Annexes to stage 4

Measure of accuracy of the estimate to the true value of the indicator

Accuracy refers to the closeness of estimates to the true value of the indicator. The gender statistics generated are estimates that may or may not give the true value of the indicator. However, if all the possible estimates using a particular estimator were considered then the average value provides an idea of the true value of the indicator. Mathematically, the expected value or the long run average value of the estimator is equal to the parameter or indicator being estimated, that is:

$$\text{Expected Value of the Estimates} = \text{True Value of the Indicator being Measured}$$

Being accurate means having an estimator that is unbiased. The difference between the average value of the estimates and the true value is called bias. Bias is equal to zero when the estimator is unbiased. Mathematically, it is expressed as:

$$\text{Bias} = \text{Expected Value of the Estimates} - \text{True Value of the Indicator being Measured}$$

There is an overestimation when Bias is positive or Bias > 0. This means that the Expected Value of the Estimates is greater than True Value of the Indicator being Measured.

There is an overestimation when Bias is negative or Bias < 0. This means that the Expected Value of the Estimates is less than True Value of the Indicator being Measured.

A large bias may be due to sampling error, non-sampling-error, or both. Non-sampling errors cover all types of errors from all sources such as response errors, coverage errors, and errors linked to data collection and processing.

Accuracy or unbiasedness is a property of the estimator itself. In the formulation of the estimator based on the sampling design of the survey, accuracy or unbiasedness was already considered. Generally, the estimators which are formulated during the design stage of the survey are expected to generate unbiased or accurate estimates. Applying it to the generation of disaggregated gender statistics, the direct approach of estimation will lead to accurate or unbiased estimates.

Measure of precision of the estimate to the true value of the indicator

Like accuracy, precision is a measure of closeness. Unlike accuracy, it is a measure of closeness of the estimates to each other. In the image below, the “center” of the circles represents the true value of indicator being measured while the “marks” represent the estimates. The first figure shows the estimates close to each other but are far from the true value. In the second figure, estimates are close to the true value but are far from each other. Disaggregated gender statistic should be aimed to be both precise and accurate.

Precision of the estimator is measured through its standard error which is a function of the square root of the ratio of the variance of the estimates and the number of observations used in the estimation. Since the standard error is a measure of error in estimation, it can be expressed that on the average, the larger the value of the standard error, the bigger is the error in estimation. On the other hand, the smaller the value, the smaller is the error in estimation.

Standard error is inversely related to the number of observations used in estimation. Thus, a greater number of observations used in estimation would mean smaller standard error of the estimate which means that it is more precise. In the generation of disaggregated gender statistics, the estimates are expected to be less precise as more disaggregation is applied in the data. The lower level of disaggregation results to a smaller subdomain with fewer observations or in extreme cases no observations at all. This usually happens in direct estimation since the sample size was computed for large domain of estimation, like national or at least regional level.

Estimation of the Variance/Standard Error of the Estimates

Theoretically, the direct subdomain estimates are generally said to be unbiased or accurate but these estimates are not precise or the estimates are expected to have large standard errors due to few observations obtained for that particular subdomain. As the disaggregation level becomes deeper it is expected that the standard error of the estimates to increase because of the decreasing number of observations in the subdomain formed due to the disaggregation. This can be seen in the following relationship between the standard error and number of observations,

$$standard\ error\ of\ the\ estimate = \sqrt{\frac{variance\ of\ the\ estimate}{number\ of\ observations\ used\ in\ the\ estimation}}.$$

The standard error of the estimates are computed in order to assess the precision of the generated disaggregated gender statistics as explained further in the next section. A small standard error means that the estimates are close to each other and hence, the estimates are more precise while high value of standard error implies the opposite.

Most nationwide surveys follow a stratified multi-stage sampling design and under this design, the expansion estimator for a total as given by Rao and Molina (2015) [53] and when applied in the direct estimator of subdomain total, $\hat{Y}_i$, given earlier can now be expressed as

$$\sum_{j \in S} w_{ihlk} y_{ihlk}$$

where w_{ihlk} is the design weight associated with the k^{th} secondary sampling unit in the l^{th} primary sampling unit (cluster) belonging to the h^{th} stratum, y_{ihlk} is the associated y-value, and $\sum_{j \in S}$ is the summation over all elements $j = (hlk) \in s(h=1,2,..L;\ l=1,2,..n_{(ih)};\ k=1,2,..n_{(ihk)})$. $n_{(ih)}$ and $n_{(ihk)}$ represent the total number of primary sampling units and secondary sampling units, respectively, that are included in the subdomain of interest. It was also reported that the sample is commonly treated as if the clusters are sampled with replacement, and subsampling is done independently each time a cluster is selected. Such action leads to overestimation of the variance, but this makes the variance estimator greatly simplified as shown in the following expression:

$$v(\hat{Y}_i) = \sum_{h=1}^{L} \frac{1}{n_{(ih)}(n_{(ih)} - 1)} \sum_{l}^{n_{(ih)}} (y_{ihl} - \bar{y}_{ih})^2$$

where $y_{ihl} = \sum_{k=1}^{n(ihk)} \left(n_{(ih)} w_{ihlk}\right) y_{ihlk}$ are weighted sample cluster totals and $\bar{y}_{ih} = \sum_{l=1}^{n_{(ih)}} y_{ihl} \big/ n_{(ih)}$.

