March 24, 2022

Technical audits of four tools measuring women's economic empowerment at the country-level

By: James C. Knowles

Introduction

This paper reports the findings of technical audits of the methods used in four population monitoring (PM) tools for measuring and monitoring women's economic empowerment (WEE) outcomes at the country level.[1] WEE-focused PM tools involve the calculation of a "composite variable" (index) that is used to compare the status of individual countries (or groups of countries) with respect to women's economic empowerment or related outcomes, such as gender inequality or discrimination against women. In recent years, the number of WEE-focused PM tools has proliferated, and they are being increasingly used by country-level and international decision-makers. The WEE Compendium (2020) previously reviewed 20 WEE PM tools, identifying some issues that warranted further review.[2]

Four PM tools designed to measure WEE related outcomes were selected for detailed audits to help identify common methodological issues as a first step toward the formulation of good practice standards. The four tools selected meet the following criteria considered necessary for a meaningful technical audit: (1) a conceptual framework clearly identifying the ultimate outcome the tool is designed to monitor as well as its dimensions and sub-dimensions, (2) a measurement model identifying the indicators and data sources used to measure the ultimate outcome, (3) a sufficiently large country sample to support multivariate analysis, (4) downloadable country-level data on all indicators, (5) a description of all adjustments made by the tool developers to the raw data (e.g., imputation of missing values, normalization, suppression of extreme values), and (6) an explanation of how the indicators are aggregated to the overall index.[3] Two of the selected tools have ultimate outcomes focused on women only while the remaining two focus on gender differences in economic empowerment outcomes.[4]

---

[1] This paper and Initial work on the audits discussed here was supported by Data2X. The author wishes to acknowledge very helpful comments received on earlier drafts of the audits from staff of the OECD's Development Centre and the Georgetown Institute for Women, Peace and Security. However, any remaining errors and omissions are the responsibility of the author.

[2] Mayra Buvinic, M. O'Donnell, J. Knowles and S. Bourgault. 2020. "Measuring Women's Economic Empowerment: A Compendium of Selected Tools." Data2X and Center for Global Development.

[3] All 20 of the PM tools reviewed in the 2020 Compendium satisfy criteria 1, 2, and 6. However, several of the tools reviewed in the Compendium do not have sufficiently large country samples to support multivariate analysis (criterion 3), some do not make downloadable country-level data available (criterion 4), and some do not describe all adjustments made to the raw data (criterion 5).

[4] The four tools audited are: (1) the Women's Economic Empowerment and Equality Dashboard (WE3), tool developed by the United States Agency for International Development (USAID); (2) the 2019/20 Women's Peace and Security Index (WPS), developed by Georgetown Institute for Women, Peace and Security and the Peace Research Institute, Oslo; (3) the 2019 Social Institutions and Gender Index (SIGI), developed by the Organisation for

This paper reports the main findings of the technical audits of the four tools with respect to their conceptual frameworks, measurement models, indicators and data sources, country rankings, adjustments to the raw data, use of multivariate analysis, aggregation formulas, external validity and transparency as well as overall conclusions and recommendations based on the findings. Sensitivity analysis is used widely in the audits to assess how the various choices and decisions made by the tools' developers affect the country rankings and their estimated "criterion validity." Estimates of criterion validity are based on the correlations between a tool's country rankings and the external indicators with which they would be expected to be most closely correlated. The seven criterion indicators used in the audits include measures of both overall development and WEE-related outcomes. They are: per capita GDP PPP, percent of the total population in urban areas, the UN's Coefficient of Inequality (CI), the UN's Human Development Index (both the overall HDI and the HDI for females), the UN's Gender Inequality Index (GII) and UN's Gender Development Index (GDI).[5] These seven indicators are used exclusively to estimate the tools' criterion validity.

Conceptual frameworks

A PM tool's conceptual framework should identify the ultimate outcome it is designed to monitor (e.g., women's economic empowerment, gender (in)equality, discrimination against women) as well as its dimensions and sub-dimensions and their relationships to the ultimate outcome. These elements of a PM tool's conceptual framework are not directly observable ("latent outcomes"), unlike the indicators that are used to measure them, which are defined by the tool's measurement model (discussed below). The conceptual framework should justify the tool's choice of ultimate outcome, dimensions and sub-dimensions, with references to the WEE literature. For example, if the tool's ultimate outcome is gender inequality, the conceptual framework should explain and justify how it measures gender gaps, for example, whether they are censored at gender equality and the rationale for doing so. The conceptual framework provides the unique opportunity for tool developers to express normative judgments about what they think is most important and why, as distinct from their judgments about how best to measure the tool's ultimate outcome validly and reliably.

