Data Processing and Associated Errors

Lars Lyberg, Stockholm University,
Department of Statistics

ITSEW 2013

Data Processing Steps

1. If necessary, entries are inspected to avoid data entry problems
2. Data are *captured* via keying, scanning or some other optical sensing device
3. Captured data are *edited*, i.e., “corrected” and “cleaned;” missing data are sometimes “imputed”

Data Processing Steps (cont'd)

4. Open-ended responses are *coded*, i.e., converted to numeric codes
5. Data are *weighted* and files are prepared for public or client use
6. As part of the *analysis,* data files are tabulated, protected from disclosure, and standard errors and other measures of uncertainty are computed
7. Data are *documented* and delivered to clients and users

Data Processing Error

- Relatively sparse literature
- Some steps are very error prone (e.g., coding and editing)
- Errors are both systematic and variable
- Correlated response variances occur whenever people are performing tasks
- Increased automation and integration reduces variable error while increasing systematic error

Data Capture Errors

- Keying errors
  - Discovered by verification keying or editing
  - Error rates usually small based on records, fields or characters
  - Studies often conducted in QC environments
  - The vital few large errors can have large effects on MSE
  - Keying not yet rare in some countries

Data Capture Errors (cont'd)

- Intelligent Character Recognition
  - Error types are substitution and rejection
  - Substitution errors can be systematic
  - Condition of incoming documents and the equipment is crucial which calls for continuing calibration
  - To estimate substitution rates manual sample checks necessary
- Error rates usually small but need to be checked

Quality Targets for Scanner Calibration Using a Truth File

Field type	Sample size	Target error rate (%)
Alpha-Numeric Handprint	50	2.5
Alpha Handprint	100	1.5
Numeric Handprint	100	1.5
OCR Typed Font	120	1.0
Tick Box	250	0.5
Bar Code	50	0.2

Statistics Canada’s ICR Capture Process

Data Capture and Paradata

Stat Can's QC reports on the data capture process

– Control charts by scanner, operator and field type

– Pareto charts showing distribution of errors by field type

– Error rates by operator and scanner

Editing Definition

- Editing is the identification and, if necessary, correction of errors and outliers in individual data used for statistics production
- The definition does not state that all errors be corrected or even identified
- Editing can be very costly, sometimes 40% of the budget

Purpose of Editing

- Essentially editing is a QC operation of the data collection
- To provide information about data quality (patterns and root causes)
- To provide information about future survey improvements
- To ”clean up” the data and get rectangular data sets
- CQI should gradually decrease the extent of editing

Different Kinds of Editing and Edits

Editing

- Micro-editing: Editing at record level
- Macro-editing: Editing at aggregate level
- Selective editing
- Output editing

Edits

- Critical edits are invalid or missing entries that must be repaired
- Query edits are suspicious entries

- Granquist and Kovar (1997) suggest that 50% of the query edits have little or no effect on the final, aggregated estimates. They advocated using "selective editing" to reduce editing costs.
- The U.S. Federal System has managed to decrease the editing budget share to about 20%. Statistics Sweden is down to 30%.
- Selective editing and editor debriefings are becoming common.
- Selective editing uses unit size, error size, survey weight, importance of the variable, and cost to decide which query edits will be followed up with respondents.

