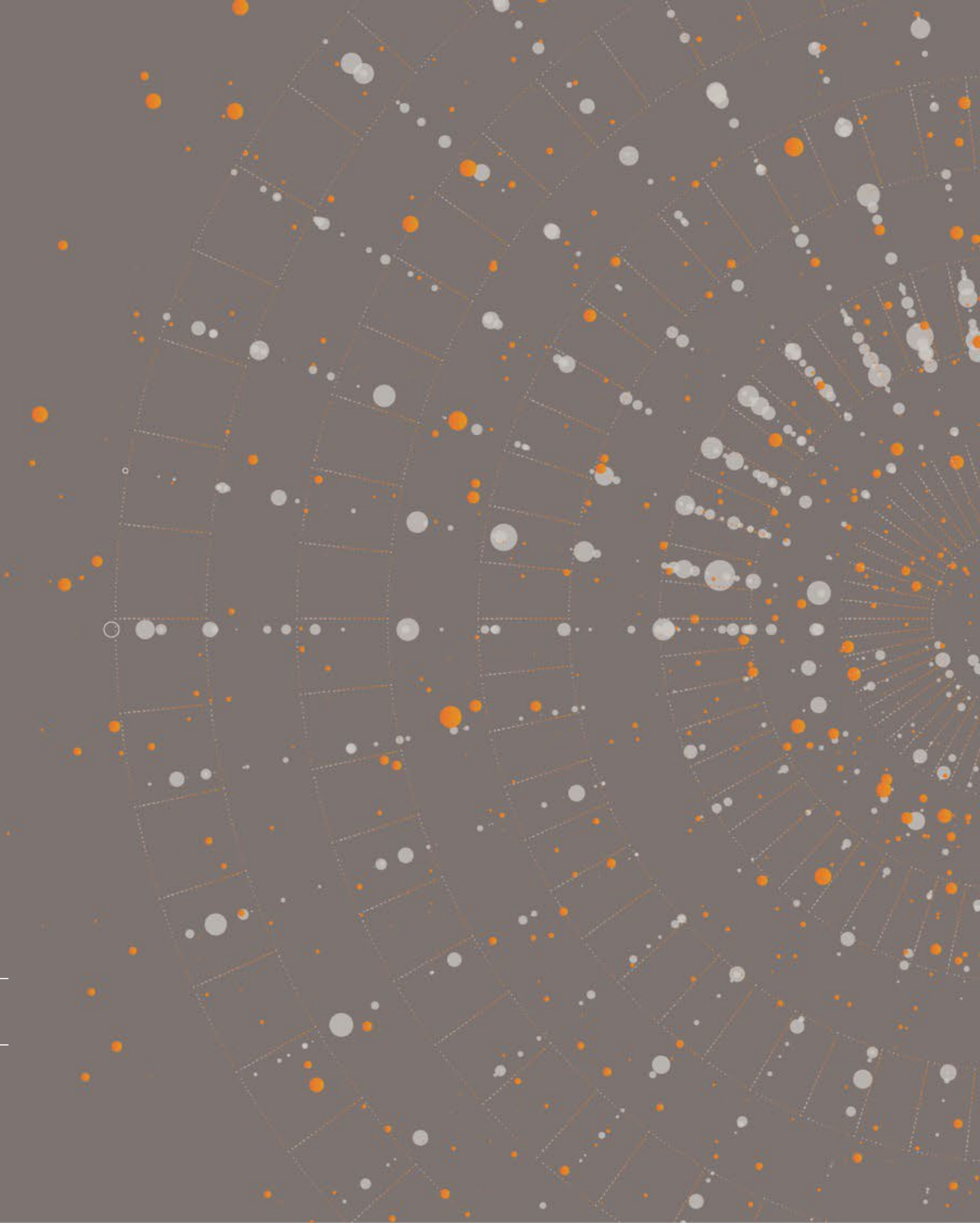
NORC at the University of Chicago

Considerations for defining a framework for reproducibility in survey data processing

FCSM Research & Policy Conference 2024

2024-10-24

Kiegan Rice, Stas Kolenikov, Amy Ihde, Quentin Brummet, Karen Grigorian, Lauren Seward



Agenda

01 What is reproducibility?

02 Reproducibility framework

03 Framework application: EPOP

04 Conclusions

What is reproducibility?

