Eurostat detailed replies on the revisions of the handbook following its review

This document presents the review of the handbook performed during November 2011 - January 2012 by the five reviewers designated by the DIME.

Before the DIME plenary meeting of 2012, Eurostat gave feedback on the comments provided by reviewers. This feedback was of three kinds:

- *feedback in red on white background* means an answer to comments to clarify misunderstandings, to point out issues already covered in the handbook etc. No further actions were foreseen.
- *feedback in red on yellow background* means that the revisions were planned to be implemented. In some cases, Eurostat specified how they would be implemented.
- *feedback in red on grey background* means that the revisions/developments were left aside for future projects. No further actions were foreseen for the present handbook.

Now, before the DIME plenary meeting of 2013, Eurostat adds *feedback on green background* on how revisions have just been performed in the handbook.

Review by Julia Aru (Statistics Estonia)

Summary conclusion: Adoption after minor revisions

I found the handbook very practical and potentially suitable to become a standard in the ESS. Although the review criteria stated that the handbook should be judged not by its statistical content but purely by its ability to become a standard, I found that some additions/revisions would contribute to this ability:

1. *Indirect sampling.* Chapter 4.1 overviews many survey design used in practice, including indirect sampling used in Estonia. Unfortunately, not much is told about variance estimation for indirect sampling. The applicability of the Handbook would increase if more details are added.

   There are few paragraphs on indirect sampling and related variance estimation in chapter 4.1. To give more practical information on how to estimate variance for indirect sampling and to consider placing this information in chapter 4.4.in order to not overload (disproportionally) chapter 4.1.

   We added two pages on variance estimation for indirect sampling in section (currently numbered) 3.4.

2. *Chapter 4.3 and Table 7.3.* The table is very practical and is valuable part of the Handbook. Although it is not clear, whether and how the methods mentioned in the table can be applied if design is more complex, e.g. systematic sampling. Also, whether some complex design can be approximated with those mentioned in the table. Is the case of unequal selection probabilities covered?

   Table 7.3 lists suitable and unsuitable methods for the main sampling designs and type of statistics. The first attempt was to list methods for more 'detailed' categories of sampling designs but we renounced. It would be too cumbersome and probably not possible to propose in the same table methods for each sampling design with different specific particularities (in terms of e.g. unequal selection probabilities, systematic sampling, rotating schemes etc.). Methods to account for unequal selection probabilities, systematic sampling (with and without implicit stratification), rotating schemes, etc are presented in chapter 4.4.

Other possible revisions are marked under points 3, 6 and 8.

1) *Is it clear what business process steps and which statistical domains are concerned by the Handbook?*
Yes

2) *Does the Handbook address current and common problems in statistical domains? Does the Handbook answer to the problems identified?*
Yes, problems addressed in the Handbook are very important. Handbook gives good overview of variance estimation methods, especially for surveys in official statistics, but minor revisions might be useful (see above).

3) *Is it easy to identify in the Handbook the distinct topics for which recommendations are formulated? Do the summaries of the chapters allow for a quick and correct understanding of the content and recommendations of the chapters?*
Summary chapter of recommendations would make them more visible. Chapter summaries are good.
In the Handbook, chapter summaries include the main recommendations. There will be an executive summary (of the Handbook) reuniting all main recommendations.

4) *Do the formulations clearly distinguish between recommendations and rules?*
Yes

5) *Are the recommendations based on the consolidated results of science, technology and experience?*
Yes, since the authors are well-known experts.

6) *Does the Handbook take account and ensure coherence with relevant existing related sources of information in the ESS and outside ESS (e.g. other handbooks, glossaries, reference textbooks)?*
As mentioned in chapter 4.6, Eurostat developed a SAS macro for variance estimation of Laeken indicators which uses a jackknife method. This macro is freely available for member states and used to compute variance estimates in some of them (e.g. in Estonia). But as stated in Table 7.3 and several other places throughout the Handbook, jackknife method is inconsistent for non-smooth statistics such as Laeken indicators (except for Gini coefficient). A comment on this contradiction would be useful.

At the end of chapter 4.6 and in chapter 5.1 (page 91), to specify that the use of jackknife method is a feasibility study and that a following step and objective will be to test other methods of variance estimation, especially bootstrapping method.

To give the following reference and to add this reference in the references chapter:

Eurostat (2011). *Working Group meeting "Statistics on Living Conditions". Processing of design variables for variance estimation.* Luxembourg, 11-13 May 2011. Done.

In sections (currently numbered) 3.6 and 4.1, we now specify that the use of jackknife for EU-SILC was a feasibility study. The following step was to test other methods than jackknife i.e. bootstrap and linearization. Comparative experiments were carried out on a limited number of countries and results of different methods are similar. The present choice is to work with linearization (ultimate cluster approximation) which was discussed at the Net-SILC2 workshop on accuracy and validated by the SILC Working Group. This approach provides acceptable results given administrative considerations.

7) *Does the Handbook also indicate bad practices (in addition to good practices)?*
Yes, this is very useful in practice.

8) *Does the standard support horizontal and vertical integration?*
Yes, Handbook supports both horizontal and vertical integration.
Horizontal (across surveys): at the moment, precision requirements and variance estimations strategies are not well harmonized across the surveys coordinated by Eurostat, and the Handbook will greatly contribute to this harmonization. Proposed metadata template will be useful for this harmonization.
Comments on metadata template:

- Consider adding a free text section for description of sample design in addition to structured content - for the reader and writer to be sure all relevant information is included;
  To add question C.20 on 'Additional information of the sampling design', with a free text section for the response. To renumber the following questions of the metadata template. Done (Appendix 7.3).

- points 15-17: is it sampling method for sampling units or for observational units? (case of indirect sampling, when these are different)
  To add in questions 15 and 17, that the sampling method is for the sampling units. Done for both questions (Appendix 7.3).

Vertical (between Member States and Eurostat):
For instance, do the recommendations on precision requirements take on board both national and European possibilities, needs and aims? Do the recommendations on calculation of national and European standard errors take account of the need to reduce duplication of work and achieve better comparability of results between Eurostat and NSIs?
Yes, proposed integrated approach serves these aims.
Reducing duplication of work and overall better harmonization of variance computations, as well as support of Eurostat are very important especially for small member states (such as Estonia) having limited human resources. Moving towards integrated approach described in the Handbook would be very welcomed, but clearly more precise guidelines than those in the Handbook will be needed taking account of complex designs used in countries.

9) Are the proposed recommendations easy to implement?
Difficult in some cases, since Handbook contains mainly references for articles and books, but academic literature is not always available for methodologists at small NSIs. Consider adding more details instead of references.

For future projects
Actually, the initial plan was to issue general recommendations by the Task Force on the basis of few available resources. But in the way, out of the desire to include useful detailed technical guidance on specific issues, the report developed extensively and became a Handbook. The members of the Task Force have contributed nevertheless on voluntary basis, unlike other initiatives, based for instance on grants of substantial amounts, that have the same result - production of handbooks. We even received strong comments that the Handbook has become a compendium, despite the initial plans.
The number of the issues covered by the Handbook is very high, and developing specific issues in future projects dedicated to those issues is more feasible.

10) What is the expected impact on the quality of statistical output?
The quality of output will definitely increase if harmonized methods are applied for variance estimation at Eurostat and NSIs. Several aspects of quality will further increase, if an integrated approach is accepted and NSIs can have a single set of resampling weights for variance estimation, which can be released to national users.

11) What is the expected impact on the efficiency of statistical production processes?
Harmonization of variance estimation methods between Eurostat and NSIs will bring substantial gain in efficiency, especially for small NSIs with limited human resources.

Review by Harm Jan Boonstra (Statistics Netherlands)

1. Summary and conclusion

The ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys discusses a wide range of topics related to precision of survey estimates. Among these topics are ways to formulate precision requirements, assessment of compliance with such requirements, variance estimation methods that account for sampling design, calibration and other processing steps, software tools for variance estimation, and approaches to increase the availability of variance estimates for EU statistics.
Some of the topics discussed are accompanied by definite recommendations, but for many topics recommendations remain very general and sometimes not very clear. This Handbook seems more useful as a broad discussion paper rather than as a *standard* on variance estimation for household surveys. Several suggestions for improvement are given below. In order to fulfill its aim of becoming a standard in the ESS the Handbook requires major revisions.

Summary: Adoption after major revisions

2. General comments with regard to the review guidelines

The Handbook can be very useful as a broad discussion paper, but in my view it is, in its current state, not a *standard* on variance estimation for household surveys. Though the Handbook provides several recommendations, many of them are not very specific. In particular, it will not be easy to implement a production strategy for variance estimation based on this Handbook. The Handbook also needs some editing. There is considerable overlap between some sections, and the level of detail varies quite a lot.

We reorganized the information on variance estimation of sections 3.2, 3.3 and 3.4 (previously numbered 4.2, 4.3 and 4.4) and removed unnecessary duplications.

Now, section 3.2 deals only with sources of variability and does not include any more a description of the main groups of variance estimation methods.

Only section 3.3 includes the description of the main groups of variance estimation methods.

Section 3.4 proposes some specific methods to account for different sources of variability. To avoid discussion on variance estimation for rotating samples both in sections 3.4 and 3.7, we now regrouped this information and we have it only in section 3.7.

We performed different other editing. Some examples:

- We now present Raj's formula only in section 3.3, and not anymore also in the Appendix 7.5.
- We now present the types of longitudinal surveys only in section 3.7 and not anymore also in the Appendix 7.2.
- We eliminated some repetition of the main recommendations in the Appendix 7.2.
- We have reduced text on SUDAAN in Appendix 7.5.

  Etc.

Mathematical notation could be more uniform as well.

We changed mathematical notations when needed for harmonization purposes.

1) *Is it clear what business process steps and which statistical domains are concerned by the Handbook?*

Yes.

2) *Does the Handbook address current and common problems in statistical domains? Does the Handbook answer to the problems identified?*

The Handbook addresses many common problems encountered in the field of variance estimation for estimates based on surveys. However, the Handbook does not address the following important issues:

a. No definite criteria for appropriateness of variance estimators are given. Most of the time (approximate) design-unbiasedness is emphasized and only at a few places the stability of variance estimators is mentioned as a concern.

   To include in chapter 4.2 a discussion about the criteria for appropriateness of variance estimators and to define them.
   To update accordingly the introduction and the summary of the chapter 4.2.

   We have worked on the comparative assessment of methods on criteria related to applicability, accuracy and administrative considerations. See the new sub-section 'Evaluation criteria of variance estimation methods and related recommendations' in section 3.3. From the point of view of accuracy, we have now three criteria named as follows: confidence interval coverage probabilities, unbiasedness and stability (according to Wolter, 2007) and defined in the same sub-section. We present results from different studies on how well replication methods and Taylor linearization (TS) perform on these criteria. The conclusion is that different studies show very good results of BRR when it comes to confidence interval coverage probabilities (the most relevant accuracy criterion). Some studies show very good results of TS when it comes to stability (MSE). However, the best variance estimation method is not obvious in terms of the stability and bias criteria. As regards GVFs, there is very little theoretical justification and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability (lower variance) to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion (Wolter, 2007, chapter 9.1).

We formulate the main findings both in the section and in the section summary.

Two footnotes on page 128 define the concepts of consistency and unbiasdness (as in Cochran, 1977). To consider therefore removing created redundant information, after addition of the definitions of the criteria mentioned above.
We removed the footnotes.
Chapter 4.4 includes the comparison of variance estimation methods in terms of consistency and stability. Additional amendments of the text will concern what is meant by consistency and stability. For more information, see answer to comment 6.10 made by reviewer Paul Smith (ONS).

b. The effect of household clustering on variance estimation. I could not find anything specific about the effects on variance estimation of observing all members of a household.

To include in chapter 4.1 a brief discussion about the effects on variance estimation of selecting all or more individuals from a household instead of one. The recent work on this issue on the European Safety Survey can be used to feed this discussion.
We added one page on this in section 3.1.
To update accordingly the summary of the chapter 4.1. Done.

c. Model-based estimation, as e.g. in small area estimation including MSE estimation and more generally the model-based literature on variance estimation (prediction approach, conditional on the realized sample). I think at least some links to the literature on these topics should be provided.

This goes much beyond the purpose of the TF. However, some references to the model-based literature on variance estimation can be added for instance in chapter 4.2. See the related comment below, where the book by Vaillant, Dorfman and Royall, which has an extensive chapter on variance estimation, is proposed. We included a reference to the book in the beginning of section 3.3.

3) *Is it easy to identify in the Handbook the distinct topics for which recommendations are formulated? Do the summaries of the chapters allow for a quick and correct understanding of the content and recommendations of the chapters?*

The recommendations given in the Handbook are often not very specific, especially concerning variance estimation methods and software tools (see detailed comments below). As a consequence, many of them are not easy to implement.
Some recommendations are specific and could easily be adopted, e.g. those on the formulation of precision requirements.
The purpose of the Handbook is to give general recommendations that should be made specific and implemented by domain specialists for particular household surveys.
Still, the recommendations have been made specific where possible in order to be of concrete help for domain survey managers and methodologists. Examples concern the use of precision thresholds in regulations, the use of specified precision measures depending on the type of indicators, the standard formulation of precision requirements etc. However other recommendations remain general. One example concerns the best variance estimation method and software tool for a particular sampling design and type of statistics. There is simply no unique appropriate method or tool. We identified a list of appropriate methods (good practices), as well as a few not suitable methods (bad practices), by sampling design and type of statistics, and we also reviewed characteristics of tools. The specific choice will be made by domain specialists in NSIs and in case an integrated approach or fully centralized approach for variance estimation will be implemented using a common replication method, the specific choice will be made jointly by NSIs and Eurostat.

4) *Do the formulations clearly distinguish between recommendations and rules?*

The Handbook contains several recommendations, but no rules as far as I can see. This is fine.
Yes, there are no rules.

5) *Are the recommendations based on the consolidated results of science, technology and experience?*

Most of the time, yes. Some exceptions are mentioned below in the detailed comments.
Your detailed comments are being addressed; see below.

6) *Does the Handbook take account and ensure coherence with relevant existing related sources of information in the ESS and outside ESS (e.g. other handbooks, glossaries, reference textbooks)?*

The Handbook contains many useful links to the literature. Personally, I would have liked to see some references to the model-based literature on variance estimation as well, e.g. to the book by Vaillant, Dorfman and Royall, which has an extensive chapter on variance estimation.
As stated above, this issue goes much beyond the purpose of the TF. However, some references to the model-based literature on variance estimation will be added for instance in chapter 4.2.
We included a reference to the book in the beginning of section 3.3.