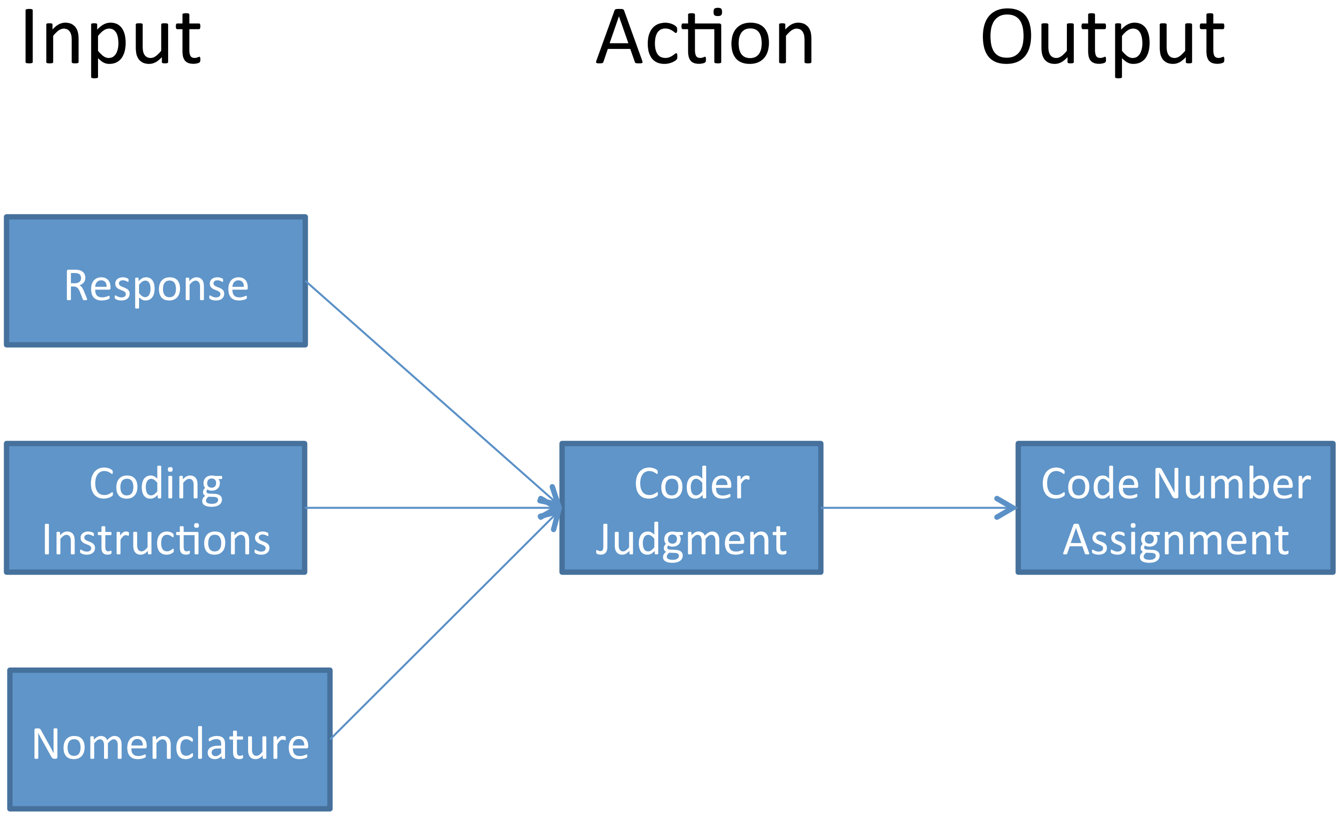
Paradata in Editing

- Edit failure rate E1
- Recontact rate E2
- Recontact productivity E3
- Correction rate E4
- Correction rate for recontacted respondents E5
- Imputation rate for recontacted respondents E6

What if E5 is much smaller than E1?

What if E6 is large?

The Generic Coding Process



Coding

- Classification process where open-ended responses are classified into coding categories
- Coding can be expensive, error-prone and boring
- Coding can be manual centralized or decentralized, automated or computer-assisted