Estimation of the Variance/Standard Error of the Estimates using STATA

Some software like STATA® generates the variance or standard error of the estimates through its *svy* commands. But again, one must be careful in this computation as the sample design might not be fully accounted for in the estimation procedure.

With the *svy* commands of STATA®, the standard error of a proportion can be generated for the proportion of child marriages using the data set of Mongolia's MICS 2018 after setting the design parameters of the survey in the command svyset. The command 'Proportion' implemented under the set of svy commands takes the Taylor linearized[1] standard error computation as default procedure. Thus the STATA command 'svy: proportion childm' will have the following output table with the estimate and its standard error based on the specified survey design parameters.

Table 1. STATA output for generating SE and CI for "proportion of women (18-49 years old) who married as children"

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
childm				
0	0.9062	0.0053	0.8953	0.9160
1	0.0938	0.0053	0.0840	0.1047

[1] Taylor linearization is also known as the delta method or the Huber/White/robust sandwich variance estimator used to derive an approximation to the variance of a point estimator, such as a ratio or regression coefficient. [R8]

The national estimate 9.38 (0.0938 on the table above) proportion of child marriage among women aged at least 18 years in Mongolia that was obtained using the 'Tabulate' command in previous section has a standard error of 0.0053 computed from 846 observations.

Doing the same for the disaggregated statistics on the '*Proportion of child marriage among the poorest women aged at least 18 years*' in Mongolia which is estimated previously to be 11.4 percent (0.1142 on the table above), the standard error of this estimate is 0.0094 based on 280 observations. The standard error was generated using the STATA® command 'svy: proportion childm_poorest' with the following output table:

Table 1. STATA output for generating SE and CI for "proportion of poorest women (18-49 years old) who married as children"

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
childm_poorest				
0	0.8858	0.0093	0.8662	0.9028
1	0.1142	0.0093	0.0972	0.1338

For the estimate of 4.9 percent (.0491 on the table below) proportion of child marriage among the richest women aged at least 18 years in Mongolia, the estimated standard error is 0.0096 based on 45 observations. This was obtained from the following output table using the STATA® command 'svy: proportion childm_richest'

Table 2. STATA output for generating SE and CI for "proportion of richest women (18-49 years old) who married as children"

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
childm_richest				
0	0.9509	0.0096	0.9281	0.9667
1	0.0491	0.0096	0.0333	0.0719

Note that there are two new variables created, namely: childm_poorest and childm_richest before the generation of standard errors. Also, notice that the estimates of the proportions did not change whether one uses the 'Tabulate' command or 'Proportion' command under the *svy* set of commands. It is the generated standard error that differs when one uses the 'Proportion' command under the *svy* set of commands compared to an ordinary 'Proportion' command.

Table 3. Comparison of generated SEs of estimates using ordinary "Proportion" command and svy "Proportion" command

INDICATOR	STANDARD ERROR (ordinary "Proportion" command)	STANDARD ERROR (svy "Proportion" command)
1. Proportion of child marriage among women ages 18-49 years old	0.0031	0.0053
2. Proportion of child marriage among *poorest* women ages 18-49 years old	0.0078	0.0093
3. Proportion of child marriage among *richest* women ages 18-49 years old	0.0050	0.0096
4. Proportion of child marriage among women residing in *urban* areas ages 18-49 years old	0.0038	0.0070
5. Proportion of child marriage among women residing in *rural* areas ages 18-49 years old	0.0057	0.0067
6. Proportion of child marriage among *poorest* women residing in *urban* areas ages 18-49 years old	0.0396	0.0346
7. Proportion of child marriage among *richest* women residing in *urban* areas ages 18-49 years old	0.0050	0.0096
8. Proportion of child marriage among *poorest* women residing in *rural* areas ages 18-49 years old	0.0079	0.0095
9. Proportion of child marriage among *richest* women residing in *rural* areas ages 18-49 years old	-	-

Some NSOs practice the release of official statistics with some measures of standard error committed in the estimation process. However, there are more who do not put this in practice. There should be advocacy for this practice as these measures provide us with guidance on assessing the quality of the generated official statistics.