7) *Does the Handbook also indicate bad practices (in addition to good practices)?*

The Handbook lists a lot of good and also some bad practices in variance estimation, which is good.

8) *Does the standard support horizontal and vertical integration?*

Concerning the formulation of precision requirements I think that this Handbook will improve both horizontal and vertical integration.
It is not clear to me that this Handbook would contribute substantially to harmonization of variance estimation across surveys and countries. The recommendations do not really point to a small set of well-established and broadly applicable variance estimation methods (see the extensive list of adequate variance estimation methods in section 7.3). Probably this is not possible anyway, given the large differences, among surveys and especially countries, in availability of auxiliary information, sampling design, imputation and calibration methods, etc.
I agree to your last phrase.
However, for the survey domains where the integrated or the fully centralized approach will be applied on the basis of a common replication method (to be decided by the specific survey managers) this harmonization will be achieved (see chapter 5).

9) *Are the proposed recommendations easy to implement?*

Since some recommendations are not very specific, they will not be easy to implement. The Handbook also lacks detail in order to serve as a reference book that can be used for implementation purposes, although it contains many links to the literature.
See the answer to your comment to criterion 3, above, about the purpose of the Handbook.
Actually, the initial plan was to issue general recommendations by the Task Force on the basis of the few available resources. But in the way, out of the desire to include useful detailed technical guidance on specific issues, the report developed extensively and became a Handbook. The members of the Task Force have contributed nevertheless on voluntary basis, unlike other initiatives, based for instance on grants of substantial amounts, that have the same result - production of handbooks or guidelines. We even received strong comments that the Handbook has become a compendium, despite the initial plans.
The number of the issues covered by the Handbook is very high, and developing specific issues in future projects dedicated to those issues is more feasible.

10) *What is the expected impact on the quality of statistical output?*

Even though not all recommendations are easy to implement, if some of the mentioned good practices are followed up, the impact on the output quality can be expected to be positive.

11) *What is the expected impact on the efficiency of statistical production processes?*

The quality of output may improve by following some of the recommendations given. However, this will require possibly substantial effort. For example, it will not generally be easy to account for multiple sources of variability of an estimator.
Chapter 4.2 describes the multiple sources of variability, but gives the following guidelines:

- Consideration of all possible sources of variability;
- **Consideration of which sources of variability can be estimated;**

- **Consideration of which sources of variability can be described with some other indicative information** (for example, level of processing errors).

3. Detailed comments

3.1 Section 2.1

- p.4: I think the last item on gross flows should be removed from this list since it is not a recommendation

  To remove the last bullet point Done (in the introduction).
- p.9, eq. (2.1.3): here and in the text below, z_{1-alpha} should be replaced by z_{1-alpha/2}

  To make this amendment Done (section 2.1).

3.2 Section 2.2

p.14, half page of text above the summary: not clear what is meant here. It seems to me that if used well, standard error and CV should lead to the same sample size requirement, whatever the value of the proportion (unless exactly zero where CV is undefined). If used well, standard error and CV should lead to the same sample size requirement, whatever the value of the proportion (unless exactly zero where CV is undefined).
That text refers to the use of a **fixed** threshold of either standard error (SE) or coefficient of variation (CV), in the regulations.
Let us consider (SRS):
a) T0: P=0.5; SE=0.0071; CV (derived from SE and P) = 1.4% => needed n=4959;
b) **T1: P changes:**

1. P=0.4; **fixed SE**=0.0071; CV (derived and different than in T0)= 1,8%=> needed n = 4761.
2. P=0.4; **fixed CV**=1,4%; SE (derived and different than in T0)=0.0057 => needed n=7387

One should fix in the regulation **either** a SE threshold or a CV threshold. One cannot fixed both, because it is not possible to meet both fixed requirements when P changes. As you see, when P changed, keeping a fixed SE required a lower sample size in T1, while keeping a fixed CV required a higher sample size in T1.

That text above the summary refers to the use of a fixed threshold of either standard error (SE) or coefficient of variation (CV), in the regulation.
The use of a fixed threshold of SE demands the highest sample size when P=0.5 and lower and lower sample size as P increases above or decreases below 0.5.
On the other hand, the use of a fixed threshold of CV demands the highest sample size for low P and lower and lower sample sizes as P increases from the theoretical 0 to 1.
As P changes over time, the choice between a fixed SE and a fixed CV therefore has a direct impact on the sample size (effort) required to attain **over time** the established **fixed requirement** in the regulation.

Figures 2.2.1 and 2.2.2 cannot be directly compared since a standard error of 0.5% is not the same as a CV of 5% at p=0.5.
The purpose of each graph is to show the change in the sample size when P changes over time, for a fixed CV (one graph) and a fixed SE (the other graph).

- p.14: The last sentence above the summary seems to contradict the next-to-last sentence in the summary box.

  To delete the last sentence above the summary Done (section 2.2).

Concerning the recommendations about using CV or standard error, it is in my opinion more important to distinguish between statistics that might be close to zero or negative and those that are not, instead of distinguishing between continuous and count variables.

The recommendations about using CV or standard error take account of both aspects. The arguments for distinguishing between continuous and count variables are in the Handbook.
Besides the '0' problem you mention, a problem specific to the use of CV for proportions (and not for continuous variables) is that the CV for a proportion is not symmetrical for P and 1-P, as the logic would require.

3.3 Section 2.3

- p.15, eq. (2.3.1): here and in the text below it, z_{1-alpha} should be replaced by z_{1-alpha/2}

  To make this amendment Done (section 2.3).

- p.16: below Fig. 2.3.1, 'There is no evidence that...'. It might be useful to distinguish between predata (design) and postdata (estimation) stages. With the data at hand, conditioning on the observed data (and in particular on observed domain sizes) is to be preferred since it yields inferences better suited to the actually observed sample. See e.g. Holt and Smith's 1979 Post-stratification paper. Before sampling, at the design stage, the unconditional variance can be useful. A reference to the Holt and Smith paper might be appropriate and perhaps to the model-based literature, e.g. the Valliant, Royall, Dorfman book, since model-based inference is always conditional on the actually observed sample including realized domain sizes.

  For future projects

- p.16/17: the next sentence ('In fact, when the relative size...') still refers to the example with sample size 8000? This may be stated more clearly.

  Yes, for the sample size 8000. To be made clearer. We did so (section 2.3).

- p.17, 2nd full paragraph: this paragraph is about allocation so I think 'unplanned domain' should be replaced by 'planned domain' (twice)

  We can have unplanned domains in terms of age, gender, education levels, etc. We do not use stratification at the design stage to fix the sample size for these domains. But we can **roughly** calculate how much sample sizes **would** result by age, gender, education, if the random overall sample structure mirrors the population structure, and we can calculate the expected precision sizes for these expected sample sizes.

- To delete the phrase 'The allocation methods used could be… Osier (2010))'.

Done (section 2.3).

- To replace the phrase 'This consists … among the domains' with: 'This consists of estimating sample sizes that would result if the random national sample structure mirrors the population structure by domain.'

We revised and now we have: “This consists of estimating the sample sizes that would result if the national sample structure mirrors the population structure by domain. Precision measures can be calculated for unplanned domain estimates in order to provide an interpretation of the expected domain sample sizes in terms of statistical accuracy. “

3.4 Section 2.5

- p.25, eq. (2.5.1): replace Z(alpha) by z_{1-alpha/2}

  To make this amendment We did so (section 2.5).

3.5 Section 4.2

- p.35, between eqs. (4.2.2) and (4.2.3): instead of calling bias and variance *measures* of accuracy it would be better to call them *components* of an accuracy measure (MSE)

  To make this amendment. When referring to variability, to replace 'is another component' with 'is the other component'. We did both changes (section 3.2).
- To the list of sources of variability might be added: errors in linking survey data to auxiliary data (population frame, registrations, ...). Missing links lead to additional missing data (or imputations) whereas wrong links are in a sense similar to measurement error and lead to a loss of efficiency in e.g. calibration.

  I think that these are included in measurement variance, sampling variance (as regards calibration), imputation variance. It is not in the purpose to detail on integration of survey and administrative data.
- p.40: Here and elsewhere it is not clearly explained what is meant by the exact analytic method. I guess that this is the case where the design variance itself has an exact closed-form expression, and an exactly unbiased variance estimator for it is available in closed form too?

  Chapter 4.4, page 56, explains the exact analytic methods, as you describe them. To add a few lines on this on page 40. We now describe analytic methods (exact and approximate analytic methods) only in section 3.3 (following reorganization of information).

3.6 Section 4.3

- some unnecessary overlap with section 4.2: e.g. classification of variance estimation methods on p.42

Chapter 4.2 contains the general classification of variance estimation methods while Chapter 4.3 groups the methods by sampling design.

We reorganized the information in the sections (currently numbered) 3.2, 3.3. and 3.4; we have now classification and description of variance estimation methods only in section 3.3.

- p.43 under stratified simple random sampling, third item, 'unequal probabilities': the text below suggests that unequal probabilities refer to selection within strata. But then it is no longer stratified simple random sampling.

  To delete the word 'simple' from 'stratified simple random sampling'
  We do not have anymore the related text, following reorganization of the information on methods.
  Moreover, the difficulty to compute exact joint inclusion probabilities is not specific to jackknife.
  Please see the generalised jackknife variance estimator, pages 5-6, from Berger Y.G. (2007). *A jackknife variance estimator for unistage stratified samples with unequal probabilities*. Biometrika, Vol. 94, No 4, 953-964
- p.43, 3rd line of last paragraph: 'first order-estimators'??

  This expression exists and can be retrieved in many documents. However, we can simply replace it by 'statistics' in the text to facilitate the understanding. We did so (section 3.3).

3.7 Section 4.4

- p.46, 1st sentence of 2nd paragraph: 'not missing at random and that the probability...' should be replaced by 'missing at random.'

  For clarity, to reverse the two parts of the first phrase from that paragraph, as follows: "Under the assumption that the values of the study variables for the non-respondents are not missing at random and that the probability of response is related to the study variable, response homogeneity groups can be formed within which the net sample size can be used for variance estimation instead of the gross sample size". We did so (section 3.4).
- p.47, two lines above Multiple listings item, '...domain estimation can also be used.' As far as I can see eq. (4.4.2) is already based on domain estimation.

  To move the two lines and place them before the text starting with 'In order to…' and to delete the word 'also' from the two lines. Done (section 3.4.)

- p.50, eqs. (4.4.16) and (4.4.17): please check; the variance expressions do not seem to scale correctly with n (consider the equal probability case p=n/N).

  To check with an external expert or with Guillaume Osier. We checked with Mr Osier and in Wolter, 2007 (page 336): the formulae are fine. We modified the notation of variance to make it conform with the other estimators for systematic samples (section 3.4).
- p.51: these equations look unnecessarily cluttered due to the double index notation for systematic sampling. Since the equations are more generally useful for unequal

probability sampling designs and to avoid confusion with second-order inclusion probabilities it might be better to use pi_j instead of p_i,j here (in line with the notation exposed in the appendix).

To harmonize notation throughout pages 48-51 and make the distinction between second—order inclusion probabilities and unequal probabilities. To note that the Appendix 7.1 defines π_{ij} as the second-order inclusion probability. To note that notation for these probabilities appear also in Appendix 7.4. Therefore to check/edit them where needed in the Handbook.
We harmonized notations.
We have now only one subscript 'i' for the inclusion probabilities in the formulae on variance estimation for unequal probability sampling (section 3.4).

- p.54, 2nd/3rd lines of 2nd paragraph and above eq. (4.4.30): the reader is warned about ignoring design covariance terms between disjoint groups, but I cannot think of a practical situation where such a term would be important since for household surveys n << N. It seems to me that the dominant covariance term comes from the panel component, and a possibly non-negligible second-order component may arise due to rotation within primary sampling units (as mentioned below eq. (4.4.30)).

The literature argues the consideration of the covariance term between the rotation groups obtained **from splitting one sample in two**.

- p.54/55 around eq. 4.4.31: this equation is not clear; is this an example of rotation within primary sampling units? But the additional term in (4.4.31) seems negligible (1/N) and based on simple random sampling. Moreover it is not clear how the first term on the rhs of this equation is defined.

To check with an external expert or with Guillaume Osier. We checked with Mr Osier: at his suggestion we removed the formula and the accompanying text (section 3.7.1).
p.56, below eq. (4.4.32): for regression estimation the 'g-weighted' variance estimator is recommended. A little more background information might be useful here, e.g. that it is an adjustment of the variance estimator based on first-order Taylor linearization, and that it better accounts for dispersion of the GREG weights (and thereby may reduce the downward bias of the naive linearization variance estimator).
To consider the incorporation of the above suggestions. See the related comment made by the reviewer Laszlo Mihalyffy (Hungarian CSO) and the planned revision.
We added these suggestions (section 3.4).
Also, it might be worth mentioning this definite recommendation in the summary at the end of the section.
But if we include this recommendation in the summary at the end of the section, then we end up with some disproportionate consideration in the summary chapter of the recommendations of specific methods (which are many in the chapter) .
We put all recommendations in italic in the text of the chapters (including this one). We put the main recommendations in italic in the summary of the chapters.

- p.59, 3rd line in 7th bullet, 'In this case, linearization can be used...followed by e.g. analytic methods'?: I do not understand this, since this is about complex non-smooth statistics, for which linearization methods would normally not be available.

  From Appendix 7.3, for non-smooth statistics (Gini coefficient, functions of quantiles etc.), we can use:
  □Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997)
  □Generalised linearisation method relying on the concept of influence function, Deville (1999), Osier (2009)
  From pages 57-58: In order to accommodate statistics with a more complex mathematical expression (e.g. quantiles, etc.), generalised linearisation frameworks have been developed:

  - Linearisation based on estimating equations (Binder, 1983; Kovacevic and Binder, 1997): this approach can for instance handle the Gini coefficient, which the Taylor method cannot;
  - Linearisation based on influence functions (Deville, 1999): this framework can encompass nearly all parameters of statistical interest. See also Osier (2009) for an intensive application of this approach to the EU-SILC indicators.

