

# **CMFB**

---

## **ELEMENTS OF A QUALITY FRAMEWORK - MANDATE FOR A TASK FORCE**

### **1. Introduction**

The 4th progress report of the Economic and Financial Committee on the Statistical Requirements in EMU refers to the quality of statistics. It states that more work is needed to operationally assess the various dimensions of quality and invites the SPC, in co-operation with the CMFB, to make proposals in this regard.

*The CMFB approved in its January 2002 meeting:*

- *The installation of a joint ECB/Eurostat task force on output quality, dealing with balance-of-payments and quarterly national accounts statistics.*
- *The mandate for this task force (see below).*

## **Mandate of a joint ECB/Eurostat task force on output quality**

The main objective of the exercise is to provide guidance to users when interpreting statistical data, from an operational point of view, while avoiding duplications with other work in this field. The CMFB proposes to install a task force that will use to the greatest extent possible the Data Quality Assessment Framework (IMF) and other work on quality conducted by the ECB, Eurostat and participating Member States. More specifically the mandate for this task force is as follows:

1. To identify a limited set of indicators to operationally measure/assess several output quality dimensions (e.g. stability against revisions, consistency) of statistics. The task force will take into account and set priorities for the quality dimensions that affect the Euro area/EU aggregates.<sup>1</sup>
2. To define and operationally assess the set of indicators in its two pilot areas: balance of payments and quarterly national accounts statistics.
3. To identify possible ways for a supplementary, more qualitative assessment of output quality.
4. To elaborate a proposal for implementing the indicators for output quality in two pilot areas in the EU Member States, including a timetable.
5. To make proposals for monitoring and communicating the operationally assessed quality dimensions in a European context (e.g. in the framework of the EMU Action Plan).

The task force will start dealing with balance of payments statistics (a report will be discussed in the June 2002 meeting of the CMFB as an input to the report to the EFC in Autumn 2002) and continue with quarterly accounts (a report will be discussed in the January 2003 meeting of the CMFB). The latter report should take into account the work done and conclusions reached by the former, e.g. on methods to assess quality indicators or on monitoring and communication. The composition of the task force may partially change, in order to deal with specific expertise in both pilot areas, while maintaining sufficient continuity.

The task force is composed of representatives coming from NCBs and NSIs. The Chairmanship is jointly ensured by the ECB (DG-Statistics) and Eurostat. The task force is likely to meet four times (twice for balance of payments in the first half of 2002 and twice for quarterly national accounts in the second half of 2002) and will be discontinued once its reports have been approved.

---

<sup>1</sup> As set out in e.g. 'Assessing the quality of the Euro area b.o.p./i.i.p. statistics (ECB ST/STC/BP/QUALIMP4), 'Quality Report' (Eurostat BP/01/45/E) and 'Some elements of a quality framework for CMFB statistics' (Algera, Keuning).

## Annex 2

# JOINT ECB/EUROSTAT TASK FORCE ON QUALITY

## List of participants

### Balance of payments

Chairmen: Jean-Marc ISRAEL (ECB) \_ [jean-marc.israel@ecb.int](mailto:jean-marc.israel@ecb.int)

- Secretaries: Carmen PICON AGUILAR (ECB) \_ [carmen.picon\\_aguilar@ecb.int](mailto:carmen.picon_aguilar@ecb.int)
- Violetta DAMIA (ECB) \_ [violetta.damia@ecb.int](mailto:violetta.damia@ecb.int)
  
- Symon ALGERA - Statistics Netherlands (previous meetings) \_ [sala@cbs.nl](mailto:sala@cbs.nl)  
Enrico VROOMBOUT, De Nederlandsche Bank \_ [e.g.a.vroombout@dnb.nl](mailto:e.g.a.vroombout@dnb.nl)  
Frank OUDDEKEN (previous meetings) \_ [f.e.m.ouddeken@dnb.nl](mailto:f.e.m.ouddeken@dnb.nl)
  
- Andrea Alivernini – Ufficio Italiano dei Cambi \_ [alivernini@uic.it](mailto:alivernini@uic.it)  
Roberto TEDESCHI - Banca de Italia (previous meetings) \_ [tedeschi.roberto@insedia.interbusiness.it](mailto:tedeschi.roberto@insedia.interbusiness.it)
  
- Marta Augusta Louro DE ANDRADE VELOSO – Banco de Portugal \_ [mveloso@bportugal.pt](mailto:mveloso@bportugal.pt)
  
- Sabine GUSCHWA - Deutsche Bundesbank \_ [sabine.guschwa@bundesbank.de](mailto:sabine.guschwa@bundesbank.de)
  
- Jorma HILPINEN - Bank of Finland \_ [jorma.hilpinen@bof.fi](mailto:jorma.hilpinen@bof.fi)  
M. SORSA (first meeting) \_ [maria.sorsa@bof.fi](mailto:maria.sorsa@bof.fi)
  
- Ingvar Karlsson- Sveriges Riksbank \_ [ingvar.karlsson@riksbank.se](mailto:ingvar.karlsson@riksbank.se)  
K. LINDELL (first meeting) \_ [kajsa.lindell@riksbank.se](mailto:kajsa.lindell@riksbank.se)
  
