

EUROPEAN CENTRAL BANK

EUROPEAN COMMISSION

EUROSTAT

Directorate C: Economic statistics and
economic and monetary convergence

Unit C-4: Balance of payments

EUROPEAN CENTRAL BANK

DIRECTORATE GENERAL STATISTICS

External Statistics Division

TASK FORCE ON QUALITY

Report on the quality assessment of balance of payments and international investment position statistics

Version 3.8

June 2004

Table of Contents

Executive summary	3
A. FINDINGS.....	3
B. RECOMMENDATIONS FOR A POSSIBLE ACTION PLAN.....	6
1. Introduction	8
2. Operational assessment of the main dimensions/elements	9
2.1. PREREQUISITES OF QUALITY.....	10
2.2. METHODOLOGICAL SOUNDNESS.....	11
2.3. INTEGRITY.....	11
2.4. ACCURACY AND RELIABILITY	12
2.4.1 Accuracy.....	12
2.4.2 Reliability.....	13
2.4.3 Plausibility.....	17
2.5 SERVICEABILITY	17
2.5.1 Relevance.....	18
2.5.2 Timeliness	18
2.5.3 Consistency.....	18
2.6. ACCESSIBILITY.....	21
2.7. THE INTERRELATION BETWEEN QUALITY DIMENSIONS AND ELEMENTS.....	21
3. Findings	22
3.1 INITIAL FINDINGS	22
3.2 SUBSEQUENT FINDINGS	25
3.2.1. Pilot exercise and feasibility study (autumn 2002).....	25
3.2.2 Empirical exercise (February-April 2003).....	26
3.2.3 Merit and cost analysis (May 2003).....	27
3.2.4 Communication test (Second half of 2003).....	27
4. Who should assess/monitor the data quality?	29
5. Communication issues	30
6. List of annexes and documents attached	32
Glossary	33
List of reference documents on quality	37

Executive summary

A. Findings

This report summarises the work completed so far by the joint ECB (DG-Statistics – DG-S)/Commission (Eurostat) Task Force on (Output) Quality (TF-QA), mandated by the Committee for Monetary, Financial and Balance of Payments Statistics (CMFB) during its meeting in January 2002, with regard to balance of payments (b.o.p.) and international investment position (i.i.p.) statistics. The TF-QA has studied the definitions of quality as described in the IMF Data Quality Assessment Framework (DQAF), and identified the dimensions/elements that are most relevant for b.o.p./i.i.p. statistics. Where appropriate, the TF-QA discussed quality indicators for these dimensions or elements which could be operationalised for b.o.p./i.i.p. outputs on the European level.

Quality is a subjective concept and encompasses all aspects of how well statistics meet users' needs. When applied to statistics as a public good, quality is linked to users' expectations about the information content of the disseminated data. Users of statistics, policymakers, market players, academics or the general public, all expect the data on which decisions are based to be sufficiently reliable. At the same time, data would not be useful if not delivered in a timely manner. Users also accept that the statistical data will be revised, notably when further breakdowns are made available, with the aim of increasing their accuracy and thereby fostering more in-depth analyses, as well as reducing noise and uncertainties in forecasting exercises.

In parallel, statisticians are expected to compile high quality statistical data in a timely manner. Therefore, they need to evaluate and improve various aspects of the data production process as well as to maintain high quality data under constraints, in particular scarce resources.

Against this background, it is not easy to measure quality. In practice, the approach selected by the TF-QA was to define a distance between given observations and reference data. For example, assuming that the most up-to-date data are the most reliable, then these are used as a reference to measure the deviation from previous assessments of the same phenomenon observed. Indicators were set up to assess the size, recurrence, bias, etc. of these deviations. For some indicators, for example on errors and omissions, it is also important to define a reference against which comparisons can be made more meaningful across countries/economic unions.

The adopted approach comprised several steps, starting from the selection of those dimensions/elements contained in the DQAF that were seen as the most relevant ones for quantitative measurement, and then moving on to the identification of possible measures, the testing and analysis of the indicators and their outcome, and the fine-tuning of various parameters/reference values, where appropriate.

The quality dimensions that could, directly or indirectly (through their elements), be measured and assessed by the TF-QA are methodological soundness, accuracy and reliability, and serviceability. These aspects of data quality could be improved, taking as a reference point a close monitoring of meaningful quantitative indicators. Integrity and accessibility, although also considered to be important, are in general well covered by the EU countries in their b.o.p./i.i.p. statistics. There are also either fewer quantitative measures for these two dimensions, or they are less appropriate, so that it therefore seems more effective to rely on periodic qualitative assessments.

A pilot exercise carried out by the ECB together with four Member States (Germany, France, Italy and the United Kingdom) in autumn 2002 suggested (i) that most of the tested indicators were relevant but may need further refinement; (ii) that appropriate parameters should be selected, tested and implemented for the resulting figures to ensure comparability across countries; and (iii) that some indicators were difficult to interpret. In the light of this pilot exercise, coupled with a preliminary feasibility assessment, the CMFB agreed in January 2003 that all Member States, Eurostat and DG-S should perform a three-month empirical exercise (lasting from February to April 2003) to further fine-tune the indicators, and to identify “key” indicators which could be relevant for public dissemination in due time, as well as “supporting” indicators which would permit internal analysis of the data quality. In actual fact, twelve Member States (Belgium, Germany, Spain, France, Ireland, Italy, Luxembourg, Netherlands, Portugal, Sweden, Finland and the United Kingdom) and the ECB participated in the empirical exercise. During this period, the selected indicators were calculated and reported to DG-S, using different parameters and time frames.

DG-S has created a “quality tool” (or “QaTool”, in short) to assist Member States and European institutions in calculating these indicators; this tool can be used for internal purposes and/or for reporting to European institutions. To initiate the empirical exercise, a workshop was organised in late January 2003 by DG-S to acquaint representatives of all Member States with the objectives of the empirical exercise, the definitions and calculation methods of the indicators, and to explain and describe the quality tool. The TF-QA members welcomed the tool as being a very helpful instrument for this exercise and encouraged its further development. A new version of this tool was disseminated by DG-S to Member States in autumn 2003.

At the end of this experimental period (in early May 2003), a new feasibility questionnaire was distributed to Member States to gather information on the findings and provide a cost-benefit evaluation by compilers for each indicator. Based on the replies from 13 countries (the 12 involved in the empirical exercise plus Austria) and DG-S, the TF-QA set up a list of key indicators for the assessment of b.o.p. data quality.

Building on these findings and on the list of indicators, the TF-QA has further elaborated to operationalise the quality indicators.

Foremost, any quality assessment should be considered as well-focused and relevant by users. As an example of how to communicate the results of the indicators to the users, a mock quality report was drafted in September 2003, focusing on data for the euro area. This mock-up could also assist Member States when designing their own quality reports, which should be tailored to their own statistics. The euro area mock report and the technical documentation about the indicators was disseminated in December 2003 and early 2004 to advanced users in a number of institutions, such as the ECB, NCBs or the European Commission, with the aim of testing adequacy and relevance, and of obtaining feedback. Comments were received from different departments within the ECB, as well as from the Deutsche Bundesbank, which will be reflected in the euro area annual reports on quality.

The indicators will be calculated and analysed by compilers twice a year (every May and November). The key indicators for the euro area (and possibly the EU) will be published at the end of the year in annual quality reports, together with a statement on the progress made and achievements so far, as well as the necessary caveats and information which will help in the interpretation of these indicators. In particular, special attention should be paid to a balanced consideration of both the quantitative as well as the qualitative indicators, noting where appropriate the trade-offs between different dimensions and elements of quality (e.g. timeliness vs. accuracy).

Benchmarks for the indicators should reflect the complex interrelation between most of the dimensions and should only be established once sufficient experience has been gained and the practice of publishing such indicators has spread over comparable economic areas and across sets of statistics. Before these prerequisites are met, benchmarks may not be easy to interpret by users or may lead to wrong incentives by compilers. Hence, the TF-QA concluded that, in the nearer future, no benchmark will be established. However, whether appropriate progress in quality levels in critical areas has been made will be closely monitored, both by Member States as regards their national data and their contribution to the euro area/EU aggregate, and by European institutions as regards the aggregates.

Once the indicators are in regular use (from 2004 onwards), DG-S and Eurostat will liaise with the Member States via the Working Groups and the Committee foreseen in the draft Regulation so as (i) to maintain (and update) the indicators, and (ii) to set up procedures for the assessment of the actual quality of regular (annual) reports and for their preparation.

Considering that measuring data quality is a complex task and given the limited experience gathered so far, the findings and recommendations presented in this report by the TF-QA should be considered as a base and contribution to the work in progress for the development and use of adequate indicators for assessing and measuring different dimensions/elements of (b.o.p. and i.i.p.) statistics.

As a follow up of the work done by the TF-QA, a one day workshop was organised in January 2004 by DG-S to acquaint representatives of acceding countries with the objectives and findings of the TF-QA, the definitions and calculation methods of the indicators, and to explain and describe the quality tool.

This report of the Task Force on Quality is available on the CMFB's website (www.cmfb.org).

B. Recommendations for a possible action plan

The Task Force recommends that the following issues be included in an action plan:

(a) Start calculating on a regular basis the key indicators¹ identified for the assessment of output aspects of b.o.p./i.i.p. data quality (stability² and consistency) for Member States (MSs), DG-S and Eurostat. These key indicators are included in the annual report on the European aggregates. As set out in the mandate, the operational indicators are intended to complement a qualitative assessment of the data, not as a substitute for it.