Coding Errors

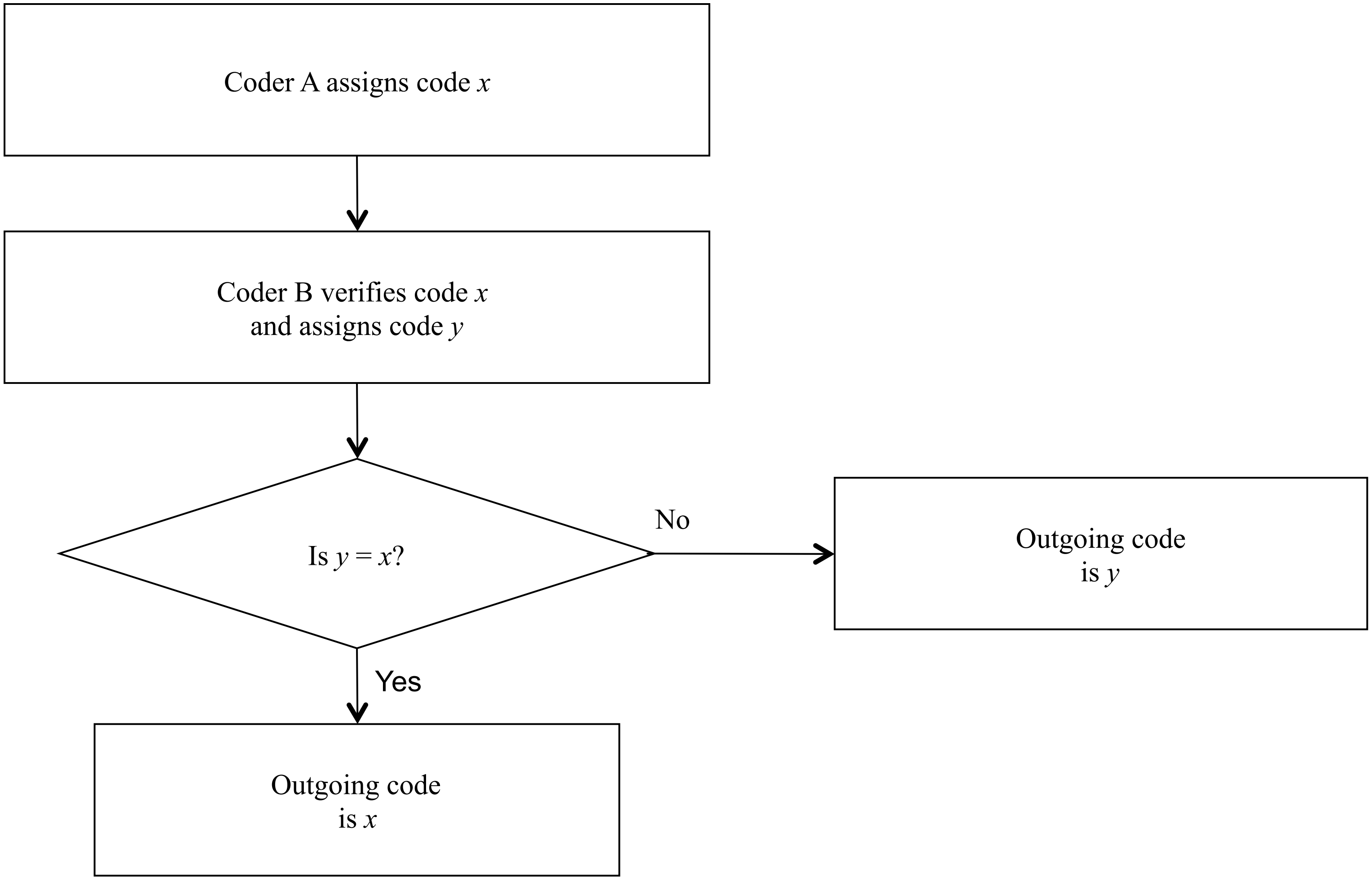
- Coding is subjective in nature
- Error rates and variability rates can be large
- Coding error occurs when there is a deviation between the assigned code number and the true code number
- Coding errors are identified by verification
- Coding rules and nomenclatures may be incomplete
- Errors are controlled by automation, dependent, and independent verification