- Alexandros MILIONIS – Bank of Greece \_ [AMilionis@bankofgreece.gr](mailto:AMilionis@bankofgreece.gr)  
Vera Rodis (previous meetings) \_ [vrodیس@bankofgreece.gr](mailto:vrodیس@bankofgreece.gr)
  
- Juan Manuel MONJAS LESAGA – Banco de España \_ [bal.pagos@bde.es](mailto:bal.pagos@bde.es)
  
- Michele MUREZ – Banque de France \_ [michele.murez@banque-france.fr](mailto:michele.murez@banque-france.fr)  
Françoise DRUMETZ (previous meetings) \_ [francoise.drumetz@banque-france.fr](mailto:francoise.drumetz@banque-france.fr)
  
- Nollaig Griffin – ONS attended the second meeting \_ [nollaig.griffin@ons.gsi.gov.uk](mailto:nollaig.griffin@ons.gsi.gov.uk)  
M. POWELL (first meeting) \_ [matthew.powell@ons.gov.uk](mailto:matthew.powell@ons.gov.uk)
  
- Diarmuid REIDY- Central Statistics Office \_ [Diarmuid.Reidy@cso.ie](mailto:Diarmuid.Reidy@cso.ie)
  
- Matthias Ludwig -EUROSTAT \_ [Matthias.LUDWIG@cec.eu.int](mailto:Matthias.LUDWIG@cec.eu.int)  
Stylianios Pantazidis (previous meetings) \_ [Stelios.pantazidis@cec.eu.int](mailto:Stelios.pantazidis@cec.eu.int)
  
- Rodrigo OLIVEIRA SOARES – ECB \_ [Rodrigo.oliveira@ecb.int](mailto:Rodrigo.oliveira@ecb.int)
- Jorge Diz Dias - ( previous meetings) \_ [Jorge.Diz\\_dias@ecb.int](mailto:Jorge.Diz_dias@ecb.int)
- Luca Buldorini (previous meetings)

## Annex 5: Methodological documentation for indicators

### A. Reliability/Stability

In the IMF's terminology, the study of revisions is normally referred as *reliability*, while some quality works at the European level refer to this as *stability*. The underlying concept is however the same and can be defined as referring 'to the closeness of the initial estimated value(s) to the subsequent estimated values. Assessing reliability involves comparing estimates over time. In other words, assessing reliability refers to revisions'<sup>2</sup>.

#### A.I. Simple measures of revisions

There is no doubt that revisions can always be measured. The number of subsequent assessments for any set of statistics depends on the revision policy applied by the statistical agency, which normally decides beforehand (sometimes in collaboration with the users) how many times and when the estimates should be revised and communicated to the public.

As an example, with reference to the generic variable or series X, the statistical agency can decide to publish it  $k$  times at predefined time lags  $\{l_1, l_2, \dots, l_k\}$ , where the time lag indicates the time elapsed between the reference period and the publication period (e.g. if a June publication refers to a revision of January data, the time lag is thus five months). Hence  $k$  different sets  $\{X(l_1), X(l_2), \dots, X(l_k)\}$  of the same variable will be available.

From the previous  $k$  sets of data, revisions can be easily derived, normally as the difference between two subsequent assessments. Therefore a revision variable or series can be defined as

$$R(l_i, l_j) = X(l_j) - X(l_i),$$

where  $i$  and  $j$  identify two specific time lags, with  $l_j > l_i$ .

#### A.II. Relative measures of revisions

The simple calculation of revisions expresses the changes in original units of the variable X and depends on its magnitude, often hampering comparability across time, across different variables and across the same variables in different countries. Therefore, it is often useful to provide a relative measure, which relates the revision to some dimensional measure of the variable.

---

<sup>2</sup> IMF working paper – Statistics Department: “Assessing accuracy and reliability: a note based on approaches used in National Accounts and Balance of payments Statistics”. Carson, Carol S. and Lucie Laliberté. February 2002.

## **Gross data**

In the case of gross data (data which express positive quantities), it is straightforward to devise a relative measure of revisions expressed in terms of percentage changes from the initial assessment with the formula  $[X(l_j) - X(l_i)] / X(l_i)$ , which is called the *percentage error*. In the usual case that X is a time series, an average can be taken across time, hence calculating a mean percentage error, with the formula  $\frac{1}{N} \sum_{t=1}^N \frac{X_t(l_j) - X_t(l_i)}{X_t(l_i)}$ , where  $i$  and  $j$  identify two specific time lags, with  $l_j > l_i$ , and  $t$  is a time indicator identifying the reference period of the series X.