Reproducibility is the ability to **complete a task**, achieve an **output** from that task, **recomplete the steps** involved in that task, and achieve **the same output**.

What is an analytic task?

- **Any work involving the creation, manipulation, analysis or presentation of data or summaries of those data.**
- **Use of statistical software (SAS, R, Stata, Python, etc.) to act upon data files to produce output.**
- **Examples of analytic tasks in survey data processing**
  - Creating a sampling file
  - Creating survey weights
  - Creating a derived variable from a survey extract
  - Performing disclosure review
  - Calculating crosstabulations for a codebook

To achieve reproducibility of an analytic task:

Can questions about the methods used and steps completed in the analytic task be **answered quickly and accurately** by any member of the analytic team using the documentation?

Can the analytic task be **recompleted by another analyst** (or at a later date, or on a different machine) and the same result be achieved?

Is there a clear **link between data, code, and resulting output** for any given task?

Reproducibility framework

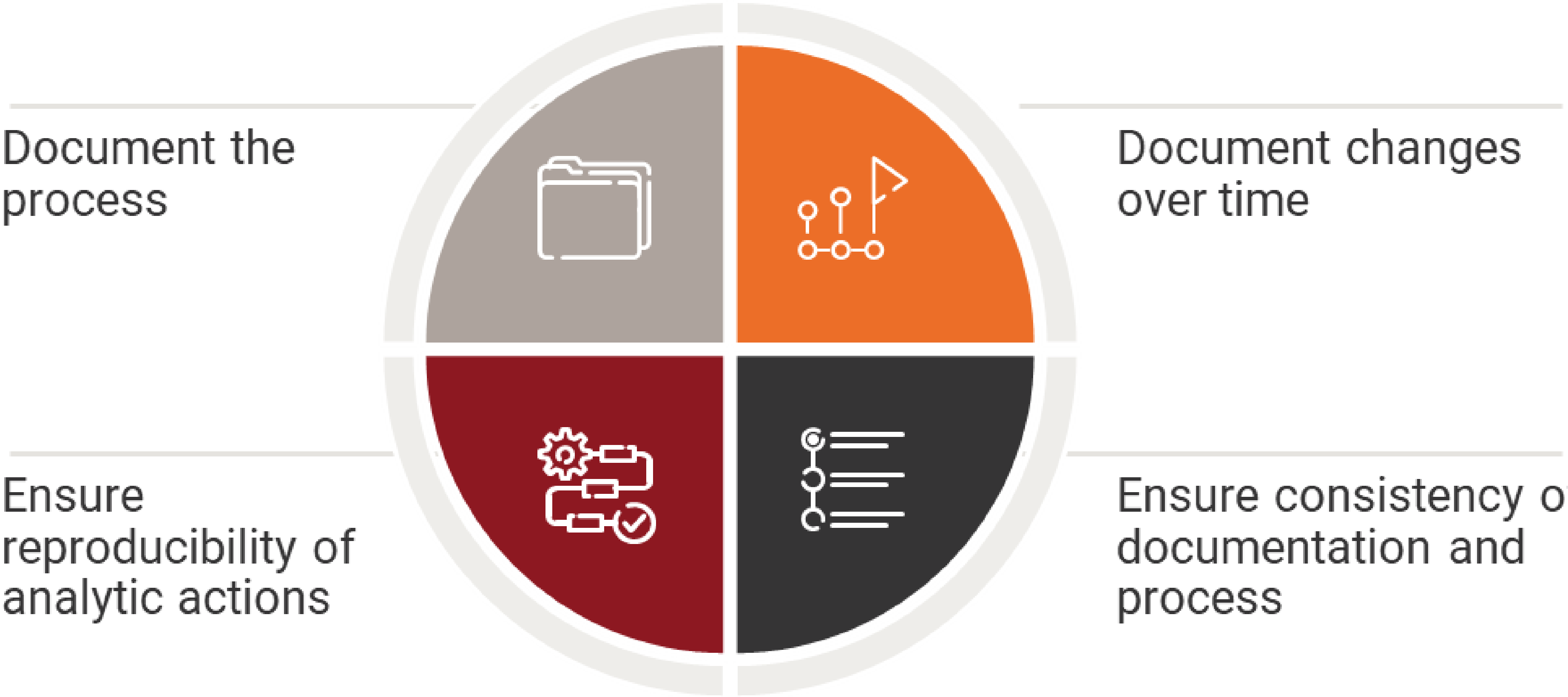
NORC at the University of Chicago

We developed a series of items aimed at determining reproducibility of a survey data production process.