- p.60, 1st bullet: 'bootstrap samples can only be generated for two selected sample sizes'; not clear to me

  To reformulate the bullet point as follows:
  "In planning a sample survey, a pilot sample is typically used for determining the required sample size to achieve a specified level of accuracy for statistical inference. Mak T. K. (2004) proposes a practical procedure based on bootstrap for estimating simultaneously the variances of a statistic for all sample sizes based on a single observed pilot sample. For an observed sample of size n_0, it is well known that the bootstrap can be used to estimate numerically the variance of a sample statistic computed from the sample. To study the variances of the statistic for other sample sizes, one can in principle generate bootstrap samples of size n for a range of values of n, and then calculate the bootstrap variance estimate for each n. This, however, will be computationally demanding and inefficient. In contrast, the method proposed requires bootstrap samples be generated for only two selected values n_1 and n_2 of n. Estimates of the variances of the statistic with small biases can then be computed for any other values of n. It is proved theoretically that these biases decrease rapidly to zero as n_1 and n_2 increase". We did so (section 3.3).

  To include the following reference in the references chapter: Mak, T.K. (2004). *Estimating variances for all sample sizes by the bootstrap*. Computational Statistics and Data Analysis, Vol. 46, Number 3, June 2004, 459-467. Done.

- p.60, 3rd and 4th bullets: first it is mentioned that BRR and bootstrap perform consistently better than jackknife and linearization (where performance implicitly seems to refer to bias of the variance estimators). Then the next bullet says that the order of performance in terms of 'stability' is just the opposite.

Overall, this does not say much about the relative performances of the methods, and the reader may wonder if anything at all can be said about total error or MSE of the variance estimators?

Please see the answer to the related comment n. 6.10, made by the reviewer Paul Smith (ONS) in his document.

We have worked on the comparative assessment of methods on criteria related to applicability, accuracy and administrative considerations. See the new sub-section 'Evaluation criteria of variance estimation methods and related recommendations' in section 3.3. From the point of view of accuracy, we have now three criteria named as follows: confidence interval coverage probabilities, unbiasedness and stability (according to Wolter, 2007) and defined in the same sub-section. We present results from different studies on how well replication methods and Taylor linearization (TS) perform on these criteria. The conclusion is that different studies show very good results of BRR when it comes to confidence interval coverage probabilities (the most relevant accuracy criterion). Some studies show very good results of TS when it comes to stability (MSE). However, the best variance estimation method is not obvious in terms of the stability and bias criteria. We formulate these conclusions both in the section and in its summary.

- Where BRR is mentioned, it is not clear what variant of BRR is considered. Is it BRR with Fay's modification, artificial strata, or grouped BRR, or some combination of these? Or is it only considered for two-per-stratum designs?

  It is the usual definition of the BRR, which is given on page 41.
- p.62, Summary, 1st open dot: see remark addressed at p.59

  See answer given above.
- p.62, Summary, 2nd open dot: as far as I know this point has not been mentioned before. (It is mentioned later, in section 4.5 on the same page).

  To mention in chapter 4.4 that the delete-1 jackknife should not be used in stratified sampling and to give the reference: Wolter, 1985, pp. 174-175.

  We added this point in the text of section 3.3 (within the sub-section on jackknife). And we have it also in Appendix 7.4.

  We made the reference to Wolter (2007) instead of 1985. We have removed Wolter (1985) from the list of references.
- p.62, Summary, 3rd open item: see earlier remark. Can any recommendation be made?

  Please see the answer to the related comment n. 6.10, made by the reviewer Paul Smith (ONS) in his document.

  We included in the summary of section 3.3: "(…) Other criteria for the choice of methods are accuracy (confidence interval coverage probabilities, unbiasedness and stability) and administrative considerations (time, cost, simplicity). With respect to accuracy, different studies show very good results of BRR when it comes to confidence interval coverage probabilities (the most relevant accuracy criterion). Some studies show very good results of TS when it comes to stability (MSE). However, the best variance estimation method is not obvious in terms of the stability and bias criteria". As regards GVFs, there is very little theoretical justification and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability (lower variance) to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion (Wolter, 2007, chapter 9.1). We put these findings in italic, as these can be taken as recommendations for the choice of methods.

3.8 Section 4.5

- p.63, item R package survey: it might be mentioned that the package can also be used for calibration and that variance estimates take that into account

  To add this We did so (section 3.5).
- p.65: variance estimation in Bascula also accounts for calibration (as is mentioned explicitly for other software tools)

  To add this We did so (section 3.5).

- p.65/66: the `GREG variance estimator' is mentioned several times, and it would be nice to link to the discussion in section 4.4 (p.56), although that discussion is currently very brief.

  To make the link
  We renounced to make the link as that discussion is very brief.

3.9 Section 4.7

- p.72-76, subsection 4.7.1, starting from Case 1: this part is not in balance with the rest of the handbook in that it is very detailed. Moreover, the notation is not in line with that in other places, e.g. the 'g-weights' are here denoted by k_{hqi}. The method described here is presented as an alternative to equation (4.7.1.2 and 3) on p.71. But it is not clear to me how it deals with correlation among the panel part of the sample.

  This detailed part is however helpful practically.
  I understand that your 'g-weights' are the w_i from page 56.The w_i from page 56 are the w_{qhi} from Case 1 (different from k_{qhi}).
  The w_i (from section 3.4, calibration) are sampling weights after calibration. The k_{qhi} (from section 4.7.1) are adjustment weights which are applied on the design weights to get the sampling weights after calibration. So the different notation is justified by these different things.
  I now understand that you refer to the g-weights as being the adjustment weights which are applied on the design weights to get the sampling weights after calibration (the $k_{qhi.}$).But we do not use other notation than $k_{qhi.}$ (we do not use the notation g in section 3.4.).
  To clarify how to the method deals with the covariance of the overlapping part of the sample, with an external expert or with Mr Ioannis Nikolaidis.
  We are still waiting for clarification from Mr Nikolaidis on how the formulae deal with correlation among the panel part of the sample.
- Section 4.7.2: again the GREG/calibration 'g-weights' are incorporated in the variance formulae, but oddly enough the y's are not replaced by the residuals

arising from the GREG/calibration. It is ok to focus on the estimation of change in this section, but then it might be better not to introduce the k_{thij} at all.

I am not sure if I understand. The 'g-weights' are calibrated weights. According to formula 4.4.32 they are used with 'y's. So in section 4.7.2.
I now understand that you refer to the g-weights as being the adjustment weights which are applied on the design weights to get the sampling weights after calibration (the $k_{qhi.}$).
But in section 3.4. we say “For regression estimation we recommend the 'g-weighted' variance estimator $\hat{V}\left(\sum_{i\in s} w_i u_i\right)$

For calibration (in general) (different methods exist, as presented in 3.4), we have

$$\hat{V}(\hat{Y}_w)=\hat{V}\left(\sum_{i\in s} w_i y_i\right)\approx\hat{V}\left(\sum_{i\in s} d_i u_i\right)$$

:
In section 3.7.2 we have *w* used with *y* as for calibration (in general).

- Some references to the literature on composite estimation of change might be given in section 4.7.

For future projects

3.10 Section 5.1

- The integrated approach to computing standard errors for national and European statistics has a number of advantages and disadvantages as discussed in this section. A disadvantage not mentioned is that datasets including replication weights could become very large. Also, it seems to me that the burden on NSIs and the required expertise is not less than in the decentralized approach, whereas for Eurostat it is higher. Replicate weights can be calibrated, but I'm not sure how e.g. imputation can be accounted for in the replicate weights. Some more research may be required.

Imputation can be incorporated in replication methods if for each replicate the imputed value is recalculated using random imputation methods. We added the formulae for estimating the variance including the imputation effect, under the use of replication methods (see section 3.4.)
To include a bullet point on the burden of NSIs as disadvantage under the use of replication methods Done (section 4.1).
As the aim is to have available the variance estimates for all relevant point estimates and breakdowns, the data transmission would be very cumbersome also for the decentralized approach and the exchange of information and clarification on particularities of all (different) sampling designs from NSIs would be difficult for the full centralized approach.
To delete from page 90 the last sentence Done (section 4.1).

- The GVF approach requires one to fit a GVF model on a set of well-estimated variances, so some full variance estimates have to be computed anyway. Therefore, besides being less accurate, GVF seems advantageous only in the situation that variance estimation is not properly automated. I'm not sure that computation time should be a major issue here, even if variances must be estimated for thousands of statistics.

To include a bullet point on the need to calculate anyway by NSIs the variance using direct methods as disadvantage under the use of GVF. But that however the parameters can be carried over from one data collection to another with similar features (in terms of sampling design, survey variables, etc.). Done (section 4.1).

3.11 Appendix

- p.131/132: text on longitudinal surveys not needed here; is already covered in section 4.7

  It is useful to have the classification and brief description of longitudinal surveys in the glossary.
  We removed it and we added repeated panels to section 3.7.
- Meta-data template: this could be useful for Eurostat to assess compliance with precision requirements. However, using the meta-data template for all household surveys would incur substantial additional overhead for NSIs. Therefore it would be better that Eurostat coordinators fill out the template, as suggested on p.29. Of course, this is only possible if NSIs provide adequate quality reports for each survey.

  Page 29 mentions that Eurostat coordinators for specific domains/survey can decide on how to detail the relevant part of the quality report according to the metadata template (in order to prevent NSIs from reporting the same information twice), but not to fill it in instead of NSIs. Experience with EU-SILC, that tried to implement the fully centralized approach using Jackknife, demonstrated that the quality reports collect description of for instance sampling designs which is very variable from one NSI to another. Many clarifications and exchanges were needed with NSIs on e.g. how many stages does the survey have after all? are there self-representing PSUs or not? is systematic sampling with implicit stratification or not? Etc. The use of the metadata template avoids this problem.
  To mention in chapter 5.1., under the fully centralized approach, that the metadata template is particularly useful for this approach under the use of replication methods, in order to collect clear and detailed information on the sampling designs. To add the above statements on the experience in EU-SILC Done (section 4.1).
  To add on page 133, before the template, that it is particularly useful when a fully centralized approach is implemented in Eurostat using a common replication method. Done (Appendix 7.3).

Review by László Mihályffy (Hungarian Central Statistical Office)

The structure of these comments is as follows:
I. Summary of Conclusions – The Outcome
II. Compliance with the Review Criteria
III. Comments Related to Conceptual Issues
IV. Comments Related to Minor Errors

I. Summary of Conclusions:

Adoption after minor revisions is recommended

The revisions are specified in Sections III-IV. It should be emphasized that in certain cases the formulation and visibility of the recommendation in the Handbook will be concerned by the revision not directly, but through some restructuring and amendment in the text, and correcting some errors (mainly typing errors).

II. Compliance with the Review Criteria

The comments follow the order of the criteria as given in the Guidelines for the reviewers. Comments marked with an asterisk are based on the assumption that the revisions recommended in Sections III-IV have been carried out.

1. In terms of GSDPM and SDMX,
the business process steps 2.4 (Design frame & sample methodology), 2.5 (Design statistical processing methodology), and quality management as overarching statistical process and
the domains 1.2 (Labour statistics), 1.5 (Income and consumption) and 3.3.3 (Information society)
are concerned by the Handbook.

2. The Handbook focuses on current and common problems in the precision issues of household surveys, and provides proper information on how to treat those problems.

3. The topics for which either rules or recommendations are formulated are easily identified. However, distinguishing between rules and recommendations is not always easy, see also next comment. The summaries provide proper information on the contents of the chapters which they belong to.

4. *Do the formulations clearly distinguish between recommendations and rules?* For the members of the TF, the most informative answer to this question would be obtained by summarising or aggregating the findings of the five reviewers in terms of the rules identified by them. The present reviewer thinks that the following guidelines are rules:

- Use minimum effective sample sizes as precision thresholds for leading indicators (Chapter 2.1). The precision thresholds are survey-specific (Ch. 2.4)

Chapter 2.1 states that in regulations 'it is recommended' to express requirements by defining minimum precision thresholds that must be met by the few leading indicators (rather than minimum effective sample sizes). As 'it is recommended' is used, it is a recommendation (not a rule).

To reformulate the first phrase of the last para on page 9, by replacing 'are the prioritised option' with 'are recommended'.

Done. The recommendation is made explicit both in the text of section 2.1 and in its summary.

The fact that precision thresholds are survey-specific is a finding.

- Use *cv* as precision measure for totals and means and use standard errors for ratios, proportions and changes (Chapter 2.2)
  Chapter 2.2 states that 'recommended precision measures' to are the *cv* for totals and means of continuous variables and the standard error for ratios, proportions and changes. Therefore, this is a recommendation (not a rule).
- Put no requirements for unplanned reporting domains (Chapter 2.3)
  Chapter 2.3 states that the survey manager 'should' avoid putting requirements for unplanned (reporting) domains. The guidelines for the reviewers argue that the standard should contain rules ('must to') and/or recommendations ('should do'). Thus, the use of 'should' indicates a recommendation.

  This Handbook contains only recommendations. To check for possible 'must' words in phrases that instructs the survey manager/methodologists on what to do, and replace with 'should'. Done.

  For clarity, maybe it is better that phrases instructing the survey manager/methodologists include also the expression 'it is recommended' besides 'should'. To reformulate therefore the above recommendation and say 'It is recommended that the survey manager should avoid putting requirements …'. The same for other similar cases in the Handbook. This is not always done, because it would add unnecessary redundancy to the phrases. See below what is done.

  We plan to document in the ESS terminology that the difference between 'must' and 'should' distinguishes between a rule and a recommendation.

  The handbook comprises only recommendations (and no rules). 'Should' and 'it is recommended' identify the recommendations.

  But in the handbook we have left words 'should' and 'must' which are not recommendations/rules of the handbook. For example, from section 2.4: "In the EU-SILC, a methodological document (Eurostat, 2001) presents the use of the ‘compromise power allocation’ method to allocate the EU sample size (which **should** not exceed 80 000-100 000 sample households) to countries." This is not a recommendation of the handbook. Another example, from section 3.1: "Invariance means that every time the i^{th} PSU (primary sampling unit) is included in the first stage sampling, the same subsampling design **must** be used". This is not a rule of the handbook.

  Therefore, we put recommendations in italic in the text of the chapters, and the main recommandations in italic in the summaries of the chapters. This would clearly identify the recommendations of the handbook.