As revisions can be positive or negative, it is usually appropriate to take them in absolute value, in order to avoid a situation whereby revisions of opposite sign cancel each other in the results of the indicator. The expression becomes therefore a *mean absolute percentage error (MAPE)*:

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{X_t(l_j) - X_t(l_i)}{X_t(l_i)} \right| \quad (1)$$

## **Net data**

In the case of net data, the revision data cannot be related to underlying quantities, and alternative dimensional measures of the variable X need to be used. A solution to this problem can be provided by any measure of the variability of the variable (series) X, which serves as a reference point for assessing the relative size of the revision. The relative error (relative revision) then becomes  $[X(l_j) - X(l_i)] / \text{var}[X(l_k)]$ , on which an average can also be taken across time. The TF-QA has termed this, given its similarities with the MAPE shown before, as the *mean absolute relative error (MARE)*. Its formula is

$$MARE = \frac{1}{N} \sum_{t=1}^N \left| \frac{X_t(l_j) - X_t(l_i)}{\text{var}[X(l_k)]} \right| \quad (2)$$

There are several ways of calculating the variability of X. The first issue to be decided is on which of the  $k$  different sets of X should the variability be calculated. The TF-QA suggests that it should be calculated on the last assessment  $X(l_k)$ , because this is potentially the most accurate. A second issue is what kind of indicator should be used to measure the variability. Here three possibilities are suggested:

1. *Standard deviation*. The standard deviation is the classic measure of variability of a series, its

formula is  $\sqrt{\frac{1}{N} \sum_{t=1}^N \left( X_t - \frac{1}{N} \sum_{t=1}^N X_t \right)^2}$ , and it represents the square root of the average of the quadratic distances from the mean.

2. *Average distance from the mean*<sup>3</sup>. Its formula is  $\frac{1}{N} \sum_{t=1}^N \left| X_t - \frac{1}{N} \sum_{t=1}^N X_t \right|$  and it represents the average of the distances from the mean in absolute value. It has the advantage of expressing the variability in original units (not distorted by the application of the squares).
3. *Median of distances from the median*. Its formula is  $\text{median}(|X_t - \text{median}(X_t)|)$ . It is similar to the formula in point 2, where the averages are replaced by medians owing to the fact that medians are extremely robust to the presence of outliers.

### **Treating outliers**

When outliers (i.e. exceptionally high revisions) need to be treated to reduce their influence on the results of the indicators, the MAPE and MARE equations can be modified by replacing the averages with medians. The MAPE and MARE are thus formulated as follows:

$$MAPE = \text{median} \left( \left| \frac{X_t(l_j) - X_t(l_i)}{X_t(l_i)} \right| \right)$$

$$MARE = \text{median} \left( \left| \frac{X_t(l_j) - X_t(l_i)}{\text{var}[X(l_k)]} \right| \right)$$

Some problems in the calculations may derive from the use of the median as the results only rely on one or two middle-ranked observations. Therefore, the indicators are resistant to outliers but extremely sensitive to the observations that define the median.

The replacement of the median with the trimmed mean<sup>4</sup> would be preferable, as it would still allow the treatment of outliers, thereby defining an appropriate cut-off percentage. As the cut-off percentage increases, fewer observations are considered in the calculation and the result becomes less sensitive to outliers, but more sensitive to middle-ranked observations. However, the TF-QA decided to apply a zero cut-off percentage, preferring to take into account the whole set of data. This therefore yields the average for these indicators.

### **Dispersion of the revisions**

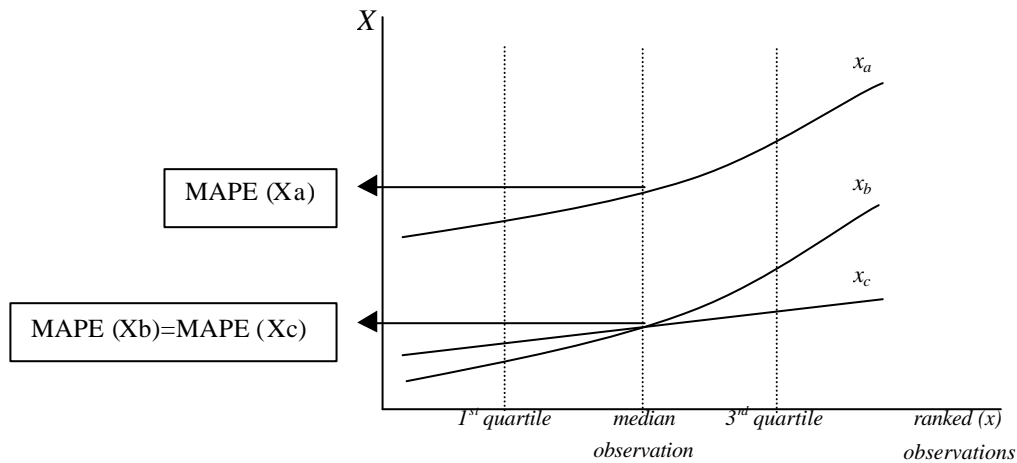
The value of the MAPE/MARE can usually clearly discriminate between two revision series, but in certain cases additional information, such as the dispersion of the revisions, is needed to fully discriminate between two revisions series. The next picture illustrates both cases for gross flow series, which are assessed using the MAPE indicator.

---

<sup>3</sup> The measure finally used in the calculations.

<sup>4</sup> The trimmed mean (?%) is the mean calculated after removing a ? percentage of the smallest and largest values of the series, so as to exclude the outliers.

1.