TF-QA: Agreement on key indicators				
Element / Indicator / Focal issue	Indicator's name	Formula	Variables / Parameters	B.o.p. items to be applied
ACCURACY & RELIABILITY DIMENSION				
Stability (Revision studies)	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 $	$X(l_i)$ - early estimation of X $X(l_j)$ - later estimation of X N - time frame data Θ - average for $X(l_j)$	Total current account, Goods, Services and Income: Debits and Credits
	Root Mean Square Relative Error (RMSRE)	$RMSRE = \sqrt{\frac{\sum_{t=1}^N (X_t(l_j) - X_t(l_i))^2}{\sum_{t=1}^N (\Theta - X_t(l_j))^2}}$		Direct Investment, Portfolio Investment, Other Investment: Assets, Liabilities, Net assets Reserve Assets Errors and omissions
SERVICEABILITY DIMENSION				
Internal consistency	Root Mean Square Errors (EO)	$RMSE(EO) = \sqrt{\frac{\sum_{i=t-a}^t (EO)^2}{a+1}}$	EO - Errors and omissions a - time-frame	Errors and omissions corrected by the average of the addition of the current account subitems (CR+DB)

(b) Start using and assess the additional information obtained from a qualitative indicator on methodological soundness³, and two quantitative indicators on external consistency, one between the financial account of b.o.p. and MFI balance sheet statistics, and a second one between b.o.p. and international trade statistics.

¹ See Annex 3 for more details of the recommended key indicators as well as the potential ones, in addition to the RMSRE and RMSE decompositions.

² *Revision studies* in DQAF terminology.

³ See Annex 4.

TF-QA: Potential key indicators				
Element / Indicator / Focal issue	Indicator's name	Formula	Variables / Parameters	B.o.p. items vs. other statistics
SERVICEABILITY DIMENSION				
External consistency Net flows	Root Mean Square Relative Error (RMSRE)	$RMSRE = \sqrt{\frac{\sum_{i=1}^N (X_i - Y_i)^2}{\sum_{i=1}^N (\Theta - X_i)^2}}$	<i>Y_t - external set of data to compare</i> <i>X_t - bop item</i> <i>N - time frame data</i> <i>Θ - average for Y</i>	<i>Other investment net and direct investment other capital vs. deposits and loans (MBS)</i>
External consistency Gross flows	C _{t,a}	$C_{t,a} = \frac{\sum_{i=t-a}^t \Delta x_i - \Delta y_i }{\sum_{i=t-a}^t (x_{i-1} + y_{i-1}) / 2}$		<i>Goods credits and debits vs ITS</i> <hr/> <i>Goods+Services vs. National accounts</i>

(c) The indicators should be used for the euro area/EU aggregate; their compilation and dissemination will be ensured by DG-S and Eurostat. Member States will retain responsibility for disclosing quality reports for their national b.o.p./i.i.p. data (and contributions to the aggregate). The Eurostat B.o.p. WG (or alternatively the future “B.o.p. Committee” to be installed by the new Regulation) and/or the WG-BP&ER will undertake reviews of the overall quality process and periodically report to the CMFB and STC on the relevance and accuracy of the quality indicators. A first review is planned by end-2005, and will additionally seek convergence with the findings and conclusions of the TF-QA on national accounts data.

(d) Statistics compilers may monitor data quality, also as the best way to raise awareness. Complementary procedures, involving compilers and users, may eventually be developed to monitor the integrity of the procedures and soundness of the underlying methodology for the indicators, consider the outcome, and provide guidance where appropriate.

(e) The following technical recommendations should be applied in Member States, DG-S and Eurostat:

- the indicators for the aggregates and country contributions will be calculated twice a year (May and November); these dates are related to the common revision practice of DG-S and Eurostat. The calculations will be based on monthly and/or quarterly data, using the last 36 observations (three years) for the monthly⁴ data⁵; and a longer period for the quarterly series, so as to ensure that sufficient statistical observations have been considered;
- the necessary data, in particular the first release of the data, will be kept available in the databases of each compiler to enable the calculation of the “stability indicators”.
- in the short and medium term, the use of the quality tool will reduce the cost and burden of calculating the indicators (a new version has been delivered in October 2003).

(f) Based on the mock up for annual quality reports commented by advanced users, the first publication by the ECB of a euro area quality report is foreseen in early 2005.

⁴ Where monthly data are published, users may wish to know the reliability of the first assessment of the data.

⁵ For the stability indicators, the time frame to be used in the calculations will be the previous three full calendar years.

1. Introduction

The quality of statistics encompasses all aspects of how well statistics meet users' needs as well as their expectations with regard to the information content of the disseminated data.

Users of statistics, such as central bank policymakers, governments and international organisations, as well as investment fund managers, financial market rating agencies and academics, all expect the data on which decisions are based to be sufficiently reliable. At the same time, the data would not be useful if not delivered in a timely manner. One of the clear trade-offs when assessing data quality is between accuracy and reliability on one hand, and timeliness on the other.

Users also expect that the statistical data will be revised so as to increase their accuracy. In particular, researchers and economists need to know the probability, direction and magnitude of any subsequent revisions in order to enhance their analyses and forecasting exercises.

In parallel, statisticians need to evaluate and improve various aspects of the data production process as well as to maintain high quality data under constraints, in particular scarce resources.

The fourth progress report on the EMU action plan, which was endorsed by the ECOFIN Council in November 2001, invited the SPC, in close co-operation with the CMFB, to make proposals enabling the various quality dimensions to be operationally assessed. To deal with this request, the CMFB approved in its January 2002 meeting the following: (1) the installation of a joint ECB/Commission (Eurostat) Task Force on (Output) Quality (TF-QA), dealing with balance of payments (b.o.p.) and quarterly national accounts statistics, and (2) the mandate for this task force⁶.

The IMF Data Quality Assessment Framework (DQAF), complemented by some quality assessment reports and studies⁷ produced by DG-S and Eurostat, was used as the reference framework for the TF-QA's work. This framework is based on (a) a common understanding of the quality dimensions and elements, and a shared approach to assessing and measuring quality, allowing the international comparability of data and metadata; and (b) a common vision as regards the priorities to be given to quality standards, especially taking into account the possible trade-offs. Furthermore, the use of this complemented DQAF would foster coherence in international organisations' requirements from MSs.

The TF-QA took stock of the concepts and definitions in the DQAF, complemented as mentioned above, with the aim of identifying and assessing, within the short time limits, a set of operational

⁶ See Annex 1.

⁷ See reference documents.

indicators on output quality that could be applied both to the euro area/EU aggregates and to contributory b.o.p./i.i.p. data from MSs.

A European Parliament and Council regulation on b.o.p. statistics that includes provisions on data quality is in preparation. The ECB has adopted a Guideline (No. ECB/2003/7) which also sets out provisions in this field. One objective of the work undertaken by the TF-QA is to identify key quality elements as an input for quality reports.

In its July 2002 meeting, the CMFB welcomed the TF-QA preliminary report on the quality of b.o.p. statistics, which was also used as input for the fifth progress report on the EMU action plan in autumn 2002. The report outlined some proposals for monitoring and communicating the operationally assessed (output) quality dimensions and the corresponding indicators. Finally, the CMFB mandated the TF-QA (1) to further test and fine-tune the proposed indicators in autumn 2002 in the so-called “pilot exercise”; (2) to assess the technical feasibility and related cost of developing and calculating these indicators with all MSs, if possible; and (3) to further elaborate on communication issues. In December 2002, the TF-QA was invited to make some more concrete proposals on operational measures of output quality at the June 2003 CMFB meeting, based on the results of a three-month empirical exercise plus a merit and cost analysis. In addition, the TF-QA was asked to prepare an example showing how the operationally assessed quality indicators could be communicated in a European context, and to consult with advanced users on this mock-up for annual reports on quality.

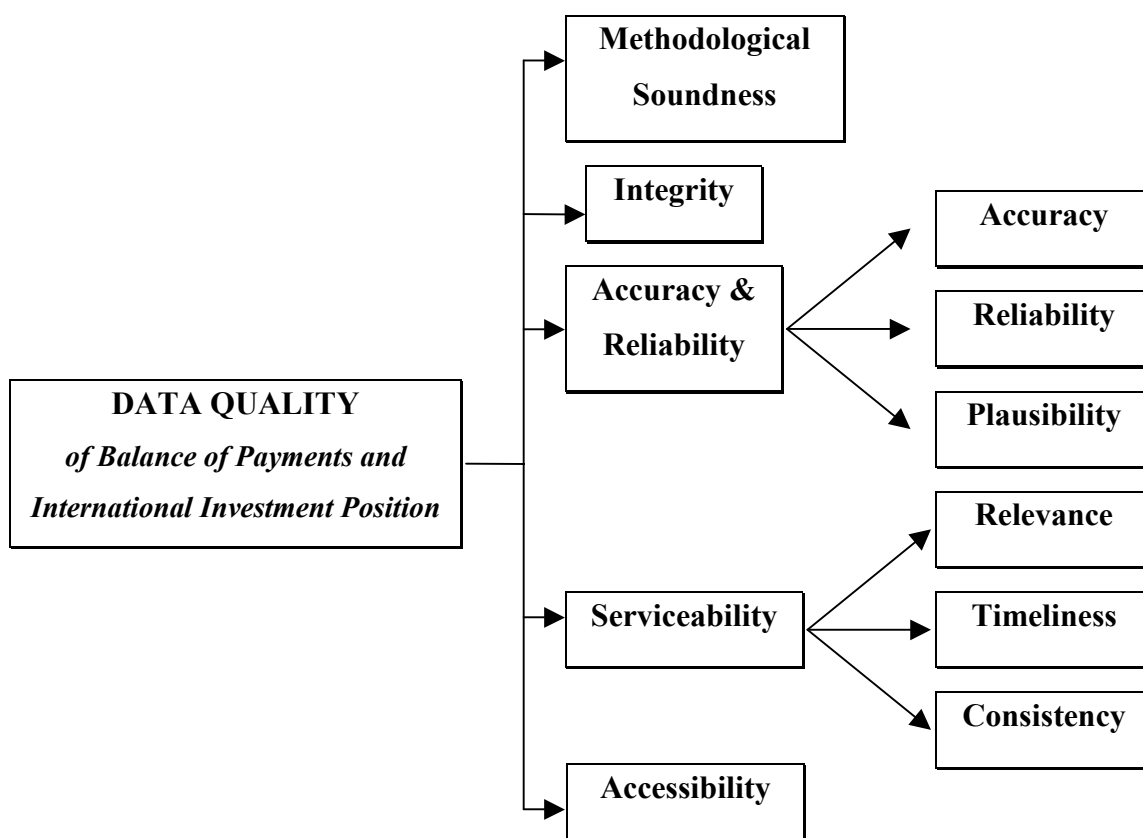
The work carried out by the TF-QA is described in this report in the following sections. Section 2, which focuses on the DQAF as the main reference framework, explains the operational assessment of the main quality dimensions/elements. Section 3 describes the steps followed by the Task Force and its consequent findings. Section 4 discusses who should monitor and assess data quality, while Section 5 illustrates a proposal of how and when to communicate the quality assessment.