Example items

- Can we tell what actions are taken on the data at each step?
- Can we tell what files are involved in each step?
- Do we know what changes have been made in the code, when they were made, and why they were made?
- Can a specific step of the process be identified and checked for consistency with documentation?
- Can the code/scripts to complete each step be re-run without any adjustment?

We organize these items into four main principles of reproducibility on survey data production.



Document the process

Document changes over time

Ensure reproducibility of analytic actions

Ensure consistency of documentation and process

Principle	Scorecard item	Item met?
Document the process	Can we tell what actions are taken on the data at each step?	
	Can we tell what files are involved in each step?	
	Can the documentation and process be discovered without intervention from the task lead?	
Ensure reproducibility of analytic actions	Can we reproduce the computing environment the code was originally run in?	
	Can the code/scripts to complete each step be re-run without any adjustment, or if adjustments are required, are they documented?	
	Are all input and code file dependencies reachable by another team member or reviewer and over time?	
Document changes over time	Do we know what the process was on a specific date?	
	Do we know what input, code, and output file versions were used on a specific date or for a specific version of the process?	
	Do we know what changes have been made in the code, when they were made, and why they were made?	
Ensure consistency of documentation and process	Can a specific step of the process be identified and checked for consistency with documentation?	
	Has the process gone through a QA/QC process by another team member or reviewer?	
	Are there summaries produced that can be checked for consistency with the goals of each step in the process?	

REPRODUCIBILITY FRAMEWORK : **SCORECARD ITEMS**

Principle	Scorecard item	Item met?
Document the process	Can we tell what actions are taken on the data at each step?	✓
	Can we tell what files are involved in each step?	✓
	Can the documentation and process be discovered without intervention from the task lead?	
Ensure reproducibility of analytic actions	Can we reproduce the computing environment the code was originally run in?	✓
	Can the code/scripts to complete each step be re-run without any adjustment, or if adjustments are required, are they documented?	✓
	Are all input and code file dependencies reachable by another team member or reviewer and over time?	✓
Document changes over time	Do we know what the process was on a specific date?	
	Do we know what input, code, and output file versions were used on a specific date or for a specific version of the process?	
	Do we know what changes have been made in the code, when they were made, and why they were made?	
Ensure consistency of documentation and process	Can a specific step of the process be identified and checked for consistency with documentation?	✓
	Has the process gone through a QA/QC process by another team member or reviewer?	✓
	Are there summaries produced that can be checked for consistency with the goals of each step in the process?	✓

REPRODUCIBILITY FRAMEWORK : **SCORECARD ITEMS**

Principle	Scorecard item	Process 1	Process 2	Process 3
Document the process	Can we tell what actions are taken on the data at each step?	✓	✓	
	Can we tell what files are involved in each step?	✓		✓
	Can the documentation and process be discovered without intervention from the task lead?			
Ensure reproducibility of analytic actions	Can we reproduce the computing environment the code was originally run in?	✓		
	Can the code/scripts to complete each step be re-run without any adjustment, or if adjustments are required, are they documented?	✓		
	Are all input and code file dependencies reachable by another team member or reviewer and over time?	✓	✓	✓
Document changes over time	Do we know what the process was on a specific date?			
	Do we know what input, code, and output file versions were used on a specific date or for a specific version of the process?			
	Do we know what changes have been made in the code, when they were made, and why they were made?			
Ensure consistency of documentation and process	Can a specific step of the process be identified and checked for consistency with documentation?	✓		
	Has the process gone through a QA/QC process by another team member or reviewer?	✓	✓	✓
	Are there summaries produced that can be checked for consistency with the goals of each step in the process?	✓	✓	✓