Key findings: None of the four audited tools cite relevant WEE or broader gender equality literature supporting their conceptual frameworks. Although one tool has a clearly defined and carefully considered conceptual framework, the remaining three tools do not. One tool reports that it identified its dimensions after first selecting the indicators, while a second reports that its dimensions focus on "outcomes" instead of "inputs," but without clearly distinguishing the two. The third tool's ultimate outcome is unclear, with several indicators measuring women's outcomes without any reference to men's, while others measure gender differences in outcomes and one measures an outcome based on the total population.

Measurement models

A PM tool's measurement model provides the functional link between its ultimate outcome and the indicators used to measure it. The tool's technical documentation should describe this functional link

---

European Economic Co-operation and Development; and (4) the Global Gender Gap Index (GGGI), developed by the World Economic Forum. The detailed audit reports are available at James C Knowles - Academia.edu

[5] Most of the criterion indicators refer to the year 2018. The exception is the CI, which refers to the period 2015-2020. The data for all seven criterion indicators were obtained from the UN Human Development Report website (http://hdr.undp.org/en/content/download-data).

clearly and explain any differences between the tool's measurement model and its conceptual framework.

Key findings. All four of the audited PM tools include measurement models that describe the functional relationship between the indicators and the tool's ultimate outcome. However, all four measurement models assign *distinct* subsets of indicators to the dimensions, making the strong assumption that the ultimate outcome is not *directly* related to the indicators but is instead *indirectly* related to the dimensions. None of the tools describes the criteria and procedures used to identify its indicators and their links to specific WEE dimensions. For example, none of the tools includes any correlation analysis demonstrating that the indicators linked to a given dimension are all highly correlated with that dimension. The measurement models also include widely varying numbers of indicators (from 11 to 47 indicators to measure three to five dimensions). Differences between the measurement model and the conceptual framework are present in only one tool and are clearly explained as due to current data limitations.

Indicators

The indicators and their data sources should be clearly identified and justified as both valid and reliable measures of the outcomes they are intended to measure. Indicators that include imputed values or that are based on sample survey estimates should be clearly identified.

Key findings. All four of the audited tools clearly identify their data sources. Most are standard international sources. However, all four audits question the reliability and/or validity of some indicators. For example, several of the tools include indicators that are survey estimates based on small country samples (e.g., the Gallup World Poll, which is usually based on country samples of 1,000 adults with sampling errors of ±4%) but are treated statistically as population values. Even when the indicators are drawn from standard international sources (e.g., UN agencies), the values reported for low-income countries are likely to be estimates based on small, infrequent surveys or unreliable official data. For example, X tool includes indicators that draw on ILO data. The ILO reports annual (and even monthly) *estimates* of many labor force indicators for 189 countries (https://ilostat.ilo.org). All of the estimates are based on econometric models using a variety of data sources.[6] Specially designed labor force surveys, which are conducted annually by most high and higher middle-income countries, are the preferred source for reliable labor force data. But many low-income countries have never conducted a national labor force survey, while many lower middle-income countries conduct them at infrequent intervals.[7] In spite of these country differences, the PM tools using ILO labor force indicators treat them as though they are population values.

The multivariate analyses conducted in all four audits identify at least some indicators as clearly invalid and/or unreliable measures of the outcomes they are intended to measure. For example, one of the tools uses the sex ratio at birth as an indicator of health outcomes, whereas the demographic literature and the remaining two tools use abnormally high values of the sex ratio as an indicator of gender discrimination.[8] Two of the tools recode the raw data on sex ratios such that absolute differences from

---

[6] https://www.ilo.org/ilostat-files/Documents/TEM.pdf.

[7] For example, according to ILOSTAT, the most recent national labor force surveys conducted in Lao PDR (a lower middle-income country with a population of more than 7 million) were conducted in 2010 and 2017.

[8] Fengqing Chao and others (2019) "Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels" PNAS 116(19): 9303-11 (https://www.pnas.org/content/116/19/9303); Hannah Ritchie and Max Roser (2019). "Gender Ratio". *OurWorldInData.* (https://ourworldindata.org/gender-ratio).

their assumed normal ratios (1.05 or 1.06) are treated symmetrically, whereas the third tool recodes all values *below* its assumed normal value (1.05) to the assumed normal value. The multivariate analysis conducted in the audits consistently indicates that all three of these recoded sex ratios are insignificantly related to the outcomes they are intended to measure.