- Use the standard formulation of precision requirements (Chapter 2.5)

Chapter 2.5 states that 'it is recommended' to follow a standard formulation of precision requirements for regulations which aims at a uniform and unambiguous understanding within the ESS. As 'it is recommended' is used, it is a recommendation (not a rule).

- Use the metadata template as an element of compliance assessment strategy (Appendix, 7.2)
  Chapter 3 states that the use of the metadata template in Appendix 7.2 of the handbook 'is recommended' when assessing whether the variance estimation method and tool used are appropriate in relation to sampling design used and type of indicators. Therefore, this is a recommendation (not a rule).

All other guidelines strongly recommended or just recommended are recommendations. A source of misunderstanding by the reader maybe that *rules* are also *recommended*.
Remarks. (1) It is not always easy to distinguish rules and recommendations in the Handbook; e.g. some staff members of the Methodology Unit, Hungarian CSO think that the guideline considered in Chapter 2.3 is a recommendation rather than a rule. *With some minor change in the wording this ambiguity could be eliminated.*
Please see the answers above.
Please see the answers above (on green background).

(2) The present reviewer regards the exclamation marks in the matrix (Appendix, 7.3, pp. 145-151) as rules, saying "You must not use bad practices".
At the beginning of Appendix 7.3, where it is mentioned:
" □□Suitable method
   ! Unsuitable method "
to add in parenthesis for both suitable and unsuitable method that these are recommended for use, respectively not for use.

Done (now Appendix 7.4).

5. *In general, the recommendations of the Handbook are based on consolidated results of science and experience. For the sake of completeness, however, insertion of the amendments suggested in Part III is important.
Your comments from section Part III are being addressed.
However, something is worth mentioning. As planned, precision thresholds are set up for the estimated standard errors. The recommendations and rules do not reflect the fact that estimated standard errors (being estimates) have some variability. Therefore, when establishing precision thresholds, the variability of estimated standard errors should be taken into account.

Chapter 2.5, page 21, states that: "The concept of standard error is closely related to survey design since it reflects the expected variability of the parameter estimator (the parameter in this case is the population percentage). Typically, the standard error remains an unknown value which in itself has to be estimated, by using an appropriate estimator (which is called 'estimator of the standard error'). Consequently, 'standard error' should be used in conjunction with 'estimator'. For determining a particular sample and a particular estimated percentage, one can calculate an estimate of the standard error by using an appropriate estimator. Hence, as the requirements are made for the survey output, the formulation refers to the 'estimate of standard error' rather than just to 'standard error'."

Chapter 3, page 29, states that: "some ***tolerance*** should be accepted when comparing the estimated standard errors with benchmarks indicated in the precision requirements. Tolerance should be accepted for the following reasons:

- (…)
- what can be computed is not the 'true' standard error for a given estimate but only an estimated standard error, which in turn has its own variance. Thus, the estimated standard error can be higher than the true one (which is unknown);
- (…)."

6. *The question of coherence with relevant sources of information in the ESS and outside ESS is related to the previous one, thus the answer is analogous.

7. The Handbook definitely indicates bad practices (e.g. in Appendix 7.3)

8. The Handbook properly serves the purpose of the horizontal and vertical integration.

9. Implementing the recommendations of the Handbook at the NSIs is an issue of the available hardware-software environment on the one hand and of the professional skill of the technicians in the staff on the other. At some NSIs probably some extra resources will be needed to set up such an environment and the properly skilled team.

10. As a result of the Handbook, some improvement in measuring the precision of the data of household surveys can be expected.

11. The Handbook provides an overview of the current best practices in terms of software tools of variance estimation, and this enables the NSIs to replace some of their old and less efficient tools by new and more efficient ones.

III. Comments Related to Conceptual Issues

III.1. The word "calibration" occurs on thirty pages of the Handbook, the number of occurrences amounts to about 50[1], yet the information provided on variance estimation in the presence of calibrated weights is not complete. The reader will probably not be able to decide on the proper method and the software if he / she wants to estimate the variance of a quantile in the case where the latter is a calibrated estimate, though the list of the recommended software products probably contains programs suitable for that purpose. To eliminate this deficiency, the following amendments are recommended.

- When introducing the replication methods (Chapter 4.2, pp.40-41), it should be stressed that
  - – the replicates are to be determined by the same algebraic expression as the original estimate, and
  - – in case of calibrated estimates this means that the sample weights should be re-computed with calibration whenever a new replicate is determined.

  Ok. To be done.

  Done (section currently numbered 3.3).

[1] These numbers do not include those in the References and the metadata template (Chapter 7.2)

Besides, the sections describing these methods ought to be more informative; inserting the basic formulae in the simplest case is recommended.
Ok.
To insert the formula for each replication method.
Done. We present the replication methods, their formulae and characteristics in section 3.3. We reorganized the sections 3.2, 3.3. and 3.4 and eliminated redundant information.

Moreover, to replace on page 101 the general formula for replication methods (which is now particularized for bootstrap only) with the formula from: http://www.ecb.europa.eu/home/pdf/research/hfcn/varianceEstimation.pdf?2f26564a9ef2da62419ded2167470ee7 (page 6), to mention this document as reference and add it in the references chapter.

Done (section 4.2 and section 6 on references).

- The section *Methods to account for variability caused by use of calibration* (Chapter 4.4, p. 55-56) should be augmented in the following way.
  – A concise but more detailed presentation of the calibration procedure as a problem of minimisation is needed. The concept of the distance function is important and note that the generalised regression estimator (GREG) associated with the distance function $\Sigma(w_i – d_i)^2/d_i$ is not the only calibration method (raking, etc.).
  – When referring to the Deville-Särndal principle, emphasize that it applies to the case where the original statistic is linear (totals, means, etc.)[2]. Replacing the nonlinear expression $\Sigma w_i y_i$ in
  $$\hat{V}(\hat{Y}_w) = \hat{V}(\sum_{i \in s} w_i y_i) \approx \hat{V}(\sum_{i \in s} d_i u_i) \qquad (4.4.32)$$
  by the linear expression $\Sigma d_i u_i$ comes from linearising the regression estimator obtained as the result of calibrating with the GREG; linearising here means that the regression coefficients are replaced by their sample estimates.
  – If the statistics is nonlinear and smooth (ratio estimator, regression coefficients, etc.) the estimator should be linearised with Taylor series approximation. In the next step, the Deville-Särndal principle can be applied to the linearised estimator. If the estimator of interest is just the regression estimator, linearising with Taylor series reduces to replacing the regression coefficients by their sample estimates, implying 4.4.32 immediately.
  – In case of non-smooth and nonlinear statistics (Gini coefficient, quantiles, etc.) linearisation based on influence functions or estimating equations can be used instead of Taylor' linearisation. Hereafter the procedure is the same as the preceding one.
  If some replication method is used, variance estimation based on the Deville-Särndal principle and linearisation uses by orders of magnitude less computing time than full reweighting in any of the above three cases.
  – Variance estimates of calibrated estimates determined by methods other than those described above may have considerable upward bias.
  – Using calibration as a means to compensate for nonresponse can be recommended only if rate of nonresponse amounts to a few percentages or if response probability of the units in the universe can be modelled (see Särndal, C.-E. (2007): *The*

[2] This condition is mentioned on p.154 (Chapter 7.4, Adequacy of sofwate products…)

calibration approach in survey theory and practice. Survey Methodology, Vol. 33, No. 2, pp. 99-119, and Bethlehem, J. and Schouten, B. (2004): *Nonresponse adjustment in household surveys*. Discussion paper 04007, Statistics Netherlands). Though these conditions are rarely fulfilled in our days, this kind of using calibration is quite common (.”For example, calibration estimators /Deville and Särndal, 1992/ are often used in practice to deal with unit non-response, thereby increasing sample efficiency…” Chapter 7.1, p.129.).– This application of calibration is probably not a specimen of best practices, yet it is something that should be accepted, since stopping it would result in even larger bias than the current practice.

To make these amendments, add the references and include the references in the references chapter.
Done (section 3.4 and section 6 on references).

III. 2. Part One of the Handbook[3] (Chapters 1-6) contains the Introduction and the presentation of topics such as precision limits in household surveys, ways of assessing compliance with those limits, variance estimation methods and suitable software that can be recommended, etc. Part Two of the Handbook[4] is the Appendix (Chapter 7) containing a Glossary of statistical terms and a so-called matrix providing information on the adequacy of different estimators and software tools in an environment determined by the sample design and the type of parameters to be estimated. Part One resembles a *concise* textbook and Part Two resembles a directory providing quick information on the units of some set or system which is basically known. Most of the concepts and methods – called henceforth *entries* – considered in Part One have their counterparts in Part Two, and this is the point where the wording becomes disproportionate, which is disadvantageous when standards are to be established. For instance, the space assigned to the software SUDAAN in the matrix (i.e. in the Appendix) ought not to be greater than that in Part One, yet this happens in the case of the current version of the Handbook. The following (probably not exhaustive) list is aimed at eliminating disproportionate wording in such and similar cases.

- *Linearisation based on influence functions* (Chapter 4.2, p. 40) – a brief sketch of the basic idea of the method is recommended.
  To be done
  Done (section currently numbered 3.3).
- *Linearisation based on estimating equations* (Chapter 4.4, p. 57) – see the previous comment.
  To be done in an additional bullet point on page 40.
  In pages 57-58, to add a reference to chapter 4.2
  Done (section currently numbered 3.3).

We reorganized the sections 3.2, 3.3. and 3.4 and eliminated redundant information. Linearisation is now only presented in section 3.3.

- . Chapter 4.5 (pp. 62-66) listing the software products available for variance estimation contains less information on the methods and the options than the corresponding entries in the matrix (pp. 151-156), though the other way of organising

[3] This is an ad hoc term.
[4] See footnote 3.

the text would be obviously the usual way, making also the matrix more transparent. The current disproportionate allocation is most conspicuous in the case of SUDAAN. SUDAAN is presented among the software listed on page 64 and mentioned to be used in Latvia on page 67.

Chapter 4.5 makes a general presentation of software tools, while the matrix takes a very structured approach indicating the suitability of tools to be used for defined sampling designs and their capacity to deal with specific issues (implicit stratification etc.).
We have reduced text on SUDAAN in Appendix 7.5.

- In the Glossary, pages 123-130 on the design effect constitute one of the most useful parts of the Handbook, but in this form they do not fit into the structure of the Glossary (which ought to be similar to a directory). It is straightforward to make an extra chapter of these pages in the Appendix.
To make a separate chapter on design effect, within the Appendix.
Done, we created Appendix 7.2.
- *Intra-cluster correlation coefficient*, page 131. Inserting a formula is recommended.
To be done.
It is done and now the concept (with the formula) is presented in Appendix 7.1 (the glossary).

IV. Comments Related to Minor Errors

- P. 47, formula 4.4.2. This cannot be found in Cochran 1977.
We find in Cochran 1977 on page 38, the formula (2.65) which gives the reduction of variance when N_i is known. This is equivalent to the inverse of $1+\frac{Q}{CV_y^2}$ which as stated in the Handbook gives the increase of variance due to the random size of the final sample.
- P. 71. Replace “unemployment rate” by “total of unemployed” 2 times
To replace .To replace also 'annual rate'.
Done (section 3.7.1).
- P. 126.”mean sample of observations” should read as “mean of sample observations”
To replace with 'mean number of sample observations per cluster'
Done (Appendix 7.2).
- P. 72. Replace “inversion of the quarterly” by “inverse of the quarterly”
To replace
Done (section 3.7.1).
- P. 74. Replace “element of order *j*” by “element *j*”, “cluster of order *i*” by “cluster *i*”
To replace. Similar issue for pages 72, 75, 76, 82 (3 times), 83, 84 (2 times)
Done (section 3.7.1).
- P. 74. Replace “inversion of” by “inverse of ” 2 times
To replace
To make the same replacement on page 82 (2 times)
Done (section 3.7.1).

Review by Karim Moussallam (INSEE, France)

My technical remarks:

- Balanced sampling is only mentioned in the metadata template (p 138). Most French household samples are drawn with balanced designs. A reference to the method of variance approximation for balanced sampling (Deville & Tillé 2005) would help widen the methodological review.

To mention at the end of chapter 4.1, that in some NSIs household survey samples are drawn with balanced designs. To further include the definition of balanced sampling (which can be retrieved in the metadata template, page 138) and to include a brief description of the method of variance approximation for balanced sampling (Deville & Tillé, 2005).

We added a paragraph at the end of chapter (currently numbered) 3.1.

To add this reference in the references chapter: Deville, J.-C.; Tillé, Y. (2005). *Variance approximation under balanced sampling*. Journal of Statistical Planning and Inference, 128:569_591.

We did so.

- Also in the theoretical background, asymptotics is essential to understand variance estimation for calibrated or non-linear estimators (or under balanced sampling), with the crucial notion of asymptotic variance. In such an asymptotic framework, there is an alternative definition of 'consistency' (note 53 page 129) .

For future projects

- It might be too theoretical a concept for the purpose of the handbook, but the idea that variance is a quadratic form helps the analytic construction of variance estimates. For multi-phases sampling, a simple formula gives the estimate of a quadratic form defined on lower level units.

For future projects

- On the same subject, the Raj formula for variance estimation of multistage sampling is given in a table item page 152. Page 44 states that 'variance estimation [for multi-stage sampling] only take account of the first stage of sampling'. Perhaps it would be more complete at this stage of the handbook to provide the Raj and Rao formulas, and to explicit that the latter formula shrinks to a single term when (1) the level estimate is Horvitz-Thompson (2) the variance estimate is unbiased (3) the first stage inclusion probabilities are 'small' (and the second stage variances not too big, as stated elsewhere in the handbook (p57)...). The simplification might not be worth doing systematically, without prior control of whether the ignored term is indeed negligible for a parrticular survey.

To include on page 57 the Raj formula (which is described on page 152) and the Rao formula and to make the above specifications for the Rao formula.

On page 44, to make a reference to page 57. To consider placing the most detailed information on page 57.

On page 50, to delete the redundant paragraph on multi-stage sampling (the first paragraph).

We present Raj and Rao formulae in section 3.3 under analytic approximate methods for multi-stage sampling. We added the conditions under which the Rao formula can be shrunk, and the caution about the simplification.

In Appendix 7.5, we left only the reference to Raj formula (we deleted duplicate information – the formula).

We reorganized the text of sections (currently numbered) 3.2, 3.3 and 3.4 and eliminated much of the redundancy of discussion about variance estimation under multi-stage sampling.