Examples of Coding Error Rates

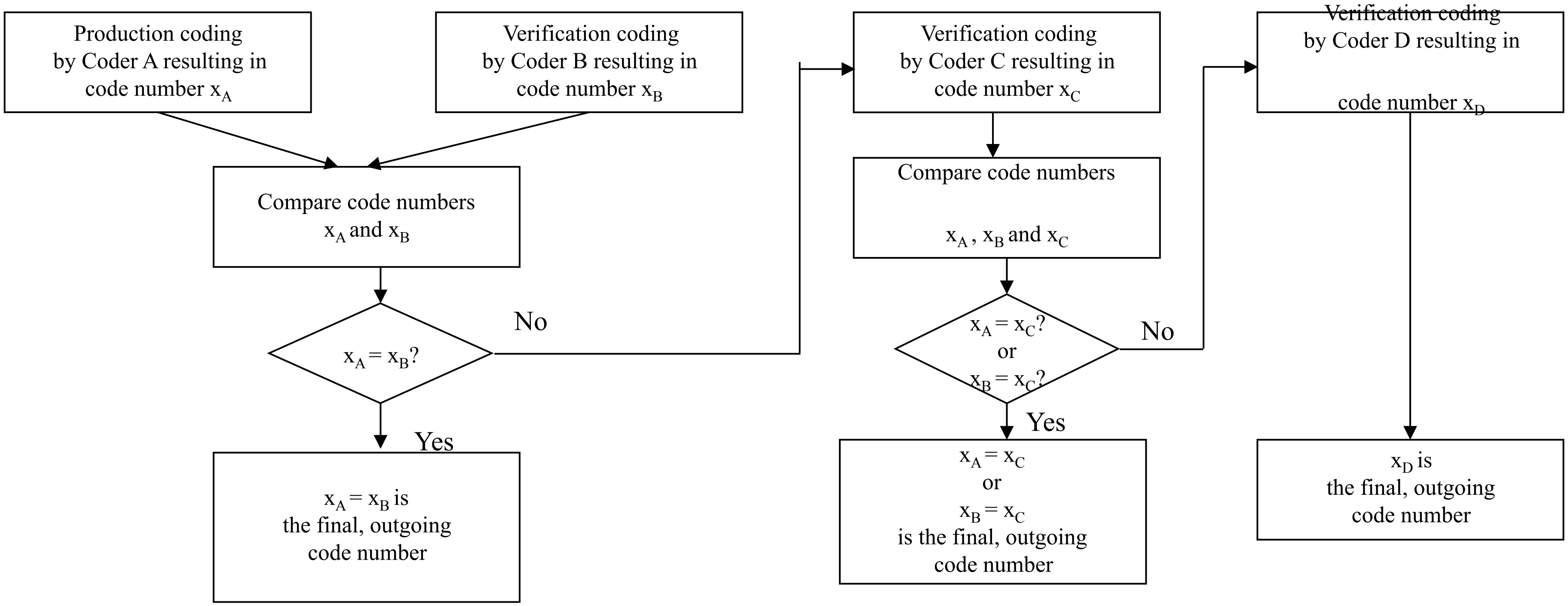
- 1970 Swedish Census
  - Occupation 13.5 %
  - Industry 9.9 %
- 1970 US Census
  - Occupation 13,3 %
  - Industry 9.1 %
- 1991 RTI
  - Occupation 21%
  - Industry 17%
- 1996 Canadian Census
  - Occupation 20% (4 digits), 9% (1 digit)
- 2001 Canadian Census
  - Occupation 22% (4 digits), 11% (1 digit)
- 2012 Swedish LFS
  - Occupation 12.1% (1 digit), 13.4% (2 digits) (large variation between occupation groups)
- 2011 Swedish crime statistics
  - Crimes 12.4 %

Quality Control Methods

- Statistical process control (SPC)
- Acceptance sampling
  - Dependent verification
  - Independent verification
  - Elaborate systems for moving coders between 100% control and sampling control
- Continuous quality improvement
- Weighting systems depending on seriousness of errors
- A certain lack of rigor in the methodology used



Dependent Verification of Coding


Production coding by Coder A resulting in code number x_A
Verification coding by Coder B resulting in code number x_B
Compare code numbers x_A and x_B
$x_A = x_B$?
No
Yes
$x_A = x_B$ is the final, outgoing code number
Verification coding by Coder C resulting in code number x_C
Compare code numbers x_A , x_B and x_C
$x_A = x_C$? or $x_B = x_C$?
No
Yes
$x_A = x_C$ or $x_B = x_C$ is the final, outgoing code number
Verification coding by Coder D resulting in code number x_D
x_D is the final, outgoing code number

Two-way Independent Verification with Adjudication

Theory of Independent Verification

- Choose system that provides highest probability for true outgoing code
- Stratify coders according to coding skills
- Give appropriate roles to the different coder pools