2. Operational assessment of the main dimensions/elements

Measuring all aspects of quality is a demanding task. While there is extensive documentation on quality in statistics, only a few current quantitative measurements can be applied. In practice, the approach followed by the TF-QA was (i) to use the concepts and definitions provided by the DQAF to foster common understanding and comparability; (ii) to select those dimensions/elements in the DQAF seen as being most relevant for quantitative measurement, via the identification of possible measures; (iii) to test and analyse the indicators and their outcome; and (iv) to fine-tune various parameters/reference values, where appropriate. The indicators were defined as a distance between given observations and reference data. For example, assuming that the most up-to-date data are the most reliable, then these are used as a reference to measure the

deviation from previous assessments of the same phenomenon observed. Indicators were set up to assess the size, recurrence, bias, etc. of these deviations. For some indicators, for example on errors and omissions, it is also important to define a standard against which comparisons can be made more meaningful across countries/economic unions.

Following the above-mentioned framework, the work of the TF-QA is reflected in the Table included in Annex 3. The table summarises (a) the list of indicators tentatively identified as “key indicators” and selected by the Task Force to be disseminated, and (b) the list of indicators considered as “supporting indicators”, to be used to internally analyse the data quality, and to assist with the interpretation of the underlying phenomena.



2.1. Prerequisites of quality

In respect of the *prerequisites of quality* – as set out in the DQAF⁸ and referring to (1) the legal and institutional environment, (2) resources and (3) quality awareness – it seems that these are broadly met in the MSs and European institutions. Some developments may be welcomed, for example where economists analyse the recent statistical observations and challenge their plausibility as a contribution to the quality management process (as done in some MSs and in the United States). However, improvements may be needed in this context. For example, the number of specialised staff members as well as the computing resources are not always adequate to perform the required tasks. Moreover, quality awareness should always be reinforced by ongoing

⁸ In the July 2003 version of the IMF DQAF, “relevance” is included as belonging to “prerequisites of quality”.

processes that have been put in place to monitor the quality of the collection, processing and dissemination of statistics.

Another important issue is revision practices. Traditional revision policies are mainly based on national requirements; the new requirements of the ECB and the European Commission for Europe-wide aggregates have created an incentive to review this situation. Taking on board an inventory of MSs' revision policies (as shown in the "B.o.p. Book"), the TF-QA welcomed the work undertaken by DG-S and Eurostat in clarifying their needs and disseminating an advanced release calendar based on common reporting requirements for MSs on new observations and revisions to data published. This approach will foster new practices across MSs that are more appropriate to the new situation.

2.2. Methodological soundness

This dimension covers several aspects of the methodological basis for statistics: concepts and definitions, scope, classification and sectorisation and, finally, the basis for recording. The Task Force suggests that all these aspects should be treated through qualitative indicators (i.e. as "yes/no" questions, to identify whether the Member State in question is following the agreed methodology⁹). A review of the major methodological deviations observed in practice by MSs is largely covered in the country chapters of the B.o.p. Book¹⁰ or the IMF/OECD Survey of Implementation of Methodological Standards on Direct Investment (SIMSDI)¹¹. The methodological work on the current account¹² is being reviewed at Eurostat. It could be complemented by an assessment of the possible impact of these methodological deviations on the accuracy of b.o.p./i.i.p. statistics.

The Task Force considered¹³ that changes relative to this dimension could be better reflected through a qualitative assessment and commented, as such, in the quality annual reports.

2.3. Integrity

Users' confidence in the *integrity* of ESCB statistical authorities is enhanced by the following three measures: i) the dissemination of the terms and conditions under which statistics are produced, including those relating to the confidentiality of individually identifiable pieces of information; ii) the absence or disclosure of any access by a central government agency to data before release; and iii) the disclosure of ministerial commentary prior to statistical releases.

⁹ See questionnaire in Annex 4.

¹⁰ In addition, a comprehensive report indicating the main deviations in the current account was submitted to the BOP WG (October 2002).

¹¹ The deviations indicated by the SIMSDI at the initiative of the IMF and the OECD will be reviewed by MSs in co-ordination with Eurostat and the ECB's DG-S.

¹² In the light of the report by the Eurostat's Task Force on the current account (1995).

¹³ A minority within the Task Force indicated that this indicator should be compiled, and moreover as a key one.

As a recommendation regarding *integrity*, MSs and European institutions are invited to periodically check at their level if they fit the DQAF¹⁴ in accordance with this dimension.

2.4. Accuracy and reliability

Although the DQAF includes *accuracy* and *reliability* in a single dimension, it makes a clear distinction between the concepts of *accuracy*, the inherently unknowable difference between an estimate and the unknown true value of the concept being measured, and *reliability*, which is defined as the difference between the first and “final” estimates of a concept. Statistics data tend to be accurate and reliable when the source data and statistical techniques are sound and when the data sufficiently depict reality. There is a certain trade-off between the two aspects of quality. Later estimates incorporating more information and statistical analysis, as well as involving more effort, could be assumed to increase *accuracy*, but at the same time subsequently diminish the *reliability* of the early estimates given fixed resources. Each aspect is considered separately below.

2.4.1 Accuracy

As its definition suggests, the *accuracy* of any statistic is an inherently unmeasurable concept, as it requires knowledge of a “true” underlying value. This does not mean, however, that we cannot gather important information relevant to accuracy assessment. For instance, although Eurostat’s Leadership Group, which is responsible for measuring the quality of sample surveys, has accepted that important aspects of non-sampling errors cannot be measured, it has suggested a whole range of quantitative and qualitative information on data collection procedures, sampling frames (the actual population may, of course, not be known), sampling errors, rates of response, exemption thresholds and grossing-up methods. By assessing the accuracy and quality of inputs and processing, it seeks to draw inferences about the quality of the output. In principle, a similar approach is possible with measures that form part of a statistical framework, such as the balance of payments or national accounts. However, the problem is more complicated here for a variety of reasons: inputs are much more diverse; processing several orders of magnitude is more difficult; the modelling and projection of base year ratios are sometimes used; while additional complexity is added by the “balancing” of information from different sources. The TF-QA took the view that the specification of accuracy indicators required in-depth investigation that fell outside its mandate. In order to identify best practices, the reports generated by the various ECB/Commission task forces in the field of b.o.p./i.i.p. statistics may be useful and provide some recommendations (e.g. TG “Direct reporting and surveys” or TF “Direct Investment”) as well as, for example, the Eurostat’s Task Force on assessing accuracy in national accounts.

¹⁴ In the July 2003 version of the IMF DQAF, the name of this dimension was modified to “Assurances of integrity”.

2.4.2 Reliability

Reliability is, strictly speaking, a broader concept than revision studies. The DQAF states that: “*Reliability refers to the closeness of the initial estimated value to the subsequent estimated value. Assessing reliability involves comparing estimates over time*”. In practice, measuring reliability consists in examining revisions in time series (e.g. their magnitude, standard deviation and bias) between two defined assessments of published data.

Although measuring the size and direction of revisions is relatively easy, a variety of possible indicators and analyses can be performed. For example, the analysis can be limited to simple descriptive statistics or, alternatively, more sophisticated statistical/econometric calculations can be applied. The choice may depend on the characteristics of the data (e.g. b.o.p. or national accounts, national data or supranational aggregates) and on the phenomenon the compiler wishes (or is able) to measure; e.g. a simple assessment of the size of revisions; a comparison of revisions across items and/or countries and lifecycles, and analysis of these revisions (and their process) over time; and the identification of systematic distortions and correlation with other variables.

The TF-QA has selected two types of indicators: (i) simple descriptive statistics, which aim at assessing the size of revision (absolute measures) and at allowing comparability of the indicators across items and/or countries (relative measures), and (ii) more sophisticated approaches (detection of biases and correlations - i.e. the detection of persistent patterns in the revisions).