- The use of generalized calibration to deal with non response is broached on page 45. Some interesting theoretical elements are given in a paper (in French) by JC Deville 'la correction de la non-réponse par calage généralisé', Insee Méthodes Actes des journées de méthodologie statistique 2002, including a hint on variance estimation with instrumented (?) residuals.

For future projects

- The prescription on page 8 that the Deff is to be computed with the sample size excluding non respondents could be repeated on page 123, as this is the part of the handbook detailing how to compute the Deff.

To be done

Ok.We again specified the recommendation in Appendix 7.2.

- There is some contradiction between the advice given on page 58 'when domain sizes are unknown... variance approximation of a ratio type estimator should be used' and page 17 'the so-called conditional approach is recommended'.

To make changes in the chapter 2.3 and in the chapter summary: say that the 'conditional approach' is not recommended and that exceptional cases are those justified by very high relative size of the domain.

Done (in section 2.3 and summary).

Chapter 2.3 argues that when the relative size of the domain is high enough, the coefficient of variation of the domain sample size is limited (around 1% when Pd = 0.5, for n=8000) and there is no much impact on the variance if we take the conditional approach.

- page 9 : z(1-alpha) should be replaced by z(1-alpha/2) in formula 2.1.3 (as well as in 2.3.1), to be coherent with the formula for the confidence interval page 121.

To be done Done (sections 2.3.1 and 2.1.3).

- On page 4, Fay's approach is described as the best to deal with unit non-response. If this is the case, maybe it would be worth describing the method in the handbook, at least the conditions on the sampling design to apply it.

To be done on page 45 Done (section 3.4).

- page 55 : 'a conservative approach in approximation' is a bit hard to understand. Moreover, at other places in the handbook a principle of simplicity in the formulas seem to prevail.

The conservative approach should be promoted. But in practice it is not always easy.

The recommendation on the 'conditional approach' is to be changed, the simplistic approach is not recommended, except in cases justified by a high size of domains.

We eliminated that sentence on the 'a conservative approach in approximation'.

- page 101 : In the first paragraph, the word 'individuate' stands probably for 'evaluate'.

To make the change Done (section 4.2).

- glossary page 117 : A multi-stages sampling design is a particular case of multi-phases sampling. To differentiate the first notion more clearly, the definition might be restricted with conditional independance of samples between lower levels units (as in stratfication). (This additional condition is necessary to apply the Raj formula.)

To add the conceptual difference between multi-stage and multi-phase sampling in the Glossary and in Chapter 4.1.

Done in section (currently numbered) 3.1 and in Appendix 7.1.

The description of the Raj formula is made for multi-stage sampling.

We added in section 3.3 the conditions of applicability of Rao and Raj formulae.

- page 119 : On the definition of variance : the given formula doesn't seem to incorporate non-sampling errors. Maybe what is meant is that the estimates of the variance are affected by non sampling errors ?

To add a clarification

We inserted the word 'random' before '… variability from non-sampling errors' in the note accompanying the definitions of variance and of standard error (Appendix 7.1).

- page 128 : Maybe it would prevent some puzzlement to indicate that the formula neglects a term in Vhat/n where Vhat is the variance estimator. Besides, the formula is correct even for complex sampling (multi-phases).

?? We added that the formula is valid even for multi-phase sampling (Appendix 7.2).

- An index would be very helpful for operational use of the handbook.

To develop an index

Done.

- The part on Deff is already long and informative. However, it might be useful to add examples of Deff sor simple designs, such as :

for a stratified SRS with proportional allocation, the Deff is approximately the share of intra-strata variance. So it captures nicely the efficiency of this design.

for a SRS followed by Poisson response at constant rate, the formula illustrates that the Deff increases with the non-response rate.

To be done

We inserted the presentation of the two cases (Appendix 7.2).

We interpreted SRS followed by Poisson response as the regular case of each selected sample unit having the same probability *p* of being a respondent.

(In the case of the French LFS, the quarterly correlations measured between the Deff and the number of respondent dwellings or the unemployment rate are much lower than with either the standard error or the coefficient of variation of this variable. This result suggests that the Deff is more informative of the quality of the sampling design in this particular case.)

- The handbook already contains a lot of formulas, that are indeed useful to provide a common theoretical background. Maybe it would not stretch it exceedingly to expand the paragraph (p57) about the collapse of strata with single unit samples to include the formula of the overestimation, as it guides the process of collapse, and proves that it leads to overestimation of the variance.

To tackle the implications of the strata collapse on variance estimation.

We added the collapse stratum variance estimator in section (currently numbered) 3.3.

Review by Paul Smith (Office for National Statistics, UK)

My response is in three parts – a summary of the adoption status, the responses to the review criteria, and a collection of more detailed comments and queries.

Summary

Most of the report is satisfactory from a methodological perspective, subject to amendment in line with the comments made elsewhere in this response. Therefore my summary advice is “adoption after minor revisions”, including consideration of specific issues with

- the recommended precision measures in chapter 2.2, where the summary is adequate, but the supporting arguments in the text are not (see my comment 6.1 below)
- the recommendation in chapter 2.5 should work for types of estimates other than percentages (my 5.2)
- The recommendation of the integrated approach in the introduction to chapter 5 but not in the summary of chapter 5.1. This needs more negotiation, and there may be mixed models which will work better (my 5.13).

Additionally, the benefits to the ESS of the developments described are not evaluated against the costs. It would be good to do this to enable DIME to decide whether this is a suitable use of scarce resources.

We are addressing your comments provided in this document, see below the answers.

Responses to Review Criteria

1) Is it clear what business process steps and which statistical domains are concerned by the Handbook?

Yes. The handbook concerns variance estimation in household surveys. This is GSBPM 2.4 and possibly 5.7, and the overarching quality management process. It potentially applies across Domain 1 of the SDMX classification.

2) Does the Handbook address current and common problems in statistical domains ? Does the Handbook answer to the problems identified?

The handbook seems to have two purposes - one to document acceptable practises for calculating and monitoring variance estimation in household surveys, and a second to make recommendations for changes to the ESS approach, which seems like a policy issue which might have been better dealt with separately. In this second respect, a summary of the recommendations would have been useful.

In the Handbook, chapter summaries include the main recommendations. There will be an executive summary (of the Handbook) with the main recommendations.

3) Is it easy to identify in the Handbook the distinct topics for which recommendations are formulated? Do the summaries of the chapters allow for a quick and correct understanding of the content and recommendations of the chapters?

In general the summaries are a good overview of the chapters, and the important recommendations are identified in them. In places the summaries repeat statements that were not backed by evidence in the chapters themselves, but I have identified these in my various comments.

We are addressing your comments

Not all recommendations are easily identified:

- The first paragraph of 2.3 apparently contains a recommendation on when precision requirements should be formulated.

  To reformulate by saying 'it is recommended that precision requirements should be formulated ….'. And to repeat the recommendation in the summary of the chapter.

  Done in both indicated locations (section 2.3).

- p17 line 3 contains a recommendation to calculate variance estimators for domains by the conditional approach (which is inconsistent with other parts of the handbook see point 7)
  See the answer to your detailed comment 'i' which is similar to your comment 6.2
- p56 para 3 contains a recommendation on the form of the variance estimator for a regression estimate.
  Yes
- p89, line 2 recommends the integrated approach, but this is not reflected in the following text or the summary on pp94-5.
  It is mentioned in the text above the summary and in the summary that the integrated approach tends to be the most feasible option. To reformulate and use the word 'recommended'.
  Done (section 4.1).

I am happy to agree with this implicit recommendations:

- tolerance should be taken into account in the compliance assessment strategy (p30)

  On page 29 and in the chapter summary, to add the word 'recommended' and reformulate the phrase as follows: "It is recommended that the compliance assessment strategy should be based on principles of transparency and tolerance".

  Done (chapter 5).

4) Do the formulations clearly distinguish between recommendations and rules?
This is a Handbook, and gives guidance on variance estimation methods. Therefore I don't really find any rules in it. Some recommendations are clear and others not (see q3).
It is right that there are no rules.

5) Are the recommendations based on a proper and sufficient analysis of the issues ?

The Handbook provides a large amount of useful information, and seems to have done a good job of identifying a wide range of issues which are of interest in the ESS. Perhaps inevitably, I have quite a number of specific comments and queries:

5.1 Costs are not widely accounted for:

- I am encouraged that the cost of calculation is considered briefly in the summary on p19 and on p39, but concerned that the need to either estimate or "describe" sources of variability will have substantial additional cost.
  It is outside the scope and very difficult to estimate quantitatively the cost on a general basis. The decision (benefits balanced by cost) will be made by domain-specific managers on the basis of known elements (precision thresholds linked to sample sizes, sources of variability which can practically be estimated linked to available tools etc.)
- total cost to the European Statistical System of implementing the recommendations relative to the benefits gained, in order to be able to make a decision over whether this is sufficiently important to warrant expenditure.
  See the above answer. We mentioned the pros and cons (including the burden) for the three options for estimation of standard errors for national and European estimates.
- first paragraph of 2.3 makes a sweeping recommendation on when precision requirements should be formulated without considering whether the costs of meeting the requirements are justified

1) If the variance is calculated for a certain indicator at national level using a certain method and tool (under a specific sampling design), the cost to additionally 'ask' the software tool to calculate the variance for the same indicator at the domain level, would not be in principle significant in comparison with the benefits. Of course, concerns arise if the domain is unplanned (as extra variability should be accounted for) but, the recommendation of the chapter is to avoid putting requirements for unplanned domains. Anyway, the evaluation of cost is to be done by domain/survey and it is very difficult to assess costs against benefits at general level.

2) Precision requirements should be formulated for planned domains if the regulation stipulates that reliable estimates for those domains should be ensured. But, the domain specialists can anyway decide the precision thresholds/sizes for those domains. Different precision thresholds imply different amounts of cost.

- the metadata template (p28 and Appendix 7.2) is very long and burdensome. I am encouraged that it is to be "adapted top the specific features of [a] statistical domain", but also concerned that a request to complete it will result in either (a) an excessive resource to enter sufficiently detailed descriptions or (b) insufficient detail for it to be usable.
  I think that it is sufficiently detailed to be useful.

  About the excessive resources: domain managers can evaluate the cost-effectiveness in order to reduce its content. Experience with EU-SILC, that tried to implement the fully centralized approach using Jackknife, demonstrated that the quality reports collect description of for instance sampling designs which is very variable from one NSI to another. Many clarifications and exchanges were needed with NSIs on e.g. how many stages does the survey have after all? are there self-representing PSUs or not? is systematic sampling with implicit stratification or not? Etc. The use of the metadata template avoids this problem.

5.2 The recommended wording formulations in chapter 2.5 need to be amended because they should work for any type of estimate, not just percentages (suggest adding standard words for other types of estimator) - eg for income in SILC there is no obvious "percentage" formulation. The first bullet in the centre of p21 should be amended in the same way. Further, it would seem to be sensible to relax the provisions for estimates near 0 or 1 (for some suitable rule of thumb governing "near", see detailed comment h. below).

We decided with the Task Force to issue standard formulation of precision requirements for proportions, as these are mostly used in household surveys and in particular in LFS and ICT (the needs in these 2 domains were the mobile to set up the Task Force). Moreover, the formulation of precision requirements is made specific to the particularities of proportions (use of standard error and not coefficient of variation, option to use a model function for the standard error threshold depending on the value of proportion etc.).

See answer to your comment h.

5.3 The approach to compliance assessment in point 2 on p28 is not dependent on the "the assumption that a common practice for estimating standard errors exists", but could be implemented with a range of approved methods (see also comments under point q9 below)

To delete the words "that a common practice for estimating standard errors exists and furthermore" We did so (chapter 5).

5.4 The "Gentleman's agreement" for the LFS described on p30 is unclear - does this propose that the elements of the agreement are specified centrally, or that each Member State provides each element? The "Gentleman's agreement" will be established jointly by Eurostat and NSIs and will include those elements (criteria). NSIs will transmit the information referred in those criteria.

To clarify in the text that 'Eurostat and NSIs will spell out the compliance assessment strategy in a gentlemen's agreement' We did so, in the last paragraph before the bullet points before the summary (chapter 5).

The first two bullets seem to refer to the first case, and the last one to the second (why?). The metadata template (second bullet point) can be adjusted jointly by Eurostat and NSIs. *Some* recommended methods are presented in the matrix (first bullet point), but this is not an exhaustive list of methods, Eurostat and NSIs may identify others which are used for the particular survey and are adequate. It is additionally not clear what the "reference on methods and tools" element in the first bullet on p30 will achieve. Merely providing a reference or giving the software that is used from the table will provide no guarantee that the methods are suitable for the design in question.

Appendix 7.3 comprises few bad practices that can signal inappropriate methods for the sampling design used and type of statistics. The use of Appendix 7.2 provides detailed clear info on the sampling design and variance estimation methods and tools which can show if there are main sources of variability not accounted for or sampling design characteristics not taken into account in the variance estimation methods.

5.5 Design effects as discussed in section 7.1 are based on design, calibration and non-response. But if coverage, measurement, processing etc errors are also to be accounted for, it is not clear whether the theory continues to hold. Particularly, for example, processing errors are unlikely to be clustered, so may reduce the deff, whereas coverage errors are possibly clustered, so may increase the deff. Does the balance of these errors contribute to a single definition of deff accounting for all variance elements?

The paper Liberts, M. (2012). *Precision Requirements in European Safety Survey (SASU).* January 20, 2012, illustrates the concept of design effect as product between *deff*1 as the effect of sampling design, *deff*2 as the effect of estimator, *deff*3 as the effect of non-sampling errors (non-response, **over-coverage**, etc.). For *deff*3, the formula is proposed.

This can enrich the section on design effect and can be incorporated on page 130, after the discussion of the concept of design effect (introduced by Loredana di Consiglio and Stefano Falorsi) as product between *deff*1 (the estimator effect) and *deff*2 (the sampling design effect). We did so (Appendix 7.2).

This paper by Liberts should be given as reference and added in the references chapter. We did so.

About measurement and processing errors a phrase can be included.

We did so, right below the previous additions (Appendix 7.2). but I think it is difficult to quantify their impact on the design effect. Besides, for the variance estimation itself, their effect is in practice not accounted for. A qualitative assessment is rather given.

