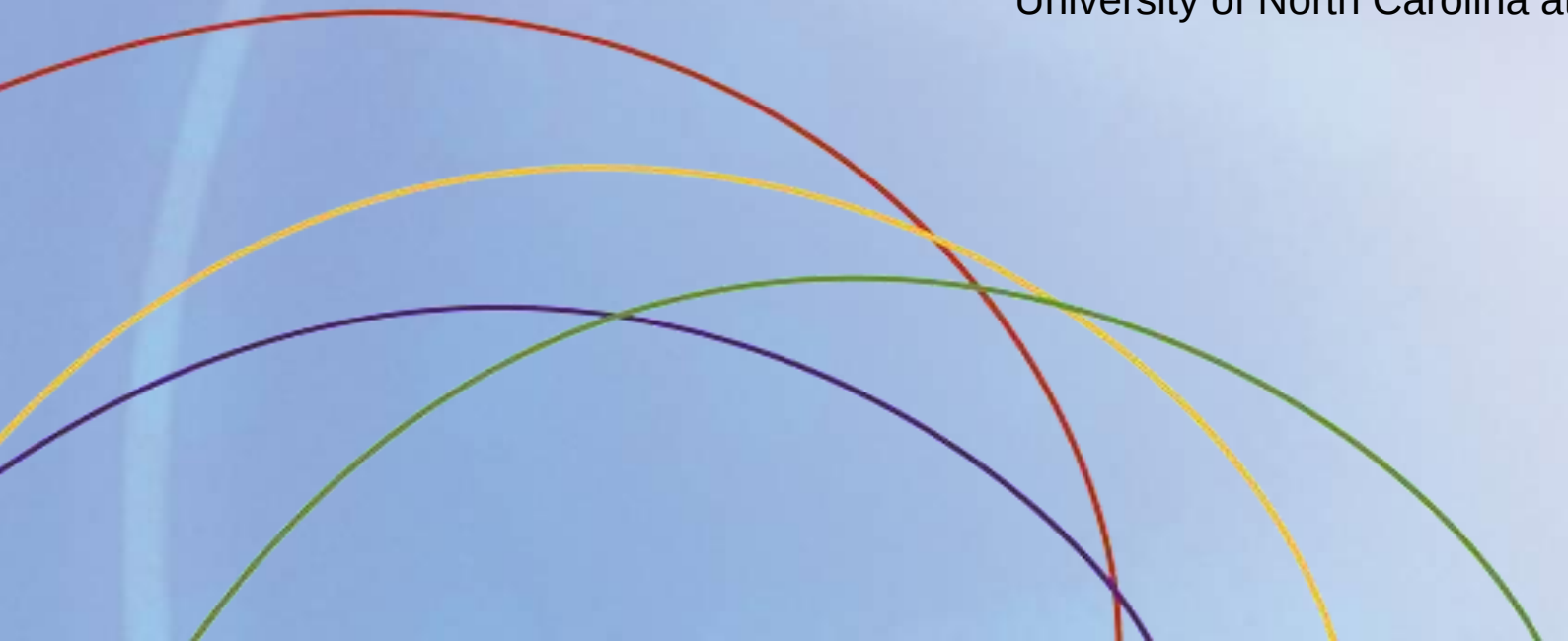
ASPIRE: A System for Product Improvement, Review, and Evaluation

Paul Biemer
RTI International and
University of North Carolina at Chapel Hill



Co-developers

Heather Bergdahl, Lilli Japec, Åke Pettersson

Statistics Sweden

and

Dennis Trewin

Australian Statistician

What is ASPIRE?

