_impact	3/5	U[1/3,4/5]

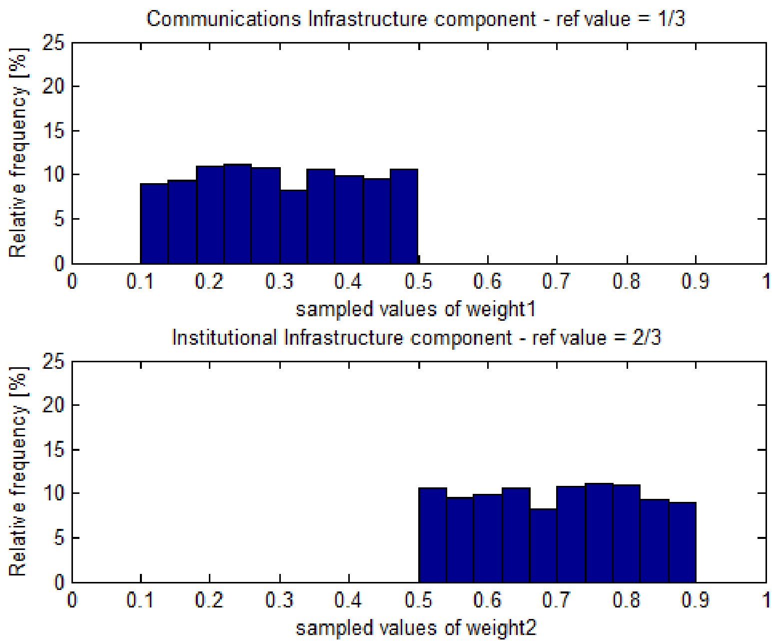


Figure 23: Weight distributions: weight1 (Comm. Infrastructure) and weight2 (Inst. Infrastructure)

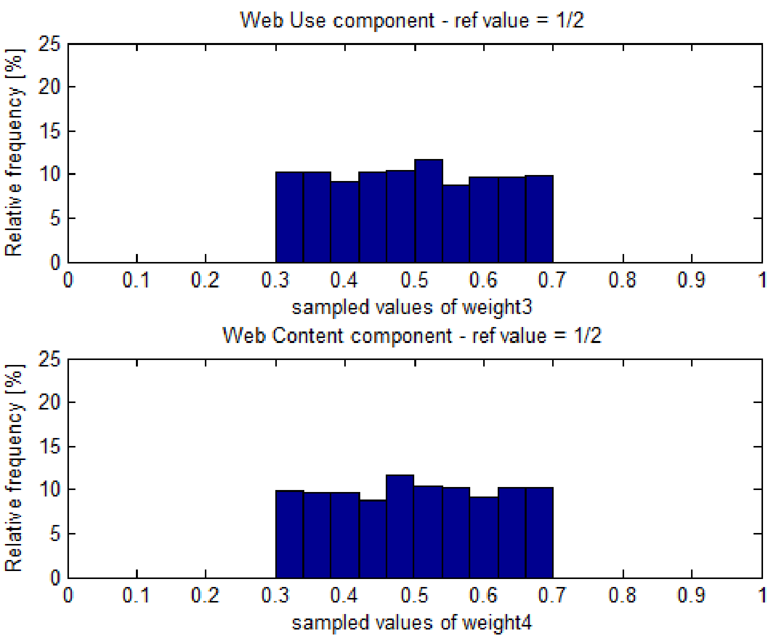


Figure 24: Weight distributions: weight3 (Web Use) and weight4 (Web Content)

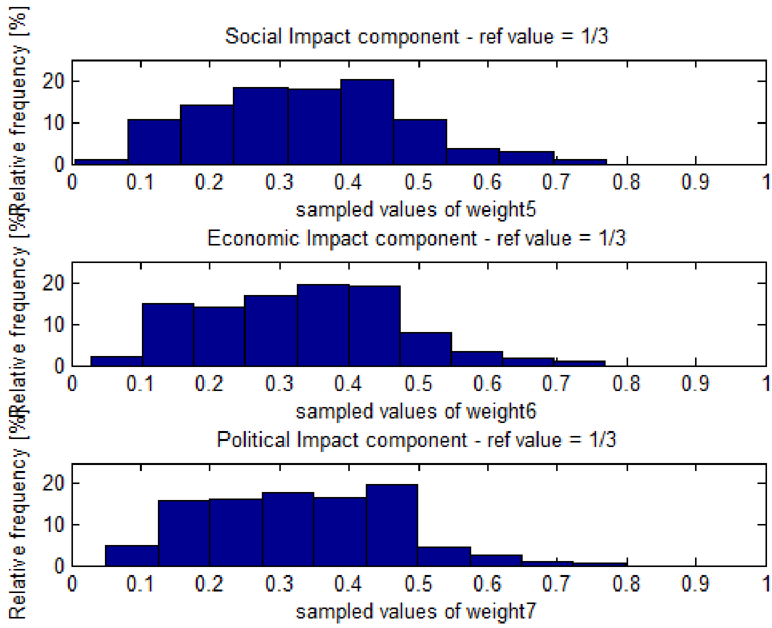


Figure 25: Weight distributions: weight5 (Social Impact), weight6 (Economic Impact) and weight 7 (Political Impact)

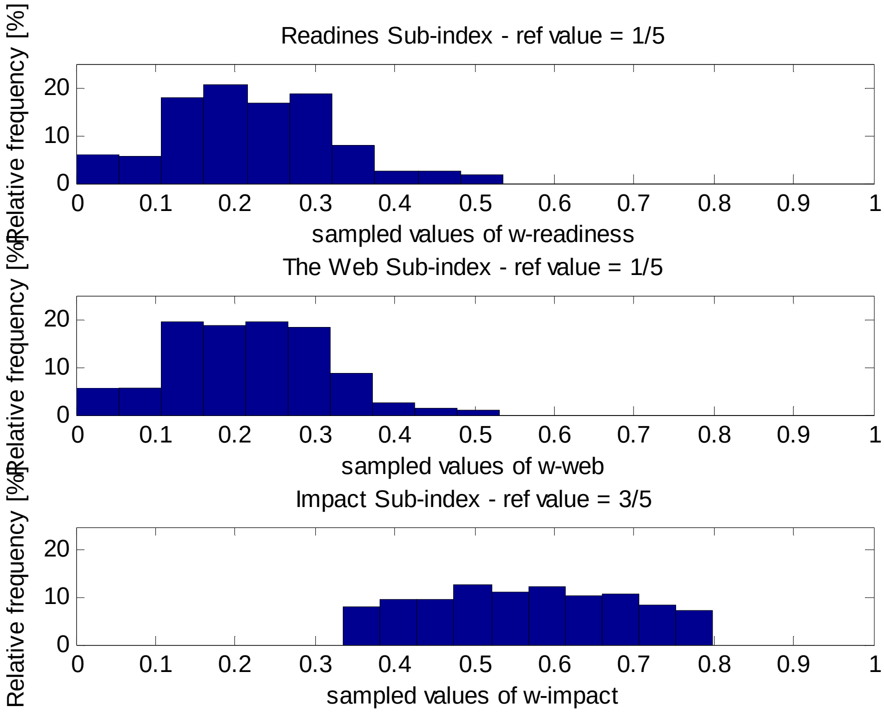


Figure 26: Weight distributions: w_readiness, w_web, w_impact

The main outcome of the uncertainty analysis is presented in Figure 27, Figure 28, Figure 29 and Figure 30. They show the distribution of the differences between the reference rank and the modified rank, computed with the simulated set of weights, for the three sub-indexes and the WI. In all the Figures, the median shift in rank (across 1200 simulations) is the red segment. The vertical boxes show the 75% of the distributions (percentiles P25 and P75 are the horizontal edges of the boxes) and vertical lines extend from minimum to maximum values.