An ECB proposal on *stability* has already been discussed in the WG-BP&ER and was presented to the IMF B.o.p. Committee in October 2001. The proposal consists of the use of (1) the mean absolute percentage error (MAPE) for the b.o.p. current account gross flows, and (2) the Standard Measure of Stability (SMS) for the b.o.p. financial account series. Both indicators belong to the first group of statistics mentioned above. This proposal was further improved by the TF-QA subgroup¹⁵, leading to the proposed indicators contained in Annex 3.

The simple calculation of revisions using the differences is often unsatisfactory. This expresses the revisions in original units of a variable ‘X’ and depends on its magnitude, often hampering comparability across time, across different variables and across the same variables of different countries. Therefore, it is often useful to provide a relative measure that links the revision to some dimensional measure of the variable.

In the case of **gross data** (data which express quantities), a relative measure of revisions can be expressed as percentage changes from the initial assessment according to the formula $[X(l_j) - X(l_i)] / X(l_i)$ ¹⁶, 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

¹⁵ See Annex 5: Methodological documentation for indicators.

¹⁶ Where X is a generic variable or series, and (l_i, l_j) the predefined time lags. The time lag indicates the time elapsed between the reference period and the publication period (i.e. in case of publication in June of data referring to January, the time lag is five months). Hence k different sets $\{X(l_1), X(l_2), \dots, X(l_k)\}$ of the same variable will be available.

$\frac{1}{N} \sum_{t=1}^N \frac{X_t(l_j) - X_t(l_i)}{X_t(l_i)}$, 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 any revisions can be positive or negative, it is usually appropriate to take them in absolute value. The expression becomes a *mean absolute percentage error (MAPE)*, which is the indicator finally selected for gross data.

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

In the case of *net data*, the application of percentage changes is not meaningful. For example, most of the b.o.p. series in the financial account are expressed in net terms, which are the results of the difference between (part of) the amount invested in and (part of) the amount disinvested from the same item. As the revision to these net data cannot be meaningfully related to underlying quantities, alternative dimensional measures of the variable X need to be used. Hence any measure of the variability¹⁷ of the variable (series) X can serve 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 to produce an expression that we here term (given its similarities with the MAPE shown before) the *mean absolute relative error (MARE)*.

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

where $\text{var}[X(l_k)]$ is the measurement of the variability of the last assessment (k months time lag) of the series X .

Following the literature on measures of forecast quality, additional indicators¹⁸ were also considered for assessing stability in net flows. The indicator finally selected is a ratio between two different mean square errors¹⁹ (MSE), making it a relative measure:

- the numerator uses the MSE applied to the difference between two assessments (revision measure)
- the denominator uses the MSE applied to the difference between variable X and a reference value for X .

The proposed reference value for X was its average²⁰, yielding the variance of X in the denominator. The advantage of using the average is that it can be decomposed²¹ into three components which have some interesting applications for the study of the revisions:

¹⁷ See various variability measures in Annex 5.

¹⁸ See Annex 5 for details of the study on measures of forecast quality.

- 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 *conditional or regression component* is another systematic component that reflects whether the overall pattern of the series with the early estimates was close to that of the series with the later estimates;
- the *unsystematic or disturbance component* is the variance of the residuals obtained by regressing the early estimates data on the later estimates. This component can be assumed to have a random nature without any predictable pattern²².

Owing to its similarities with the previous indicators, this indicator is called the *root mean square relative error (RMSRE)*, and is expressed as a percentage of series volatility:

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

where Θ_t is the reference value for X.

The last two indicators (MARE and RMSRE) are highly correlated, consisting of the difference explained by the weights the indicators put on revision observations. As already mentioned, the advantage of the latter is that it can be decomposed into three components which have some interesting applications for the study of the revisions, such as bias or correlation identification. A general formula, called the *mean relative error (MRE)*, condenses these two indicators as follows:

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

where ρ is the power parameter.

¹⁹ The mean square error is defined as: $MSE = \frac{1}{N} \sum_{i=1}^N (X_i - Y_i)^2$

²⁰ Other measures of distribution location, such as the median and the trimmed mean, were also tested. Assuming that 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 by using the first difference of the series itself.

²¹ See the algebra and further explanations in Annex 5.

²² This indicator only accounts for linear relationships. The unsystematic part could still contain non-linear patterns.

In addition to these two key indicators (MAPE and RMSRE statistics), the TF-QA has developed some supporting indicators that can help in interpreting the previous one, namely *upward revisions* and *directional reliability*²³.

For these calculations, the TF-QA considered it important that both Eurostat and DG-S (for the aggregates) and the MSs (for contributions and national data) keep a repository of revisions for some partners (world, extra-EU, extra-euro area) and components (current account, goods, services, income, direct, portfolio and other investment).

It was agreed to compare the first assessment against the latest available release of the data, though by construction this leads to different assessments of the data across time series, i.e. the latest available series will contain both data revised up to seven times as well as data only revised once. To moderate the risk of introducing some noise into the analysis, the time frame for data to be used in the calculations will be the previous three²⁴ full calendar years. This means that, considering the current revision practice for the euro area/EU aggregate, the observations to be included in the study will have been revised at least twice in the May calculations, and four times in November.

On the specific issue of calculating the selected indicators, it is proposed that MSs compile stability indicators for national contributions and for national data referring to revisions to their publications on a voluntary basis, and that the Commission (Eurostat) and the ECB calculate them for the aggregates available to European institutions at the time of compiling the aggregate.

Moreover, a qualitative assessment may reflect whether revision practices are made public – e.g. advanced release calendar, number of revisions (compliant with SDDS requirements) – and also whether major revisions (expected ones, e.g. for methodological reasons or because of availability of further data, such as the results of a specific survey) are explained or not.

²³ See Annex 3 for indicators and Annex 5 for further explanations.

²⁴ 36 observations are considered the optimum balance between a data frame which contains the recent developments of a series without giving too much importance to previous practices, and the volume of data needed to assume and test normality.

Additionally (where possible and appropriate), the lifecycle of revisions²⁵ may be analysed by the authorities compiling the statistics, e.g. the ECB for the euro area aggregate, as a complement to the selected key indicators, so as to identify any possible significant structure in the revision process of statistics which would indicate areas for further improvement in terms of the quality of the produced statistics. This primarily pertains to the timeliness and reliability, in particular the stability, of the series. Such studies may, in turn, help to anticipate at least some future revisions, and thus produce more unbiased estimations as part of the regular data quality process.

2.4.3 *Plausibility*

Although not included in the DQAF, *plausibility* is considered to be a significant element of the dimension of accuracy. Plausibility describes the likelihood of the data. It may be assessed over time (trend) or in comparison with related (non-statistical²⁶) series. Although it may be difficult to predefine a stable and constant or merely “plausible” pattern in statistics, and it is unlikely that the relative structures of the b.o.p. of different countries/zones will remain unchanged over time, in principle unexpected sizeable outliers nonetheless deserve attention.²⁷ Significant outliers or sudden and unexpected changes in the trend need to be investigated, especially when there is virtually no economic and/or methodological explanation for them. In practice, the assessment of plausibility takes into account the fact that certain statistics are by nature more predictable (and within a shorter time horizon) than others. This assessment is particularly difficult to apply to the b.o.p. financial account, owing to the size and volatility of the flows. However, data which are “implausible” in the sense that they cause the most surprise to analysts are precisely those which have the highest analytical value, which is linked to their lack of predictability. Even though not fully discussed by the Task Force, it was remarked that plausibility should be assessed by the level of detail at which the integration has taken place. It was therefore decided that further work via the WG-BP&ER should be encouraged so as to increase the accuracy of the statistics produced.

2.5 **Serviceability**

As regards the dimension of *serviceability*, the TF-QA decided to focus on the following three elements: relevance, timeliness and consistency.

²⁵ See “Lifecycle analysis of revisions in the euro area balance of payments”, which was presented to the WG-BP&ER in May 2003.

²⁶ See also the *consistency* element below.

²⁷ As an example, the simplest way to formalise a plausibility check is to standardise the variable x under observation. The average (\bar{x}) and the variance (σ_x^2) of the variable can be calculated over a specific historical time range. The plausibility check is therefore performed on the observation x_i with the formula $|s_i| = \frac{|x_i - \bar{x}|}{\sigma_x} > c$, where c is the limit over which the observation is considered implausible.

2.5.1 *Relevance*

The *relevance* element²⁸ refers to the degree to which the statistics produced continue to meet users' needs. Since macroeconomic statistics are a public good, the Task Force considered the possibility of constructing, applying and updating indicators that would assess the degree of relevance. In the case of negative signals, further improvement or even the re-construction of all processes may become necessary. The number of hits on the website or feedback on whether users are ready to pay for the data are some suggested indicators that could disclose the level of the users' interest in the statistics provided. Moreover, a "yes/no" question on whether there should be a periodic review of users' needs would indicate the existence of an established process of review, as well as the organisation's interest in satisfying these needs.

2.5.2. *Timeliness*

The TF-QA considered that timeliness is well covered and monitored by defining and publishing an advanced release calendar for data dissemination (including a contribution to the euro area/EU aggregate). In addition, simple indicators concerning deviations of* the established timeliness, where relevant, can easily be constructed (for example, the number of delays and average/total days of delay with regard to reporting/dissemination timetables). An annual report on data quality needs to reflect this element.

2.5.3. *Consistency*

Concrete indicators measuring overall consistency across statistical series²⁹ are broken down in Annex 3 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 (for example, in the case of methodological or institutional – i.e. enlargement – changes, 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). Conceptual consistency, as highlighted by the IMF, fosters the international comparability of statistics, even when compiled by different institutions.³⁰ In addition, different measurements of the same phenomenon should not result in unreasonably different data³¹.