From chapter 4.2: The guidelines for constructing a suitable variance estimator under fixed survey conditions can be summarised as follows:

- Consideration of all possible sources of variability;
- Consideration of which sources of variability can be estimated;
- Consideration of which sources of variability can be described with some other indicative information (for example, level of processing errors).

5.6 The discussion of replication methods is often not balanced. Pp43-44 for example says "replication" but talks almost exclusively about the jackknife, whereas on p101 "replication" concerns only the bootstrap. It would be nice to have a clearer discussion of the different options for replication-based variance estimation, when they are appropriate, and under what conditions one might be preferred over another.

- Following a related comment made by László Mihályffy (see the document with his comments and Eurostat feedback), it is planned to insert on pages 40-41 the formula for each replication method.
We present the replication methods, their formulae, characteristics and applicability in section 3.3. We reorganized the sections 3.2, 3.3. and 3.4 and eliminated redundant information.
- Moreover, it is planned to replace on page 101 the general formula for replication methods (which is now particularized on page 102 for bootstrap only) with the formula and indications from:
http://www.ecb.europa.eu/home/pdf/research/hfcn/varianceEstimation.pdf?2f26564a9ef2da62419ded2167470ee7 , (page 6,) to mention this document as reference and add it in the references chapter. This will adjust the balance on replication methods on page 101.
Done (section 4.2 and section 6 on references).

- What remains is to give more examples about other replication methods than jackknife for the sampling designs listed on page 43-44. We dropped this sub-section, following reorganization of information on methods (see above).
- Some findings on comparative assessment on replication methods are in chapter 4.4.
We have developed the comparative assessment of the methods on criteria related to applicability (to the sampling design used and the type of statistics), accuracy (confidence interval coverage probabilities, unbiasedness and stability) and administrative considerations (cost, timeliness and simplicity). See section 3.3

5.7 The methods for adjustment of variances for overcoverage on p47 assume that overcoverage is random. This may be useful in adjusting the reference srs variance as part of the design effect. What method should be used to adjust the effect on a clustered survey of possibly clustered overcoverage?
For future projects
A remark: Each non-sampling error is composed by random error (variance) and bias. In the Handbook we discuss the random effect of non-sampling errors and not the bias effect. For over-coverage (see page 37) we discuss the effect on variance because of the reduced and random obtained sample size (after removing the out of scope units).

5.8 p57, second bullet. Having n=2 certainly makes variance estimation more practical than when n=1, but it would be good to comment on the likely accuracy of the variance

estimator in this case! More generally, there is little guidance on what to do and which estimators might be satisfactory when the sample size is small throughout the document.
To include a comment

We added the collapse stratum variance estimator in section 3.3.

Page 22 presents the practice in EU-SILC to refrain from publishing estimates or to flag estimates obtained from too low number sample units.

5.9 The typology (classification) for surveys on pp69-70 (and summary p88) is not very clear. A Venn diagram may help to show the relationships between the different terms? At least approximately, rolling samples are a subset of rotating panels are a subset of longitudinal surveys are a subset of repeated surveys?
Reading their description, rolling samples are a separate category from longitudinal surveys. While rolling samples are taken by moving to different PSUs each wave, longitudinal surveys include units of the initial sample in the new sample.
Longitudinal surveys are divided into 3 separate categories (panel surveys, rotating panel surveys and split panel surveys) which are described. We added the category of repeated panels (section 3.7).
So far everything is distinct.
On the other hand, repeated surveys may or may not be rolling samples and may or may not be longitudinal surveys. Repeated surveys are a general term.
To add a clarification or a Venn diagram dedicated to the distinction between these concepts, on the basis of the definitions in the Handbook.
We added a paragraph (section 3.7).

5.10 Section 4.7.1 seems rather specific relative to the rest of the handbook, contains a number of implicit assumptions which are not true in general, and needs to be clarified. In particular (a) equation 4.7.1.4 assumes that the Response Homogeneity Groups are strata, but this is generally not the case (if non-response weighting is by age x sex, for example) (thhis also applies on p82); (b) equation 4.7.1.5 gives unequal weights by quarter if m_{hq} is variable, and therefore potentially biased estimators for annual averages especially for seasonal variables; (c) equation 4.7.1.8 and 4.7.1.14 only work if an ultimate sample unit appears in exactly one quarter (that is, they are not valid for rotating panel designs).
To reflect on this and make clarification in the text by e.g. adding the assumptions (consider both the text of stratified single-stage sampling and stratified multi-stage sampling from annual averages and estimates of change).
We added the assumption about strata being response homogeneity groups in section 3.7.1 (for both stratified single-stage sampling and stratified multi-stage sampling) and in section 3.7.2 for stratified multi-stage sampling. About whether the formulae do actually work for rotating panel designs, we’ve had some exchange of opinions with the external experts from Agilis and we are waiting for Mr Nikolaidis’ reply.
I find your point a) clear, but I do not understand the concerns under points b) and c). Because the author proposed those formulas particularly for rotating panel designs.

In addition, para 2 on p76 and the last para on p83 probaby should say that ultimate units *may be* taken from the same PSUs; then the next two sentences appear to contradict each other over whether the ultimate units overlap (I guess no overlapping is intended).
Please see answer to your detailed comment 'll'.

5.11 Equations 4.7.2.19 to 4.7.2.23 seem to calculate variances from the totals of the ultimate clusters, ignoring the component of variance within these clusters. Is this intended? Is it reasonable to assume that the within-cluster variance is negligible? If not then the estimates of variance and covariance will be biased.

To check with an external expert or with Mr Ioannis Nikolaidis. Similar issue for the formulae from the chapter on variance estimation for annual averages. To at least include some statements about this.
We inserted a paragraph on the assumptions in both sections 3.7.1 and 3.7.2 (for the multi-stage stratified sampling case).

5.12 Section 4.7.2 rather stops at the stage when analytic variance estimation becomes intractable. In designs where the PSUs change, where the ulimate units are on a rotating panel design, and so on, what are the recommendations then? The claim in the summary (p88) that "This handbook provides analytic methods..." is true, but only in a few special cases and therefore rather misleading by omission. The more general description at the top of p88 is too vague to provide any useful guidance in any given situation (rather understandably - the problem is a difficult one for any specific situation, and a general theory has not (yet!) been developed.

Regarding your comment: "In designs where the PSUs change, where the ultimate units are on a rotating panel design, and so on, what are the recommendations then?" :
In section 4.7.2, going through the analytic methods proposed for stratified multi-stage sampling, I understand that we are in the case of a rotating panel design and there are implicitly common PSUs between periods (because there are common ultimate sampling units between periods). So common PSUs always exist between periods and they are clearly accounted for in formula 4.7.2.18 (which is developed further with 4.7.2.19 until 4.7.2.23).
Then, formulae 4.7.2.24 and 4.7.2.25 concern the case where **all** PSUs are common (we keep selecting the new part of the sample of ultimate units from the same PSUs, we do not change PSUs between periods).

To state in the summary chapter on page 88 that analytic methods are proposed for specific cases and to specify these cases (from 4.7.1 and 4.7.2) (as identified above).
We reformulated the paragraph in the summary chapter to illustrate the cases for which analytic methods are provided.

5.13 The options for how to do variance calculation in section 5.1 are all valid, but do not tell the full story. At the bottom of p90 the "high burden on NSIs" is actually disproportionately high on smaller NSIs. This might argue for a centralised service for variance calculation to support smaller NSIs, but this option is not considered.
To add to that paragraph that the burden is especially or disproportionally high on smaller NSIs.
We did so (section 4.1).

The second para of p91 considers that replication methods can be applied "without knowing the original sampling design", but this is only true if the full replicate information is provided by NSIs; otherwise the details need to be known for a proper selection of replicates.

To delete the sentence 'These enable us to analyse data without knowing the original sampling design' We did so (section 4.1).

At the bottom of p91 it is suggested that Eurostat can include random imputations, but this would seem to require duplication of MS's imputation systems.
It is true that the imputation done by Eurostat results in a duplication of imputation work by NSis and this sentence can be added. We did so (section 4.1, footnote).
However, the random imputation done by Eurostat is done only with the purpose to incorporate variability of imputation into the whole variance and the imputed values are not actually used for other purpose i.e. the estimation of point estimates. While there are probably NSIs that simply use deterministic imputation methods (instead of random imputation methods) that are anyway unhelpful for the estimation of variability due to imputation. This can be also added. We did so (section 4.1, footnote).

Pp93-94 suggest that microdata transmission is needed but that this does not "put huge pressure on data transmission by NSIs". I think this makes transmission out to be much easier than it is in practice - there is a significant cost in preparing data in a suitable format.

To delete 'without putting huge pressure on data transmission by NSIs' from page 94
We did so (section 4.1).

5.14 Table 7.4 seems rather general, and, perhaps understandably, to cover largely the software in use by members of the Task Force which wrote the Guidelines. But if it is to be a template for responses to the metadata questionnaire, it should be expanded to include all the software tools used in MS's?

Table 7.4 offers guidelines on characteristics of some software tools. Having in mind the available resources (see Eurostat summary paper, specification of cost effectiveness), it would have been difficult to include all software tools used in MSs. This can be done in future projects.

5.15 Section 7.5 is unclear; after reading it I do not know which of $n_{\min}^{(\text{long})}$ and $n_{\min}^{(\text{cross})}$ is fixed. And then for example at the end of the para below equation 7.5.2 the text says "minimum required sample size" without qualifying whether this is cross-sectional or longitudinal.

To make the necessary reformulations to make clear what sample sizes are fixed in the different parts of text and formulae on page 158.
- Formula 7.5.1. assumes that $n_{\min}^{(\text{cross})}$ is fixed and derives the $n_{\min}^{(\text{long})}$ .To avoid confusion, the sentence "We may desire a minimum number of $n_{\min}^{(\text{long})}$ units in the panel sample (2,3,4)" can be deleted. We did so (Appendix 7.6).
- Below formula 7.5.2, the text can be replaced as follows: "In panel surveys, r is generally high so β should be close to α . Thus, by rotating out less sample units than initially planned, we can allow for a certain degree of non-response and then ensure that we achieve the minimum required longitudinal sample size that would have been achieved under full response, when a proportion α of the sample rotates out at each wave." We did so (Appendix 7.6).

- One line above formula 7.5.3, the sentence can be replaced by: "Thus, the minimum longitudinal sample size between *t* and *t*+1 is given by:" We did so (Appendix 7.6).

6) Does the Handbook take account and ensure coherence with existing related sources of information in the ESS and outside ESS (e.g. other handbooks, glossaries, reference textbooks)?

The list and scope of references is very good, and the synthesis of the information is helpful. Again I have quite a number of specific comments:

6.1 The comments on sample sizes for small proportions in 2.2 are misguided.
Cv's for proportions are unhelpful because (a) they're not symmetrical when measuring the complement.
This remark can be added on page 12, at the end of paragraph starting with "Coefficients of variation (relative standard errors) are generally…". We did so (section 2.2).
However this remark is strongly linked to your following one.

and (b) inflated for rare characteristics (*p* close to 0).
This remark is well developed in the para starting with "Coefficients of variation (relative standard errors) are generally…" (page 12), in figure 2.2.1 and in the first para on page 13.

Near 0 and 1 confidence intervals for proportions are not well approximated by a Normal distribution, and should be asymmetric. I would be concerned if accuracy requirements were based on proportions very near 0 or 1.
We do not refer to confidence intervals. Unlike confidence intervals, the variance, standard error and coefficients of variation do not need to rely on the normality assumption.
It would be appropriate to add exemptions based on the estimated *p* to the exemptions already given in section 2.5. [It is unclear how this interacts with the sampling procedure; in some cases it may be possible to oversample either elements with the rare characteristic directly or areas where such elements are more prevalent. This should improve the precision for some p's at the expense of others.]
We do not recommend CVs for proportions. If however these are used in a specific case by domain specialists, a recommendation to deal with the shortcoming of using the CV is given (second bullet point on page 14).

The cv should never be used for changes because of the risk that the estimated change is close to 0. This is mentioned, standard errors are recommended for changes close to 0.

As mentioned in a reply to a similar comment made by the reviewer Harm Jan Boonstra (CBS), we will delete the last phrase above the summary on page 14.

Why not use cv targets for totals and means of categorical variables (NB mean of categorical variable = proportion). Need redraft of typology for section 2.2.
We do not recommend CVs for proportions for the reasons presented in the chapter!

6.2 p17 line 3 contains a recommendation to calculate variance estimators for domains by the conditional approach, mainly for simplicity reasons. I agree that such variance estimators are simpler. However, they ignore variation due to domain size variation, and therefore sample sizes set in this way will tend to have larger than expected variances. The Handbook recommends using the ratio approach to calculating the variance of a domain estimator where the denominator has strictly non-zero variance (p11).

See answer to your detailed comment 'i'

6.3 Gabler, Häder & Lynn (2005) discuss the problem of combining design effects from surveys covering different domains (based on the European Social Survey). It would have been nice to see some consideration of whether this approximate approach could be useful more generally in a European context, perhaps around the comment on the need for fast EU data on p89. (Gabler, Siegfried, Häder, Sabine and Lynn, Peter (June 2005) 'Design effects for multiple design samples', Working Papers of the Institute for Social and Economic Research, paper 2005-12. Colchester: University of Essex, http://www.iser.essex.ac.uk/pubs/workpaps/).

For future projects

6.4 When calculating variances using a bootstrap procedure it is vitally important to use the sampling weights so that the with-replacement sampling is from a reconstructed version of the population (Canty & Davison 1999,section 3.4). Suggest replacing "from the original sample" with "from a population reconstructed from the sample using appropriate weights, see Canty & Davison (1999)" in the first sentence of the "Bootstrap" section on p41. A similar amendment needs to be made to the second bullet on p59, where it needs to be emphasised that there are techniques for doing it properly!

To make these amendments
We did so, in section 3.3, when we present the bootstrap. Following reorganization of information on methods, we now have presentation of the bootstrap only in one place in section 3.3.

6.5 The description of random groups variance estimation on p41 is not adequate. Even the first sentence of Wolter (1985 [1st edition!] section 2.1) has a much better summary.

To make a better description
We developed and improved descriptions of all replication methods, including independent and dependent random groups (section 3.3).
See also the answer to your comment 5.6, above.