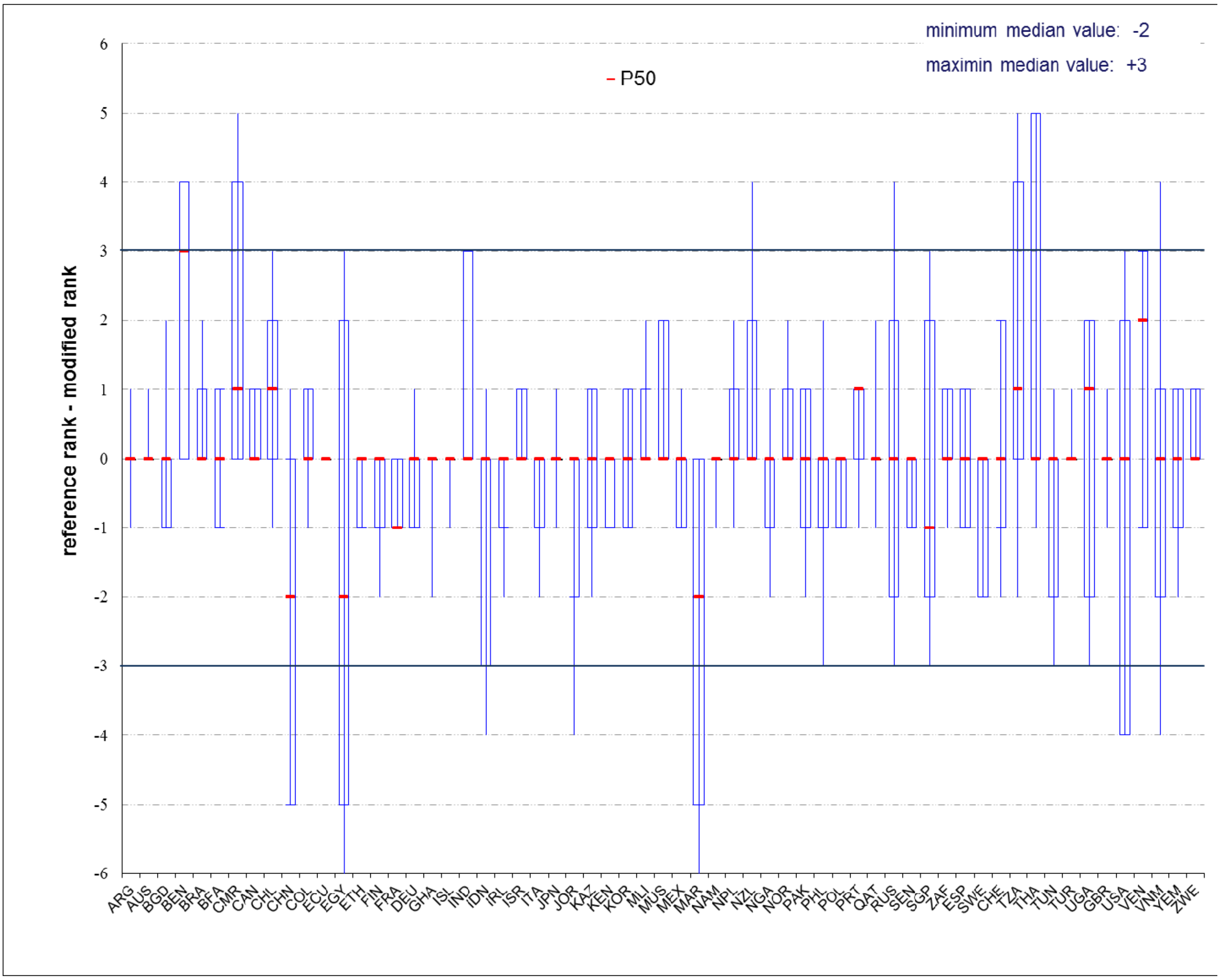


Figure 27: UA on weights: effects on country ranking for the Readiness sub-index

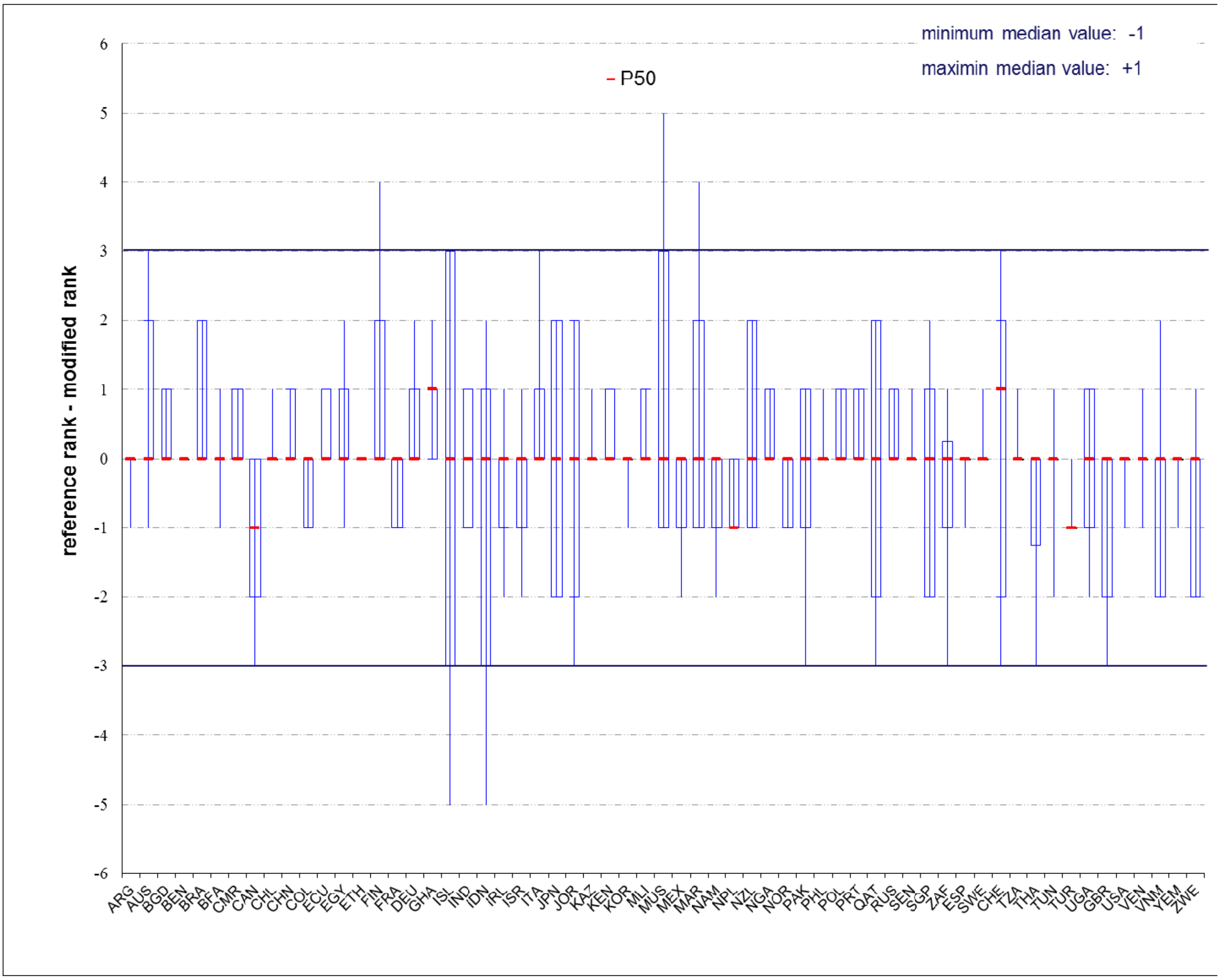


Figure 28: UA on weights: effects on country ranking for the Web sub-index

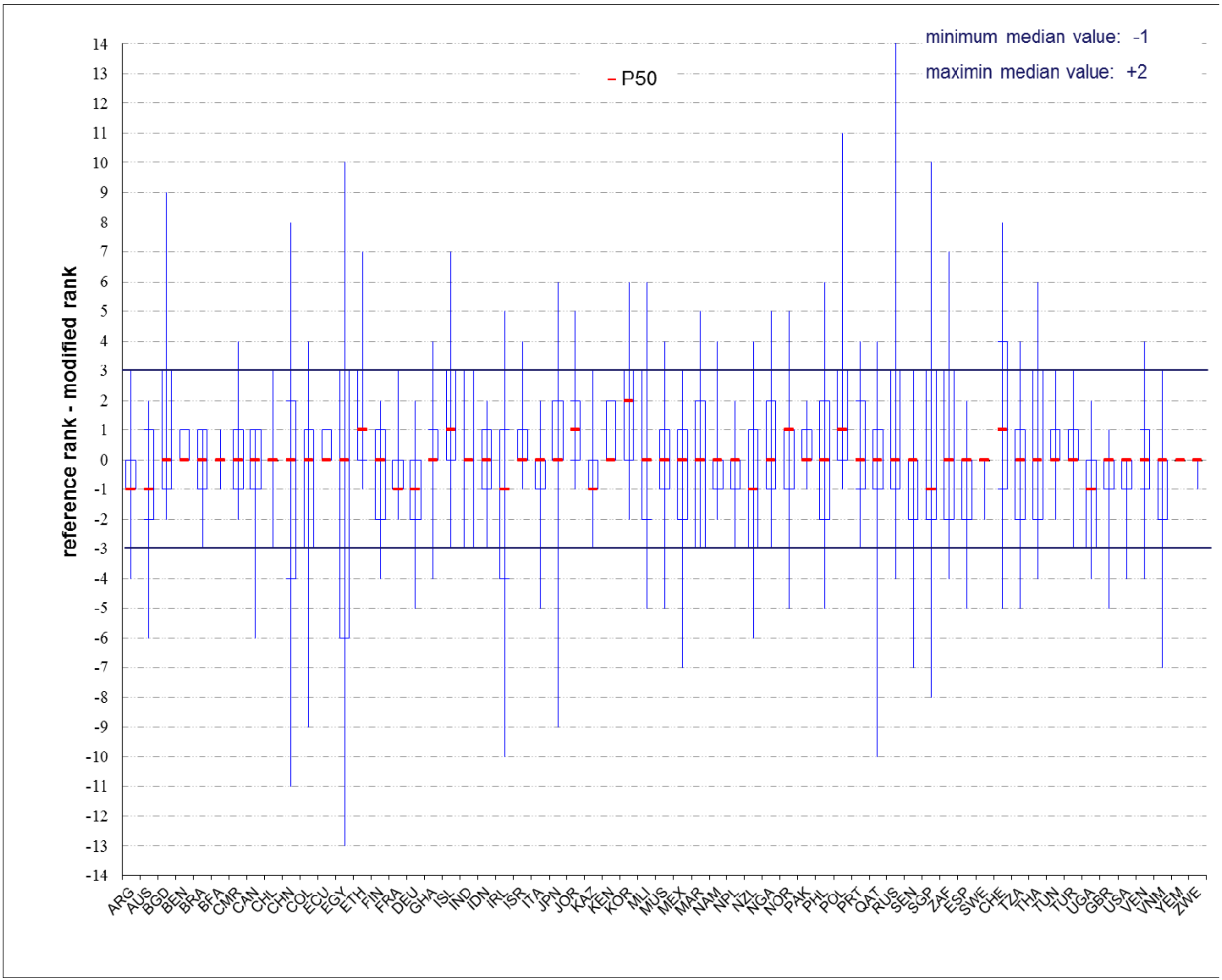


Figure 29: UA on weights: effects on country ranking for the Impact sub-index

The ranking proves to be rather robust. The median shift in rank always lies within the interval [-1,+2] meaning that on average the shift in country rank is up to 2 positions for all simulations considered (each corresponding to a different set of weights). The band ±3 (marked by horizontal red lines in the Figures) is considered as a tolerance band of rank variation and indicates about 5% of rank shift. This band is considered here as a rule of thumb to highlight volatile countries. In most cases boxplots lie within this band with some few exceptions: eight countries for the Readiness sub-index (BEN, CMR, CHN, EGY, MAR, TZA, THA, USA), four countries for the Impact sub-index (CHN, EGY, IRL, CHE) and one country for the WI (RUS). This means that at least 50% of the times the absolute shift in rank is not higher than 3 positions for most of the countries.

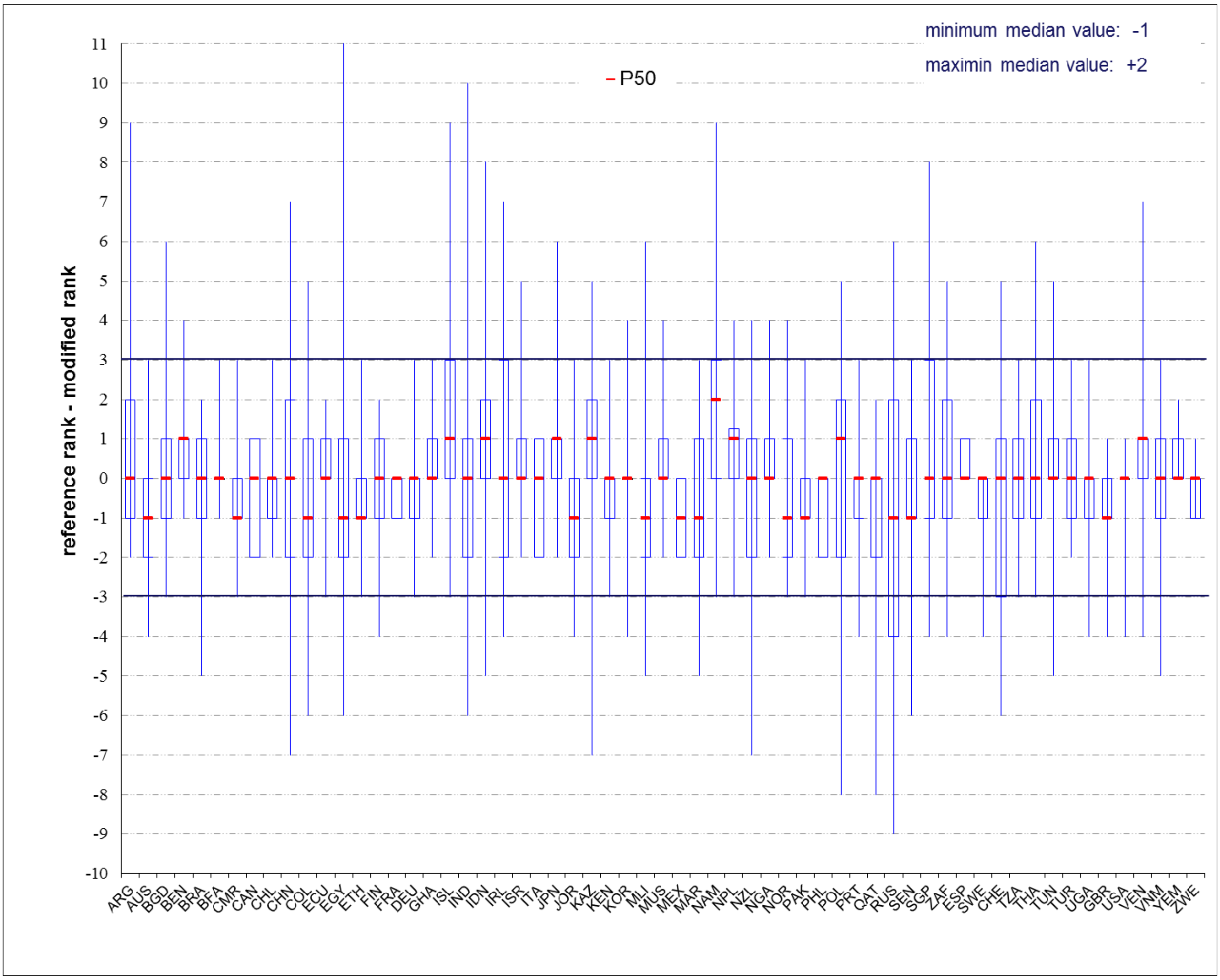


Figure 30: UA on weights: effects on country ranking for the WI

Table 15 shows the frequency matrix of modified ranks. This matrix displays for each country the percentage of times the country ranks within a certain rank interval calculated over all the 1200 simulated scenarios. The frequency matrix shows most and least stable countries and provides a synthesized picture of the overall country ranking. Frequency distribution is classified into 12 intervals 5 ranks wide, with the exception of the last interval which is 6 ranks wide ([1,5], [6, 10], …, [56, 61]). Countries are reordered from best to worst according to their original WI rank. A country is considered ‘stable’ if its rank frequency is higher or equal to 95% (highlighted in blue in Table 15). ‘Volatile’ countries are instead those with rank values spanning at least three rank intervals (highlighted in yellow in the Table).

Top and stable countries in the Table are Switzerland and United States, which are always among the top five, whatever the weighting scheme adopted. They are the top performers in terms of the overall *state and value of the Web* as measured by the WI. At the lower, right end of the frequency matrix, the following countries are found: Benin, Burkina Faso, Zimbabwe and Yemen. These countries are stable, low performers in all simulations as they rank among the worst six for all the 1200 different choices of WI weights. Most unstable ranks are mostly associated to middle performing countries which are probably characterized by very similar scores: for these countries a small variation in the score produces a remarkable variation in the rank.

Figure 31 shows the median rank of the countries and the associated 90% confidence interval – 90% CI - computed across all the 1200 scenarios. Countries are reordered from best to worst performers according to their reference WI rank (in red). Median ranks are shown as blue dots while error bars represent 5^{th} and 95^{th} percentiles of the rank distribution for each country. Differences between median and reference ranks are almost negligible, indicating that there are no ‘hotspot-countries’ in terms of volatile ranks. Countries with an estimated confidence interval higher than 6, meaning an oscillation of the country rank 6 positions wide (representing about 10% of shift in rank) are: Switzerland, Ireland, Singapore, Colombia, Poland, China, Russia, India and Egypt. It is Egypt which has the widest 90% CI = 12.

The picture given by the uncertainty analysis on the WI weighting scheme indicates that the Index is rather robust with respect to weight changes.

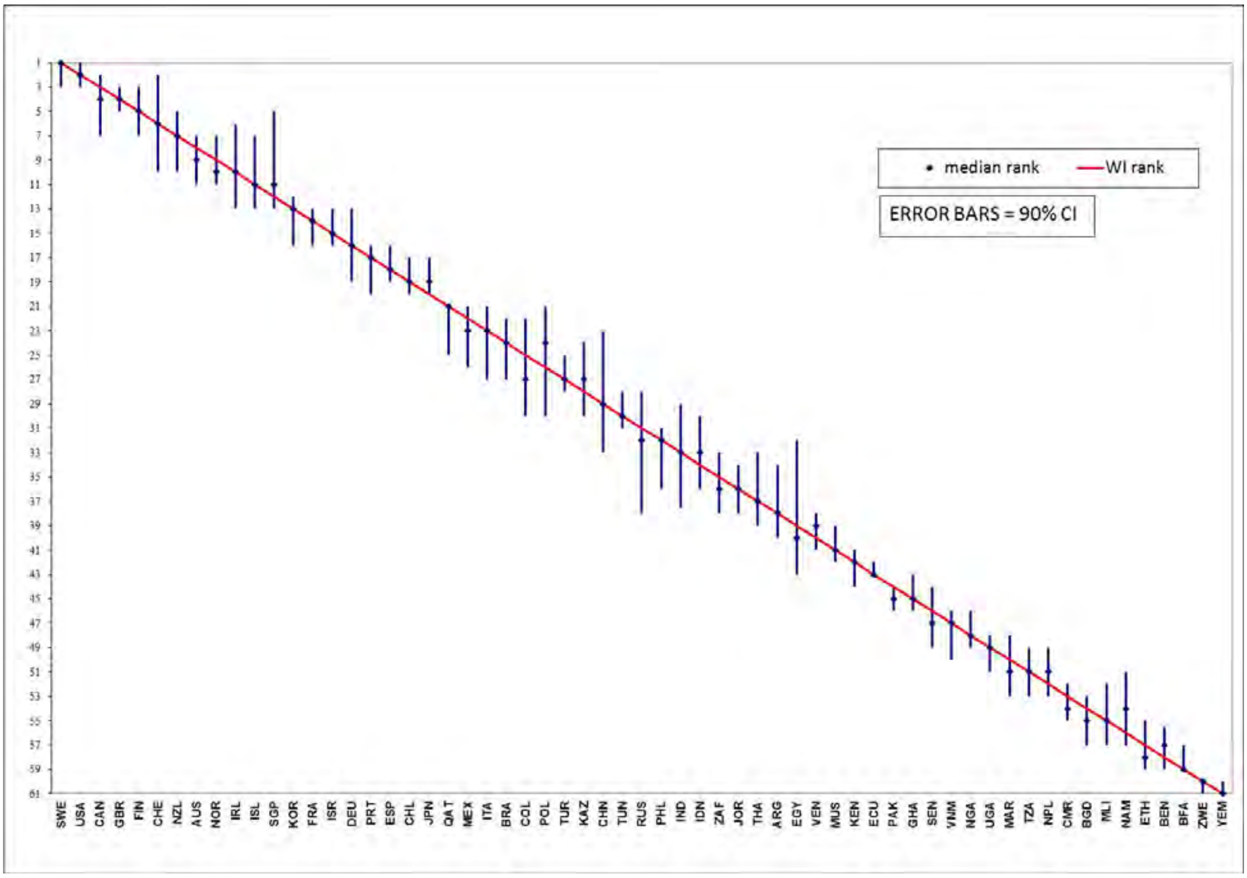


Figure 31: Median and 90% CI of rank distributions

Table 15 Frequency matrix of the country rank for the WI (only frequencies >5% are displayed).

Country	WI rank	[1,5]	[6,10]	[11,15]	[16,20]	[21,25]	[26,30]	[31,35]	[36,40]	[41,45]	[46,50]	[51,55]	[56,61]
SWE	1	100											
USA	2	100											
CAN	3	99											
GBR	4	72	28										
FIN	5	65	35										
CHE	6	39	57										
NZL	7	10	89										
AUS	8		86	14									
NOR	9		71	29									
IRL	10		48	47									
ISL	11	8	40	51									
SGP	12		44	54									
KOR	13			92	7								
FRA	14			94	6								
ISR	15			88	12								
DEU	16			24	76								
PRT	17				97								
ESP	18				100								
CHL	19				100								
JPN	20				98								
QAT	21					95							
MEX	22					89	9						
ITA	23					84	16						
BRA	24					78	22						
COL	25					62	36						
POL	26					30	68						
TUR	27					15	85						
KAZ	28					24	72						
CHN	29					15	58	26					
TUN	30						81	19					
RUS	31						34	44	22				
PHL	32							89	9				
IND	33						7	69	21				
IDN	34						6	88	6				
ZAF	35							43	57				
JOR	36							46	54				
THA	37							33	68				
ARG	38							17	82				
EGY	39							16	48	35			
VEN	40								91	8			
MUS	41								38	62			
KEN	42									96			
ECU	43									100			
PAK	44									83	17		
GHA	45									83	17		
SEN	46									28	69		
VNM	47										95		
NGA	48										98		
UGA	49										90	10	
MAR	50										49	51	
TZA	51										34	66	
NPL	52										25	75	
CMR	53											98	
BGD	54											56	43
MLI	55											52	47
NAM	56											70	27
ETH	57											14	86
BEN	58											5	95
BFA	59												100
ZWE	60												100
YEM	61												100

3.1.2 Scenarios 1.2 & 1.3: Equal weighting

As discussed in Section 1, the reference WI assigns different weights to sub-indexes, giving more relevance to the Impact sub-index while equally weighting the other two. Also two components in the Readiness sub-index, Communications Infrastructure and Institutional Infrastructure, are assigned different weights. An equal weighting scheme is tested first at the sub-index level, then at both the component and sub-index level. Figure 32 and Figure 33 display the results. In both cases no country shifts its rank on average by more than 2 positions. The highest shift is for Iceland which improves 5 positions in the rank and Argentina whose improvement is of 4 positions under both hypotheses of equal weighting (see Table 16). Namibia scores 4 positions higher only when both components and sub-indexes are equally weighted. Overall, the hypothesis of equal weighting either at the sub-index level or at both component and sub-index level does not imply major changes in ranks indicating a substantial robustness of the WI. Indeed as we will see later in Scenario 6 components are so highly correlated among themselves which generally implies that the choice of the weighting scheme plays a minor role in defining the ranking (Hagerty and Land, 2007; Michalos, 2011).