Measure of reliability of the estimate to the true value of the indicator

Reliability of the estimates are measured using coefficient of variation (*CV*). It is a measure of variability relative to the value of the estimates. Mathematically, it is expressed in percent and computed as ratio of the standard error of the estimate and value of the estimate as seen in the following formula for computing the *CV* of an estimate for a subdomain total:

$$CV(\hat{Y}_i) = 100\% \times \frac{\sqrt{v(\hat{Y}_i)}}{(\hat{Y}_i)}.$$

Relative to the value of the estimate, CV measures the spread of the estimates, so that the bigger the value of the *CV*, the less reliable the estimate is.

In principle, preferred estimates are those with *relatively* small values of CV. It is important to note that there are no internationally agreed standards or recommendations as to the “acceptable” values of CV for a certain type of estimator. In practice, CV thresholds vary country to country and in some cases, from surveys to surveys. Some literature regards a measure of CV less than 10 percent as highly acceptable while a CV with value between 10 and 20 percent as still acceptable. For CV values ranging between 20 and 33 percent, estimates are regarded as less but still sufficiently reliable but should be used with caution. For those greater than 33 percent, caveats should be provided in terms of the level of reliability of these estimates. More detailed discussion on the practical application of this general principle (or more appropriately “general rule of thumb”) as well as empirical basis of the cut-offs used is provided in Section H.4.1 and H.4.2 presenting the Philippines’ and Canadas’s experiences, respectively.

Estimation of the Variance/Standard Error of the Estimates using STATA

Applying these measures in the nine disaggregated gender statistics generated using Mongolia’s MICS 2018, the computed estimates and corresponding characteristics are summarized in the following table:

Table 1. Indicators’ Disaggregating Variables, Number of Observations, SE, and CV

INDICATOR	ESTIMATE	DISAGGREGATING VARIABLE/S	NUMBER OF OBSERVATIONS	STANDARD ERROR	CV (%)
1. **Proportion of child marriage among women aged at least 18 years**	0.0938	AGE OF WOMAN	846	0.0053	5.61
2. **Proportion of child marriage among *poorest* women aged at least 18 years**	0.1142	AGE OF WOMAN and WEALTH INDEX QUINTILE	280	0.0093	8.14
3. **Proportion of child marriage among *richest* women aged at least 18 years**	0.0491	AGE OF WOMAN and WEALTH INDEX QUINTILE	45	0.0096	19.51

INDICATOR	ESTIMATE	DISAGGREGATING VARIABLE/S	NUMBER OF OBSERVATIONS	STANDARD ERROR	CV (%)
4. **Proportion of child marriage among women aged at least 18 years and residing in *urban* areas**	0.0906	AGE OF WOMAN and LOCATION (URBAN/RURAL)	388	0.0070	7.78
5. **Proportion of child marriage among women aged at least 18 years and residing in *rural* areas**	0.1006	AGE OF WOMAN and LOCATION (URBAN/RURAL)	399	0.0067	6.69
6. **Proportion of child marriage among *poorest* women aged at least 18 years and residing in *urban* areas**	0.1141	AGE OF WOMAN, WEALTH INDEX QUINTILE and LOCATION (URBAN/RURAL)	14	0.0346	30.36
7. **Proportion of child marriage among *richest* women aged at least 18 years and residing in *urban* areas**	0.0492	AGE OF WOMAN, WEALTH INDEX QUINTILE and LOCATION (URBAN/RURAL)	45	0.0096	19.51
8. **Proportion of child marriage among *poorest* women aged at least 18 years and residing in *rural* areas**	0.1142	AGE OF WOMAN, WEALTH INDEX QUINTILE and LOCATION (URBAN/RURAL)	266	0.0095	8.34
9. **Proportion of child marriage among *richest* women aged at least 18 years and residing in *rural* areas**	-	AGE OF WOMAN, WEALTH INDEX QUINTILE and LOCATION (URBAN/RURAL)	0	-	-

Out of these nine indicators, the first one makes use of only one disaggregation variable, while the next four have two disaggregation variables and the remaining four have three disaggregation variables. Expectedly, indicator 1 above, which has the most number of observations, has the most precise estimate. On the other hand, indicator 6 above, which has the least number of observations – due to multiple disaggregations – would produce the least precise estimates. There is no estimate for indicator 9 as the small subdomain formed has no observations captured in the disaggregation process.

In terms of reliability as measured by the CV of the estimates, only five of eight estimates are relatively reliable (indicators 1, 2, 4, 5, and 8 above). Two estimates with CVs greater than 10 but less than 20 percent are relatively sufficiently reliable but should be used with caution (indicators 3 and 7 above). On the other hand, a caveat should be provided when publishing indicator 6 above given a CV of 30 percent.

In summary, it is almost always computationally possible to generate disaggregated gender statistics from existing household surveys. However, given the nature of the statistical exercise of producing statistics with multiple disaggregations – that is, effectively decreasing number of observations – quantitative assessment of these disaggregated estimates is an imperative. In this light, generation of measures of accuracy, precision, and reliability should be promoted, observed, and institutionalized as a best practice of NSOs and other data producers.