Country rankings

Individual country/territory rankings based on a tool's overall index are of greatest interest to most tool users. Some tool developers report that even shifts over time of as little as one position in the ranking of individual countries are a source of concern to both country-level and international decision-makers.

Key findings. The country rankings for three of the tools present few surprises. However, the country rankings of one tool include four countries in their top 10 whose *median* rankings among the four tools range from 48th to 68th.[9] Although the country rankings of all four PM tools are significantly related at the 0.05 level or lower with the expected signs to the country rankings of all seven "criterion" indicators, estimates of overall criterion validity (as measured by the R^2s obtained from multiple regressions of the tools' country rankings on all seven criterion indicators) range from moderate (0.582) to high (0.924).[10]

One audited tool does not provide country rankings because its indicators have too many missing values even after some imputation. The audit obtains country rankings for this tool by imputing values based on linear regression models with the explanatory variables limited to dummy variables representing four income groups and seven regions (and their significant interactions). The resulting country rankings achieve the highest estimated criterion validity (0.924) among the four tools audited. None of the country rankings reported by the four tools reflects the uncertainty introduced when missing values are imputed or when indicators are based on survey estimates (as discussed above under "Indicators"). This can be very misleading, especially in the case of low-income countries with limited and/or unreliable data.

Adjustments to the raw data

Adjustments to the raw data include the treatment of missing (unreported) values, normalization of indicators to comparable scales, recoding of extreme values to prevent them from distorting country rankings, or recoding indicators to restrict them to certain ranges (e.g., censoring values that exceed gender equality thresholds).

Key findings. The way missing values are treated has by far the biggest impact on PM tools and their country rankings, as discussed in the Box ("Treatment of missing values"). In addition, all four tools normalize their indicators in different ways. The normalization methods are not described in one tool, while another tool censors its indicators (ratios of women's to men's outcomes) at values of 50% (signifying gender equality) without any assessment of how this affects the country rankings or their criterion validity. A third tool combines non-standard normalization with special weights that are used in aggregating distinct sets of indicators to the dimension values. The audit of this tool found that use of a more standard normalization method (standardization followed by min-max normalization) and an unweighted arithmetic mean aggregation formula significantly increase the estimated criterion validity of the country rankings. Other adjustments to the raw data (recoding extreme values, restricting values

[9] The four countries are: Nicaragua (5), Rwanda (6), Philippines (8), and Namibia (10).

[10] Because the sizes of country samples vary substantially between the four tools, it is more meaningful to compare the estimates for the common country sample of 113 countries. Within this common country sample, however, the corresponding estimates are quite similar (i.e., from 0.539 to 0.941).

to certain ranges) were found to have large effects on the country rankings in some cases, but without significantly affecting their estimated criterion validity.

Box. Treatment of missing values

Reflecting the limited availability of gender data, the treatment of missing values in a PM tool's indicators is one of the biggest challenges facing tool developers. The four tools audited address this challenge in one (or a combination) of the following ways: (1) by dropping *countries* with missing values on the included indicators from the country sample, (2) by dropping *indicators* with missing values, and/or (3) by imputing the missing values. Use of any of these options comes with limitations. For example, dropping countries with missing values reduces a PM tool's geographical coverage and, depending on the characteristics of countries dropped, can result in a biased country sample. One of the tools measuring "discrimination against women" dropped all countries with missing values from its country sample, which included most countries with majority-Muslim populations. Dropping indicators with missing values is a particularly non-transparent option because no information is typically provided on the indicators that were dropped. Missing values are imputed by three of the four audited tools without any effort to assess their effects on the country rankings. To address this limitation, all four audits used multiple imputation (MI) to estimate confidence intervals for the predicted country rankings based on imputed values.[11] The results indicate that the widths of the predicted confidence intervals vary considerably across countries, depending mainly on the number of missing values imputed in each country but also on the distributions of their reported indicator values.

Multivariate analyses

Multivariate analysis should be used to assess whether the assumptions in the measurement models are consistent with the tools' actual databases of indicators, and if not, how this affects the estimated criterion validity of the tool's country rankings.

Key findings. None of the audited PM tools reports any multivariate analysis (MVA) of its indicators. However, the multivariate analysis conducted in all four audits found that the actual databases of indicators are inconsistent with the assumptions in their measurement models. For example, the exploratory factor analysis (EFA) of all four tools obtained patterns of loadings of the individual indicators on rotated factors that are inconsistent with the assignment of the indicators to dimensions in their measurement models. Similarly, estimates obtained from both restricted and unrestricted structural equations models (SEMs) used in the audits of all four tools indicate that the assignment of distinct sets of indicators to dimensions in their measurement models significantly reduces the models' ability to explain the variation in their ultimate outcomes.