- General system evaluating the quality of a statistical product
- Currently focuses on Accuracy, but has been also been applied to other quality dimensions
- Based upon a TSE decomposition of product quality
- Assigns quality ratings to products based upon
  - The intrinsic risk of each error source to quality
  - Risk mitigation activities by error source under five criteria
  - A weighted average of ratings over TSE components
- Presents visual summaries of the results accessible by upper management

TSE Decomposition

NOTATION:

$\hat{Y}_P$ = preliminary published estimate

$\hat{Y}$ = final published estimate

Y = error-free parameter under the operational specification

X = true parameter or preferred specification

$$\underbrace{\hat{Y}_P - X}_{\text{total error}} = \underbrace{(\hat{Y}_P - \hat{Y})}_{\text{(revision error)}} + \underbrace{(\hat{Y} - Y)}_{\text{(sampling and other nonsampling errors)}} + \underbrace{(Y - X)}_{\text{specification error}}$$

Sampling and Other Nonsampling Errors

$$\hat{Y} - Y = \varepsilon_1 + \cdots + \varepsilon_6$$

where the ε's correspond to these six error sources:

1. Sampling error
2. Frame error
3. Nonresponse error
4. Measurement error
5. Data processing error
6. Model/estimation error

Products Reviewed

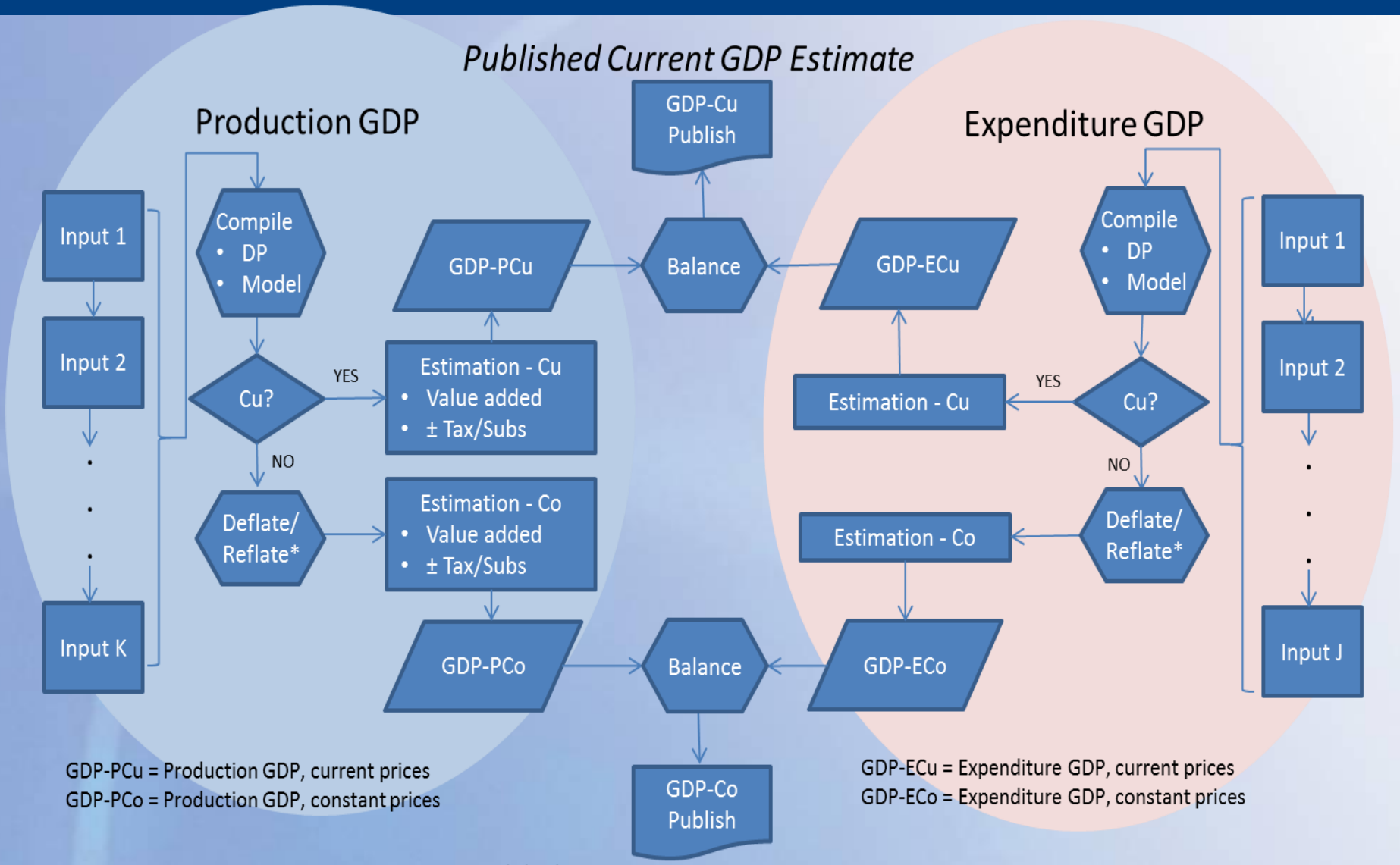
Survey Products	Error Sources
Foreign Trade of Goods Survey (FTG) Labour Force Survey (LFS) Annual Municipal Accounts (RS) Structural Business Survey (SBS) Living Conditions Survey (LCS) Consumer Price Index (CPI)	Revision error Specification error Sampling error Other nonsampling errors • Frame error • Nonresponse error • Measurement error • Data processing error • Model/estimation error