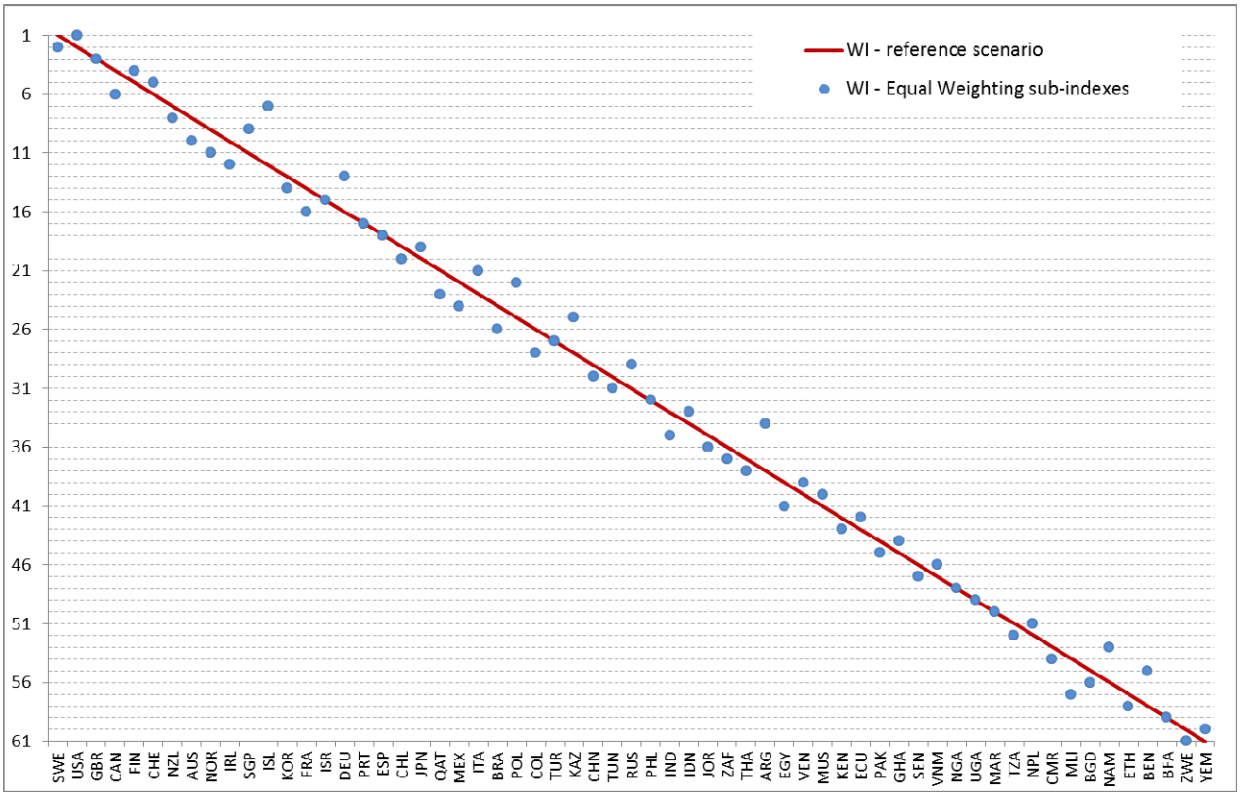


Figure 32: Ranking of countries in WI and the scenario with equal weighting of sub-indexes

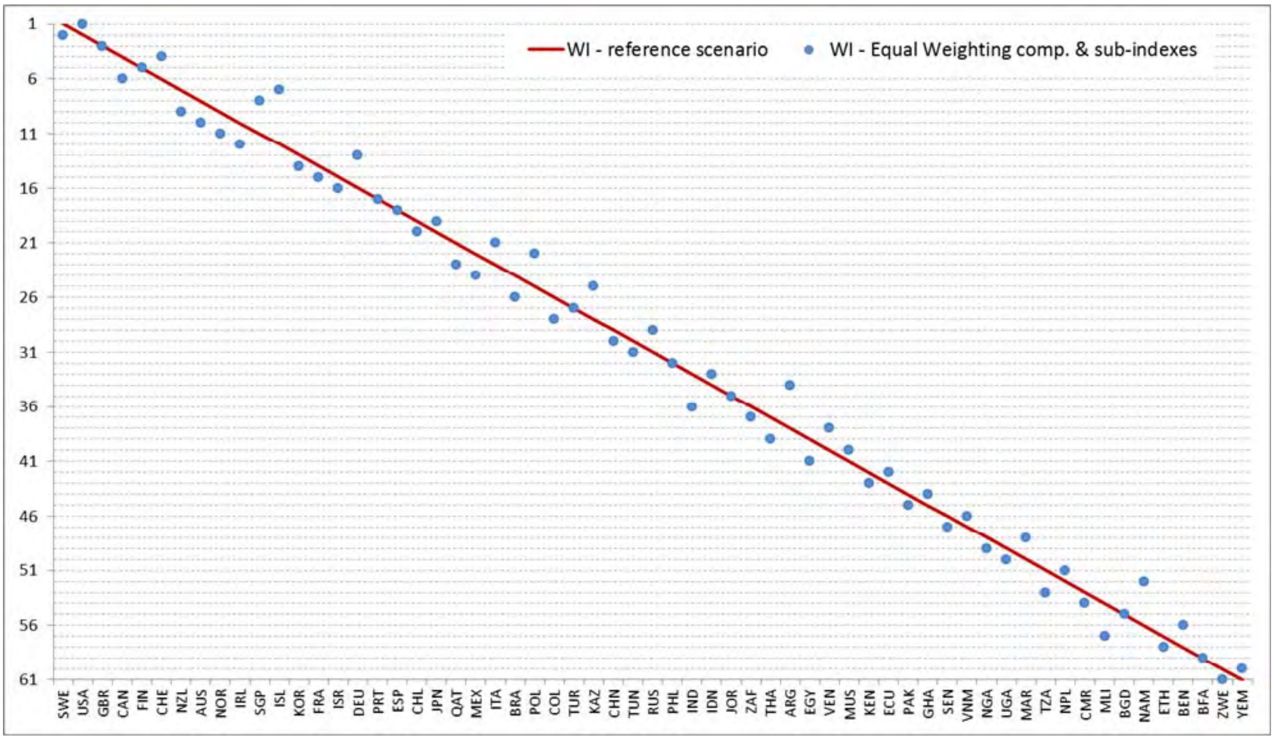


Figure 33: Ranking of countries in WI and the scenario with equal weighting of components and sub-indexes.

Table 16: Equal Weighting: summary of rank shifts

Equal Weighting components & sub-indexes				Equal Weighting sub-indexes			
Country	*reference rank-modified rank*	Country	*reference rank-modified rank*	Country	*reference rank-modified rank*	Country	*reference rank-modified rank*
SWE	-1	IND	-3	SWE	-1	IND	-2
USA	1	IDN	1	USA	1	IDN	1
CAN	-2	ZAF	-1	CAN	-2	JOR	-1
CHE	2	THA	-2	FIN	1	ZAF	-1
NZL	-2	ARG	4	CHE	1	THA	-1
AUS	-2	EGY	-2	NZL	-1	ARG	4
NOR	-2	VEN	2	AUS	-2	EGY	-2
IRL	-2	MUS	1	NOR	-2	VEN	1
SGP	3	KEN	-1	IRL	-2	MUS	1
ISL	5	ECU	1	SGP	2	KEN	-1
KOR	-1	PAK	-1	ISL	5	ECU	1
FRA	-1	GHA	1	KOR	-1	PAK	-1
ISR	-1	SEN	-1	FRA	-2	GHA	1
DEU	3	VNM	1	DEU	3	SEN	-1
CHL	-1	NGA	-1	CHL	-1	VNM	1
JPN	1	UGA	-1	JPN	1	TZA	-1
QAT	-2	MAR	2	QAT	-2	NPL	1
MEX	-2	TZA	-2	MEX	-2	CMR	-1
ITA	2	NPL	1	ITA	2	MLI	-3
BRA	-2	CMR	-1	BRA	-2	BGD	-1
POL	3	MLI	-3	POL	3	NAM	3
COL	-2	NAM	4	COL	-2	ETH	-1
KAZ	3	ETH	-1	KAZ	3	BEN	3
CHN	-1	BEN	2	CHN	-1	ZWE	-1
TUN	-1	ZWE	-1	TUN	-1	YEM	1
RUS	2	YEM	1	RUS	2		

3.2 Scenario #2: different aggregation for 3 indicators

Among WI indicators describing the Communications Infrastructure component three variables describe the modality people use to access the Web: ITUB – Broadband subscribers per 100 population, ITUC – % of households with a personal computer, ITUD – Mobile phone subscriptions per 100 population. In the reference scenario they are included as separate variables and averaged with the other variables of the component. One could argue that the important thing to capture is the access per se, no matter how it is obtained. Then it is not proper to penalize a country on the fact that people access the Web mainly by only one of these modalities. Therefore a new variable is computed - $\max\{ITUB, ITUC, ITUD\}$ - to replace the original ones. The modified ranks are then compared to those of the reference WI for the overall index, the sub-index Readiness and the component Communications Infrastructure (Figure 34, Figure 35, Figure 36).

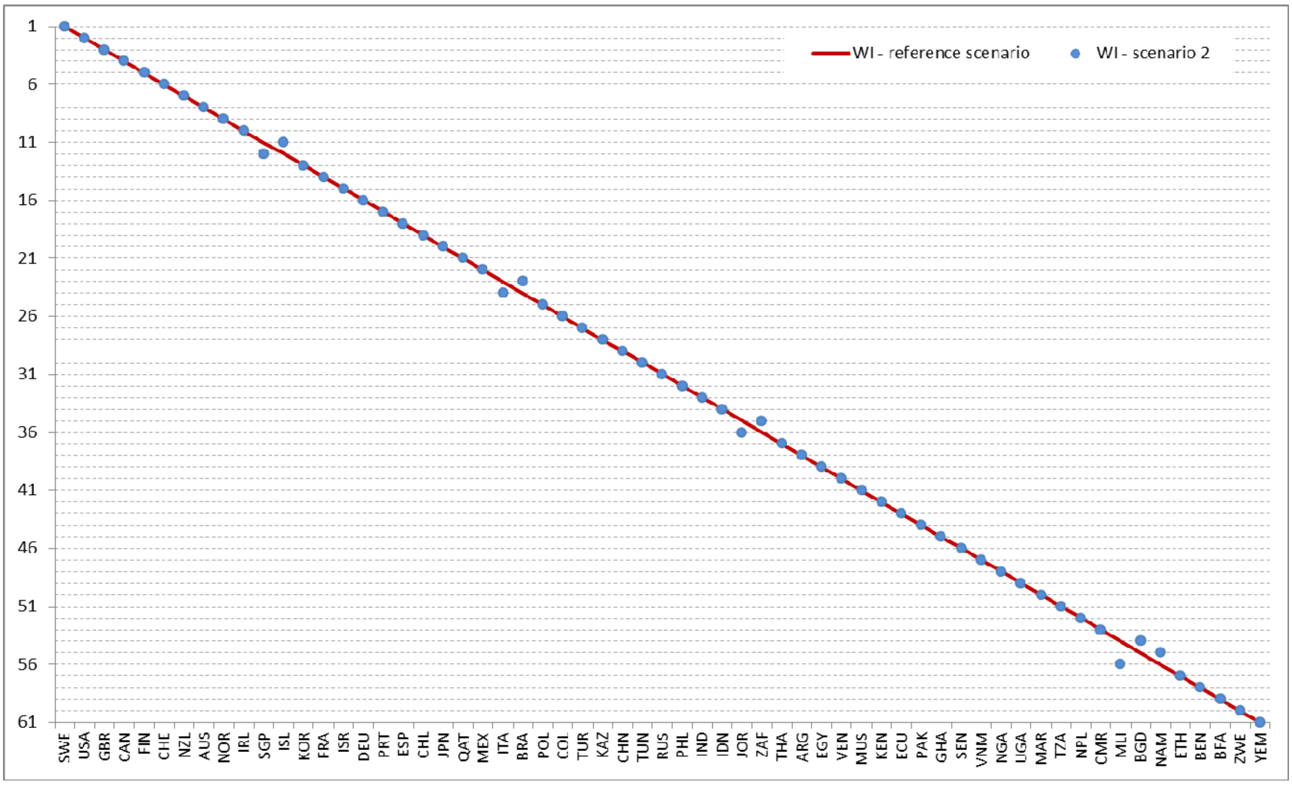


Figure 34: Ranking of countries in WI – scenario #2

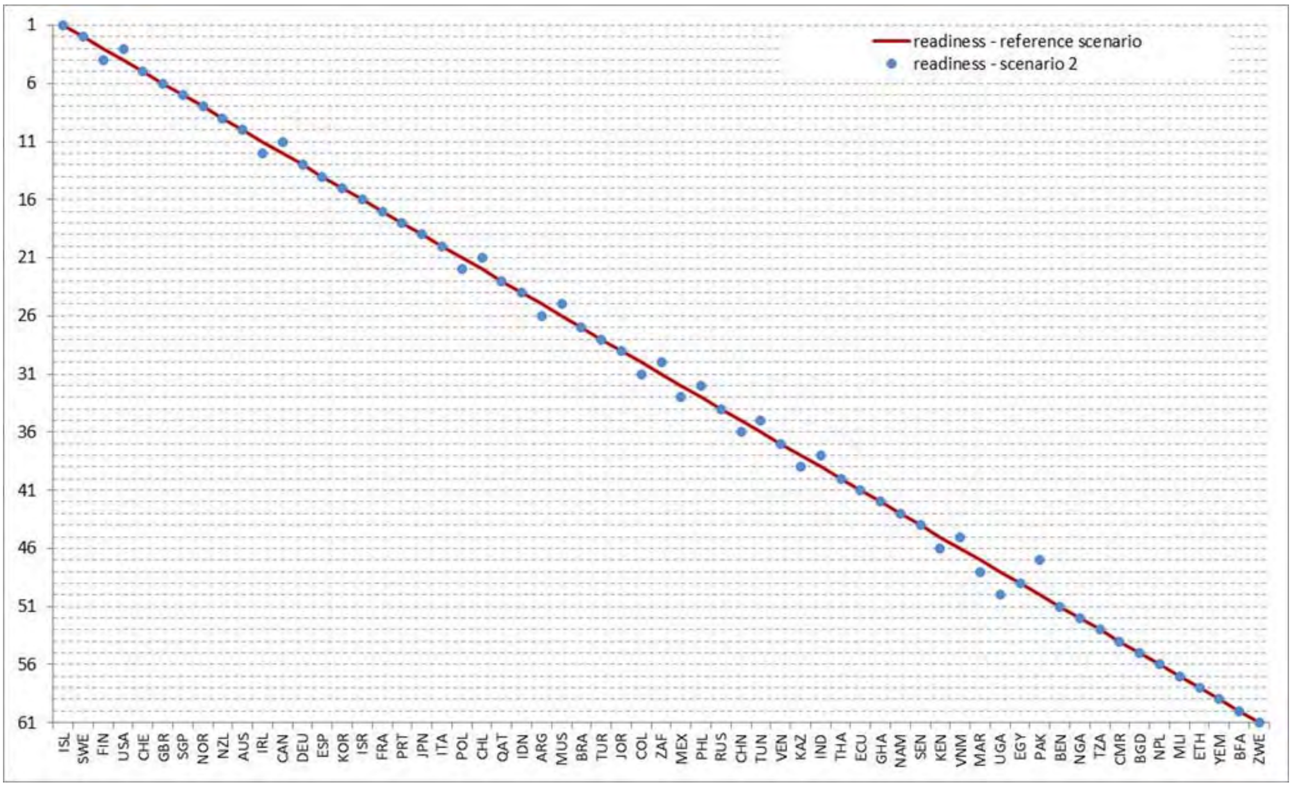


Figure 35: Ranking of countries in Readiness sub-index – scenario #2

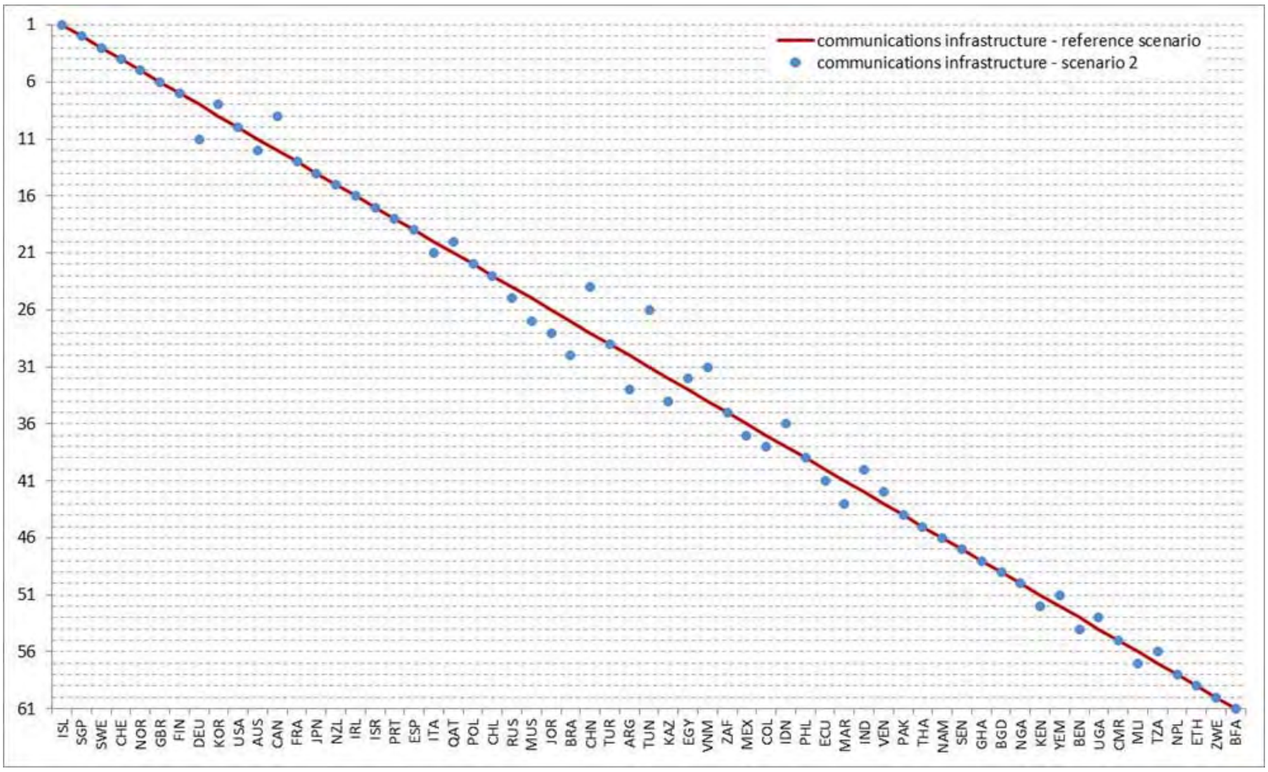


Figure 36: Ranking of countries in Communications Infrastructure component – scenario #2

The results show that the WI is not affected by the different aggregation method for clustering three indicators describing web access. There are almost no differences in the ranks of countries. A slightly higher volatility is observed for the Readiness sub-index and the Communications Infrastructure component. In the case of Readiness the differences in ranks are generally of 1 position with the exception of Uganda (2 positions upward) and Pakistan (3 positions downward). In the case of Communications Infrastructure the shift amounts to maximum 5 positions (Tunisia upward) and 4 positions (China upward). Table 17 shows the countries with a shift in rank for the Readiness sub-index and the Communications Infrastructure component (countries are ordered from the best to the worst). Overall, it can be said that the alternative way to account for web access influences the Web sub-index only marginally.