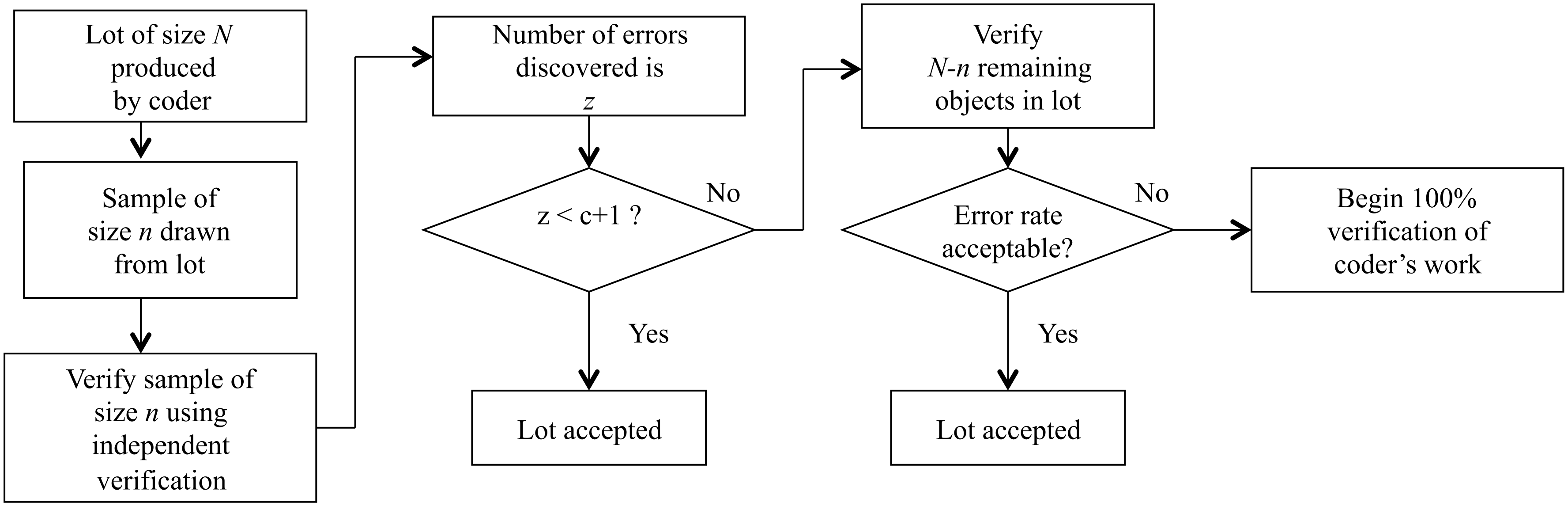
The general situation with dependent verification

- Dependent verification is inexpensive and ineffective
- Independent verification is expensive but discovers lots of errors
- Transition to independent verification is backed by evidence on error rates and vague notions such as "the dependent verifier is influenced by the value already assigned"
- But before we rule out dependent verification we should know more about what's going on cognitively. Can the dependent verification process be adjusted so that it becomes more effective? After all, it is relatively inexpensive.

Acceptance Sampling

- Divide production in lots of size N
- Sample n out of N for inspection
- Count number of defects z
- Compare z with acceptance number c
- c and n are chosen according to an AOQL plan (quality level desired)

Acceptance sampling criticized by Deming and others (Biemer, Caspar and yours truly) but sometimes the process is too chaotic for SPC due to, for instance, high operator turnover


Lot of size *N* produced by coder
Sample of size *n* drawn from lot
Verify sample of size *n* using independent verification
Number of errors discovered is *z*
z < c+1 ?
No
Yes
Lot accepted
Verify *N-n* remaining objects in lot
Error rate acceptable?
No
Yes
Lot accepted
Begin 100% verification of coder's work

Acceptance Sampling for Samples of Size *n* from Lots of Size *N* Using Acceptance Number *c*

Problems with Inspection

- Costly
- Inspection must be nearly perfect
- 100% inspection required for control at very small error levels
- Responsibility for improving quality given to inspectors
- Implies operators are responsible for all errors
- Feedback is usually ineffective