Products to be Reviewed (cont'd)

Registers	Error Sources
Business Register (BR) Total Population Register (TPR)	Specification error Frame: Overcoverage Undercoverage Duplication Missing Data Content Error
Compilations	**Error Sources**
Gross Domestic Product (GDP) • Annual GDP (production) • Quarterly GDP (production)	Customized error profile

7

Process for Estimating GDP by Current and Constant Price Approaches

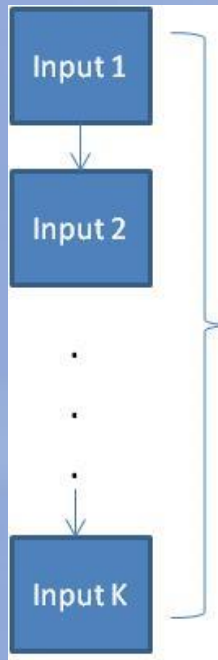


*NOTE: Some items follow the deflation process in the opposite direction and are complied starting with information on volume change from the previous year. The volume estimate i s then reflated with the price index in order to come to the current price estimate. Items within the Energy sector is one such example.

GDP Error Sources

- Input data error (up to four sources)
  - Subject to all sampling error as well as the all relevant nonsampling error sources
- Compilation error
- Data Processing Error
- Model/Estimation Error
- Deflation/Reflation Error
- Balancing Error
- Revision Error

Effects of Input Data Errors on GDP Accuracy



The GDP input data sources give rise to a set of variables:

$$x_1, x_2, \ldots, x_P$$

These input variables are used in the calculation of GDP and are subject to error; i.e.,

$$x_p = x'_p + \varepsilon_p$$

where x'_p is the true value and ε_p is the error.

The true GDP has the form $GDP' = g(x'_1, x'_2, \ldots, x'_P)$

Therefore, the estimate of GDP can be written as

$$GDP = f(x_1, x_2, \ldots, x_P) = GDP' + e$$

The ASPIRE project is particularly interested in how the errors, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_P$ affect the magnitude of the error *e*.

Quality Criteria were Applied to Each Error Source

Intrinsic Risks to Data Quality by Error Source
High, Medium, Low

Criteria by Error Source	Ratings by Criterion
1. Knowledge of risks 2. Communication with users 3. Compliance with standards and best practices 4. Available expertise 5. Achievement toward risks mitigation and/or improvement plans	Poor (●) Fair (◒) Good (○) Very Good (◒) Excellent (◎)

An Example of the Rating Guidelines – Knowledge of Risks

Poor	Fair	Good	Very Good	Excellent
Internal program documentation does not acknowledge the source of error as a potential factor for product accuracy.	Internal program documentation acknowledges error source as a potential factor in data quality.	Some work has been done to assess the potential impact of the error source on data quality.	Studies have estimated relevant bias and variance components associated with the error source and are well-documented.	There is an ongoing program of research to evaluate all the relevant MSE components associated with the error source and their implications for data analysis. The program is well-designed and appropriately focused, and provides the information required to address the risks from this error source.
	But: No or very little work has been done to assess these risks	**But:** Evaluations have only considered proxy measures (example, error rates) of the impact with no evaluations of MSE components	**But:** Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses	