²⁸ In the July 2003 version of the IMF DQAF, "relevance" is included as an element of "prerequisites of quality".

²⁹ See also *plausibility* above.

³⁰ Differences may still arise from different practices regarding the publication of revisions. To the extent that different institutions aim at *integrity* and *accuracy*, these differences should not be so high as to produce a different picture of the reality described.

³¹ For instance, consistency between aggregated b.o.p. statistics compiled by different international organisations.

Annex 3 provides one key indicator and two potential key indicators that have been selected to measure the first and third sub-categories of the consistency element.

Regarding *internal consistency*, the indicators are based on the errors and omissions series. The errors and omissions indicators of volume should be interpreted taking into account an important feature of the nature of the errors and omissions item: as it is a sum of items, its actual size may vary according to the following factors:

- mutual cancellation of estimation/sampling errors, each of them being of a large magnitude;
- many items containing errors that may or may not point in the same direction.

While a small net residual cannot be taken as an indicator of the relative consistency of the b.o.p., a large, persistent residual is a clear indicator of inaccuracy.

A measure of volume can be provided by the *average absolute error*³² of the errors and omissions (EO):

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

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

An alternative measure of volume is provided by the *root mean square error of net errors and omissions (RMSE(EO))*, which can be decomposed into the bias and the variance components³³:

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

To make these volume 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 as GDP + imports or i.i.p. assets will be used in the supporting indicators.

Regarding *external consistency*, the reconciliation of b.o.p. with MFI balance sheet statistics is deemed important and sensitive. The quality indicator should compare “other investment” in the financial account of b.o.p. and the deposits/loans of monetary financial institutions (MFIs). These series move somewhat randomly, so the consistency indicator should consequently take into account both the magnitude of the differences and the volatility of the original series.

The RMSRE statistic, which is used as a stability measure for the financial account items, has also been considered sufficiently efficient to assess the external consistency of net flows. For more explanation concerning the properties of this indicator, refer to Section 2.4.2 or Annex 5.

³² The internal consistency indicators have been built with the assumption that the true value of errors and omissions is nil.

This indicator is only relevant for countries that do not obtain their b.o.p. financial account directly from MFI balance sheet data. Where appropriate, the Task Force recommends to use it as a key indicator..

Increasing consistency may lead to some errors and omissions. This shows a possible trade-off between reconciliation and internal consistency and, to some extent, with the accuracy of the data.

A second indicator, relative to *external consistency*, assesses the consistency between goods credits and debits from b.o.p. statistics, and export and import trade for international trade statistics. This indicator has been constructed taking into account the nature of the two analysed series. Both series are gross flows with a positive trend, and in which methodological and valuation differences are more or less constant during the period under consideration.

A simple indicator based on the difference between both series will show a constant bias that does not allow direct comparability between items and/or reporting areas, and is based on dissimilar conceptual/compilation methods that are already known. To overcome this bias, the first difference of the series is analysed instead of the raw data. Moreover, consistency is related to the magnitude of the discrepancies and not to their direction; therefore, absolute values of the differences are used. Finally, to remove units and make the indicator comparable across items and/or reporting areas, the statistic is divided by the average of both series:

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

where y is b.o.p. data series, x is the series under comparison, Δ is the 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 (perfect match) to plus infinity (no match possible).

As mentioned above, the interpretation of this indicator should take into account the methodological differences between these two data sets. Apart from the different coverage which affects both data exports and imports, the different valuation of the imports in b.o.p. and external trade statistics (f.o.b. versus c.i.f.) steers the value of the indicator towards a positive constant amount.

This indicator was found difficult to interpret by advanced users. The TF-QA recommendation is to develop indicators to capture the external consistency between gross flows.

In addition to the indicators that have already been explained, the TF-QA considered three supporting indicators: *count positive errors and omissions*, *directional consistency*, and *dispersion measures*³⁴. They will also be developed and checked with users, in due course.

³³ See the algebra in Annex 5.

³⁴ See indicators in Annex 3 and further explanations in Annex 5.

2.6. Accessibility

From the users' point of view, accessibility reflects the ease of obtaining the information disseminated by a statistical agency, the suitability of the form in which it is shown, the media of dissemination and the availability of metadata.

In this context, accessibility is defined as a direct dimension of quality. Information that is not accessible to intended users is thereby assumed to be of poor quality, regardless of its accuracy.

The dissemination of statistics in electronic form via the ECB's and NCBs' websites, or any websites of statistical agencies, complete with a user-friendly browser and interfaces, is a way of improving accessibility. A printed publication may no longer suffice when markets need easy and equal access to information. The accessibility of data can be viewed as a dynamic dimension, given the developments in information systems and technologies. The interaction with users should also be given extensive consideration, in particular as regards feedback on data presentation and content quality. The ECB's DG Statistics is currently engaged in a project designed to improve the accessibility of data on the ECB's website.

For the dimension of *accessibility*, a series of quantifiable and non-quantifiable indicators are proposed as supporting indicators.

2.7. The interrelation between quality dimensions and elements

While the quality dimensions have so far been considered in isolation, in practice these dimensions are interrelated; in some cases, moreover, there is a clear trade-off between them. It is important to obtain a clear picture of this interrelation, in order to approach the issue of quality in a systematic manner. Some trade-off cases are listed below.

1. Between *timeliness* and *accuracy/reliability*: it is commonly understood that the shorter the deadline, the more challenging it is to achieve accuracy. Up to a certain point, timeliness may be improved without (substantially) reducing accuracy. After this point, however, it is no longer possible. This occurs when, to reduce timeliness, the producer is forced to compile data from incomplete source data. As more data become available afterwards, important revisions will be published which will damage the overall reliability of the data. An optimum balance should therefore be achieved.

2. Between *stability (reliability)* and *accuracy*: although reliability and accuracy are listed in the DQAF as one single dimension owing to the high level of correlation between them, it is possible to see a potential trade-off or negatively interpret the relationship. Users appreciate stable data, yet stability could indicate that additional (more comprehensive) information is not being used to enhance the picture given in the first assessment. Furthermore, it could suggest that deficiencies in the first compilation of the observation tend to remain undisclosed for some time. In this sense, stable data might not reflect reality. An additional scenario is when a methodological change is made but back observations are not revised, thus providing little information on the

trend. On the other hand, a high level of instability (whether disclosed or not) is an indication of the potential inability of the data collection and compilation procedure to cope with the required timeliness.

3. Between *stability (reliability)* and *integrity*: any statistical agency which publishes data is expected (particularly those countries subscribing to the IMF's Special Data Dissemination Standard) to deliver an accurate picture of recent as well as previous periods, according to an advanced release calendar. Despite all efforts to produce relevant statistical information, it is unlikely that data are compiled during the first assessment using the complete set of information needed to provide the most accurate figures. Thus, the more timely the data, the more they are subject to subsequent revisions. However, revisions of a large magnitude would indicate a lack of accuracy in the data collection and/or compilation process, and hence would question their integrity. In special cases, significant revisions may lead to the data collection and compilation system being reviewed and improved. A balance is needed between artificially frozen data and a Brownian motion; this may rely on professional standards, experience gained and a proactive attitude (for example, vis-à-vis users) on the part of the statistical agency³⁵.

4. Between *consistency (serviceability)* and *accuracy/reliability*: an example of a trade-off between these dimensions is if the errors and omissions item is, partly or fully, hidden or incorporated elsewhere in the accounts. This could lead to a low value being erroneously interpreted as representing a high level of internal consistency, which would damage the accuracy and reliability of b.o.p. statistics.

3. Findings

3.1 Initial findings

The Task Force initially dealt with (1) the identification of the users, (2) the description and classification of their needs and (3) a definition of what is meant by “good” quality.

- There are three major groups of users of statistics which are interested in quality assessments:
 - policymakers, governments, the ECB and NCBs within the Eurosystem/ESCB, the Commission and international organisations such as the IMF, OECD or BIS; in all these institutions, the data analysis and study work is performed by staff in specific business areas (e.g. economics departments) who also need to know the quality of the data they are examining;
 - “advanced” external users, e.g. investment funds, financial market rating agencies, academics, economists using econometric models to assess or forecast economic

³⁵ “Harmonisation of revision practices for the euro area/EU b.o.p and i.i.p.”, May 2003.

developments, etc.; the general public usually has access to data and the quality assessment of these data through these advanced users or through the media;

- the statistics producers themselves at the European and national level.
- These users need comprehensive quality measures and indicators for statistics that are capable of assessing the quality of the statistics produced and disseminated, so that:
 - the first group can base its decisions on timely and reliable data, enabling it to understand the developments at stake and to anticipate the effects of any changes in the policy instruments, such as interest rates; ideally this group would need to know the probability, direction and magnitude of any subsequent revisions;
 - the second group may wish to accurately assess the business cycles and structural developments; this would help diminish the level of uncertainty, e.g. in producing macroeconomic analyses or making investment decisions;
 - the third group will be able to evaluate and improve the various aspects of the data production process and to maintain high data quality under constraints.

The communication process with these groups of users is further developed in Section 5.

In dealing with these various aspects of the data and their quality assessment, the Task Force's report focuses mainly on b.o.p. statistics. However, some of the report's recommendations also refer to related statistics such as the i.i.p.

Notwithstanding the importance of the whole compilation process for a comprehensive data quality assessment, the TF-QA has restricted its work to indicators based on output. The TF-QA realised that the accuracy dimension, in particular, is inseparably linked to the production process, and has therefore decided not to develop specific indicators for the time being.