Table 17: Scenario #2 - Summary of rank shift

WI		Readiness		Communications Infrastructure	
country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank
SGP	-1	FIN	-1	DEU	-3
ISL	1	USA	1	KOR	1
ITA	-1	IRL	-1	AUS	-1
BRA	1	CAN	1	CAN	3
JOR	-1	POL	-1	ITA	-1
ZAF	1	CHL	1	QAT	1
MLI	-2	ARG	-1	RUS	-1
BGD	1	MUS	1	MUS	-2
NAM	1	COL	-1	JOR	-2
		ZAF	1	BRA	-3
		MEX	-1	CHN	4
		PHL	1	ARG	-3
		CHN	-1	TUN	5
		TUN	1	KAZ	-2
		KAZ	-1	EGY	1
		IND	1	VNM	3
		KEN	-1	MEX	-1
		VNM	1	COL	-1
		MAR	-1	IDN	2
		UGA	-2	ECU	-1
		PAK	3	MAR	-2
				IND	2
				VEN	1
				KEN	-1
				YEM	1
				BEN	-1
				UGA	1
				MLI	-1
				TZA	1

3.3 Scenario #3: inclusion of 4 additional indicators

The Institutional Infrastructure component includes all gender related variables of the WI. In particular variables Q9a, Q9b, Q9c and Q9d, referring to computer training and science and technology, are included in the Index in a combined way in order to measure a sort of distance between female and male situation (gender bias). Variables Q9a-b and Q9c-d are then calculated using the following formula:

$$Q9a - b = 10 - |Q9a - Q9b|$$

$$Q9c - d = 10 - |Q9c - Q9d|$$

where:

Q9a – to what extent are boys trained in the use of computers

Q9b – to what extent are girls trained in the use of computers

Q9c – to what extent are girls encouraged to focus on science and technology

Q9b – to what extent are boys encouraged to focus on science and technology

With this construction only the magnitude of the difference between girls and boys is measured. In order to take into account also the level of these variables, and not only the “distance” between girls and boys, four additional variables - Q9a, Q9b, Q9c, Q9d - are added to the Institutional Infrastructure component. The arithmetic mean of the 6 variables (Q9a, Q9b, Q9c, Q9d, Q9a-b and Q9c-d) is then calculated. The results are displayed in Figure 37, Figure 38, and Figure 39.

The inclusion of the four variables describing levels has almost no influence on the WI (the highest rank shift is of 1 position). Volatility increases at the sub-index and component level remaining however much below 10% of the maximum possible shift in rank. The biggest differences in the sub-index Readiness are of 4 (Morocco upward) and 3 positions (Benin downward). In the Institutional Infrastructure component the highest differences are of 5 positions for Ecuador (upward) and China (downward). Countries with any shift in rank for the WI, the Readiness sub-index or for the Institutional Infrastructure component are listed in Table 18.

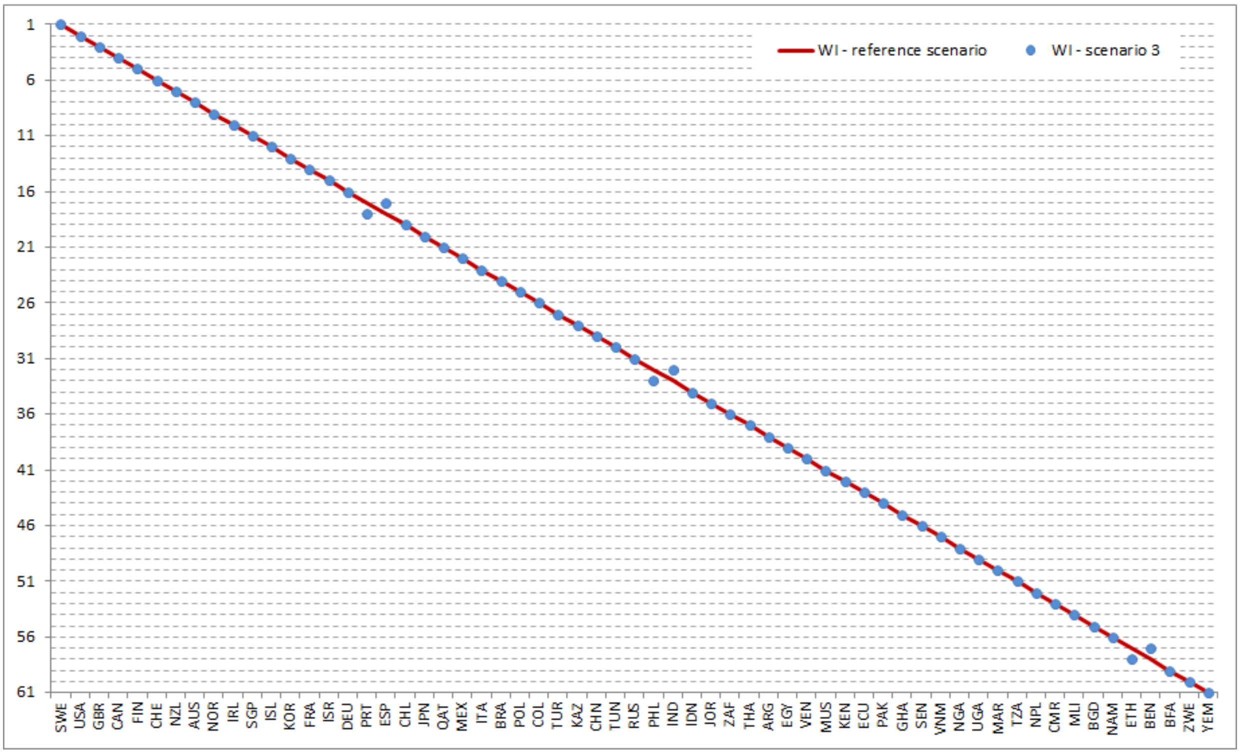


Figure 37: Ranking of countries in the WI – scenario #3

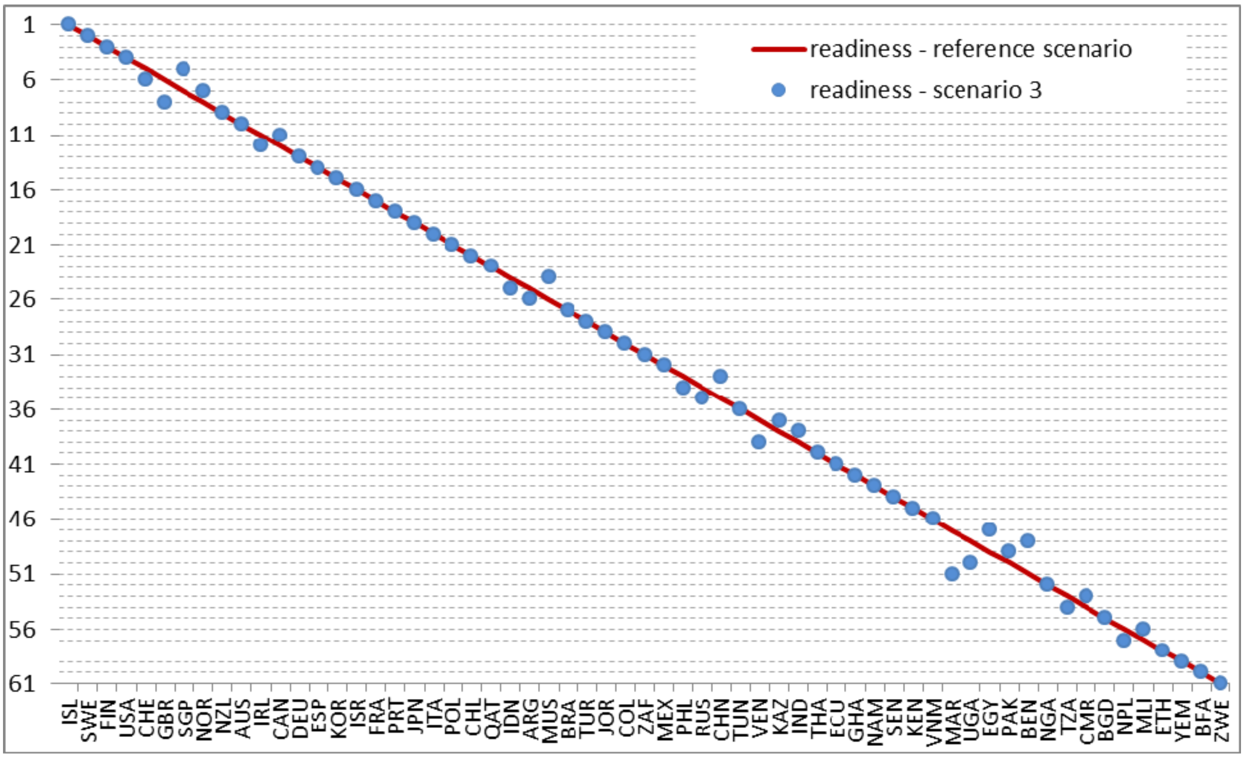


Figure 38: Ranking of countries in Readiness sub-index – scenario #3

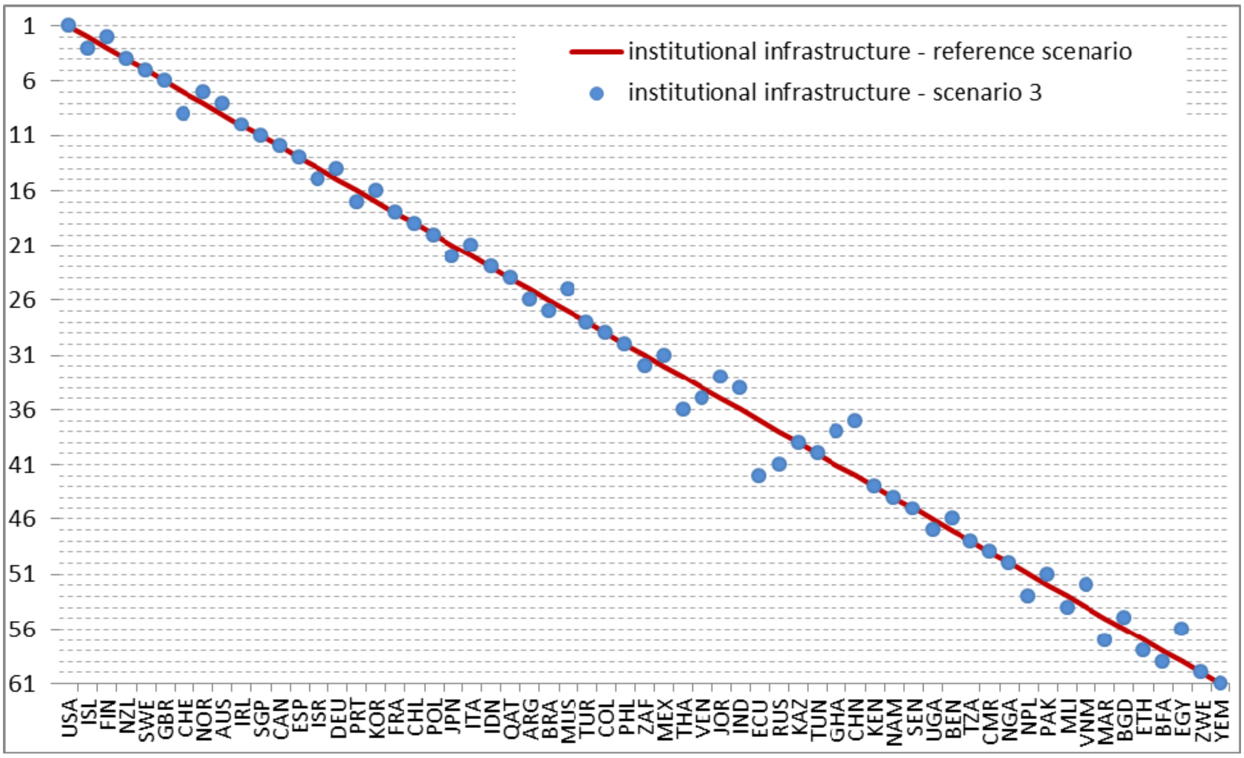


Figure 39: Ranking of countries in Institutional Infrastructure component – scenario #3

Table 18: Scenario 3 - Summary of rank shift

WI		Readiness		Institutional Infrastructure	
country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank
PRT	-1	CHE	1	ISL	1
ESP	1	GBR	2	FIN	-1
PHL	-1	SGP	-2	CHE	2
IND	1	NOR	-1	NOR	-1
ETH	-1	IRL	1	AUS	-1
BEN	1	CAN	-1	ISR	1
		IDN	1	DEU	-1
		ARG	1	PRT	1
		MUS	-2	KOR	-1
		PHL	1	JPN	1
		RUS	1	ITA	-1
		CHN	-2	ARG	1
		VEN	2	BRA	1
		KAZ	-1	MUS	-2
		IND	-1	ZAF	1
		MAR	4	MEX	-1
		UGA	2	THA	3
		EGY	-2	VEN	1
		PAK	-1	JOR	-2
		BEN	-3	IND	-2
		TZA	1	ECU	5
		CMR	-1	RUS	3
		NPL	1	GHA	-3
		MLI	-1	CHN	-5
				UGA	1
				BEN	-1
				NPL	2
				PAK	-1
				MLI	1
				VNM	-2
				MAR	2
				BGD	-1
				ETH	1
				BFA	1
				EGY	-3

3.4 Scenario #4: different quantification for survey data

This scenario is meant to assess the impact of the use of Rasch scores instead of original survey variables in the computation of the WI. As discussed in Section 2 Rasch analysis allows for quantifying survey data by ad hoc statistical models. So far in this study Rasch analysis has been used mainly to assess the quality of the questionnaire of the expert assessment survey. In this scenario final country Rasch scores replace the sub-set of primary variables in the WI components. The strategy employed is the following:

1) Communications Infrastructure component:
   - since there are only two primary indicators in the component (Q18 and Q20) Rasch model cannot be employed so the reference scenario is left unchanged.
2) Institutional Infrastructure component:
   - All primary variables included in the component are replaced by their overall Rasch score as described in Section 2;
   - no aggregation of WEFG, WEFH (secondary variables) and Q9e and Q9g (primary variables) can be applied in this case because all information embedded in primary data is captured by the Rasch score used in this scenario.
3) Web Use component:
   - All primary variables included in the component are replaced by their overall Rasch score as described in Section 2.
4) Web Content component:
   - Three Rasch score variables instead of 22 primary variables are used as three separate Rasch models are run for the following groups of primary variables: (Q5a, Q5b and Q5c), (from Q23a to Q23j) and Q3, Q8a, Q8b, Q8c, Q9k, Q22, Q24, Q26) (see Section 2).
5) Political Impact component:
   - since there are only two primary indicators in the component (Q1 and Q2b) Rasch model cannot be employed so the reference scenario is left unchanged.
6) Economic Impact component:
   - All primary variables included in the component are replaced by their overall Rasch score as described in Section 2.

7) Social Impact component:

— All primary variables included in the component are replaced by their overall Rasch score as described in Section 2.

Differences in ranks for the WI and the three sub-indexes are shown in Figure 40 to Figure 43 and in Table 19. Figure 44 to Figure 48 and Table 20 display the results at the component level. The largest changes are those for Australia and Philippines with an improvement of 4 positions in the WI ranking and by Singapore, Iceland and Benin which decline by 4 positions. A much higher ranking volatility is present at the sub-index and component level especially for the Web Content where Indonesia could drop 14 positions in the WI ranking while Bangladesh and Ecuador would climb by 16 positions and South Africa by 13.

As expected, the inclusion of Rasch scores instead of primary variables induces some rank volatility at all the WI levels. Two factors influence these results: the replacement of most of original primary data with Rasch scores and the changes in the structure of the WI due to the use of Rasch scores.

Let us first focus on the first reason. Rasch scores, as discussed in Section 2, are derived from a statistical method which takes into account the overall pattern of answers and the difficulty of the questions. Furthermore, Rasch scores aim at capturing the latent underlying variable hence they do not explain the whole original data variability but only the part linked to the latent phenomenon under measurement (see Section 2). This implies that the part of the information included in original data, which is connected to some other phenomena or to noise, is discarded in the Rasch outcomes.

The second factor influencing the result is the impossibility to preserve the reference structure of the WI. It results from both the necessity of discarding the misfitting (according to Rasch requirements) questions and the changes in the structure of some components due to the necessity of inclusion of all primary data into Rasch modelling instead of averaging them with secondary data.

So why are we testing this scenario? Primary data are used in the computation of WI in a very straightforward way. We want to assess the influence on country ranks of a sounder statistical model for survey data. The analysis of this scenario (Figure 40 to Figure 48) shows that some volatility is indeed present and it is higher at the component level. As far as the final WI ranks are concerned, the maximum rank shift is 4, well below 10% of maximum possible shift in rank.