Example: LFS Accuracy Ratings for 2011

	Quality criteria						
Error sources	Average Score	Knowledge of risks	Commun-ication to users	Available expertise	Compliance to standards& best practices	Plans to mitigate risk	Risk to data quality
Specification error	66%					○	L
Frame error	58%					○	L
Non-response error	66%		○			○	H
Measurement error	50%	○	○	○			H
Dataprocessing error	54%	○				○	M
Sampling error	70%						M
Model error	46%	○	○		○	○	M
Revision error							N/A
Total score	**58%**						

13

The Evaluation Process

- Pre-interview activities
  - Background reading by the two evaluators
  - Self-assessments by each program area
- The Quality Interview
  - ½ day sessions involving 4-5 key product owners
  - Overview discussions of product processes
  - Detailed assessment of each of the 5 criteria
- Post-interview activities
  - Review of and comment on ratings by product owners
  - Ratings adjustments by evaluators to achieve equity

Example: LFS Accuracy Ratings for 2012

Error source	Score round 1	Score round 2	Knowledge of Risks	Communication to Users	Available Expertise	Compliance with standards & best practices	Plans towards mitigation of risks	Risk to data quality
Specification error	66	70						L
Frame error	58	58						L
Non-response error	56	52						H
Measurement error	50	56						H
Data processing error	54	62						M
Sampling error	70	78						M
Model/estimation error	50	60						M
Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Total score	**56,4**	**60,9**						

Example: LFS Accuracy Ratings for 2012

Error source	Score round 1	Score round 2	Knowledge of Risks	Communication to Users	Available Expertise	Compliance with standards & best practices	Plans towards mitigation of risks	Risk to data quality
Specification error	66	70						L
Frame error	58	58						L
Non-response error	56	52						H
Measurement error	50	56						H
Data processing error	54	62						M
Sampling error	70	78						M
Model/estimation error	50	60						M
Revision error	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Total score	**56,4**	**60,9**						

Blue indicates deterioration in quality

Pink indicates improvement in quality

LFS Change Ratings between Round 1 and Round 2

Error source	Score round 1	Score round 2	Knowledge of Risks	Communication to Users	Available Expertise	Compliance with standards & best practices	Plans towards mitigation of risks	Risk to data quality	*Correction from 2011 rating* *Improvement from 2011 rating* *Deterioration from 2011 rating* **Comments on changes**
Specification error	66	70	7	7	7	7	5→7[1]	L	[1]Planning cognitive lab work to reduce specification error. Reinte being planned will also help in this regard.
Frame error	58	58	7	7	7	3	5	L	
Non-response error	56	52	7→6[1]	5	~~9~~ 5[2]	~~7~~6→5[3]	5→5[4]	H	[1]Knowledge of the causes of nonresponse have deteriorated. Altho theories, the true causes of the increases in both intrinsic and resi be sorted out. [2]Corrected due to level of expertise in data-collection [3]This is both a correction to the Round 1 ratings and a deteriorati for telephone panels is to use face to face interviewing for Wave 1 reasons but foremost is to reduce nonresponse bias. There are oth practices as well. [4]Despite the considerable planning effort, this rating stayed at "Go mitigation activities have been slow to materialize while residual to a "critical" or "crisis level. This actually represents somewhat thus we note it even though there was no change in the rating.
Measurement error	50	56	5	5	5	3→5[1]	7→8[2]	H	[1]Monitoring of TIs has commenced and further cognitive testing is questionnaire. However, to achieve compliance with best practices examination of measurement error is needed; for example, to bette causes and effects of rotation group bias, and removal of the facto a large extent aware of which calls are being monitored. [2]Plans are in place to conduct reinterview survey; however, more i measurement errors in the labour force estimates.
Data processing error	54	62	5	3→5[1]	7	7	5→7[2]	M	[1]QD documents data editing and provide information on coding er planned in conjunction with ISO standards work. [2]Plans to review the automated coding quality are in place.
Sampling error	70	78	7	7→9[1]	7	7→9[2]	7	M	[1]QD documents sample design and sampling error. [2]Work on sampling error is well regarded and is consistent with t
Model/estimation error	50	60	5	5	~~3~~ 5 →6[1]	3→7[2]	5→7[3]	M	[1]Error corrected in last year's evaluation of seasonal adjustment [2]Work on time series adjustment regarded as state of the art. Also estimation is very good. [3]Plans in place to revise estimation approach have been approved implementation is underway.