The TF-QA also noticed that there is a trade-off between some dimensions, and that the importance of the defined indicators can vary depending on whom (i.e. compilers and different users) they are addressed to or what the indicators are used for (e.g. improving data production, monetary analysis, structural analysis, forecasting, etc.). The discussion on the MSs' practices for assessing and ensuring the quality of b.o.p. data revealed that, although much work is currently done by most MSs on checking the quality of their b.o.p. data, the procedures followed are, in many cases, somewhat informal. Moreover, from the answers to a questionnaire distributed by the ECB to MSs, it seems that: (1) although there are similarities in dealing with "integrity", "methodological soundness" and "accessibility" dimensions, the treatment of these dimensions needs to be periodically monitored, and (2) that MSs are using different methods to assess and measure the elements/indicators within the "accuracy and reliability" and "serviceability" dimensions. These dimensions therefore need to be the focus of the TF-QA (without necessarily neglecting other categories).

The work of the TF-QA was organised based on the following principles:

1. MSs are currently engaged in assessing b.o.p. data quality in several ways. Considerable harmonisation has been achieved so far in concepts and definitions, as shown in the ECB's "EU b.o.p./i.i.p. statistical methods" (the "B.o.p. Book") of which Chapter 3 is the Compilation Guide (see the ECB's website), in the Eurostat Vademecum, in the draft regulation (in particular the definition of services items), and in the Manual on Trade in Services.
2. The DQAF format provides a suitable framework for assessing data quality, and is designed so that it can be applied to all countries regardless of their statistical development. In the European context, the dimensions of "integrity", "methodological soundness" and "accessibility" are broadly covered by MSs (but may need to be periodically monitored).
3. Accordingly, the TF-QA considered that the main focus should be on the "accuracy and reliability" and "serviceability" dimensions. In this respect, definitions were refined, and indicators that need to be operationally assessed were identified.
4. The TF-QA felt that the list of indicators should be limited to ensure that they are operational and useful. For some categories, quantitative indicators could be developed, whereas for others, qualitative indicators seem to be more appropriate.
5. The relevance of the quality indicators may differ regarding the frequency of data dissemination and the priorities of the users. These should be complemented with an appropriate commentary.
6. The indicators may need improvement and fine-tuning depending on users' needs and economic developments. Once the indicators are in regular use (in 2004), they should be frozen for a couple of years, and then reviewed by end-2005.
7. The focus of the TF-QA's work was on euro area/EU aggregates, including the MSs' contributions to the aggregates. National b.o.p. data may be subject to a similar quality assessment, as their quality is inextricably linked to the MSs' contributions (e.g. errors and omissions).
8. The TF-QA also discussed who should monitor/assess the data quality.
9. The participants of the TF-QA were acting in their capacity as experts, not as representatives of their national institutions³⁶. Accordingly, action can only be taken by DG-S, Eurostat and Member States, where appropriate with the assistance and co-ordination of the relevant bodies and committees (i.e. working groups, STC, CMFB).

³⁶ The list of participants is provided in Annex 2.

Moreover, as shown in Section 2 and in the annexed tables, the Task Force has made its best effort to do the following for each indicator:

- to clarify and sort concepts/definitions;
- to define for each indicator
 - (1) *what* purpose it should serve (i.e. focus on euro area/EU aggregate and national contributions; a similar approach is encouraged for national b.o.p./i.i.p. data),
 - (2) *to whom* these criteria may be of use (e.g. compilers, internal users, advanced external users and the general public). Key indicators are deemed relevant for users as well as for compilers, whereas supporting indicators are primarily used by compilers, though in case of specific features highlighted in key indicators, they may also be shown to users.
 - (3) *ways* of assessing it,
 - (4) the *frequency* of calculation and reporting, and
 - (5) *where* quantitative indicators would be relevant (i.e. whether they are applicable to all items of the b.o.p./i.i.p., or to some items only);
- to prioritise indicators and decide which require further examination;
- to define if and when these indicators are going to be published, to whom they are intended, and who will be their final assessor.

3.2 Subsequent findings

As requested by the CMFB, a pilot exercise was run during autumn 2002 to test, assess and, where relevant, modify the proposed indicators. Four Member States participated in this exercise: DE, FR, IT and UK, as well as the ECB.

The ECB's Balance of Payments Statistics and External Reserves Division (S/BOP), with the aim of facilitating not only the calculations of the proposed indicators, but also their reporting, has developed a toolkit called "QaTool" that could be used either for the pilot exercise or for internal use. QaTool is a user-friendly application developed by S/BOP in the Microsoft Office environment that allows easy and quick calculations on the Data Quality Indicators proposed by the TF-QA.

The TF-QA members welcomed the toolkit as being a very helpful instrument for this exercise, and encouraged its further development. An updated version, which includes some fine-tuning made to the indicators, was available in autumn 2003.

3.2.1. Pilot exercise and feasibility study (autumn 2002)

The outcome of the pilot exercise and subsequent feasibility study was as follows:

1. most of the tested indicators were deemed relevant, but need further refinement (i.e. simplification of the formulae, a study of the properties of the indicators, the presentation of supporting graphs),
2. the appropriate parameters should be selected, tested and implemented to ensure that the resulting figures are comparable across countries,
3. some indicators were deemed difficult to interpret,
4. the number of indicators for dissemination (beyond the compiling community) should be restricted.

Overall, the feasibility study was deemed less relevant, as the QaTool reduces the cost and burden of calculating the indicators. Additionally, the initial loading of back data implies that European institutions should provide the adequate series to MSs, where no longer available.

As a result, and to overcome the above mentioned problems, the TF-QA decided that

1. the indicators should be distinguished into key and supporting ones: the first group as homogeneous and simple to interpret as possible, and calculated by general, precise and straightforward formulae; whereas the second group would be used to assist the interpretation of the underlying phenomena;
2. the proposed indicators would need to be supplemented by other instruments such as reconciliation tables or graphs;
3. the underlying methodology should be developed and documented to be as user-friendly as possible.

All MSs together with the ECB's S/BOP and Eurostat agreed in the January 2003 CMFB to run a three-month empirical exercise during the first half of 2003. The aim of this overall test was to refine, assess and implement the selected indicators, and to fine-tune the QaTool.

A new questionnaire was designed by the Task Force and circulated to MSs in April 2003 with the aim of collecting the results of the empirical exercise and conducting a cost-benefit analysis for each indicator.

3.2.2 Empirical exercise (February-April 2003)

The three-month empirical exercise agreed in the January 2003 CMFB ran from February to April 2003 and involved twelve MSs and DG-S. To initiate the empirical exercise, a workshop was organised in late January 2003 by the ECB. The aim of this workshop was to acquaint representatives of all MSs with the objectives of the exercise, the definitions and calculation methods of the indicators, and to demonstrate how the QaTool could facilitate the exercise.

Twelve Member States (Belgium, Germany, Spain, France, Ireland, Italy, Luxembourg, Netherlands, Portugal, Finland, Sweden and the United Kingdom) plus the ECB (DG-S) participated in this exercise. During this period, all the initially selected indicators were calculated and reported to DG-S up to three times, using different parameters and time frames, solving any

technical problems (for example, the unavailability of some data in certain MSs), and applying the calculation to the country contribution to the euro area as well as to national data.

The results were analysed and interpreted, and made it possible to complete the exercise, i.e. by refining and classifying most of the initially selected indicators as either key or supporting ones (prioritisation). However, the experience gathered with the exercise showed that some of the indicators would benefit from further refinement and/or need to be accompanied with caveats. In particular, some isolated results could lead to misinterpretation without an appropriate explanatory commentary. Further consideration to discriminate across key or supporting indicators will be subject to further experience and discussion across compiling agencies, and will additionally depend on the on-going feedback to be obtained from users, which are important addressees of any quality assessments.

3.2.3 Merit and cost analysis (May 2003)

In early May 2003, around the end of the experimental period, a questionnaire was disseminated to MSs to gather information on the findings and on a ‘merit and cost’ analysis for each indicator. Based on the 14 replies received from the 12 MSs and DG-S that participated in the three-month empirical exercise, plus Austria, those indicators considered the most relevant for the compilers were identified, as well as their cost of calculation and data availability.

The TF-QA agreed a tentative set of indicators which are considered key for the assessment of data quality and are thus subject to regular (annual) publication. The remaining indicators were considered to be supporting indicators³⁷ and could be useful for the compilers by assisting in the interpretation of the set of key indicators. These could also occasionally be shown to users.

3.2.4 Communication test (Q4 2003 and Q1 2004)

To further assess the relevance and appropriateness of the selected indicators for users, advanced users were consulted at the end of 2003 and in early 2004, which led to some fine-tuning of the indicators. These users were provided not only with the indicators themselves, but also with a corresponding commentary on their shortcomings and on the existing trade-offs between them.

Suggestions received

Input was received from advanced users at the ECB (Directorate General Statistics and Directorate General Economics) and at the Deutsche Bundesbank (International Relations Department and Economics Department). Basically, they suggested the following actions:

³⁷ The list of supporting indicators is available on request.

- a) To clarify the nomenclature used within the formulae for the indicators in the annex (sometimes the nomenclature could be simplified or become unambiguous³⁸),
- b) To stress the exceptional nature of initial euro area b.o.p. and i.i.p. statistics in 1999 and 2000 (resulting in larger revisions than for the following years),
- c) To mention the availability and publication of seasonally adjusted data for the euro area current account as adding analytical value,
- d) To add a simple explanation on the interpretation on errors and omissions (in which significant discrepancies could be netted out),
- e) To include an analysis of mirror data with major counterpart countries once the geographical details become available for the euro area,
- f) To focus also on indicators for the net items of the current account,
- g) To include information about the “speed of convergence” of each time series.