The picture above shows three lines representing the revision series  $x_a$ ,  $x_b$  and  $x_c$ . The vertical axis shows the value of the absolute relative revision ( $X$ ) for each series observation. On the horizontal axis, the observations have been ranked by the  $x$  values of each revision series. The revision series  $x_a$  and  $x_b$  are correctly discriminated by the MAPE value (which is the median value)<sup>5</sup>, i.e. series  $x_b$  will be considered more stable than  $x_a$ . Conversely, series  $x_b$  and  $x_c$  have the same MAPE value but the distribution of the revisions is rather different; up to the median value, values of series  $x_b$  are smaller, whereas after that point values of series  $x_b$  are much higher than  $x_c$  revisions. To avoid major revisions,  $x_c$  seems preferable.

Dispersion measures can be used to complement the first indicator so as to discriminate between series with the same MAPE/MARE but with different distributions. These dispersion measures will only be focused in the upper tail (largest revisions) e.g. the 3<sup>rd</sup> quartile, the 90<sup>th</sup> percentile and the maximum.

### Upward revisions

In principle, positive and negative revisions should be equally possible in a series. If the revisions are systematically positive or negative, it will be necessary to analyse the underlying reasons for this, e.g. a lack of coverage in early estimates, and there should be an attempt to correct this systematic bias. This simple indicator is the ratio of upward revisions and the number of observations ( $N$ ).

$$\text{upward revisions} = (\# \text{ upward revisions}) / N \quad (3)$$

---

<sup>5</sup> To simplify the example, normality is assumed.

## **Directional reliability**

To assess if the information contained in the earlier estimates has not been altered by the revisions, a 2 x 2 contingency table can be set up. In this contingency table the columns consist of positive and negative first difference in the early estimates ( $X_t(l_i)$ ), while the rows consist of positive and negative changes in the observed values ( $X_t(l_j)$ ).

*Contingency table for directional reliability*

	D $X(l_i) > 0$	D $X(l_i) \leq 0$	Subtotal
D $X(l_j) > 0$	n <sub>11</sub>	n <sub>12</sub>	n <sub>11</sub> +n <sub>12</sub>
D $X(l_j) \leq 0$	n <sub>21</sub>	n <sub>22</sub>	n <sub>21</sub> +n <sub>22</sub>
Subtotal	n <sub>11</sub> +n <sub>21</sub>	n <sub>12</sub> +n <sub>22</sub>	N

The directional reliability indicator (Q) is built as follows:

$$Q = \frac{n_{11} + n_{22}}{N} \quad (4)$$

This coefficient (Q) is equal to 1 when the early estimate changes and the observed values changes follow the same pattern ( $n_{11} + n_{22} = N$ ), while it is equal to 0 when there is a total dissociation ( $n_{11} + n_{22} = 0$ ).

The directional reliability indicator (Q) expresses the percentage of cases in which earlier and later assessments move in the same direction. High values are optimal in terms of increasing the reliability of the data.

### **A.III. More sophisticated approaches**

More complex measures are used to detect possible bias or persistent patterns in the revisions. Following the literature on forecast quality measures, the most appropriate one for the analysis of the revisions has been built based on mean square errors (MSE), as it can be decomposed into bias and variance.

The use of this kind of measure is based on the following interpretation of the revisions. The early assessment ( $X_t(l_i)$ ) is considered the best forecast of the series, estimated with the information available at that moment. The latest assessment ( $X_t(l_j)$ ) is assumed as the most accurate estimation and thus the closest to the real observation. Consequently, the revision is considered a forecast error, and the indicator based on MSE seeks to measure the quality of the forecast.

The indicator which was finally considered the most appropriate by the TF-QA is a ratio between two different mean square errors, making it a relative measure:

- the numerator uses the MSE applied to the difference between two assessments (revision measure):

$$MSE = \frac{1}{N} \sum_{t=1}^N (X_t(l_j) - X_t(l_i))^2$$

- the denominator uses the MSE applied to the difference between variable X and a reference value for X. The MSE for such a reference forecast would be:

$$MSE = \frac{1}{N} \sum_{t=1}^N (\Theta - X_t(l_j))^2$$

This indicator is a relative measure of the revisions expressed as a percentage of series volatility. Owing to its similarities with the previous indicators (MAPE and MARE), this indicator is thus called the *mean square relative error (MSRE)*, and is expressed as:

$$MSRE = \frac{\sum_{t=1}^N (X_t(l_j) - X_t(l_i))^2}{\sum_{t=1}^N (\Theta - X_t(l_j))^2}$$

In order to use the original units and additionally to build an indicator comparable with previous ones, the square root is applied to the ratio. The *root mean square relative error (RMSRE)* is expressed as

$$RMSRE = \sqrt{\frac{\sum_{t=1}^N (X_t(l_j) - X_t(l_i))^2}{\sum_{t=1}^N (\Theta - X_t(l_j))^2}} \quad (5)$$

where  $\Theta$  is the reference value for X.

RMSRE's value is 0 when the forecast is perfect, 1 if the forecast is only as accurate as the reference forecast, and greater than 1 when the forecast is less accurate than the reference forecast.