6.6 Section 4.4 gives a long list of methods suitable for different situations, but does not say whether methods which do involve the joint inclusion probabilities (see p51) are worthwhile in any common situations - for example where there is an equal probability survey.
To add a remark on the applicability of the methods on page 51
This comment is not clear to us, unless you meant "… methods which do **not** involve …".
Joint inclusion probabilities are required in order to obtain an unbiased estimate of the variance. Unfortunately, they cannot be calculated in most common situations (except under simple random sampling). That's why we have no other choice but to resort to approximate formulas which do not involve joint probabilities.
We developed an explanatory paragraph at the beginning of the subsection in section 3.4.

6.7 The "Taylor series expansion" paragraph on p58 gives some basic advice on the consequences of approximating variances for without replacement designs by with-replacement variance estimators. It would be nice to expand this and give some more general guidance.

For future projects

6.8 I disagree that the jackknife is easier than the bootstrap (p59 bullets 4 and 5).

To add the reference: Shao, J. and Tu D. (1995). *The Jackknife and Bootstrap.* Springer, p.18 , for this statement We did so (section 3.3). To add the same reference in the references chapter. We did so.

6.9 I don't understand the description of the drawback of the bootstrap in the first bullet on p60. What two sample sizes are being referred to? Is this something to do with number of replicates? Needs much better explanation and a reference.

Ok. Please see the answer to the related comment made by the reviewer Harm Jan Boonstra (CBS) in his document.
We developed that paragraph and better explained; please see the paragraph in the sub-section on bootstrap, section 3.3.

6.10 The terms "consistently" and "stability" in bullets 3 and 4 on p60 and in the summary p62 are insufficiently described. It's not clear how "consistently better" is evaluated. Perhaps this means statistical consistency? And by stability I understand the variance of the variance estimator - is that correct? If so, then the combination of consistency and variance (and any bias in the estimated variance) will interact in a way which seems to be unpredictable, and not allow any generalisation of which method is best. It would be nice to state whatever general conclusion can be drawn (including if there is none).

To reformulate bullet 3 and 4 as follows:

- Empirical studies show that ***Balanced repeated replication*** (BRR) and ***bootstrap*** perform consistently better than ***jackknife***, which in turn performs better than ***linearisation***. Consistency refers to the coverage probability of the *1-α* level confidence intervals, $\hat{\theta} \pm t_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}$ (Rust, 1985; US Census Bureau, 1993). However, the observed differences are small for relatively simple non-linear statistics such as ratios. Bootstrap performs consistently better than jackknife, particularly with non-smooth statistics (quantiles). Under unequal probability sampling, bootstrap is generally consistent when sampling design has high entropy[5] (Chauvet, 2007). Asymptotic consistency exists in the variance of the ***BRR method*** for smooth and non-smooth statistics, such as the median, when the number of strata increases ($L \rightarrow \infty$).
- Furthermore, empirical studies show that the performance of the same methods with respect to the stability of the variance estimator is in reverse order: ***linearisation*** performs

[5] Entropy is a measure of randomness and it will be large when the probability mass is well distributed over the set of possible samples.

> better than ***jackknife***, which in turn performs better than ***balanced repeated replication*** and ***bootstrap***. Stability refers to the mean square error of the variance estimator (Rust, 1985; US Census Bureau, 1993). In a Monte Carlo study, Rao and Beagle (1967) found that the stability of the jackknife variance estimator of a ratio estimate is similar to the Taylor series linearisation variance estimator.

We did not find the comparison of methods from the point of view of the total error/combination of consistency and variance.

To further investigate with external experts if a general conclusion about the comparison of methods from the point of view of the combination of consistency and variance can be made. If not, to add in a new bullet point that the combination of consistency and variance will interact in a way which seems to be unpredictable, and not allow any generalisation of which method is best. To add the general conclusion in the summary of chapter 4.4, whatever this is (even if there is none).

We have worked on the comparative assessment of methods on criteria related to applicability, accuracy and administrative considerations. See the new sub-section 'Evaluation criteria of variance estimation methods and related recommendations' in section 3.3. From the point of view of accuracy, we have now three criteria named as follows: confidence interval coverage probabilities, unbiasedness and stability (according to Wolter, 2007) and defined in the same sub-section. We present results from different studies on how well replication methods and Taylor linearization (TS) perform on these criteria. The conclusion is that different studies show very good results of BRR when it comes to confidence interval coverage probabilities (the most relevant accuracy criterion). Some studies show very good results of TS when it comes to stability (MSE). However, the best variance estimation method is not obvious in terms of the stability and bias criteria. There is very little theoretical justification for GVFs and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability (lower variance) to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion (Wolter, 2007, chapter 9.1).We formulate these conclusions both in the section and in its summary.

To add to the references chapter:
- US Census Bureau (1993). *Variance Computation by Users of SIPP Micro-Data Files*. http://www.census.gov/srd/papers/pdf/rr93-6.pdf. We did so.

- Rust, K. (1985). *Variance Estimation for Complex Estimators in Sample Surveys*. Journal of Official Statistics, 1(4):381-397, 1985. We did so.

To add other references. There are other sources such as: http://onlinelibrary.wiley.com/doi/10.2307/3315214/abstract but I cannot access them (payment is needed). We do not have access to it, so we did not make reference to it and did not add it to the reference list.

7) Does the Handbook also indicate bad practices (in addition to good practices)?

In a few places the Handbook notes that naive application of certain techniques gives the wrong answers. In view of the complexity of the topic under consideration, I think it is right to avoid trying to identify bad practice explicitly. I would however like a comment on sample substitution (p37) to be included, which I would see as poor practice.
Ok. See 'Detailed comments' point *s*

8) Does the standard support horizontal and vertical integration?
8.1 The Handbook is strong on horizontal integration (across surveys), and covers elements of the vertical integration, in particular presenting three models for variance calculation in chapter 5.1, but there are some deficiencies in the discussion:

- chapter 5.1 says several times that the "objective and challenge for Eurostat and NSIs is to converge and use the same method", but there is very little discussion of why such an objective is desirable. Earlier chapters consider the consistency of countries evaluations against the requirements of Regulations, but there is no evidence presented that different methodological approaches lead to different conclusions in such evaluations. The Handbook is self-contradictory on this matter, saying in chapter 4.4 (p57) that differences between methods are negligible for many statistics and elsewhere (penultimate bullet p90) that results obtained using different methods "lack comparability".

  References need to be given to the "empirical studies" showing that differences are negligible (p57) and for different methods "lack[ing] comparability" (p90, summary p94).
  See answer to 'Detailed comments', point pp
  To add the reference: Kish and Frankel, 1974 for the statement on page 57. We did so (section 3.3).
  We reformulated to produce a clearer understanding on the fact that the results from different applied methods and tools lack comparability if the methods and tools do not account for exactly the same sources of variability (section 4.1).
  We also say 'In the long run, the objective and a main challenge are for Eurostat and NSIs to converge and use the same (replication) method. This will allow Eurostat (and other data users) to estimate standard errors for all relevant indicators and breakdowns, including for extra/unforeseen ones. It will also prevent comparability problems when different national methods and tools do not account for exactly the same sources of variability' (section 4.1).

  Chapter 3 seems to provide a better formulation, where it says that "a minimum degree of harmonisation" is needed - this could mean using one of a range of approved methods (this could mean designating methods as A, B and C methods, as for example for National Accounts).

- there is no evaluation of the total cost to the European Statistical System of the different proposals. For example the fully centralised approach is likely to involve a lot of discussion with each Member State on its sample design, with the result that two groups of experts within the ESS will need to be trained in the design nuances, which has a cost.
  The estimation of the cost is out of scope. We mentioned the pros and cons (including the burden) of the approaches.
  The metadata template has the role to facilitate the collection of clear detailed information on sampling design from NSIs. Under the fully centralized approach

implemented with the use of replication method, Eurostat does not need to go back to NSIs *n* times to ask for clarifications.

8.2 In several places (eg p41, p44) balanced repeated replication is said to be "popular" for variance estimation. It is widely used in the United States, but is there any history of its use in Europe? I don't know of any.

Whenever this statement is done, to add that the method is popular in the United States. We added that the method is popular in the United States, in section 3.3. In the revised version of the handbook we do not have anymore this statement in different places; we reorganized the information on methods and eliminated unnecessary duplications.

8.3 Chapter 5.2 and p93 consider GVFs. I have some ideas for further research on these which might make it more integrated at a European level: (a) Although these are typically specific to particular designs, it would be interesting to examine to what extent a single GVF could cover several MSs. It shouldn't be too hard to fit a model with a parameter per MS to examine this? (b) Equation 5.2.2 gives a parametric version of a GVF, but there are now good nonparametric smoothers available (eg LOESS). Do these provide a better fit to variances? (c) p97 para 2 gives the assumption that Deff and n vary little; can you not less restrictively but just as validly (under 5.2.3) assume that $\frac{Deff}{n} \approx k$ for some constant *k*? (d) It is interesting that GVFs smooth out variation but may be biased, whereas separately calculated estimates are approximately unbiased but possibly with large variances. Therefore it is not clear whether direct variance estimates are more accurate than indirect ones (summary, p102). It would be an interesting research project to evaluate this.

For future projects

About your point d). We have in section 3.3 that: “There is very little theoretical justification for GVFs and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability (lower variance) to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion (Wolter, 2007). In conclusion, GVFs do not aim to provide the best variance estimators possible, but rather aim at providing users with a sort of ‘black-box’ from which they can get a variance estimate for any survey statistic.” This is all we can say for the moment (based on available information).

9) Are the proposed recommendations easy to implement?

Some recommendations such as accounting for *all* elements of variation are both difficult and potentially costly to implement. Requirements to use specific methods are also likely to be costly if they involve changes to systems. Many other recommendations correspond with existing practice and can be implemented.

From chapter 4.2.: The guidelines for constructing a suitable variance estimator under fixed survey conditions can be summarised as follows:

- Consideration of all possible sources of variability;
- **Consideration of which sources of variability can be estimated;**
- **Consideration of which sources of variability can be described with some other indicative information (for example, level of processing errors).**

10) What is the expected impact on the quality of statistical output?

If additional variance components (for imputation, processing error etc) are added to variance estimates, then the *measured* quality will appear worse. However it could be argued that the quality of the variance estimates will be improved.

The effect of increased harmonisation of variance estimation methods is not sufficiently demonstrated by the Handbook, which says both that different methods lack comparability, and that differences between methods are negligible for many statistics. It should additionally be borne in mind that small (and in some cases quite large) differences in variance estimates do not affect the way in which survey estimates are interpreted; so harmonisation may be less important than *facilitation* of calculation of variance estimates by an appropriate method.

See the answer and clarification to your detailed related comment 'pp'

11) What is the expected impact on the efficiency of statistical production processes?

This is an open question; if increased harmonisation or centralisation result in duplication of effort, there will be a decrease in efficiency in the ESS. However, the Handbook provides a lot of very useful information which can be expected to gradually improve efficiency of variance estimation over many years as long as it remains known and used.

Detailed comments

a. The LFS regulation requirements on p6 seem mis-specified(!). It should be the *standard error* not the *relative standard error* which is less than 8% of the population size (the rse is already a percentage, and 8% of the population size will generally be a large number, so the rse will be smaller automatically in largish populations!).
The interpretation of the current precision requirements for LFS has been debated within the dedicated Group of Experts on LFS. The interpretation is that the current precision requirements use the *relative standard error*, other name of the *coefficient of variation* (CV), as a precision measure. *Ceteris paribus*, the coefficient of variation for a proportion is bigger the smaller the proportion is. For proportions close to zero, the coefficient of variation tends to infinity. This is why the current requirements are set for a theoretical situation (a group of unemployed people representing 5% of the working age population), where the estimate is set sufficiently higher than zero. See more information in the minutes of the two meetings of the Group of Experts, uploaded on the page:

https://webgate.ec.europa.eu/fpfis/wikis/display/essmethodology/Precision+requirements+and+estimation+for+Labour+Force+Survey (scroll down to bottom page).
The DIME TF on Accuracy liaised with the LFS Group of Experts in this work.

b. Equations (2.1.3) and (2.3.1) (which are the same) account for the finite population correction. In many cases it will be negligible, so the first term in the denominator will dominate.
This specification can be included in the Handbook when the formula is presented.
We added that the equations account for the finite population correction (sections 2.1 and 2.3).

c. p9 penultimate paragraph needs clarifying. It seems to suggest setting a target according to the available budget and a srs design , and then requiring Member States to meet this target even where their survey is not SRS, thereby bearing any additional costs. Although there is also the possibility of more efficient sampling (ie deffs <1), deffs are generally >1 for household surveys, so additional costs are more likely.
If we add that the total budget at EU level should be inflated to account for national non-response rates and deffs greater than 1, we risk to no longer encouraging countries to keep under control the non-response rates and the efficiency of the sampling designs. Thus, the paragraph can incorporate the idea that adjusting the sample sizes at country level to account for non-response and design effects triggers additional costs which should be considered in the total budget at EU level (given that budgetary constraints are weaker), but on the condition that the non-response and design effects are kept under control. We did so (section 2.1).

d. Last sentence of p9 touches on outputs produced from combinations of administrative and survey sources. Need more discussion (perhaps a subsequent TF?) on appropriate quality measures in these circumstances (eg death rates).

For future projects

e. p10. Bethel (1989)'s approach to multivariate allocation is only one of a wide range on this topic. Perhaps provide more guidance on what is best practice in this area?

For future projects

f. p15, first para. Is there a general definition of "reliable", or is this up for discussion each time a Regulation is proposed? Who decides on the *requirement* for reliable estimates?

Please see chapter 2.4 that tackles the issue about which precision thresholds confer estimates the qualification of being reliable.
"Precision thresholds are generally survey-specific and depend on users' needs and on their required reliability. Furthermore, and beyond statistical concerns, the determination of precision thresholds is also a political and resource-related decision. Examples of precision sizes/thresholds used in different contexts by different institutions are provided. They are not meant to be prescriptive but rather to give some feasible benchmarks that may be used when defining precision thresholds."

Thus there is no general valid definition of 'reliable', this is discussed for every survey by the survey managers and methodologists.

g. delete leading bullets in equation (2.3.2) - they look too much like minus signs!

To be done. We did so (section 2.3).

h. It would be useful to provide a rule of thumb for when P_d is too small for the cv to be a credible measure.