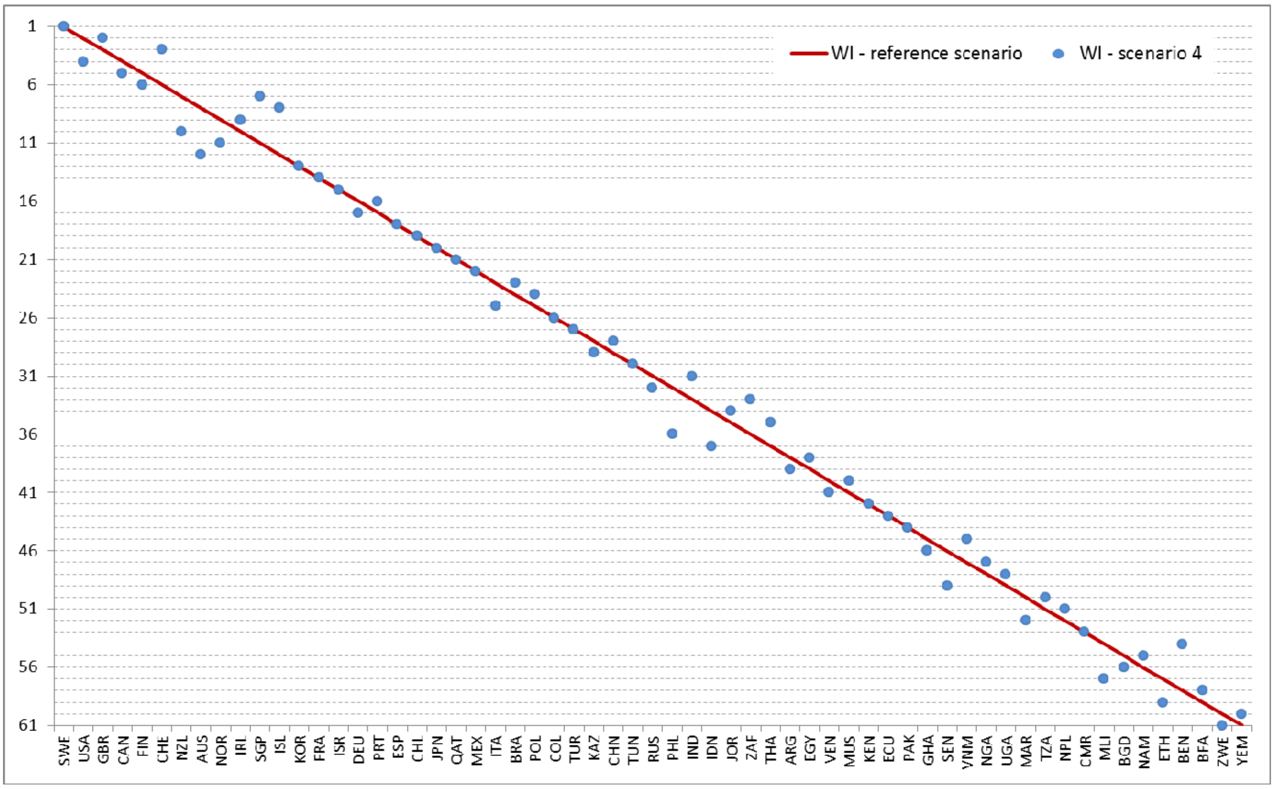


Figure 40: Ranking of countries in WI – scenario #4

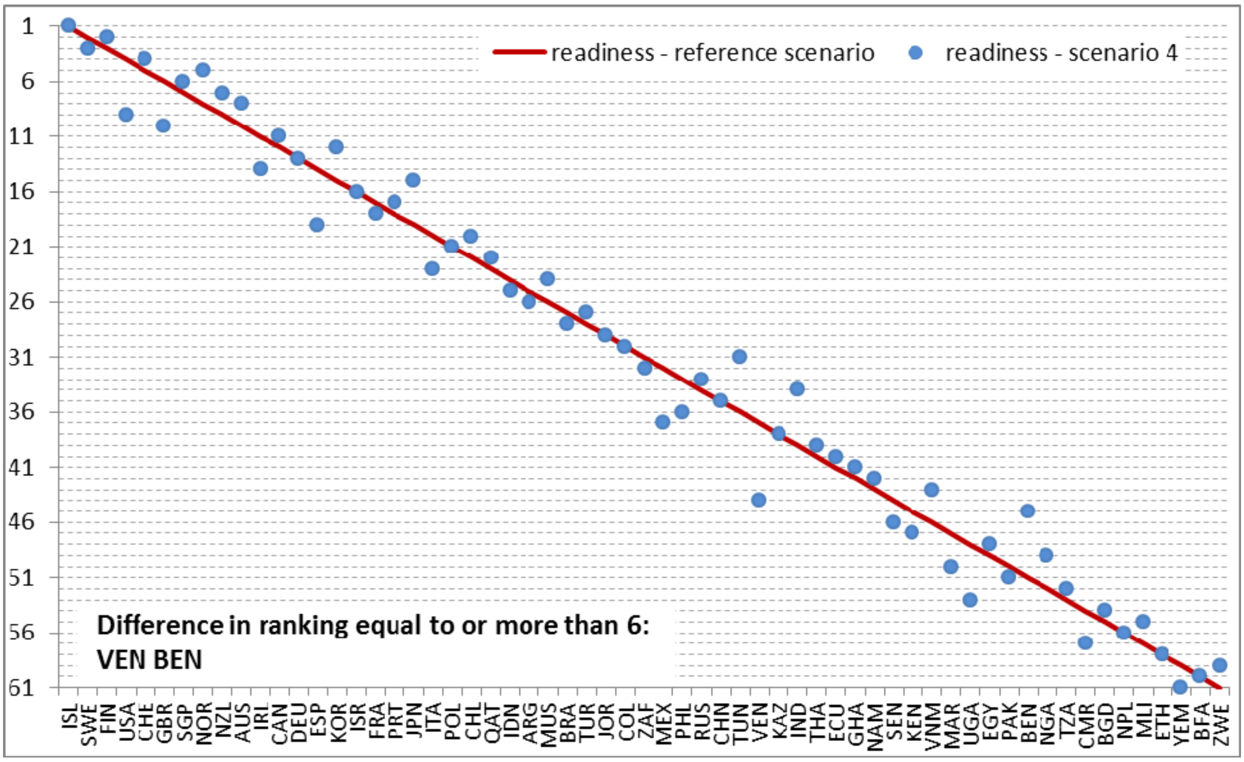


Figure 41: Ranking of countries in Readiness sub-index – scenario #4

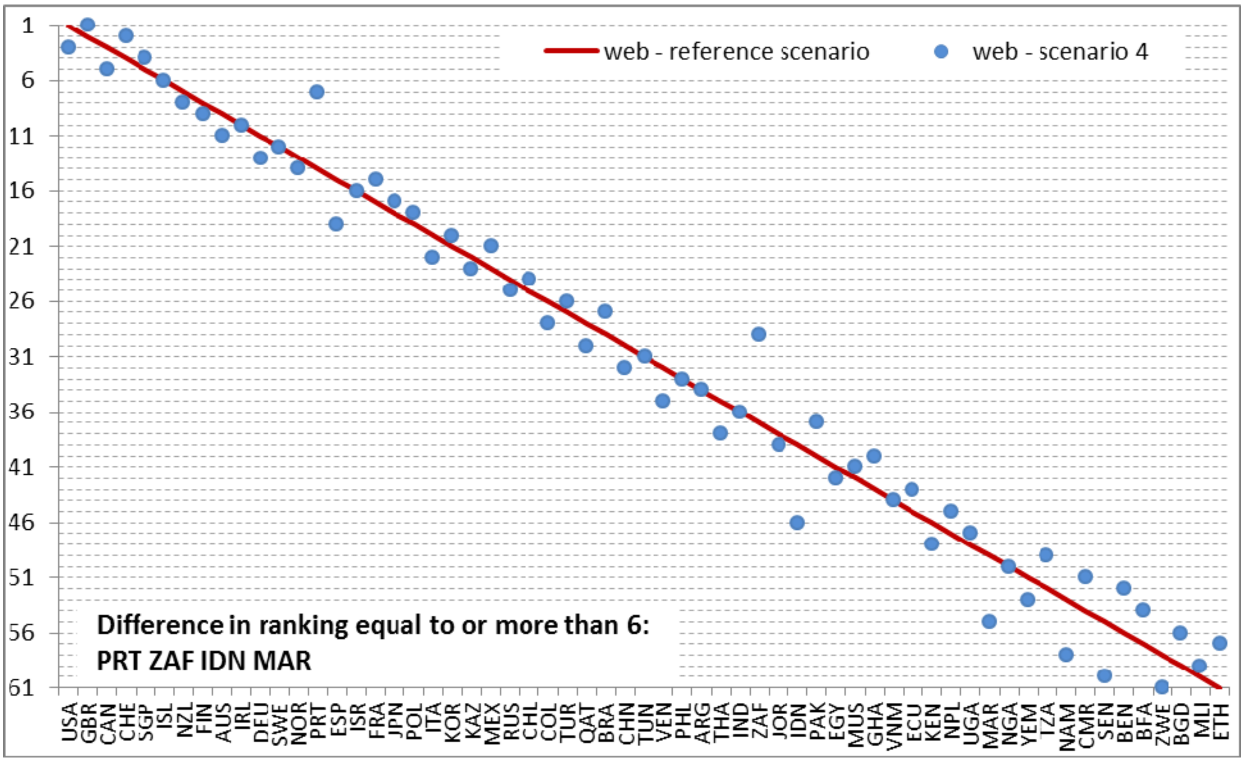


Figure 42: Ranking of countries in Web sub-index – scenario #4

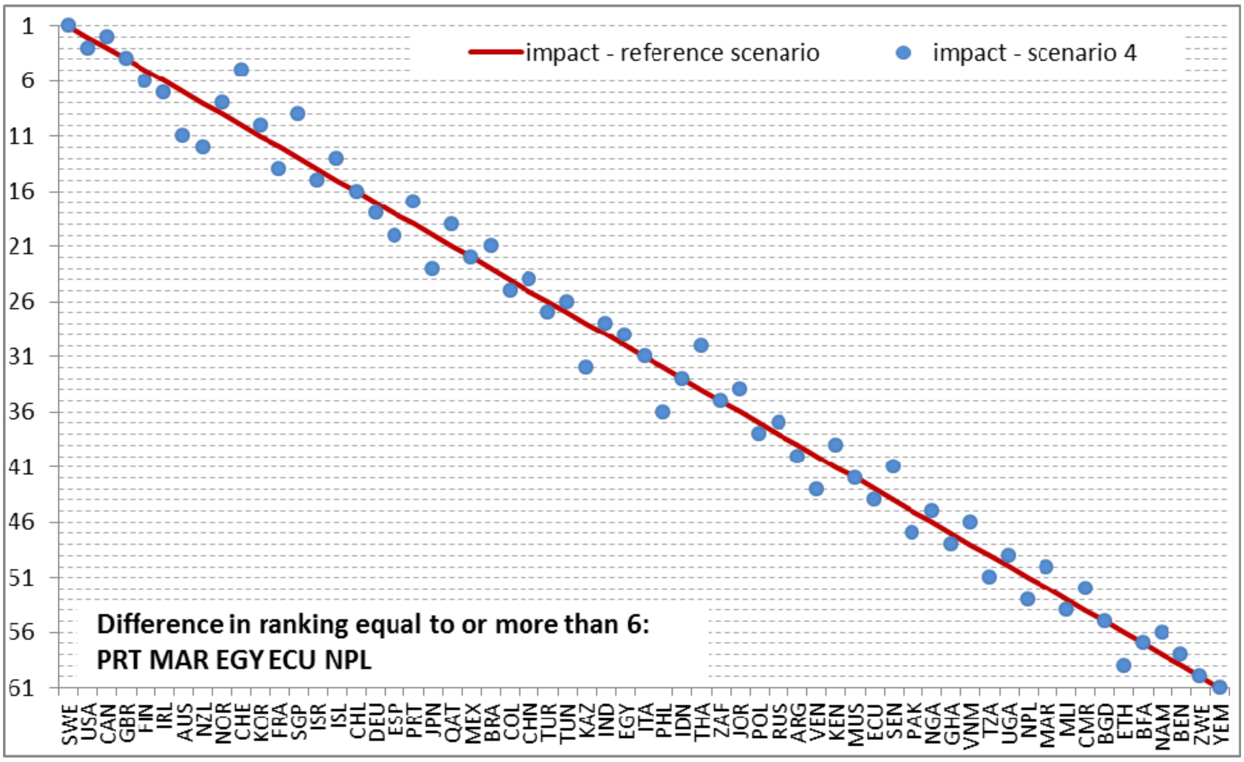


Figure 43: Ranking of countries in Impact sub-index – scenario #4

Table 19: Scenario 4 – country rank shift at the index and sub-index level

WI		Readiness		Web		Impact	
country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank
USA	2	SWE	1	USA	-2	USA	-1
GBR	-1	FIN	-1	GBR	1	CAN	1
CAN	1	USA	5	CAN	-2	FIN	-1
FIN	1	CHE	-1	CHE	2	IRL	-1
CHE	-3	GBR	4	SGP	1	AUS	-4
NZL	3	SGP	-1	NZL	-1	NZL	-4
AUS	4	NOR	-3	FIN	-1	NOR	1
NOR	2	NZL	-2	AUS	-2	CHE	5
IRL	-1	AUS	-2	DEU	-2	KOR	1
SGP	-4	IRL	3	NOR	-1	FRA	-2
ISL	-4	CAN	-1	PRT	7	SGP	4
DEU	1	ESP	5	ESP	-4	ISR	-1
PRT	-1	KOR	-3	FRA	2	ISL	2
ITA	2	FRA	1	JPN	1	DEU	-1
BRA	-1	PRT	-1	POL	1	ESP	-2
POL	-1	JPN	-4	ITA	-2	PRT	2
KAZ	1	ITA	3	KOR	1	JPN	-3
CHN	-1	CHL	-2	KAZ	-1	QAT	2
RUS	1	QAT	-1	MEX	2	BRA	2
PHL	4	IDN	1	RUS	-1	COL	-1
IND	-2	ARG	1	CHL	1	CHN	1
IDN	3	MUS	-2	COL	-2	TUR	-1
JOR	-1	BRA	1	TUR	1	TUN	1
ZAF	-3	TUR	-1	QAT	-2	KAZ	-4
THA	-2	ZAF	1	BRA	2	IND	1
ARG	1	MEX	5	CHN	-2	EGY	1
EGY	-1	PHL	3	VEN	-3	PHL	-4
VEN	1	RUS	-1	THA	-3	THA	4
MUS	-1	TUN	-5	ZAF	8	JOR	2
GHA	1	VEN	7	JOR	-1	POL	-1
SEN	3	IND	-5	IDN	-7	RUS	1
VNM	-2	THA	-1	PAK	3	ARG	-1
NGA	-1	ECU	-1	EGY	-1	VEN	-3
UGA	-1	GHA	-1	MUS	1	KEN	2
MAR	2	NAM	-1	GHA	3	ECU	-1
TZA	-1	SEN	2	ECU	2	SEN	3
NPL	-1	KEN	2	KEN	-2	PAK	-2
MLI	3	VNM	-3	NPL	2	NGA	1
BGD	1	MAR	3	UGA	1	GHA	-1
NAM	-1	UGA	5	MAR	-6	VNM	2
ETH	2	EGY	-1	YEM	-2	TZA	-2
BEN	-4	PAK	1	TZA	3	UGA	1
BFA	-1	BEN	-6	NAM	-5	NPL	-2
ZWE	1	NGA	-3	CMR	3	MAR	2
YEM	-1	TZA	-1	SEN	-5	MLI	-1
		CMR	3	BEN	4	CMR	2
		BGD	-1	BFA	3	ETH	-3
		MLI	-2	ZWE	-3	NAM	2
		YEM	2	BGD	3	BEN	1
		ZWE	-2	MLI	1		
				ETH	4		