The proposed reference value for X is a constant forecast of the mean of the variable being forecast<sup>6</sup>, yielding the variance of X in the denominator. The advantage of using the average is that the MSE can be decomposed<sup>7</sup> into three components, with interesting applications for the study of the revisions:

$$MSE = (1) \text{ Bias component} + (2) \text{ Regression component} + (3) \text{ Disturbance component}$$

---

<sup>6</sup> Other measures of distribution location, like the median and the trimmed mean, were tested. Assuming that the b.o.p. financial net flows are stationary, the average was chosen owing to its simplicity, ease of interpretation, and because it makes the indicator's decomposition possible. Although not implemented by the TF-QA, when the series are not stationary the indicator can still be applied using the previous value of the series as the reference value, or using the first difference of the series.

<sup>7</sup> See Theil, H. (1966) and Murphy, A. H. (1988).

Applying this decomposition to the mean square relative error (MSRE) results in:

$$MSRE = \left[ \frac{\overline{X_{(l_j)}} - \overline{X_{(l_i)}}}{S_{X_{(l_j)}}} \right]^2 + \left[ r_{X_{(i)}, X_{(j)}} - \frac{S_{X_{(l_i)}}}{S_{X_{(l_j)}}} \right]^2 + [1 - (r_{X_{(i)}, X_{(j)}})^2]$$

where  $X_t(l_i)$  is the *forecast* (early assessment),  $X_t(l_j)$  is the observed event (latest assessment),  $r_{X_{(i)}, X_{(j)}}$  is the correlation between the two series,  $s_{X_{(i)}}$  and  $s_{X_{(j)}}$  are the standard deviations and  $\overline{X_{(l_i)}}$  and  $\overline{X_{(l_j)}}$  the means of  $X_t(l_i)$  and  $X_t(l_j)$ , respectively.

These three components can also be presented as proportions of the RMSRE, and their addition will equal 1.

$$1 = \frac{RMSRE^2}{RMSRE^2} = \frac{\left[ \frac{\overline{X_{(l_j)}} - \overline{X_{(l_i)}}}{S_{X_{(l_j)}}} \right]^2}{RMSRE^2} + \frac{\left[ r_{X_{(i)}, X_{(j)}} - \frac{S_{X_{(l_i)}}}{S_{X_{(l_j)}}} \right]^2}{RMSRE^2} + \frac{[1 - (r_{X_{(i)}, X_{(j)}})^2]}{RMSRE^2}$$

- 1) The *unconditional or bias component* is an indication of systematic error (revision), since it measures the extent to which the average values of the early and later assessment series deviate from each other. The revisions can be considered biased if the mean revision is significantly different from zero<sup>8</sup>.
- 2) The *conditional or regression component* is another systematic component which reflects whether the overall pattern of the series with the early estimates was close to that of the series with the later estimates. If the forecast correctly reflects the pattern/variability of the later estimate series, the correlation between both series will be quite high and the component close to zero.
- 3) The *unsystematic or disturbance component* is the variance of the residuals obtained by regressing the early estimates data on the later estimates. It can be assumed to have a random nature and without any predictable pattern<sup>9</sup>.

#### General indicator

The two indicators explained for net flows, (2) MARE and (3) RMSRE can be expressed in a general formula as follows:

$$MRE(r, k) = \left[ \frac{\sum_{t=1}^N |X_t(l_j) - X_t(l_i)|^r}{\sum_{t=1}^N |\Theta_t - X_t(l_j)|^r} \right]^{\frac{1}{r}}$$

<sup>8</sup> Normality is assumed for revisions in order to apply the t test.

<sup>9</sup> This indicator only accounts for linear relationships. The unsystematic part could still have non-linear patterns within it.

where  $\rho$  is the power parameter ( e.g.  $\rho=1$  for MARE and  $\rho=2$  for RMSRE) The sum function can be a trimmed one with a cut-off percentage (?) previously defined. The ? parameter can be used to exclude an increasing number of observations (making the measure more robust) and, to the limit, it can approximate the median. However, the TF-QA decided to apply a zero cut-off percentage, preferring to take into account the whole set of data.

The  $r$  value changes the weight of the argument in the sum. The  $\Theta_j$  can be any forecast that one wishes to test against the early assessments. For instance, it could be the forecast result of a benchmark model used to evaluate the time series. As explained before, the use of the average is very useful from a analytical point of view, but in the case of non-stationary series, its average and the MSE decomposition become meaningless, making it more convenient to use a different reference forecast.

#### **A.IV. Qualitative information about revisions**

To ensure comparability over time and across reporting areas, the series must not have any breaks in methodology and/or compilation procedures. To assess this degree of comparability, the following qualitative question about revisions should be introduced and explained in any quality assessment:

*Was there any significant change in concepts or methodology for the period under evaluation? (Yes/No. Explain).*

#### **B. SERVICEABILITY/CONSISTENCY**

In the IMF's DQAF, *consistency* is defined as (i) over time; (ii) between data collected at different frequencies; (iii) internationally; (iv) across variables, either vertically (across transactions), horizontally (across institutional sectors)<sup>10</sup>, and/or between flows and stocks. The TF-QA decided to follow the IMF's DQAF for b.o.p. 2001 principles by mostly concentrating on this element and categorising it into the following sub-categories:

- Internal consistency (e.g. within the integrated statistics (b.o.p./i.i.p. or national accounts))
- Consistency over time (e.g. in the case of methodological or institutional changes, such as enlargement, historical data are reconstructed as far back as is reasonable)
- External consistency (between different sources of data and/or different statistical frameworks, including mirror statistics - international statistics should be comparable even when compiled by different institutions or by different units of the same institution).