Aggregation methods

A key decision by PM tool developers is how to aggregate the indicators to their overall indexes. The choice of aggregation formulas (e.g., linear versus non-linear, weighted versus unweighted) should be based on which aggregation formulas lead to an overall index that yields country rankings with the highest estimated criterion validity rather than on normative criteria.

---

[11] Michela Nardo, Michaela Saisana, Andrea Saltelli & Stephano Tarantola. 2005. "Tools for Composite Indicators Building." Joint Research Centre, European Commission (https://publications.jrc.ec.europa.eu › JRC31473).

Key findings. The four PM tools audited use a variety of formulas to aggregate their indicators to the overall index, with the formulas selected in some cases on the basis of normative considerations. For example, one tool introduced new aggregation formulas to replace principal components analysis (PCA)-based formulas in which "the importance that the different variables received … depended on ethically irrelevant considerations." The audits found that varying aggregation formulas can have important effects on the country rankings, and in some cases, significant effects on their estimated criterion validity. For example, the estimated criterion validity of one tool's country rankings increased significantly when an alternative nonlinear aggregation formula was used instead of an unweighted arithmetic mean to aggregate its dimension values to the overall index.

All four audits also found that predicted scores obtained from the PCA and the unrestricted SEMs that do not use the dimensions as mediators (unlike the scores obtained from the EFA and restricted SEMs) can be used directly to obtain country rankings that have significantly higher criterion validity than country rankings based on the tools' measurement models. For example, the estimated criterion validity of one tool's country rankings (i.e., the R^2 in the multiple linear regression of the country rankings on all seven criterion indicators) increases significantly (p=0.000) from 0.582 to 0.756 (or from 0.582 to 0.817) when countries are ranked directly on the basis of PCA-predicted scores (or the unrestricted SEM-predicted scores).

External validity

The external validity of the PM tools is assessed in the audits both by comparing country rankings to the country rankings of seven criterion indicators (as discussed above under "Country rankings") and by comparing the audited tools' country rankings to each other (as discussed below).

Key findings. Despite having different ultimate outcomes, the country rankings of all four tools are positively and significantly correlated at the 0.05 level or below with the country rankings of the other PM tools, with ρ's ranging from 0.646 to 0.930 in a comparable sample of 113 countries. However, the country rankings of one tool have consistently lower correlations with the country rankings of the *other* three tools (i.e., ρ's=0.646 to 0.721, compared to ρ's=0.850 to 0.930 between the other three country rankings). Although the country rankings of all four tools are significantly correlated, the rankings of some individual countries vary widely across the tools. For example, whereas Germany has a median ranking of 14 with an average absolute difference of 1.33 positions across the four tools (based on six comparisons), France has the same median ranking but with an average absolute difference of 8.00 positions, while Austria has a median ranking of 10.5 with an average absolute difference of 21.17 positions.

Transparency

An independent and credible assessment of the data and methods used by the PM tools and of the validity and reliability of their country rankings requires PM tool developers to make their raw data available, preferably in a form that is convenient for analysis, and to provide complete and accurate technical documentation.

Key findings. Although all four of the audited PM tools deserve credit for making their data available and for providing some technical documentation, all four audited tools have some transparency gaps. For example, none of the audited tools provide their downloadable data in a format that is ready for use by analysts. Instead, they provide their downloadable data in Excel files that do not include descriptive variable labels or value labels. One tool provides data only to one or two decimal places, which makes it impossible to obtain the correctly normalized values of some indicators. Another tool provides only

censored values (not raw data) in its downloadable data set so that it is not possible to assess the effect of the data censoring on the country rankings.

The technical documentation of all four audited tools is also incomplete and/or inaccurate in some respects. For example, the technical documentation of one tool does not mention that its country rankings are obtained by imputing a value of zero to two indicators with numerous missing values. One tool uses complex normalization procedures that are not explained in its technical documentation, while another tool shows the special normalization procedure it uses for one indicator only in the small print of an illustrative example. Another tool fails to mention that the extreme values of several indicators are "winsorized" (recoded to the next highest or lowest non-extreme value).