Chapter 2.2 recommends not using CVs for proportions. This is a recommendation. The end of the chapter states that for specific cases, experts may decide to use CV. But if so, "one should pay attention to the huge increase of sample size whenever proportion approaches 0 and should furthermore set up a low threshold of proportion, under which the requirement does not apply".

i. p17, para 2, sentence 2. No, the sample size has been selected according to the design of the survey (for example, if the unplanned domain contains elements from two strata, they will likely have different selection probabilities). Assuming srs will not be appropriate in some (many?) cases.

Ok. To modify the sentence, that sentence is valid when for instance the whole sample size is selected with SRS; therefore formula (2.3.1) cannot be applied if the selection probabilities are different for the units included in the same unplanned domain.
We made a revision in the paragraph (section 2.3)

Please see the answer given to the related comment made by the reviewer Karim Moussallam (INSEE). We will make changes in the chapter 2.3 and in the chapter summary by saying that the 'conditional approach' is not recommended and that exceptional cases are those justified by very high relative size of the domain. (Chapter 2.3 argues that when the relative size of the domain is high enough, the coefficient of variation of the domain sample size is limited (around 1% when Pd = 0.5, for n=8000) and there is no much impact on the variance if we take the conditional approach.)

j. p17, para 3. Bankier (1988) also gives methods for a compromise allocation. (Bankier, M.D. (1988) Power allocations: determining sample sizes for subnational areas. *The American Statistician* **42** 174-177.)

Following a comment made by the reviewer Harm Jan Boonstra, this para will be reformulated and the last phrase deleted.

k. p17, para 4. Although the precision can be improved by poststratification, poststratification is biased which should be mentioned - the effect on the mse should be considered before poststratifying.

It is not clear in fact if poststratification (calibration in general) reduces or increases bias. I learnt that calibration reduces bias and with respect to variance, it can increase it if by calibration the dispersion of weights increases. To find the correct idea and include it in the paragraph. We modified the last paragraph before the summary (section 2.3).

l. p21, para 2. replace "statistically significant" by "not significantly different from 0 at the p = alpha level".

To be done. We did so (section 2.5).

m. p21, para 3. I agree that additional provisions may be provided - I would say "should" rather than "may".

To be done. We did so (section 2.5).

I believe that NUTS provides for areas of approximately similar size across the EU, so this is protection for all MS's?

Yes, a protection for all MSs is intended, when it comes to the NUTSs.

n. p24, para below square bullets: If the estimated percentage is quite low, an increase in the percentage may cause an increase in clustering and therefore increase the design effect. If the estimated percentage is quite high, a further increase is likely to reduce the clustering and therefore reduce the design effect.

To reflect on this idea to see if to incorporate it in the paragraph as an assumption for the LFS case.
We dropped the related paragraph (section 2.5). The LFS group of experts did not provide a theoretical justification for the finding.

o. p27, last para. The expression "closed and ad-hoc formulae" is not very clear and seems too prescriptive. The next paragraph gives a 'for instance' of using the design effect and srs variance, but it's not clear whether other approaches are possible - for example Kish (1995) uses the rate of homogeneity (roh) as a means to adjust design effects for changes in average cluster size (this is the precursor of the model on p125). Could this be used as an approximation allowing design effects from one design to be used in a similar but different one?

The example given in the Handbook is clear. Your example is interesting but I find it difficult to apply as monitoring mechanism on regular basis.

p. p30, para 2. Explain more clearly what is meant in "Tolerance may be possibly indicated by a confidence level". Does this mean confidence that the true value of the variance is less than the target based on a test using the estimated variance? Or something else?

To remove sentence ' tolerance may be possibly indicated by a confidence level'
We did so (chapter 5).

q. p35, first equation: S_0 is undefined.

To insert the definition of S_0 as the set of all possible samples of the population (it is defined in the glossary). We did so (section 3.2).

r. p37, para 1: 'deterministic' imputation provides no more protection against informative non-response than random imputation. Random is the standard for household surveys. Suggest this part is redrafted.

To delete the following part of the phrase: "and such cases …. are used (Ardilly, 2006)". And to delete the following part of the subsequent phrase: "when assuming that imputed values are exact".

We did so (section 3.2). We also deleted "Variance is underestimated particularly when the non-response mechanism is informative (that is, the probability for a unit to respond depends on the value of the study variable)" because informative non-response produces bias.

s. p37, bottom: substitution is bad practice if there is any risk that the process by which units to be substituted are identified is informative. This *is* likely to be a risk in practice, so should be mentioned.

To add a brief para on this.

We added the sentence: "Substitution is bad practice if there is any risk that the process of identification of units to be substituted is informative" (section 3.2).

t. p38, table, third row: Need a reference for "Fay's approach"

Add the references mentioned on page 45 (in the last but one para).

We did so (section 3.2).

u. p41, line 6: it is not possible to generate an estimate of the bias from a jackknife procedure.

There are many sources that argue the contrary. One example: http://www.scss.tcd.ie/Rozenn.Dahyot/453Bootstrap/2007Jacknife04.pdf , slide 79.

v. p44: should "customary" in bullets 4 and 5 be "customised"?

It is 'customary', I checked in Berger Y.G. (2007). *A jackknife variance estimator for unistage stratified samples with unequal probabilities*. Biometrika, Vol. 94, No 4, 953-964.

w. p46, line 2 up: replace "hot-desk" with "hot-deck"

To replace. Done (section 3.4).

x. p48, equation 4.4.5 and below: it would be much better (and more standard) to use E_M for model expectation, avoiding confusion with the error term.

To replace, and be sure to do the replacement in all cases. We did so in all places where models are used (section 3.4).

y. p51: B in equation 4.4.21 is the same as C in equation 4.4.18 - could rationalise the notation?

To check and rationalise the notation. See also page 155. We dropped the last formula for variance estimation under unequal probability sampling, due to its equivalence to the first one (section 3.4).

z. p52, 3 lines above equation 4.4.23: what is it that y_i is assumed not to be independent of?

To reformulate: "The model assumes that the responses given by respondents y_i and y_j $(i \neq j)$ are not independent …"

We reformulated as follows: "The model assumes that the responses given by respondents i and $j, (i \neq j)$ are not independent (…)" (section 3.4).

aa. p56: the estimator in 4.4.32 uses $d_i u_i$ and the recommendation below is for $w_i u_i$. I could not find the right place in Sarndal, Swensson & Wretman to check whether the difference in weights is intentional - a page reference would help!

Please ask Yves Berger for the page: Y.G.Berger@soton.ac.uk
I copy a relevant part of one of his messages: "If you want to refer to calibration in the general sense, you leave as it was (sum(Wi ui) & Sum(di ui)), and refer to the 1992 paper. You could say that for regression estimation sum(Wi ui) is recommended (Särndal, Swensson, Wretman, 1989)."

We checked in Särndal, Swensson, Wretman, 1989 (accessible in JSTOR) and we read that the sampling weight adjusted for calibration is used with the residuals. See for instance formula 4.6, page 532.

bb. p57, last black bullet: change "expressed" to "expressible"

To replace. We did so (section 3.3).

cc. p58, last bullet, second sentence. I find this confusing...is the "researcher" the person producing the data file with replicate weights on it, or is it the person using this file for their own research, in which case they may redo their regression (eg) for each replicate?

To reformulate the second sentence: "The replicate weights created by the survey methodologists can quite simply be included in the data file and be used by users of secondary survey data (e.g. researchers) to estimate the variance." The paragraph can add at the end "However, the release of replicate weights with the public use data files may still raise confidentiality issues. See chapter 5.2 for a possible solution to this problem

We reformulated as follows (section 3.3):

"The replicate weights created by the survey methodologists can quite simply be included in the data file and be used by users of secondary survey data (e.g. researchers) to estimate the variance. This is especially useful when there are confidentiality issues involving sample units and there is need to prevent dissemination of any information that identifies the sample units. However, the release of replicate weights with the public use data files may still raise confidentiality issues. See section 4.2 for a possible solution to this problem."

dd. p60, last bullet line 2: change "TLS" to "TSL

To replace. We changed the acronym to TS in all places.

ee. p65: is CLAN really "small"?

To delete "small" We did so (section 3.5).

ff. p69, line 3: suggest replaci9ng "are not desirable be accumulated over time" with "increase the variance of the sum of estimates over time" in line with the text on the following page.

To replace the phrase with: "The variances of changes and sums in periodic surveys should take into account any overlapping correlations; the positive correlation between periods increases the variance of the sum of estimates over time but reduces the variance of changes". We did so (section 3.7).

gg. p71, section 4.7.1, line 1: Some LFS's in the EU are monthly or pseudo-monthly.

To delete 'like that of the EU-LFS' Done (section 3.7.1).

hh. p72, case 1, line 2: something is missing after "values" - perhaps Y_h?

No, the values refer to the index of the strata themselves

We reformulated as follows: "Let h be the final strata of the surveyed units (ultimate units: e.g. households or individuals), $h = 1,2,...,H$." (section 3.7.1).

ii. p72, case 1: the order of the subscripts is not consistent in this section.

To adjust the order of subscripts

We made the order to subscripts consistent throughout the sections 3.7.1 and 3.7.2.

jj. p73 & p75: replace "nominator" with "numerator"

To replace. We deleted those paragraphs as they repeat the issues (section 3.7.1).

kk. p76 General remarks, line 3. Need reference to evidence for BRR being less stable (and clarity over what "stable" actually means - the variances of the variance estimate?)

See the answer to the comment 6.10. Add references: Rust, 1985; US Census Bureau, 1993. AGILIS: we did so, see response to your earlier comment 6.10.

ll. p76 General remarks, para 2. This doesn't contain quite enough information to understand what is being suggested. Expand slightly?

The paragraph suggests that first we can produce point estimates and variance estimates for each of the 13 weeks, and on the basis of these, we can derive point estimates and variance estimates for the quarter. Instead of deriving directly point estimates and variance estimates for the quarter (section 3.7.1).

Another issue:

To reformulate:

"In order to reduce costs of quarterly surveys, the ultimate units (e.g. individuals) of all time points (e.g. quarters) are taken from the same PSUs. Only partial overlapping exists in the ultimate units, between time points. In this case …"
Done (section 3.7.1).

To reformulate also the last para on page 83. Done (section 3.7.2).

mm. It is not clear why $\hat{Y}_t$ in equation 4.7.2.16, as described in the text, needs a hat. Is it an estimator of the population total?
To remove the hat of the parameter $\hat{Y}_t$ for its appearance in the text, after the word 'Moreover'. We did so (section 3.7.2).

nn. p85, last three lines: $\bar{C}$ is undefined.

To add that "$U_{t,\bar{C}}$ and $U_{t+1,\bar{C}}$ may be the subset of people who are employed at the corresponding time points". We did so (section 3.7.3).

oo. p89, line 8 up: I don't see how users could "deal with variance estimation" in the absence of numerical information?
It is 'almost no access to numerical information'. The phrase refers to the absence of access to microdata. For instance generalized variance functions do not need microdata. One needs the point estimates and the parameters.
We reformulated the sentences (section 4.1).

pp. p90, penultimate bullet: I understand (but don't necessarily agree) how the results from different methods could lack comparability, but I would be concerned if estimates from different *tools*, correctly applied, lacked comparability. Experience in ONS is that calculating the same variances in different tools gives differences only after many decimal places.

To replace the bullet point by:
"The use of different variance estimation methods leads in principle to negligible differences in the results. However, the results from different applied methods and tools lack comparability if the methods and tools do not account for exactly the same sources of variability." Done (section 4.1).
To do a similar replacement whenever the Handbook states that the methods and/or tools lack comparability. Done in the summary of the same section.

qq. p91, last two bullets - do these refer to cross-sectional estimates, or also longitudinal ones?

It is a general specification, for both.
We dropped them.

rr. p94, summary: replace "requisite" with "situation"

Requisite stands for 'need'/'requirement' (i.e. the need to provide standard errors for more and more policy making indicators). We replaced with ‘requirement’ (summary of section 4.1)

ss. p95, penultimate para, line 2 up: replication methods can do many statistics simulatneously.
To modify the paragraph and state '(…) without resorting to direct variance computations, which, as far as analytic methods are concerned, is more difficult and complex task. Besides, direct variance computations (analytic and replication methods) usually need microdata files with variables (…). We did so (section 4.2).

tt. p96, fourth bullet: tables of variances can be presented as web versions, in which case neither size nor readability is important. What is the loss in quality from using a GVF?
It is about that there is no need to cover pages of documents with variance estimates when these can be easily calculated for the point estimate(s) needed using the function.
GVFs is mainly empirical especially when it comes to quantitative variables.
The GVF procedure should be flexible enough to fit most of the commonly used sample strategies; there is however no scientific evidence for this claim.
There is very little theoretical justification for GVFs and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability (lower variance) to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion (Wolter, 2007, chapter 9.1). This is included in section 3.3.

uu. p97, several places: I don't like " 'design effect' factor". There is a design effect which is what is used here, and there is a design factor, which is something different (as on p125 - I think this has been widely adopted since 1997!). Please just use *design effect.*
To delete 'factor' from appearances on page 97 (2), on page 100 and on page 8. We did so.

vv. p99, para 3: I don't follow this. 5.2.2 and 5.2.7/5.2.9 are not the same, so it seems that the Poisson case is not consistent with 5.2.2?
To clarify the paragraph with external experts or with Guillaume Osier/ Stefano Falorsi
We discussed this with Mr Osier and following his suggestion we deleted the last sentence of the paragraph (section 4.2).

ww. p99, penultimate para: need to give conditions under which 5.2.2. + WLS is a good choice.
To ask external experts or Guillaume Osier/ Stefano Falorsi
We added reference to Valliant (1987) (section 4.2). Mr Osier confirmed that this is sufficient. These conditions are set on page 501 of Valliant’s article. They are very technical so we don’t think we should include them in the handbook (which should be kept readable).

xx. p101, Replication Methods, para 3. Are there any European examples of releasing files with replicate weights?
To investigate and include information on this
We did not find any reference of European examples.

yy. p103, last para: Does the averaging of weights work only for the bootstrap (as in the reference given earlier ion the text), or is it general for all replication methods?
To investigate and include information on this
We found reference only for bootstrap. We developed and improved the clarity of the text (section 4.2).

Other:

- To replace third para, page 22, with

"The use of the margin of error in the formulation of precision requirements assumes a normal distribution of the sample means across all possible sample realisations. However, for instance, bootstrap confidence intervals are not based on the normality assumption"
We revised in line with this (section 2.5).

Because the standard error does not need to rely on the normality assumption.