---

<sup>10</sup> B.o.p. statistics can be defined as horizontally consistent when the differences between identical data items coming from diverse sources are minor or even non-existent, i.e. different measurements of the same item related to the same sector do not result in unreasonably different data.

## B.I. Internal consistency (consistency within the dataset)

The b.o.p. statistics have a *natural* indicator for internal consistency: the net errors and omissions series (EO). The principle of double-entry bookkeeping used in b.o.p. implies that the sum of all international transactions should be equal to zero. Nevertheless, "*Data for balance of payments estimates often are derived independently from different sources; as a result, there may be a summary net credit or net debit (i.e., net errors and omissions in the accounts). A separate entry, equal to that amount with the sign reversed, is then made to balance the accounts. Because inaccurate or missing estimates may be offsetting, the size of the net residual cannot be taken as an indicator of the relative accuracy of the balance of payments statement. Nonetheless, a large, persistent residual that is not reversed should cause concern. Such a residual impedes analysis or interpretation of estimates and diminishes the credibility of both. A large net residual may also have implications for interpretation of the investment position statement*" (IMF, BPM5, 1993, p.17).

According to the IMF's DQAF for b.o.p. 2001, internal consistency implies checking that "over the long run [the] errors and omissions item (1) has not been large and (2) has been stable over time".

A measure of the size can be provided by the *average absolute error of the errors and omissions (AAE)*:

$$AAE(EO) = \frac{\sum_{i=t-a}^t |EO_i|}{a+1} \quad (6)$$

where  $t$  is the period for the last observation and  $a$  is the time frame for the analysis.

As with the MSRE index of the revisions, an alternative (additional) measure of the size can be provided by the *mean square error of the net errors and omissions (MSE(EO))*.

$$MSE(EO) = \sum_{i=t-a}^t (EO_i)^2 / (a+1)$$

The advantage of this measure is that it can be decomposed into its bias and variance components<sup>11</sup>:

$MSE(EO) = (1) \text{ bias component} + (2) \text{ variance component}$

$$MSE(EO) = \left[ \frac{\sum_{i=t-a}^t (EO_i)}{a+1} \right]^2 + \frac{\sum_{i=t-a}^t (EO_i - \overline{EO})^2}{a+1}$$

where  $\overline{EO}$  is the average of EO between  $t-a$  and  $t$ .

---

<sup>11</sup> Following the most simple MSE decomposition. See "Elements of Forecasting", Diebold, Francis X. 2001

A possible time frame for its calculation could be three years for long term and one year for short term; the TF-QA agreed to use the former. To assess the recent dynamics of EO, it is possible to use the previous yearly time frame.

In order to use the original units and to ensure comparable and harmonised indicators across all elements, the square root is applied to the ratio resulting in the *root mean square error of the net errors and omissions (RMSE(EO))*, which can be expressed as

$$RMSE(EO) = \sqrt{\sum_{i=t-a}^t (EO_i)^2 / (a+1)} \quad (7)$$

Following the methodology presented in Section A.III regarding revision studies, the components of the MSE(EO) can also be presented as proportions of the RMSE(EO):

$$1 = \frac{RMSE(EO)^2}{RMSE(EO)^2} = \frac{\left[ \frac{\sum_{i=t-a}^t (EO_i)}{a+1} \right]^2}{RMSE(EO)^2} + \frac{\sum_{i=t-a}^t (EO_i - \overline{EO})^2}{(a+1) RMSE(EO)^2}$$

Further to the previous indicator, the TF-QA proposes to use the number of positive EO during the period under study divided by the number of observations so as to assess the relative frequency of positive EO:

$$CP(EO) = \frac{Count(EO_t > 0)}{N} \quad (8)$$

where N is the time frame data.

As presented in the report, to make these indicators comparable, the TF-QA agreed that the series used in the key indicators should be scaled by the total gross flows (half-sum of debits and credits) in the current account. Other scales for the errors and omissions *GDP + imports* or *i.i.p. total financial assets* will be used in the supporting indicators.

## **BII. External consistency (consistency with other data sources and/or statistical framework)**

Although minor discrepancies arising from methodological differences can be present in two sets of data stemming from different sources of data and/or different statistical frameworks<sup>12</sup>, the comparison can still provide a useful measure of consistency and contribute to the overall quality increase.

---

<sup>12</sup> E.g., the comparison between the euro area goods item (b.o.p.) and Eurostat's external trade data, or the comparison between the b.o.p. flows of the monetary financial institution (MFI) sector and flows derived from the consolidated MFI balance sheet from money and banking statistics.