Automated Coding

- There should be a computer-stored dictionary
- Responses are entered online or via some other medium like scanning or keying
- Responses are matched with dictionary descriptions and based on that matching the responses are coded by the software or transferred to manual coding
- By collecting and analyzing process data the system is continually improved

Levels of Automation

- Computer Assisted Coding
- Automated
- Matching can be exact or inexact
- Coding degrees obtained:

  -Purchases 73% (Sweden)

  -Industry and occupation 63% (US)

  -Occupation 75% (Sweden)

Paradata in Coding

- Coding degree in AC and MC
- Effects in coding degree by updates of dictionary
- Coding degree by category, AC and MC
- Coding error rate by coders, categories, digit-level, coding mode and dictionary update version
- CAC consultation degree by category and coder

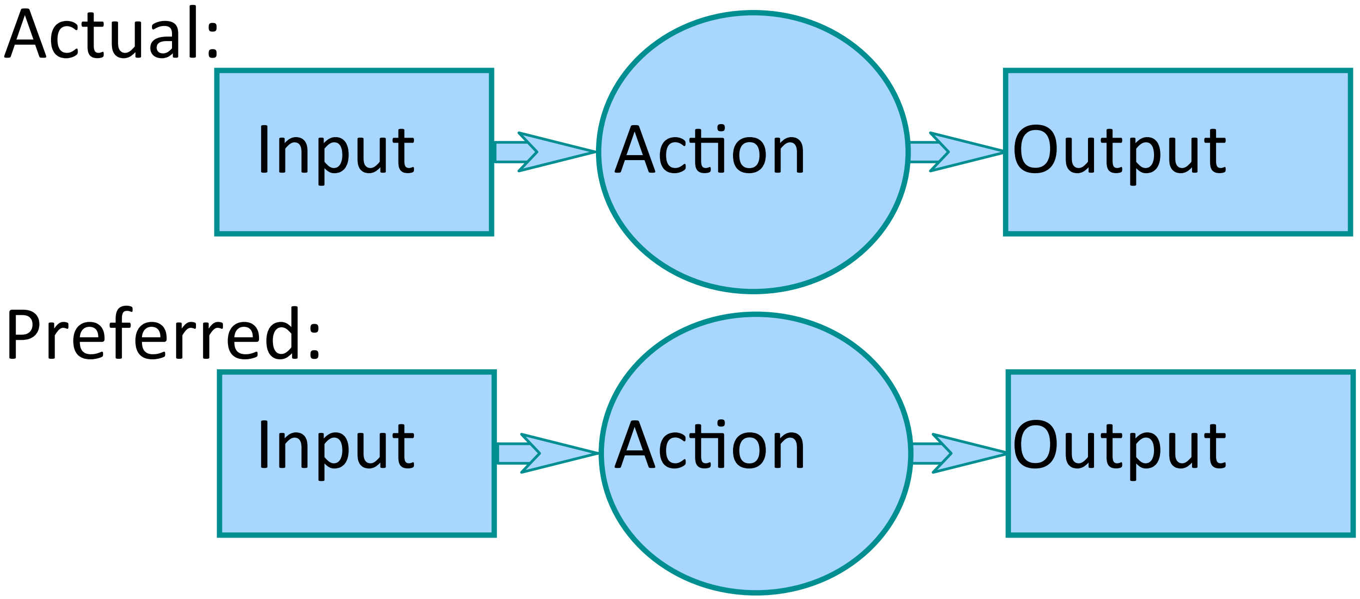
File Preparation

- Attaching weights to each unit
- Final weight is a product of base weight and adjustment factors for nonresponse and noncoverage
- Computation can be difficult
- Application of disclosure avoidance techniques, macrodata and microdata

Paradata in file preparation

- Standard errors themselves
- Time spent checking and correcting weights
- Number of cases where SDC methods failed to protect the data
- Indicators of suitability of disclosure-limited data products

The Quality Improvement Model for Survey Operations Revisited



Actual:

Input → Action → Output

Preferred:

Input → Action → Output

Noncomformity = Actual − Preferred