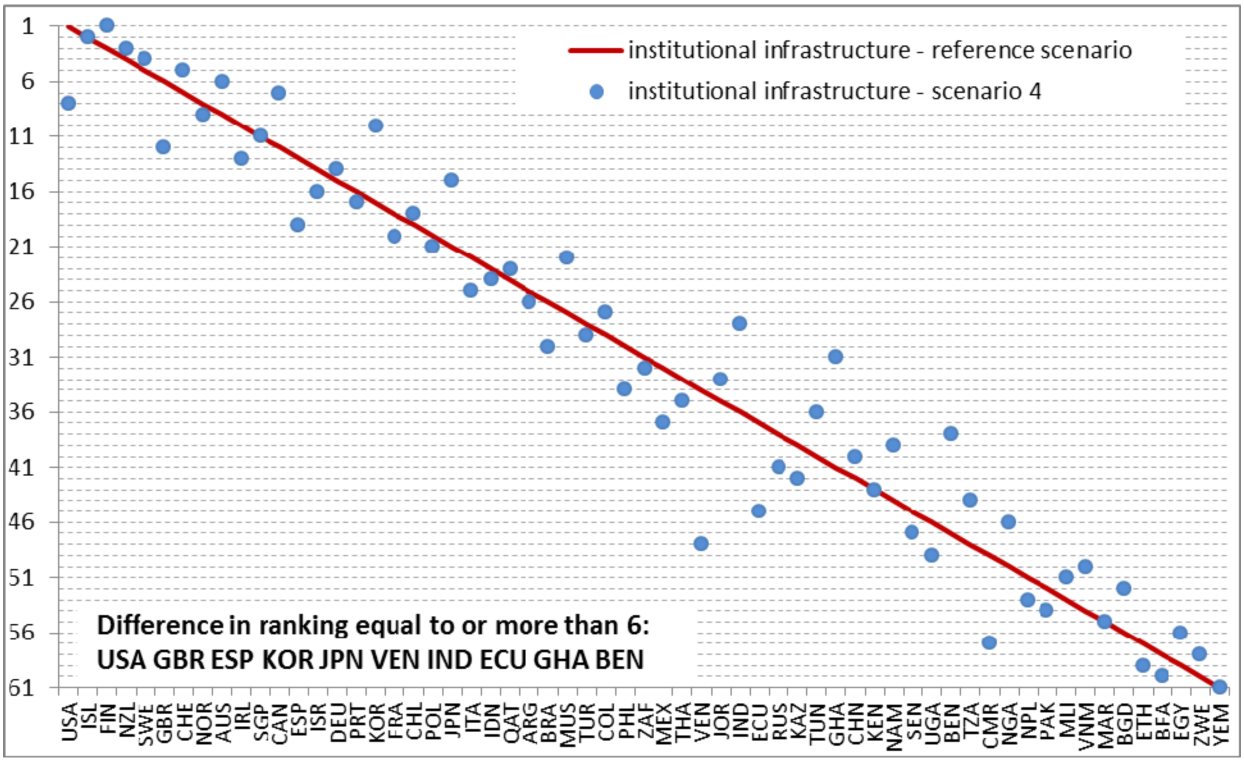


Figure 44: Ranking of countries in Institutional Infrastructure component– scenario #4

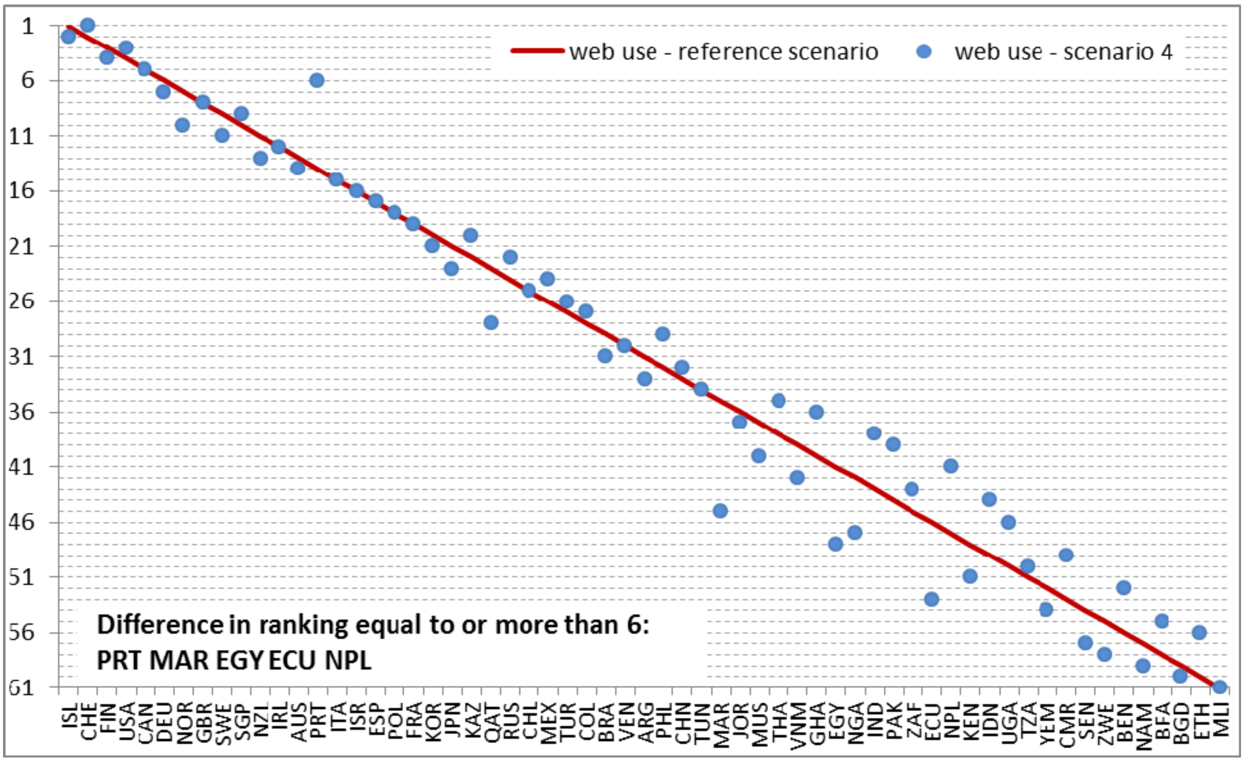


Figure 45: Ranking of countries in Web Use component– scenario #4

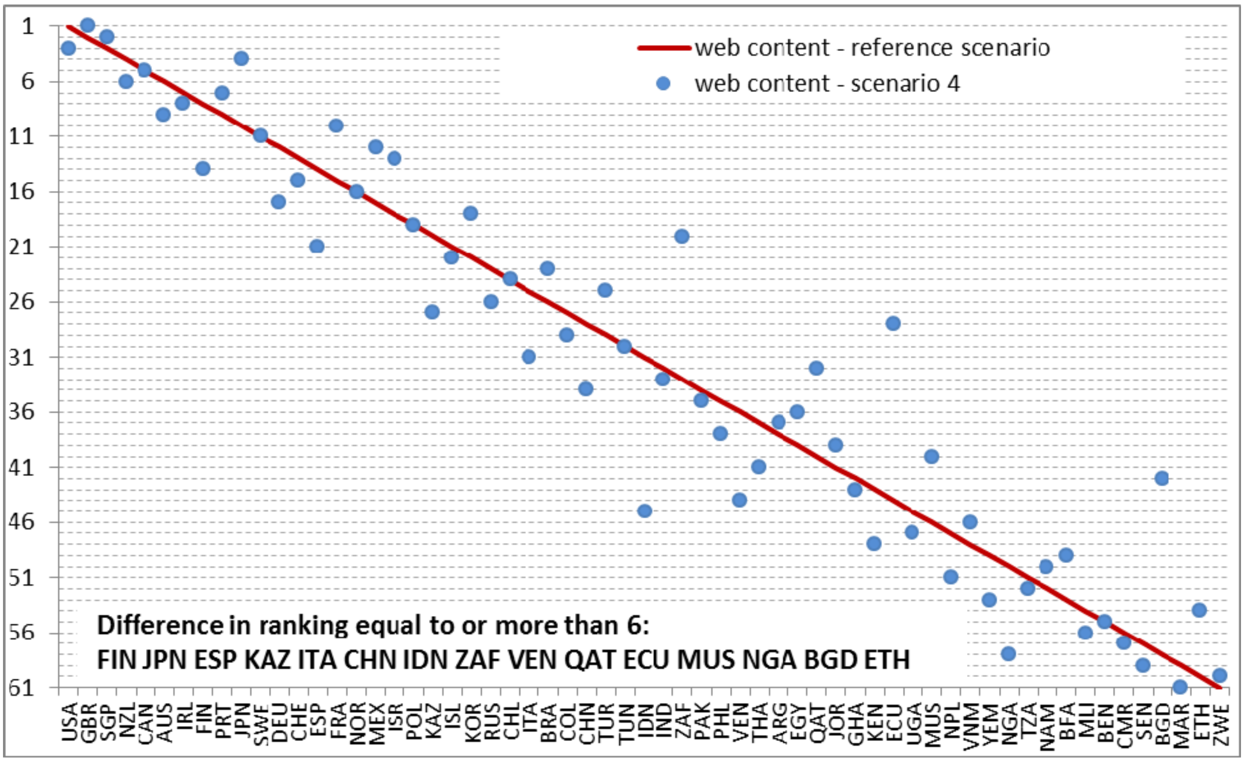


Figure 46: Ranking of countries in Web Content component– scenario #4

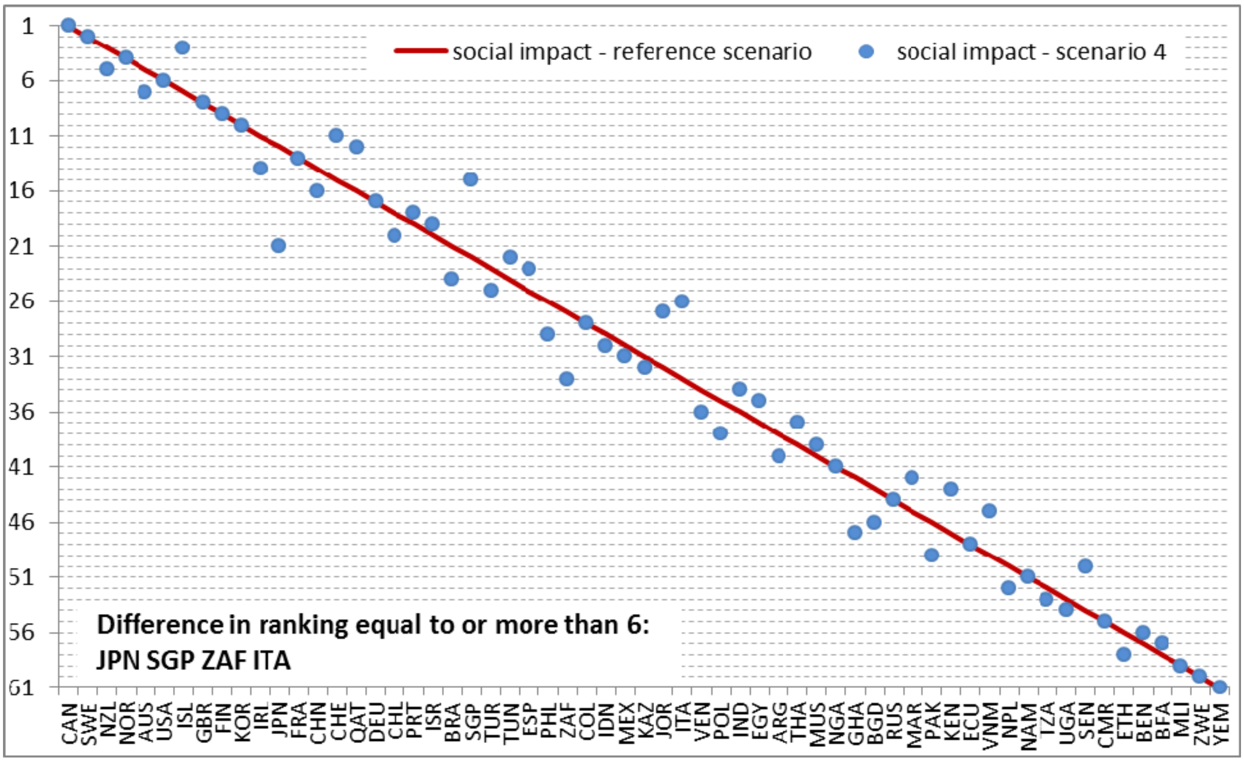


Figure 47: Ranking of countries in Social Impact component– scenario #4

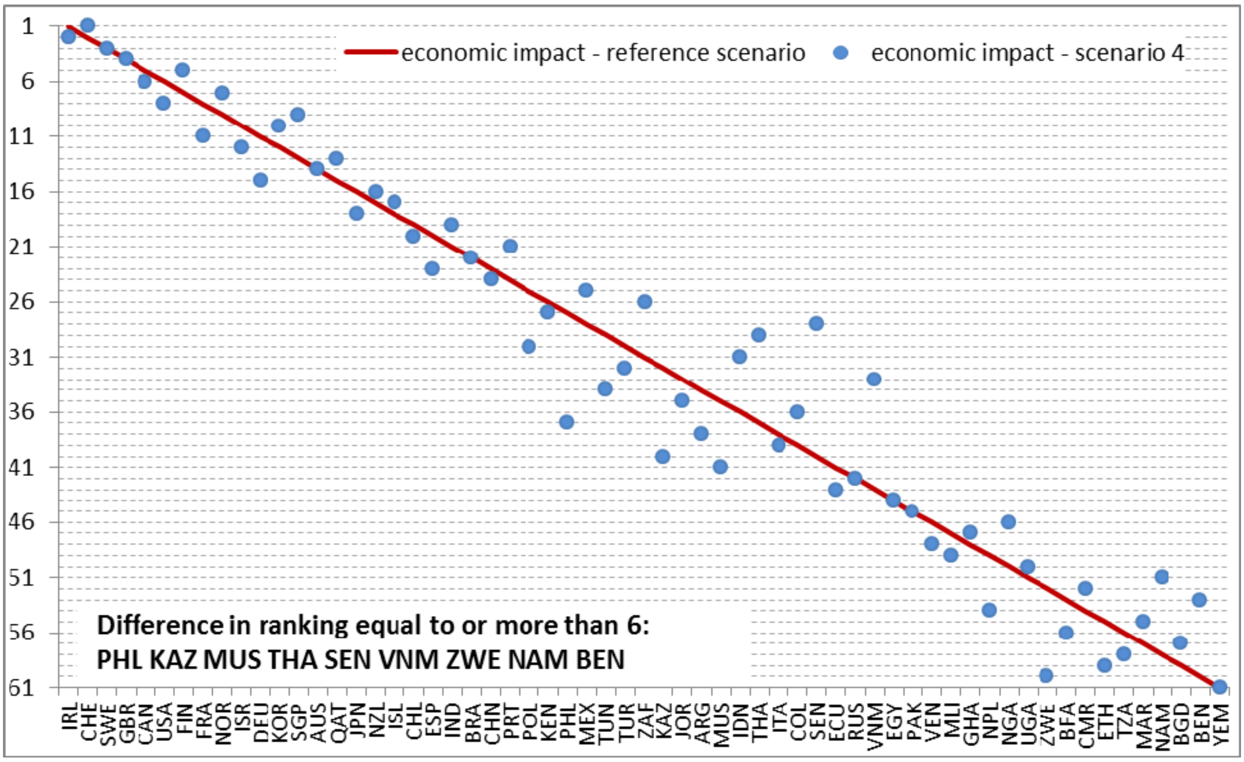


Figure 48: Ranking of countries in Economic Impact component– scenario #4

Table 20: Scenario 4 - Country shift in rank at the component level

Institutional Infrastructure		Web Use		Web Content		Social Impact		Economic Impact	
country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank	country	reference rank - modified rank
USA	7	ISL	1	USA	-2	NZL	-2	IRL	-1
FIN	-2	CHE	-1	GBR	1	AUS	-2	CHE	1
NZL	-1	FIN	1	SGP	1	ISL	4	CAN	-1
SWE	-1	USA	-1	NZL	-2	IRL	-3	USA	-2
GBR	6	DEU	1	AUS	-3	JPN	-9	FIN	2
CHE	-2	NOR	3	IRL	-1	CHN	-2	FRA	-3
NOR	1	SWE	2	FIN	-6	CHE	4	NOR	2
AUS	-3	SGP	-1	PRT	2	QAT	4	ISR	-2
IRL	3	NZL	2	JPN	6	CHL	-2	DEU	-4
CAN	-5	AUS	1	DEU	-5	PRT	1	KOR	2
ESP	6	PRT	-8	CHE	-2	ISR	1	SGP	4
ISR	2	KOR	1	ESP	-7	BRA	-3	QAT	2
DEU	-1	JPN	2	FRA	5	SGP	7	JPN	-2
PRT	1	KAZ	-2	MEX	5	TUR	-2	NZL	1
KOR	-7	QAT	5	ISR	5	TUN	2	ISL	1
FRA	2	RUS	-2	KAZ	-7	ESP	2	CHL	-1
CHL	-1	MEX	-2	ISL	-1	PHL	-3	ESP	-3
POL	1	TUR	-1	KOR	4	ZAF	-6	IND	2
JPN	-6	COL	-1	RUS	-3	IDN	-1	CHN	-1
ITA	3	BRA	2	ITA	-6	MEX	-1	PRT	3
IDN	1	ARG	2	BRA	3	KAZ	-1	POL	-5
QAT	-1	PHL	-3	COL	-2	JOR	5	KEN	-1
ARG	1	CHN	-1	CHN	-6	ITA	7	PHL	-10
BRA	4	MAR	10	TUR	4	VEN	-2	MEX	3
MUS	-5	JOR	1	IDN	-14	POL	-3	TUN	-5
TUR	1	MUS	3	IND	-1	IND	2	TUR	-2
COL	-2	THA	-3	ZAF	13	EGY	2	ZAF	5
PHL	4	VNM	3	PAK	-1	ARG	-2	KAZ	-8
ZAF	1	GHA	-4	PHL	-3	THA	2	JOR	-2
MEX	5	EGY	7	VEN	-8	MUS	1	ARG	-4
THA	2	NGA	5	THA	-4	GHA	-5	MUS	-6
VEN	14	IND	-5	ARG	1	BGD	-3	IDN	5
JOR	-2	PAK	-5	EGY	3	MAR	3	THA	8
IND	-8	ZAF	-2	QAT	8	PAK	-3	ITA	-1
ECU	8	ECU	7	JOR	2	KEN	4	COL	3
RUS	3	NPL	-6	GHA	-1	VNM	4	SEN	12
KAZ	3	KEN	3	KEN	-5	NPL	-2	ECU	-2
TUN	-4	IDN	-5	ECU	16	TZA	-1	VNM	10
GHA	-10	UGA	-4	UGA	-2	UGA	-1	VEN	-2
CHN	-2	TZA	-1	MUS	6	SEN	4	MLI	-2
NAM	-5	YEM	2	NPL	-4	ETH	-2	GHA	1
SEN	2	CMR	-4	VNM	2	BEN	1	NPL	-5
UGA	3	SEN	3	YEM	-4	BFA	1	NGA	4
BEN	-9	ZWE	3	NGA	-8			UGA	1
TZA	-4	BEN	-4	TZA	-1			ZWE	-8
CMR	8	NAM	2	NAM	2			BFA	-3
NGA	-4	BFA	-3	BFA	4			CMR	2
NPL	2	BGD	1	MLI	-2			ETH	-4
PAK	2	ETH	-4	CMR	-1			TZA	-2
MLI	-2			SEN	-2			MAR	2
VNM	-4			BGD	16			NAM	7
BGD	-4			MAR	-2			BGD	2
ETH	2			ETH	6			BEN	7
BFA	2			ZWE	1				
EGY	-3								
ZWE	-2								

3.5 Scenario #5: compensability

Can high web use or a high social impact compensate for poor institutional or communications infrastructure? The linear aggregation used in the WI assumes it can, as it entails full compensability of *bads* and *goods* at the country level. To overcome this hypothesis other types of aggregation procedures must be used. Besides multi-criteria evaluation, which is not compensatory at all, the geometric aggregation could be for example applied. Let j be the number of variables to be aggregated, the weighted geometric aggregation is computed as $\left[\prod_{j=1}^{J} x_{ji}^{w_j}\right]^{1/\sum w_j}$, where w_j is the weight assigned to variable j and x_{ji} is the score of country i on variable j.