Overall conclusions

The audits of the four PM tools find that their country rankings have moderate to high estimated criterion validity. However, the audits also reveal that all four audited tools have the following problems:

- All four lack multivariate analysis supporting the strong assumptions made in their measurement models, including their choices of dimensions and the distinct sets of indicators assigned to them. Multivariate analysis conducted in the audits finds that the data structures defined in the measurement models of all four tools are significantly inconsistent with their actual databases of indicators.
- The role played by dimensions in all four PM tools needs to be reconsidered. Whereas dimensions can be very useful as elements in a conceptual framework and in identifying potential indicators, they should not be used as mediating factors between distinct sets of indicators and the tool's ultimate outcome. All four audits found that predicted scores obtained from PCA and unrestricted SEM models in which the dimensions play no mediating role can be used directly to obtain country rankings with significantly higher estimated criterion validity than the tools' own country rankings.
- The treatment of missing values is the most serious data-related problem faced by all four PM tools, reflecting persistent gaps in the availability of suitable gender data. The audited tools address this problem in different ways: by imputing values (two tools), by limiting the country sample to countries reporting the values of all indicators (two tools), by narrowing the choice of indicators (three tools), and/or by declining to calculate higher-level indexes (one tool). None of the audited tools assess the effects of these measures on the estimated criterion validity of the country rankings.
- Sensitivity analysis is not used by any of the four audited tools to assess the effects on the country rankings of the numerous decisions and choices made in developing the PM tools. The sensitivity analysis conducted in the audits finds that the country rankings are in many cases quite sensitive to these decisions and choices, although their estimated criterion validity is not always significantly affected.
- Although the audited tools deserve credit for making their data publicly available, limited transparency is still a problem for all four audited tools. Transparency gaps identified in the audits include: the omission of critical data from their downloadable data sets, inconvenient formats of the downloadable data, and omitted or misleading information in their technical reports.

Recommendations

This paper makes the following recommendations to increase the validity and reliability of PM tools measuring WEE outcomes and to bolster the utility and credibility of their country rankings. These recommendations can form the basis for best practice standards and guidelines for future PM tool development. The first three recommendations address methodological problems in connection with index construction; the last two address the problem of limited transparency:

1. PM tools should include multivariate analysis assessing whether their databases of indicators are consistent with the assumptions made in their measurement models. As the functional link between a PM tool's database of indicators and the overall index used to rank countries, even small changes to the assumptions in a tool's measurement model can significantly affect the country rankings and/or their estimated criterion validity. The multivariate analysis of PM tools should use a range of statistical models that are suitable for the analysis of latent outcomes, including (but not limited to) Cronbach's alpha, item response theory (IRT), PCA, EFA, and SEMs.

2. PM tools should include sensitivity analysis to assess the effects of decisions and choices made in the course of developing a tool on their country rankings and their estimated criterion validity, including (but not limited to) the assumptions built into measurement models, adjustments to raw data, and the formulas used to aggregate indicators to the overall index. Criterion validity can be estimated by comparing the tool's country rankings to the country rankings of external development indicators that are considered most relevant for a given tool.

3. Whenever possible PM tools should use systematic rather than ad hoc methods to adjust raw data. For example, missing values can be imputed using a method such as "multiple imputation" that provides estimates of confidence intervals for predicted country rankings based on imputed values. When the values of indicators are imputed or based on survey estimates (as opposed to population values), the country rankings should, to the extent possible, include estimates of their standard errors.

4. Raw data used by all PM tools should be readily available for downloading by interested users in a format that facilitates their analysis. In addition to a standard format (e.g., Excel, CSV), the downloadable data should also be provided in the format of a widely used statistical package that supports descriptive variable and value labels (e.g., Stata, SPSS, SAS) so that interested users can analyze the data without having to invest a lot of time converting, recoding and re-labeling the downloaded database to prepare it for analysis. Downloadable data sets should also include the ISO 3166-1 alpha-3 country/territory codes identifying the countries and territories represented in their data sets to facilitate comparisons of the country rankings of their PM tools with those of other PM tools or with those of widely used external criterion indicators (e.g., the UN's Human Development Indicators). The site providing the downloadable data should also clearly identify the person(s) to contact for questions about the data or the technical report and should also list any restrictions on the use of the data.

5. All PM tools should be accompanied by an accurate and complete technical report that clearly explains the assumptions and choices made with respect to the conceptual framework, the measurement model, data sources, adjustments to the raw data, formulas used to aggregate indicators to indexes, and the results of multivariate and sensitivity analyses. Once equipped with the technical report and the raw data, a user should be able to reproduce not only the

country rankings but also the normalized indicators and reported index values. If some of the information needed is provided in earlier technical reports, that information should be clearly referenced in the current report.