### Gross flows or stock data

The ECB proposal<sup>13</sup> to the Task Force on Quality<sup>14</sup> for measuring the consistency between b.o.p. and international trade statistics (ITS) was as follows:

$$C_i = 100 \times \frac{|x_i - y_i|}{(x_i + y_i)/2}$$

where  $x_i$  and  $y_i$  are the two series under comparison.

The indicator presented above managed to capture the magnitude of the discrepancies (since we are not interested in their direction, the differences are neutralised) but failed to do the following:

- it does not take into account the autocorrelation of the discrepancies; and
- it does not capture the dispersion of discrepancies.

Furthermore, this indicator can be continuously biased, thus preventing direct comparability between items and/or reporting areas, since its construction is based on two different sources with potentially dissimilar conceptual/compilation methods. As an example, on one hand the imports of goods are measured on a c.i.f. basis for ITS and on an f.o.b. basis for b.o.p.; on the other hand, however, exports of both are measured on an f.o.b. basis.

To deal with these shortcomings, this indicator was then refined as:

$$C_{t,a} = \frac{\sum_{i=t-a}^t |\Delta x_i - \Delta y_i|}{\sum_{i=t-a}^t (x_{i-1} + y_{i-1})/2} \quad (9)$$

where  $\Delta$  stands for series first difference,  $t$  is the period for the last observation and  $a$  is the time frame for the analysis. The values for  $C_{t,a}$  range from zero (a perfect match) to plus infinity (no match possible). Knowing the different methodology behind the construction of these statistics, the perfect match should not be considered an optimal result, if not as a lack of necessary adjustments between both series.

The proposed time frame is three years, as already mentioned under internal consistency.

As the revisions section suggests, to discriminate between series with the same indicator but different distributions, dispersion measures can be used to complement the proposed indicator. The dispersion of  $C_{t,a}$  could be evaluated by taking the difference between the 3<sup>rd</sup> and the 1<sup>st</sup> quartiles of the three-year time frame, i.e. the *interquartile dispersion* expressed as the difference of:

$$\text{Interquartile dispersion} = 3\text{rd quartile} - 1\text{st quartile}$$

Besides ITS, these indicators can be used with the rest of the world account in national accounts (one can test for both goods and services on a quarterly basis).

---

<sup>13</sup> Based on the paper “Some elements of a quality framework for CMFB statistics”, Keuning, S and S. Algera.

<sup>14</sup> See Annex 9 of the Preliminary Report: “Methodology proposed to measure consistency of balance of payments series”.

## Net flows

In the case of comparison between b.o.p. flows from the MFI sector (e.g. other investment plus direct investment -other capital and the deposits/loans derived from the consolidated balance sheet of MFIs), the two time series appear to vary considerably in terms of magnitude. These differences can be attributed to a variety of factors: dissimilar timeliness in terms of recording and reporting, different revision policies and different valuation methods.

Moreover, the initial series seem to move somewhat randomly. Consequently, the consistency indicator should take into account both the magnitude of the differences and the volatility of the original series.

The ECB's proposed indicator is:

$$C = \frac{\bar{x} - \bar{y}}{Vy} = \frac{\text{trimmean}(|x_i - y_i|)}{\text{trimmean}(|y_i - \bar{y}|)}$$

where  $x$  is the b.o.p. data series,  $\bar{x}$  is its trimmed mean<sup>15</sup>,  $y$  is the series under comparison,  $\bar{y}$  its trimmed mean and  $Vy$  a measure of variability of  $y$ . The trimmed mean cut-off percentage used should be around 5%.

This indicator is very similar to the MARE indicator, which is, as mentioned before, a stability measure. The questions posed on the C indicator are directly applicable here too; it was therefore considered efficient that the same indicators proposed for assessing stability would be used for assessing consistency between comparable net flows. These indicators, when properly adapted to the series to be analysed, can be distinguished as follows:

- the *mean absolute relative error (MARE)*, and
- the *root mean square relative error (RMSRE)* and its decomposition into *bias (unconditional)*, *regression (conditional)* and *disturbance (unsystematic) components*.

The first one can be defined as:

$$MARE = \frac{1}{N} \sum_{t=1}^N \frac{|X_t - Y_t|}{\text{var}[X_t]} \quad (10)$$

where  $X_t$  is the b.o.p. item,  $Y_t$  the external set of data to compare, and  $N$  the time frame data.

The latter item was constructed and proposed in order to monitor external consistency respecting the methodology followed in the revisions studies and in terms of internal consistency:

---

<sup>15</sup> The trim mean (%) is obtained by deleting a % percentage of the smaller and larger values from a data set, so as to exclude the outliers, and then computing the mean of the remaining values. The actual formula to calculate the observations to be excluded each side is  $\text{Int}[(1+(n-1)*\%)/2]$ . The function  $\text{Int}(\cdot)$  truncates its argument to the lowest integer.

$$RMSRE = \sqrt{\frac{\sum_{t=1}^N (X_t - Y_t)^2}{\sum_{t=1}^N (\Theta - X_t)^2}} \quad (11)$$

where  $Y_t$  is the external set of data to compare,  $X_t$  the b.o.p item,  $N$  the time frame data, and  $T$  the average for  $X$ . As shown in Section A.III., the MSE decomposition can be applied to this indicator, representing the bias, regression and disturbance components as proportions of the *root mean square relative error (RMSRE)*:

$$1 = \frac{RMSRE^2}{RMSRE^2} = \frac{\left[\frac{\bar{X} - \bar{Y}}{S_X}\right]^2}{RMSRE^2} + \frac{\left[r_{XY} - \frac{S_Y}{S_X}\right]^2}{RMSRE^2} + \frac{[1 - r_{XY}^2]}{RMSRE^2}$$

where  $\bar{X}$  and  $\bar{Y}$  are the average of  $X_t$  and  $Y_t$  respectively,  $S_X$  and  $S_Y$  the standard deviations of  $X_t$  and  $Y_t$  accordingly, and  $r_{XY}$  the correlation coefficient.