Implementation of the suggestions

1. The suggestions a) to d) have been addressed within the version of the report circulated in March 2004.
2. Regarding suggestion e), the analysis of mirror data had already been suggested by the TF-QA and is under investigation by DG-S. In practice, it will be developed when the euro area b.o.p. and i.i.p. with main counterparts is published, i.e. in early 2005.
3. As regards suggestion f), i.e. indicators for the net items of the current account, it will be necessary to analyse whether the assumptions on which the indicators are based are valid, such as the stationarity of the series.
4. Finally, DG-S has worked on the speed of convergence of euro area b.o.p. time series³⁹. This work may be reflected in a paper in the future and serve as reference documentation.

As an example of the report sent to advanced users, the following paragraphs, which comment on the *Stability* indicators for the financial account, provide a good idea of how the indicators are interpreted and what the annual report will be like.

[...] For the items of the b.o.p. financial account, another type of indicator was used: the root mean square relative error (RMSRE). This indicator measures the distance between the first assessment and the final assessment in relation to the volatility of each time series. In the financial account, the volatility is used instead of the respective series to assess the relative size of the revisions. This choice is based on the fact that the series contain net data, and thus can be either positive or negative. The volatility of each series was

³⁸ For example, there was a misunderstanding owing to the term “variability”, which was understood as “variance”.

³⁹ A draft was circulated to the WG-BP&ER: “*Life cycle analysis of revisions in the euro area balance of payments*”, May 2003.

estimated by its variance, under the assumption that the series of the financial account fluctuate around the average over a three-year period.⁴⁰

The RMSRE can be decomposed into three factors: the bias component, the regression component and the disturbance or unsystematic component. The bias reflects whether the revisions were systematic. The regression component reflects the extent to which the overall pattern/variability of the series of the first shots was close to that of the series with the final figures.

Chart 4: Revisions of the euro area financial account as % of volatility

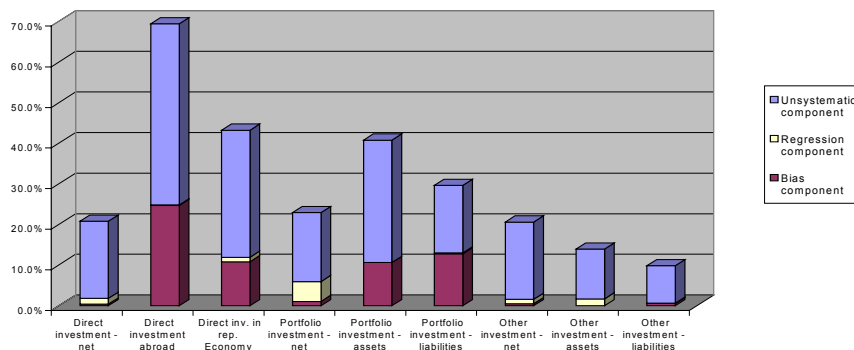


Chart 4 shows the results obtained for the euro area b.o.p. The revisions were highest for direct investment abroad. Moreover, revisions were significantly biased not only for the direct investment abroad and in the euro area but also for the portfolio investment assets and liabilities of the euro area. However, the net flows show no bias. The regression component was low for almost all series, i.e. in general the first shots have the same pattern as the final figures.

4. Who should assess/monitor the data quality?

In principle, there are three possible approaches: either the European institutions could take charge of this task, or the MSs themselves (for example, in a committee, such as the GNP Committee, which Eurostat is considering setting up in the framework of the draft Regulation), or a third party⁴¹ could be put in charge. Dealing with rather simple, straightforward indicators, and assuming the integrity of the compiling agencies, national compilers for own (national and contributing data) and European institutions for euro area/EU aggregate could establish and maintain these indicators and monitor quality reports.⁴² The B.o.p. WG (or alternatively the future

⁴⁰ The assumption of stationarity was confirmed by standard statistical tests.

⁴¹ A senior committee, possibly composed of representatives of compilers, users and academics (e.g. three times two persons), could review any possible developments and monitor the quality management for the euro area/EU aggregate. The committee could be commissioned for integrated statistics, in particular b.o.p./i.i.p. and national (economic and financial) accounts.

⁴² For some data/indicators, the quality of contributions depends on the quality of national data (e.g. the level of errors and omissions). Poor national data and good contribution data may not occur. Eurostat/DG-S should look at national data indicators as a prerequisite for evaluating the quality of the contribution to EU/euro area statistics.

B.o.p. Committee) and/or the WG-BP&ER may undertake reviews of the overall quality process, i.e. the compilation and transmission of national contributions, as well as the compilation and publication of the EU/euro area aggregate. The approach for the monitoring will be further discussed by the CMFB, also in the light of the forthcoming proposals for national (economic and financial) accounts.

As an example of the above procedure, taking into account the annual report on the quality of the data⁴³, DG-S will prepare a report on the quality of the euro area b.o.p. aggregate from December 2004. This report will largely be based on the quality indicators selected by this Task Force. The TF-QA considered that assessing/measuring the quality as proposed in this report will involve significant resources – which should remain in proportion to the relevance of the related indicators.

5. Communication issues

The operational measures of data quality, which were disseminated to advanced users in December 2003 and early 2004, are intended to complement qualitative assessments of the data, and not as a substitute. DG-S and Eurostat will prepare quality reports for the euro area/EU aggregates based on both quantitative indicators and qualitative assessments. Member States are encouraged to prepare them for their national data.

The TF-QA deems it useful to communicate a few key quality indicators to the three major user groups (see Section 3.1) after a transitional, fine-tuning period. Other indicators can only be used in the normal compilation process and would usually only be disseminated among statistics producers. For example, if data are transmitted to the European institutions with some delay, but the aggregates are released on time according to the ECB/Commission (Eurostat) release calendar, concerns would only relate to the procedure.

Communication entails two elements: data and metadata. The latter cover a wide range of information, e.g. explanations on revisions to data previously published, methodological notes, etc. In the case of major changes to data or metadata, the European institutions may disseminate this information to the public at large if the Member State(s) concerned agree.

Among the quality dimensions/elements set out in Section 2, candidates for regular assessments include stability, timeliness, consistency and methodological soundness. The proposed indicators will be calculated twice a year, preferably in May and November, for monthly data and using the

⁴³ According to Article 6 of the Guideline ECB/2003/7 of 2 May 2003 (OJ No. L 131/20), “[t]he ECB assesses [...] data relating to the euro area balance of payments, international investment position statistics and international reserves. The assessment shall be carried out in a timely manner. The Executive Board of the ECB shall report yearly to the Governing Council on the quality of the data.”

last 36 observations available⁴⁴. These results will first be discussed and interpreted among compilers.

Annual reports will be produced with the aforementioned November inputs (see Section 3.2.4), at least for the euro area/EU aggregates. The reports will be published at the end of the year together with a statement on the progress made and achievements so far, as well as the necessary caveats and information that will help in the interpretation of these indicators. In particular, special attention should be given to a balanced consideration of the quantitative as well as qualitative indicators, noting as appropriate the trade-offs between different dimensions/elements of quality (e.g. timeliness vs. accuracy). No benchmark will be considered *until sufficient experience has been gained and the practice of publishing such indicators has spread over comparable economic areas and across sets of statistics*. However, the attainment of an appropriate level of quality in critical areas will be closely monitored both by MSs for their national data and their contribution to the euro area/EU aggregate, and by European institutions for the aggregates.

These reports would be addressed to the statistical community, the STC and the CMFB, and could be published on the CMFB website, for example. In the future, the reports could be complemented by the work on quarterly national accounts, as they may be of interest for external users, as well as the international statistical community.

The TF-QA agreed on the following calendar regarding communication issues:

After a period of internal calculations and results analysis, in early 2005 the final selected indicators will be published in (annual) quality reports, together with the necessary caveats and methodological information.

The dissemination of an abridged set of indicators would relate to the euro area/EU aggregate from 2004 onwards, and would be encouraged for national data.

⁴⁴ For the revision study indicators, the time frame to be used in the calculations will be the previous three full calendar years.

6. List of annexes and documents attached

Annex 1: CMFB, Elements of a Quality Framework – Mandate for a Task Force

Annex 2: List of participants

Annex 3: List of recommended indicators

Annex 4: Draft questionnaire on the methodological soundness

Annex 5: Methodological documentation for indicators

Glossary

Accessibility: one of the five DQAF quality dimensions, which refers to the data, metadata and assistance to users. Statistics should be presented in a clear and understandable manner, forms of dissemination should be adequate, and statistics should be made available on an impartial basis. Up-to-date and pertinent metadata will be made available, accompanied by prompt and knowledgeable service support.

Accuracy: a DQAF quality dimension linked to the dimension of *Reliability*, which considers whether source data and statistical techniques are sound and whether statistical outputs sufficiently portray reality. The accuracy dimension is generally evaluated at the level of the materially significant b.o.p. data items.

Balance of payments (b.o.p.): the statistical statement that systematically summarises, for a specific time period (usually monthly, quarterly and/or annually), the economic transactions of an economy with the rest of the world. Transactions between residents and non-residents consist of those involving goods, services and income; those involving financial claims on and liabilities to the rest of the world; and those classified as transfers (such as gifts) which involve offsetting entries in order to balance – in an accounting sense – one-sided transactions.