In this scenario two different geometric aggregations are considered: (1) at the sub-index level by aggregating Readiness, Web and Impact scores; (2) at both the component and sub-index level. In the computation of weighted geometric means reference weights are always adopted.

Results show (Figure 49, Figure 50 and Table 21) minor changes in the rankings with respect to the reference index (a shift of at most 2 positions in the ranking). Compensability does not seem an issue in this case. This is due to the fact that no country scores relatively high in some components and low in some others.

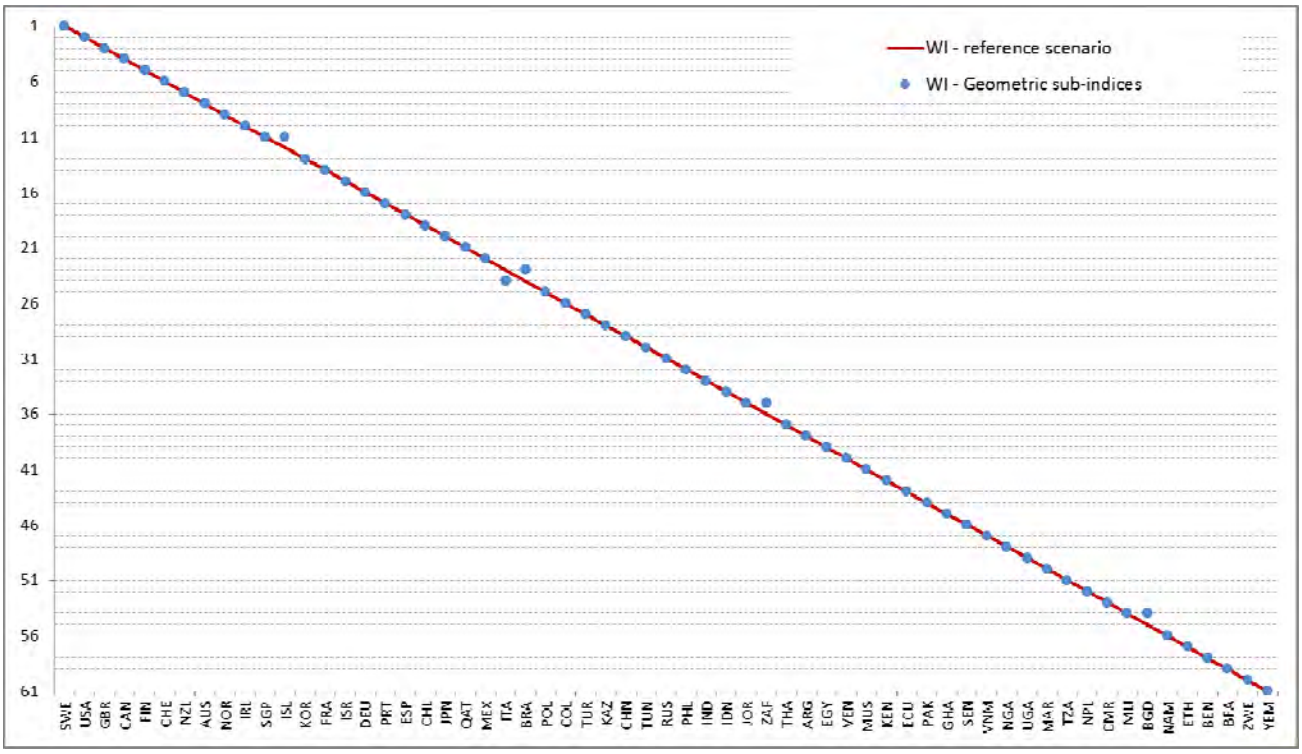


Figure 49: Ranking of countries using geometric aggregation of sub-indexes versus reference WI

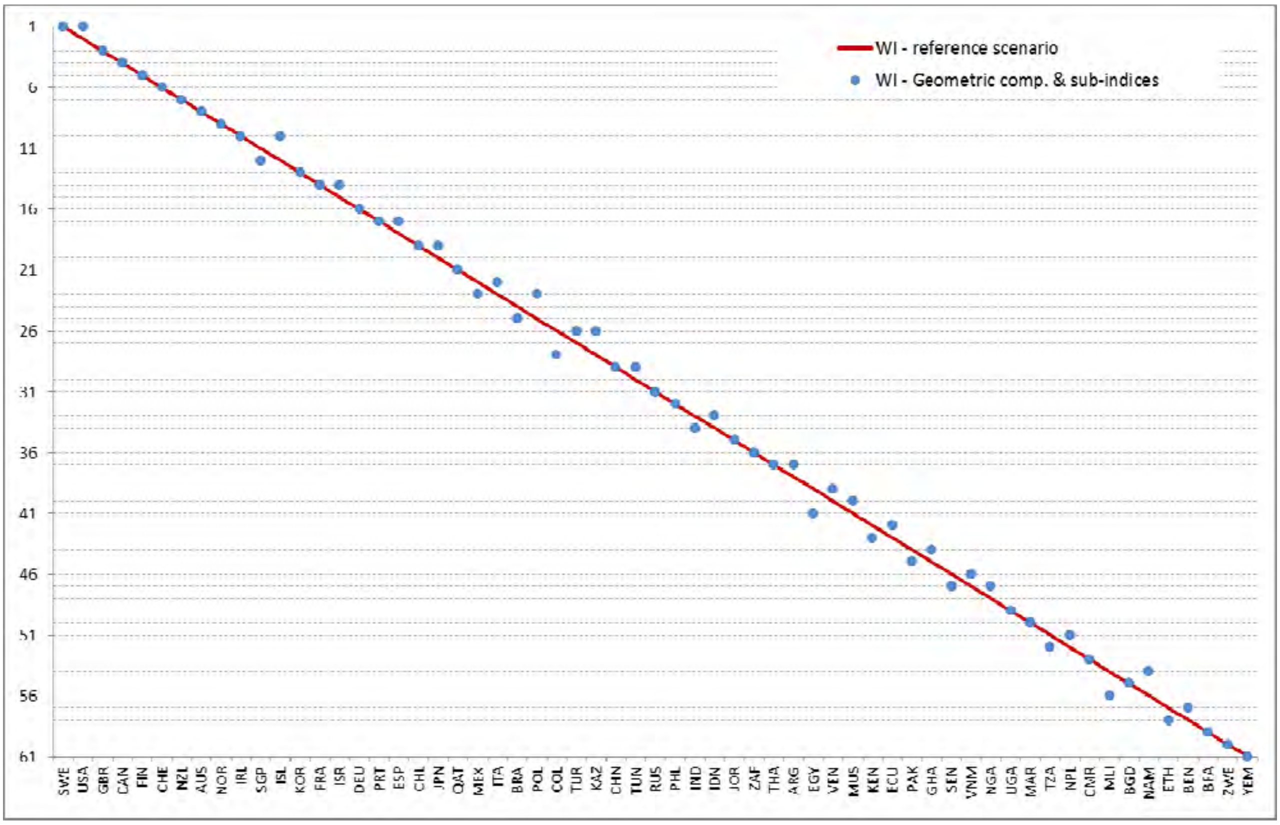


Figure 50: Ranking of countries using geometric aggregation of sub-indexes and Components versus reference WI

Table 21: Scenario 5 - country shift in rank using geometric aggregation

Geometric aggregation on sub-indexes		Geometric aggregation on components & sub-indexes	
Country	*reference rank-modified rank*	Country	*reference rank-modified rank*
ISL	1	USA	1
ITA	-1	SGP	-1
BRA	1	ISL	2
ZAF	1	ISR	1
BGD	1	ESP	1
		JPN	1
		MEX	-1
		ITA	1
		BRA	-1
		POL	2
		COL	-2
		TUR	1
		KAZ	2
		TUN	1
		IND	-1
		IDN	1
		ARG	1
		EGY	-2
		VEN	1
		MUS	1
		KEN	-1
		ECU	1
		PAK	-1
		GHA	1
		SEN	-1
		VNM	1
		NGA	1
		TZA	-1
		NPL	1
		MLI	-2
		NAM	2
		ETH	-1
		BEN	1

3.6 Scenario #6: contribution of each component and sub-index

The ‘importance’ of each single component on WI scores is assessed by setting all weights back to their reference values (Section 1) and computing country scores and ranks discarding one component at a time. Seven simulations are run each discarding one component at a time.

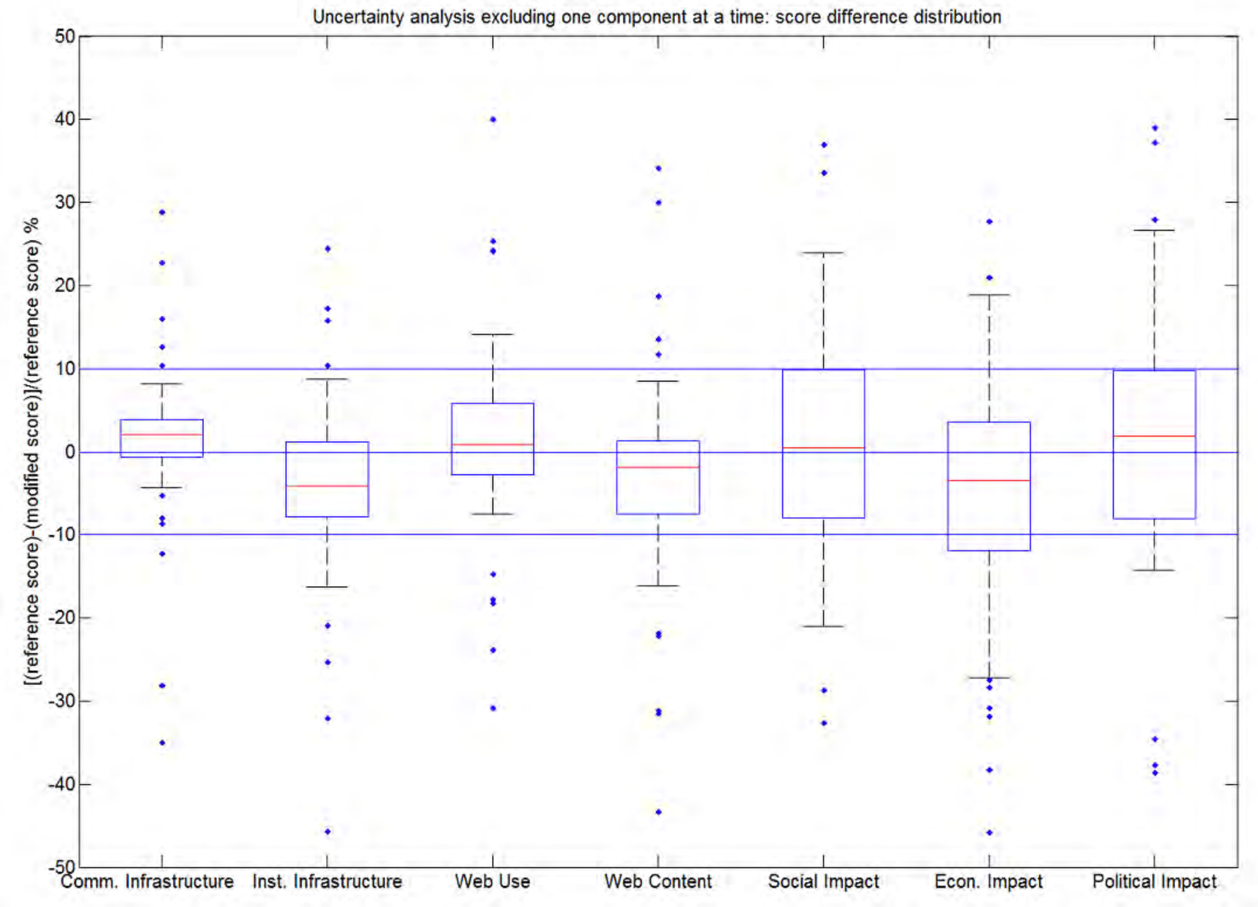


Figure 51: Effect of discarding one component at a time on scores

Figure 51 and Figure 52 summarize the outcome of this analysis. Boxplots refer to the seven different simulations. The horizontal axis shows the discarded component. Boxplot whiskers indicate 1.5*IQR (interquartile range), both upwards and downwards, whilst dots are outliers, i.e. observations (countries in our case) falling outside the interval [(P25-1.5*IQR),(P75+1.5*IQR)].

Figure 51 shows differences in scores as percentage difference with respect to the reference WI scores. The percentage difference is almost always in the range ±10%. As expected the Impact components are the most influential ones due to the higher weight assigned to the Impact sub-index (0.6), with respect to the others (0.2 each). Figure 52 shows boxplots of differences in ranks. All the interquartile ranges are within the band [-2; +2], meaning that for all the simulations the maximum shift of the country rank is up to 2 positions in at least 50% of the cases. This confirms that, on average, all the components contribute in a balanced way to the WI.

Figure 52: Effect of discarding one component at a time on ranks

The relationship between the Index and its indicators is shown in Table 22. We should expect a positive and significant correlation between each indicator and its corresponding components. The positive sign for all correlation coefficients is confirmed meaning that an increase in the indicator always implies an increase in the Index. However, some indicators are not significantly correlated with the Index. This is the case of indicators WEFF, Q9l, the cluster between Q9a-b and Q9c-d with the Institutional Infrastructure component, and Q12 with Economic Impact component.

Table 22: Pearson correlation coefficients between WI variables and their corresponding components, sub-indexes and Index.