### Mirror statistics

Although mirror statistics are being tackled by the Task Force on Asymmetries, Eurostat has proposed some potential indicators that might be calculated in a centralised environment (the whole set of data needed for their calculation is not available for the countries themselves).

### Indication of asymmetries between countries

Mirror statistics are a way of assessing the consistency and comparability of statistics in some domains. They can be obtained when different compilers of b.o.p. statistics are measuring similar statistical characteristics, and are common practice in the measurement of flows (such as the current account, goods, services, transport, travel etc.). The external consistency of a source can be indicated by looking at the size of the discrepancies. Data typically consist of two matrices with inbound flows in one and outbound flows in the other. Absolute differences of inbound and outbound flows for a pair of countries can be then summed up for each country, yielding an indicator based on discrepancies. This value may be interpreted as an indication of this country's comparability with the rest. Asymmetry analyses arise because data can be looked at from the perspective of either of the countries involved. For example, country Y's estimate of its exports to country X should be the same as country X's estimates of its imports from country Y, and vice versa. These types of checks are known as mirror statistics, while the divergences between mirror statistics are termed asymmetries. Comparisons of this type offer helpful cross-checks on b.o.p. statistics data.

The proposed indicator is defined in a two-step approach as:

$$(1) C_{ij} - D_{ji}$$

$$(2) (C_{ij} - D_{ji}) / (C_{ij} + D_{ji}) / 2$$

The indicator for the measurement of asymmetries for aggregates is expressed by:

$$(1) S / (C_{ij} - D_{ji}) /$$

$$(2) (S / (C_{ij} - D_{ji})) / S (C_{ij} + D_{ji}) / 2$$

where  $C_{ij}$  represents the credits of country  $i$  with country  $j$ , and  $D_{ji}$  defines the debits of country  $j$  with country  $i$ <sup>16</sup>. Country  $j$  is here one of the EU/euro area countries and does not equal country  $i$ . To rectify the misleading nature of absolute figures, term 2 is introduced. To compute magnitudes, the difference is divided by the sum of the two observations, and finally an average is calculated.

The indicator presents magnitudes and not absolute figures. The formula says that, when not null, the intra-EU/euro area countries balance is equivalent to the intra-EU/euro area countries asymmetry, and is thus a measure of the reliability and consistency of the data.

### Directional consistency

Last, but not less important, is the consistency about the information provided by the two sources, i.e. if the signs of the first differences coincide in both sources. A measure that can provide this consistency check is a 2 x 2 contingency table. In this contingency table the columns are the positive and negative changes for b.o.p. series (?), and the rows are the positive and negative changes for the other series (y).

*Contingency table for directional consistency*

	D X > 0	D ? £ 0	Subtotal
D ? > 0	$n_{11}$	$n_{12}$	$n_{11} + n_{12}$
D ? £ 0	$n_{21}$	$n_{22}$	$n_{21} + n_{22}$
Subtotal	$n_{11} + n_{21}$	$n_{12} + n_{22}$	N

where  $n_{11}$  is the number of cases when both ? ?<sub>t</sub> and ? Y<sub>t</sub> are positive, and  $n_{22}$  when they are negative.

For maximum directional consistency, one should expect a high sum for the leading diagonal ( $n_{11} + n_{22}$ ). The *directional consistency indicator* ( $Q_C$ ) is then built as follows:

$$Q_C = \frac{n_{11} + n_{22}}{N} \quad (12)$$

This coefficient ( $Q_C$ ) is equal to 1 when the changes in the b.o.p. series and the changes in the comparable series follow the same pattern ( $n_{11} + n_{22} = N$ ); it is equal to 0 when there is a total dissociation ( $n_{11} + n_{22} = 0$ ).

---

<sup>16</sup> If the compiler is interested in the direction of the differences, then he/she could use the actual values and not the absolute ones.

The directional consistency indicator ( $Q_C$ ) expresses the percentage of cases in which the first differences of the two time series under comparison move in the same direction. High values are optimal in terms of increasing the consistency of the data.

### **B.III. Qualitative information about consistency**

These indicators cannot deal with different revision policies. Nevertheless, they remain valid indicators if the revisions do not substantially alter the information provided by the first estimates (evaluated by the *Reliability* element for the b.o.p. source). A qualitative question could be attached to ascertain the frequency of revision practices for both sources:

*What is the b.o.p. revision frequency when compared to the other sources?  
(Higher/Equal/Lower)*