B.o.p. Book: see “European Union b.o.p./i.i.p. statistical methods”

Committee for Monetary, Financial and Balance of Payments Statistics (CMFB): this committee was established by a Council Decision in 1991 to assist the European Commission in drawing up and implementing work programmes concerning monetary, financial and balance of payments statistics. The CMFB is the forum for co-ordination of statisticians from the National Statistical Institutes and Eurostat on the one hand, and the national central banks and the ECB on the other.

Consistency: a quality element of *Serviceability*, which covers different aspects such as whether statistics are consistent within a dataset, over time, or with major datasets.

Data Quality Assessment Framework (DQAF): this framework defined by the IMF, in co-operation with a number of statistical agencies and international organisations, covers all aspects of the statistical environment or infrastructure in which data are collected, processed and disseminated, by integrating aspects of the quality of the institution and of its products. Five dimensions – *Integrity, Methodological Soundness, Accuracy and Reliability, Serviceability* and *Accessibility* of data quality – and a set of prerequisites for the assessment of data quality form the basis of the DQAF.

Economic and Monetary Union (EMU): the Treaty establishing the European Community describes the process of achieving Economic and Monetary Union in the European Union in three stages. Stage One of EMU started in July 1990 and ended on 31 December 1993. It was mainly characterised by the dismantling of all internal barriers to the free movement of capital within the European Union. Stage Two of EMU began on 1 January 1994. It provided for, inter alia, the establishment of the European

Monetary Institute (EMI), it prohibited central banks from financing the public sector, it furthermore prohibited the public sector from obtaining privileged access to financial institutions, and it sought to avoid excessive government deficits. Stage Three started on 1 January 1999 with the transfer of monetary competence to the ECB and the introduction of the euro.

European Union b.o.p./i.i.p. statistical methods (B.o.p. Book): a manual produced by the ECB which aims to provide parties interested in b.o.p. and i.i.p. statistics (i.e. as users or compilers) with information relating to all EU countries on (I) the content and structure of statistical data and (ii) the collection methods used. It also gives an overview of the compilation of the euro area aggregate figures by explaining the compilation procedures and the underlying methodological concepts agreed by the EU Member States. The “B.o.p. Book” was first issued in January 1998, and has been successively updated every year. The latest version is November 2002.

EU Council: a European Community institution that is made up of representatives of the governments of the Member States, normally the ministers responsible for the matters under consideration (it is therefore often referred to as the ‘Council of Ministers’). The EU Council meeting consisting of the finance and economy ministers is often referred to as the ‘ECOFIN Council’.

Euro area: the area encompassing those Member States in which the euro has been adopted as the single currency in accordance with the Treaty establishing the European Community, and in which a single monetary policy is conducted under the responsibility of the ECB. The euro area comprises Belgium, Germany, Greece (as from 2001), Spain, France, Ireland, Italy, Luxembourg, the Netherlands, Austria, Portugal, Finland, and for statistical purposes, the ECB as well.

European Central Bank (ECB): the ECB is at the centre of the European System of Central Banks (ESCB) and the Eurosystem and has legal personality under Community law. The ECB ensures that the tasks conferred upon the Eurosystem and the ESCB are implemented either by its own activities pursuant to the Statute of the European System of Central Banks and of the European Central Bank, or through the national central banks.

European Commission (Commission of the European Communities): the institution of the European Community which ensures the application of the provisions of the Treaty. The Commission develops Community policies, proposes Community legislation and exercises powers in specific areas. In the area of economic policy, the Commission recommends broad guidelines for economic policies in the Community and reports to the EU Council on economic developments and policies. It monitors public finances within the framework of multilateral surveillance and submits reports to the Council.

Eurostat: the Statistical Office of the European Communities. Eurostat is part of the European Commission and is responsible for the production of Community statistics. It collects and systematically processes data that are produced mainly by the national authorities.

Integrity: one of the five DQAF quality dimensions, which considers whether the principle of objectivity in the collection, processing and dissemination of statistics is firmly adhered to in terms of

professionalism, transparency and ethical standards.

International investment position (i.i.p.): the statistical statement of the value and composition of the stock of an economy's financial assets or financial claims on the rest of the world, and of an economy's financial liabilities to the rest of the world.

Methodological soundness: one of the five DQAF quality dimensions, which considers whether the methodological basis for the statistics follows internationally accepted standards, guidelines or good practices. This dimension is assessed against the balance of payments guidelines outlined in the fifth edition of the Balance of Payments Manual (BPM5). The application of these guidelines is generally evaluated at the level of materially significant balance of payments data items (e.g. goods, services, income, direct investment and portfolio investment).

Monetary financial institutions (MFIs): these comprise central banks, resident credit institutions as defined in Community law, and other resident financial institutions whose business is to receive deposits and/or close substitutes for deposits from entities other than MFIs and, for their own account (at least in economic terms), to grant credits and/or make investments in securities.

Plausibility: a quality element which describes the likelihood of the data. It may be assessed over time (trend) or in comparison with related series. Although not included in the DQAF, *Plausibility* is considered to be a significant element of the dimension of *Accuracy*.

Relevance: a quality element of *Serviceability*, which reflects whether statistics cover relevant information on the subject field. The relevance and practical utility of existing statistics in meeting users' needs are monitored. [In the July 2003 version of the DQAF, *Relevance* has been reclassified as a "prerequisite" of quality.]

Reliability: a DQAF quality dimension linked to the dimension of *Accuracy*, which refers to the closeness of the initial estimated value to the subsequent estimated value. Assessing reliability involves comparing estimates over time, and considering whether revisions have been tracked and analysed for the information they may provide.

Serviceability: one of the five DQAF quality dimensions, which considers whether statistics are relevant, timely, consistent, and whether they follow a predictable revisions policy.

Stability: a quality element of *Reliability*, which refers to the likelihood or intensity of revisions of a data item until its final value is calculated.

Statistical Programme Committee (SPC): this committee is composed of the heads of the National Statistical Institutes of EU Member States, and is empowered to amend EU regulations in the field of statistics in some specific circumstances (the 'comitology' procedure).

Task Force on (Output) Quality (TF-QA): see mandate in Annex 1.

Timeliness: a quality element of *Serviceability*, which displays the time lag between the reference

period and the data publication. Timeliness should follow internationally accepted dissemination standards.

List of reference documents on quality

ABS, Quality measures for systems of economic accounts, J. Zarb, Analytical Services Branch, Methodology Division, Australian Bureau of Statistics.

ECB, Assessing the quality of the euro area b.o.p./i.i.p. statistics, April 2001 (ECB/ST/STC/QUALIMP4.DOC).

ECB, Data quality: Work in progress on balance of payments and related statistics, January 2002 (ECB/ST/WG/BP/QDIMIMPL1.DOC).

ECB, Trade-off between timeliness and accuracy: ECB requirements for general economic statistics, ECB Monthly Bulletin, April 2001.

Eurostat, Quality measurement - Eurostat experiences.

Eurostat, Quality Report, May 2002.

IMF, Data Quality Assessment Framework for balance of payments statistics, Statistics Department, July 2001 Vintage.

SN, Some elements of a quality framework for CMFB statistics, by S. Keuning and S. Algera, Statistics Netherlands, October 2001.

On revisions studies:

ABS, "Quality of Australian balance of payments statistics". McLennan, W. February 1996. This report was presented at the BOPCOM meeting, Canberra 21-25 October 1996 under the title "Revisions in Australia's balance of payments (BOP) statistics".

BEA, "Evaluation of the GNP estimates". Young, Allan H. (Carol S. Carson, Frank de Leeuw and Robert P. Parker also contributed to this article). August 1987.

BEA, "Reliability and accuracy of the quarterly estimates of GDP". Young, Allan H. October 1993.

BEA, "Reliability of GDP and related NIPA estimates". Fixler, Dennis J. and Bruce T. Grimm. January 2002.

BEA, "Reliability of the quarterly and annual estimates of GDP and gross domestic income". Grimm, Bruce T. and Robert P. Parker. December 1998.

ECB, "Analysis of bias in the euro area balance of payments revisions", February 2002 (ECB/ST/WG/BP/BIAS.DOC).

ECB, "Harmonisation of revision practices for the euro area/EU b.o.p and i.i.p.", May 2003

(ECB/ST/WG/BP/HARMRP_CMFB_FU.DOC).

ECB, “Lifecycle analysis of revisions in the euro area balance of payments”, July 2003 (ECB/ST/BP/EUBOPREVIS.DOC).

ECB, “Proposal for measurement of stability as a quality”, December 2001 (ECB/ST/IMF/BP/STABILITY.DOC).

IMF working paper – Statistics Department, “Assessing accuracy and reliability: a note based on approaches used in national account and balance of payments statistics”. Carson, Carol S. and Laliberté, Lucie. February 2002.

ONS, “Revisions analysis of initial estimates of annual constant price GDP and its components”. Symons, Peter. *Economic Trends* No. 568, March 2001.

ONS, "Revisions analysis of initial estimates of key economics indicators and GDP components". Barklem, Adèle. *Economic Trends* No. 556, March 2000.

Statistics New Zealand, “Revisions in the New Zealand balance of payments”, Nesbit, Shirley. Paper for the BOPCOM meeting, Canberra, 21-25 October 1996.

Stewart, T. R., & Lusk, C. M. “Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts”. *Journal of Forecasting*, No. 13, 1994, pp.579-599.

World Bank, “The World Bank’s unified survey projections: How accurate are they? An ex-post evaluation of US91-US97”. Verbeek, Jos. December 1998.