Pearson correlation					
			component	sub-index	WI
Readiness	Communication Infrastructure	ITUA	0.70	0.67	0.62
		ITUB	0.94	0.91	0.90
		WBA	0.78	0.80	0.77
		WEFA	0.93	0.89	0.89
		WEFB	0.77	0.77	0.79
		ITUC	0.96	0.93	0.91
		ITUD	0.68	0.66	0.57
		IEAA	0.73	0.63	0.63
		ITUE	0.73	0.63	0.62
		ITUF	0.70	0.63	0.60
		ITUG	0.47	0.44	0.42
		Q18	0.83	0.76	0.72
		Q20	0.90	0.87	0.90
	Institutional Infrastructure	RSFA - WEFC	0.74	0.73	0.66
		WBB	0.85	0.86	0.82
		UNA	0.89	0.88	0.83
		WEFD	0.64	0.64	0.68
		UNB	0.79	0.79	0.73
		WEFE	0.79	0.83	0.86
		WEFF	0.23	0.26	0.26
		WEFG, WEFH, Q9e, Q9g	0.68	0.71	0.76
		FHA	0.72	0.71	0.61
		FHB	0.82	0.79	0.70
		Q9a to Q9d	0.29	0.26	0.22
		Q9f	0.63	0.58	0.55
		Q9h	0.64	0.62	0.62
		Q9i	0.64	0.62	0.58
		Q9l	0.31	0.28	0.19
		Q10	0.55	0.50	0.34
		Q13	0.76	0.76	0.75
		Q16	0.62	0.65	0.71
		Q25	0.63	0.61	0.69

Pearson correlation					
			component	sub-index	WI
The Web	Use	ITUH	0.94	0.91	0.91
		Q11a to Q11f	0.93	0.92	0.84
	Content	WIKIA	0.66	0.60	0.55
		UNC	0.86	0.83	0.88
		Q2a	0.66	0.63	0.65
		Q3	0.85	0.85	0.82
		Q5a to Q5c	0.85	0.81	0.83
		Q8a	0.89	0.88	0.88
		Q8b	0.60	0.59	0.55
		Q8c	0.77	0.80	0.79
		Q9k	0.73	0.69	0.63
		Q22	0.78	0.71	0.67
		Q23a to Q23j	0.89	0.85	0.84
		Q24	0.60	0.56	0.54
		Q26	0.78	0.78	0.76
Impact	Social	WEFI, Q7	0.88	0.89	0.91
		WEFJ	0.80	0.84	0.82
		Q4	0.86	0.83	0.83
		Q6	0.73	0.57	0.57
	Economic	WBC	0.43	0.24	0.22
		WEFK	0.86	0.89	0.88
		WEFL	0.88	0.89	0.85
		WEFM	0.88	0.90	0.87
		Q12	0.25	0.08	0.06
		Q14	0.82	0.88	0.90
		Q15	0.75	0.77	0.76
		Q17	0.91	0.87	0.86
	Political	UND	0.82	0.80	0.78
		WEFN	0.79	0.76	0.72
		Q1	0.88	0.87	0.88
		Q2b	0.88	0.81	0.82

red: values not significant at the 5% level

Table 23: Squared Pearson correlation coefficients between indicators and its corresponding components.

Square of the Pearson correlation				
			component	difference with max
Readiness	Communication Infrastructure	ITUA	0.49	-47%
		ITUB	0.89	-3%
		WBA	0.61	-34%
		WEFA	0.86	-6%
		WEFB	0.59	-35%
		ITUC	0.92	0%
		ITUD	0.46	-50%
		IEAA	0.53	-42%
		ITUE	0.54	-41%
		ITUF	0.49	-46%
		ITUG	0.22	**-76%**
		Q18	0.69	-25%
		Q20	0.80	-12%
	Institutional Infrastructure	RSFA -WEFC	0.55	-30%
		WBB	0.72	-9%
		UNA	0.79	0%
		WEFD	0.41	-48%
		UNB	0.63	-21%
		WEFE	0.62	-21%
		WEFF	0.05	**-93%**
		WEFG, WEFH, Q9e, Q9g	0.47	-41%
		FHA	0.52	-34%
		FHB	0.67	-16%
		Q9a to Q9d	0.09	**-89%**
		Q9f	0.40	-49%
		Q9h	0.41	-48%
		Q9i	0.40	-49%
		Q9l	0.10	**-88%**
		Q10	0.30	-62%
		Q13	0.58	-26%
		Q16	0.38	-52%
		Q25	0.40	-50%

Square of the Pearson correlation				
			component	difference with max
The Web	Use	ITUH	0.88	0%
		Q11a to Q11f	0.87	-1%
	Content	WIKIA	0.44	-44%
		UNC	0.74	-7%
		Q2a	0.43	-45%
		Q3	0.72	-10%
		Q5a to Q5c	0.72	-8%
		Q8a	0.79	0%
		Q8b	0.36	-54%
		Q8c	0.59	-26%
		Q9k	0.53	-33%
		Q22	0.61	-24%
		Q23a to Q23j	0.79	0%
		Q24	0.35	-55%
		Q26	0.61	-23%
Impact	Social	WEFI, Q7	0.77	0%
		WEFJ	0.63	-17%
		Q4	0.74	-4%
		Q6	0.53	-31%
	Economic	WBC	0.19	-77%
		WEFK	0.74	-11%
		WEFL	0.78	-6%
		WEFM	0.77	-6%
		Q12	0.06	**-93%**
		Q14	0.67	-19%
		Q15	0.56	-32%
		Q17	0.83	0%
	Political	UND	0.68	-13%
		WEFN	0.63	-19%
		Q1	0.78	0%
		Q2b	0.77	-1%

The correlation structure plays an important role in determining the real weight that an indicator (or a component) has in the composite index. As shown by Paruolo et al. (2012) the structure of correlation between indicators interacts with the weight assigned to them in determining the real importance of an indicator (component). It might be the case that in spite of a weight of e.g. 0.2 an indicator (component) is actually much more important in the composite index. The "importance" can be approximated using the square of the Pearson correlation (see Paruolo et al. for a measure of importance which captures non linearities) and comparing the values within each component. The situation depicted in Table 23 is rather heterogeneous: only the two indicators in the Web Use component have actual weight equal to the theoretical one (0.5). In the Communications Infrastructure component indicator ITUG is actually worth 76% less than ITUC in spite of having the same theoretical weight. In the Institutional Infrastructure (where all listed indicators are theoretically equally weighted) WEFF is actually worth 93% less than UNA while the cluster of Q9a-b and Q9c-d, Q9l but also Q10, Q16 and Q25 are worth at least half of UNA. The indicators Q8a contributes twice as much than Q8b and Q24 to the component Web Content, while Q12 counts 93% less than Q17 and WBC counts 77% less than Q17 in the Economic Impact component. The components Political and Social Impact are instead more balanced meaning that indicators contribute equally to their respective component.

The difference between the theoretical and actual weight of each indicator is calculated also at the component and sub-index level and presented in Table 24. In the Readiness sub-index, the Communications Infrastructure component actually weighs only 6% less than the Institutional Infrastructure component, instead of 50% less as meant by WI developers. The rest of the components are balanced. At the sub-index level, the Impact sub-index is theoretically worth 3 times more than the other two sub-indexes. Its actual importance does not reflect this scheme as all the three sub-indexes equally contribute to the WI score.

The analysis suggests a possible redundancy in the Index framework giving room to a reduction in the number of indicators for future releases of the Index.

Table 24: Pearson correlation coefficients between components and the corresponding sub-indexes and with the WI. Comparison between theoretical and real importance.

	Pearson correlation		Squared Pearson correlation		teorethical weight	difference with max
	sub-index	WI	sub-index	difference with max		
Communications Infrastructure	0.96	0.93	0.91	-6%	0.33	-50%
Instiutional Infrastructure	0.99	0.95	0.97	0%	0.66	0%
Web Use	0.98	0.94	0.95	0%	0.50	0%
Web Content	0.96	0.94	0.93	-2%	0.50	0%
Social	0.95	0.95	0.90	-2%	0.33	0%
Economic	0.94	0.93	0.89	-3%	0.33	0%
Political	0.96	0.95	0.92	0%	0.33	0%
Readiness		0.96		-3%	0.20	-67%
The Web		0.97		-2%	0.20	-67%
Impact		0.99		0%	0.60	0%

4 Conclusions

This study is an assessment of the Web Index 2011 (WI), published by the World Wide Web Foundation in September 2012. The Index, computed for 61 countries, is composed of 85 indicators and uses both survey (primary) data and hard (secondary) data. We analyse both the survey questions with the aim of checking the statistical consistency of the answers, and the WI in order to evaluate its robustness with respect to some of its main methodological assumptions.

The presence of primary data is one of the innovative aspects of the first release of the WI. They play a remarkable role in the construction of the Index as they account for about 60% of the WI indicators. The survey to collect primary data constructed *ad hoc* for the first release of the Index consists of a detailed questionnaire submitted to experts/professionals from 61 countries worldwide and assessed by national and regional reviewers. Designing questionnaires is generally a difficult task. The WI case is particularly challenging given the complex nature of the topic surveyed and the wide coverage required.

Our analysis of primary data aims at providing the questionnaire designers with some insights into possible problematic questions and/or unexpectedly behaving countries. To this purpose a specific model belonging to the family of the Rasch models is employed. Results show that the questionnaire is balanced and the response structure organised in a ten category scale is always appropriate. A few questions stand out as problematic: Q10 (To what extent does the government impose restrictions on access to websites (censorship)?), Q25 (Does the government have a specific Open Government Data initiative?), Q2a (Do the main political parties have websites?) and Q12 (To what extent do you think that the Web is making it easier to undertake criminal activities in your country?). Some of those questions are too technical for the respondents while others are not clear enough or seem counter-oriented with respect to the concept to be measured. In general, we suggest the rephrasing of the problematic questions to make them clearer. No country shows a notable unexpected pattern of answers, confirming that the questionnaire has been always scored by experts at their best. Question difficulties are almost always as expected with a strong indication that gender bias does matter. Finally, primary data from the questionnaire describe an almost unique factor in each WI component, as supported by the Rasch dimensionality analysis. This means that the grouping of the different survey indicators into different WI components is statistically appropriate.

The second part of this report contains the robustness analysis of the WI. Every composite index is the result of a number of choices on the framework, the number and identity of indicators to include, their normalization, the weights attached to each indicator and component, the aggregation method and many others. As with every composite indicator, some choices are openly normative and subjective, driven by developers' and experts' opinion, others can be justified on the basis of statistical analysis, mathematical simplicity or common practice. The uncertainty analysis presented in this study aims at assessing the extent to which some of these choices might affect the country scores and ranks based on the composite Index. To this purpose six alternative scenarios are simulated each challenging one particular assumption made in the WI, including different aggregation methods and different weighting schemes. The assessment of the scenarios is always done in comparative terms with respect to the reference scenario, that is: the WI published by the World Wide Web Foundation in September 2012.

The WI proved to be robust and consistent. For each of the six simulated scenarios, even for the most distant from the reference one, the maximum shift in WI country ranks is always in the band ± 6, which corresponds to 10% of the maximum possible shift in this case. Nevertheless, a few indicators are found to be not in line with the underlying concept, while the general high correlation across WI elements (indicators, components and sub-indexes) highlights a possible redundancy in the number of indicators included.

Overall, despite its multifaceted structure, the wide coverage of different countries and the fact that it includes both survey and hard data, from the statistical point of view the index is robust.

References

Andrich, D. (1988). *Rasch models for measurement.* SAGE Publications.

Bond, G.T., and Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in the Human Sciences*. New Jersey: LEA.

Fischer, G.H., and Molenaar, I.W. (1990). *Rasch Models: Foundations, Recent Developments and Applications*. New York: Springer-Verlag.

Hagerty, M. and Land, K. (2007) Constructing summary indices of quality of life: a model for the effect of heterogeneous importance weights. *Sociological Methods and Research*, 35: 455-496.

Linacre, M.J. (2002). *Optimizing Rating Scale Category Effectiveness*, Journal of Applied Measurement 3:1, p.85-106.

Linacre, M.J. (2012). *A User's Guide to: WINSTEPS and Ministeps: Rasch-Model Computer Programs*. ISBN: 0-941938-03-4.

Michalos, A. (2011) What did Stiglitz, Sen and Fitoussi get right and what did they get wrong? *Social Indicators Research*, 102: 117-129.

Morrison, D. (2005) *Multivariate Statistical Methods*. Thomson.

Paruolo, P., Saisana, M., and Saltelli, A., (2012), Ratings and rankings: Voodoo or Science?

Journal of Royal Statistical Society A, 176: 2, p.1-26.

Smith, E.V.Jr and Smith, R.M. (2004). *Introduction to Rasch Measurement: Theory, Models and Applications*. Maple Grove, Minnesota: JAM Press.

Wilson M. (2005), *Constructing Measures: An Item Response Modeling Approach*, Mahaw, New Jersey, London, Lawrence Erlbaum Associates, Publishers.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA PRESS.

Appendix: list of survey questions

Q1 - To what extent has the Web been used for political mobilisation in your country?

Q2a – Do the main political parties have websites?

Q2b - Do they campaign through the Web - if it is legal to do so?

Q3 - To what extent is there reliable and trusted health information on the Web, to help, for instance, identify ailments, and offer preventative or curative measures, in a language readable by the local population?

Q4 - In cases of outbreak of widespread infectious diseases or epidemics, does the government provide information to the public about disease control or prevention via the Web?

Q5a - To what extent is the local/state curriculum available on the Web: primary education

Q5b - To what extent is the local/state curriculum available on the Web: secondary education

Q5c - To what extent is the local/state curriculum available on the Web: tertiary education

Q6 - To what extent is distance learning used for the training of teachers?

Q7 - To what extent are social networking sites used in the country?

Q8a - To what extent is there relevant and useful content in the local official language of the country in: personal safety and security

Q8b - To what extent is there relevant and useful content in the local official language of the country in: general news

Q8c - To what extent is there relevant and useful content in the local official language of the country in: searching for jobs

Q9a-b – To what extent is there implicit gender bias in computer training

Q9c-d –To what extent is there implicit gender bias in focusing on science and technology education

Q9e - To what extent does the government publicize the importance of access to the Web to all the population?

Q9f - To what extent does the government publicize the importance of access to the Web specifically for women?

Q9g - To what extent are there government programmes specifically focusing on funding training for their staff in ICT use?

Q9h - To what extent are there government programmes specifically focusing on funding training for their women staff in ICT use?

Q9i - In your country, to what extent are there female "role models" in the ICT field?

Q9k - In your country, to what extent are there women's groups' Websites?

Q9l - In your country, in tertiary education, what proportion of ICT graduates are women?

Q10 - To what extent does the government impose restrictions on access to websites (censorship)?

Q11a - To what extent do the segments of society listed below have effective and useful access to the Web: elderly people?

Q11b - To what extent do the segments of society listed below have effective and useful access to the Web: illiterate people?

Q11c - To what extent do the segments of society listed below have effective and useful access to the Web: people with visual disabilities?

Q11d - To what extent do the segments of society listed below have effective and useful access to the Web: people with learning disabilities?

Q11e - To what extent do the segments of society listed below have effective and useful access to the Web: people susceptible to seizures?

Q11f - To what extent do the segments of society listed below have effective and useful access to the Web: people with hearing disabilities?

Q12 - To what extent do you think that the Web is making it easier to undertake criminal activities in your country?

Q13 - To what extent are there laws against cyber crime in your country?

Q14 - To what extent would you say that the Web is trusted as a means of buying and selling goods and services?

Q15 - To what extent do government or non-government agencies disseminate information to farmers?

Q16 - To what extent would you consider your country to be ranking amongst the World's best in training computer engineers?

Q17 - To what extent would you consider your country to have developed successful businesses based on the use of the Web?

Q18 - How reliable is the electricity supply in your country?

Q20 - To what extent would you say that Web access is affordable to the large majority of the population of your country?

Q22 - To what extent are government agencies publishing information on the Web using open licenses?

Q23a - To what extent are there Government data on the Web in the following areas: international trade data

Q23b - To what extent are there Government data on the Web in the following areas: detailed data on budgeted and actual spending on different departments

Q23c - To what extent are there Government data on the Web in the following areas: data on health sector performance (hospitals, doctors etc)

Q23d - To what extent are there Government data on the Web in the following areas: education performance data

Q23e - To what extent are there Government data on the Web in the following areas: transport data and schedules

Q23f - To what extent are there Government data on the Web in the following areas: census data

Q23g - To what extent are there Government data on the Web in the following areas: map data

Q23h - To what extent are there Government data on the Web in the following areas: information on tax returns and how to submit these

Q23i - To what extent are there Government data on the Web in the following areas: information on whom to reach for different government services

Q23j - To what extent are there Government data on the Web in the following areas: data and statistics on crime

Q24 - How easy is it to access government data on the Web in open, machine readable formats?

Q25 - Does the government have a specific Open Government Data initiative?

Q26 - To what extent are Web applications and services in areas such as health, education, security, budgets etc "built" on top of government data?

European Commission
EUR 25476 – Joint Research Centre – Institute for the Protection and the Security of the Citizen

Title: Assessment of the Web Index. Survey questionnaire calibration and uncertainty analysis.

Authors: Paola Annoni, Dorota Weziak-Bialowolska and Michela Nardo

Luxembourg: Publications Office of the European Union

2012 – 100 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424

ISBN 978-92-79-25988-3

doi:10.2788/43830

Abstract

This study contains the assessment of the Web Index 2011 (WI), published by the World Wide Web Foundation in September 2012. The Index, calculated for 61 countries, is composed by 85 indicators and uses both survey and hard data. We analyse both the survey questions with the aim of checking the statistical consistency of survey answers and the WI in order to evaluate its robustness vis à vis some of its main methodological assumptions.

Our analysis of primary data provides survey designers with some insights into possible problematic questions and/or unexpectedly behaving countries. In particular we check for category redundancy in the questions, unexpected answers in some questions and/or by some countries, the relative difficulty of the questions and the validity of the selected framework. The second part of the analysis contains the robustness analysis of the WI. To this purpose 6 alternative scenarios are simulated each challenging one particular assumption made in the WI: from different aggregation methods to different weighting schemes. Overall, despite its multifaceted structure, the wide coverage of different countries and the fact that it includes both survey and quantitative data, the index turned out to be robust from the statistical point of view.

LB-NA-25476-EN-N

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new standards, methods and tools, and sharing and transferring its know-how to the Member States and international community.

Key policy areas include: environment and climate change; energy and transport; agriculture and food security; health and consumer protection; information society and digital agenda; safety and security including nuclear; all supported through a cross-cutting and multi-disciplinary approach.

Publications Office

ISBN 978-92-79-25988-3

9 789279 259883