Summary of Results from Round 1 – Dec. 2011

Error Source	RS	CPI	FTG	LFS	NA	SBS	BR	TPR	Avg
Specification	**74**	**68**	**62**	66	**56**	**46**	62	44	60
Frame	36	**42**	**62**	58		**62**			49
Overcov.							**48**	**52**	
Undercov.							**40**	34	
Duplication							46	64	
NR/Miss. data	**62**	36	**62**	**66**	64	**74**	40	**60**	57
Meas/Content	**52**	**40**	**54**	**50**	**58**	**50**	**42**	50	50
Data proc.	**46**	**70**	**46**	**54**	**44**	**52**			52
Sampling		**54**		**72**	**44**	**80**			64
Model/est'n	**54**	**64**	**66**	**46**	**44**	**60**			56
Revision	74		62		**62**	**58**			64
Total	**57**	**56**	**59**	**58**	**51**	**59**	**45**	**52**	55

Red Bold = High Risk, Black Bold = Medium Risk, No Bold = Low Risk

Summary of Results from Round 2 – Dec. 2012

Error Source	RS	CPI	FTG	LFS	SBS	LCS	BR	TPR	Rating
Specification error	N/A	68	58	70	54	34	66	46	57
Frame error	60	62	58	58	64	42	55	62	58
Overcoverage							56	56	
Undercoverage							46	60	
Duplication							63	70	
Nonresponse error/Missing	52	55	66	52	70	40	48	66	56
Measurement error/Content	58	62	62	56	52	46	46	58	55
Data processing error	48	76	60	62	60	42	N/A	N/A	58
Sampling error	N/A	66	N/A	78	84	54	N/A	N/A	71
Model/estimation error	38	52	80	60	60	38	N/A	N/A	55
Revision error	58	N/A	76	N/A	56	N/A	N/A	N/A	63
Round 2 Mean Rating	49,6	63,9	65,8	60,9	61,4	42,1	52,2	58,0	57
Round 1 Mean Rating	46,7	60,3	57,3	56,4	59,6	N/A	47,2	52,2	54
Improvement	2,9	3,6	8,5	4,5	1,8	N/A	5,0	5,8	2,5

RED BOLD = HIGH RISK
BLACK BOLD = MEDIUM RISK
REGULAR FONT = LOW RISK
N/A = NOT APPLICABLE

Results for National Accounts from Round 2 (December 2012)

Error Source	GDP Quarterly	GDP Annual
Input data source (Average)	53	66
Structural Business Survey (SBS)	N/A	66
Index of Service Production (ISP)	58	N/A
Index of Industrial Production (IIP)	58	N/A
Merchanting Service of global enterprises	42	n.e.
Compilation error (modelling)	48	48
Compilation error (data processing)	40	35
Deflation error (including specification error)	48	48
Balancing error	56	50
Revision error	56	54
Round 2 Mean Rating	50,5	49,9

Future Work

- Work is continuing to address areas of high risk that have fair to poor ratings
- Criteria, checklists, etc for all quality dimensions will be revised and enhanced
- Criteria for National Accounts will be further revised and detailed quality criteria checklists will be developed
- Round 3 of the Accuracy assessment will begin in November, 2014
- User dimensions to be evaluated by internal quality units rather than external consultants