REPRODUCIBILITY FRAMEWORK : **SCORECARD ITEMS**

Principle	Scorecard item	Process 1	Process 2	Process 3
Document the process	Can we tell what actions are taken on the data at each step?	✓	✓	
	Can we tell what files are involved in each step?	✓		✓
	Can the documentation and process be discovered without intervention from the task lead?			
Ensure reproducibility of analytic actions	Can we reproduce the computing environment the code was originally run in?	✓		
	Can the code/scripts to complete each step be re-run without any adjustment, or if adjustments are required, are they documented?	✓		
	Are all input and code file dependencies reachable by another team member or reviewer and over time?	✓	✓	✓
Document changes over time	Do we know what the process was on a specific date?			
	Do we know what input, code, and output file versions were used on a specific date or for a specific version of the process?			
	Do we know what changes have been made in the code, when they were made, and why they were made?			
Ensure consistency of documentation and process	Can a specific step of the process be identified and checked for consistency with documentation?	✓		
	Has the process gone through a QA/QC process by another team member or reviewer?	✓	✓	✓
	Are there summaries produced that can be checked for consistency with the goals of each step in the process?	✓	✓	✓

REPRODUCIBILITY FRAMEWORK : **SCORECARD ITEMS**

Principle	Scorecard item	Process 1	Process 2	Process 3
Document the process	Can we tell what actions are taken on the data at each step?	✓	✓	
	Can we tell what files are involved in each step?	✓		✓
	Can the documentation and process be discovered without intervention from the task lead?			
Ensure reproducibility of analytic actions	Can we reproduce the computing environment the code was originally run in?	✓		
	Can the code/scripts to complete each step be re-run without any adjustment, or if adjustments are required, are they documented?	✓		
	Are all input and code file dependencies reachable by another team member or reviewer and over time?	✓	✓	✓
Document changes over time	Do we know what the process was on a specific date?			
	Do we know what input, code, and output file versions were used on a specific date or for a specific version of the process?			
	Do we know what changes have been made in the code, when they were made, and why they were made?			
Ensure consistency of documentation and process	Can a specific step of the process be identified and checked for consistency with documentation?	✓		
	Has the process gone through a QA/QC process by another team member or reviewer?	✓	✓	✓
	Are there summaries produced that can be checked for consistency with the goals of each step in the process?	✓	✓	✓

Principle	Scorecard item	Process 1	Process 2	Process 3
Document the process	Can we tell what actions are taken on the data at each step?	✓	✓	
	Can we tell what files are involved in each step?	✓		✓
	Can the documentation and process be discovered without intervention from the task lead?			
Ensure reproducibility of analytic actions	Can we reproduce the computing environment the code was originally run in?	✓		
	Can the code/scripts to complete each step be re-run without any adjustment, or if adjustments are required, are they documented?	✓		
	Are all input and code file dependencies reachable by another team member or reviewer and over time?	✓	✓	✓
Document changes over time	Do we know what the process was on a specific date?			
	Do we know what input, code, and output file versions were used on a specific date or for a specific version of the process?			
	Do we know what changes have been made in the code, when they were made, and why they were made?			
Ensure consistency of documentation and process	Can a specific step of the process be identified and checked for consistency with documentation?	✓		
	Has the process gone through a QA/QC process by another team member or reviewer?	✓	✓	✓
	Are there summaries produced that can be checked for consistency with the goals of each step in the process?	✓	✓	✓

Framework application: EPOP

NORC at the University of Chicago

EPOP is a five-year survey project focused on measuring entrepreneurship activity in the U.S.

EPOP

Understanding Entrepreneurship in the US

Survey Background

- Sponsored by the Kauffman Foundation
- 8 categories of entrepreneurship by Metropolitan Statistical Area (MSA), by State, and geographies crossed with race and gender, separately
- Survey data: AmeriSpeak panel, address-based sample, and opt-in only estimates
- Small area estimation (SAE) used
- Direct estimates combined with publicly-available sources in SAE models

EPOP is a five-year survey project focused on measuring entrepreneurship activity in the U.S.

EPOP
Understanding Entrepreneurship in the US

Components of data production

- Data processing & variable creation (initial and final)
- Disclosure review (initial and final)
- Weighting – probability samples
- Weighting – combined prob/non-prob sample
- Codebook preparation

EPOP is a five-year survey project focused on measuring entrepreneurship activity in the U.S.

EPOP
Understanding Entrepreneurship in the US

Reproducibility considerations for data production

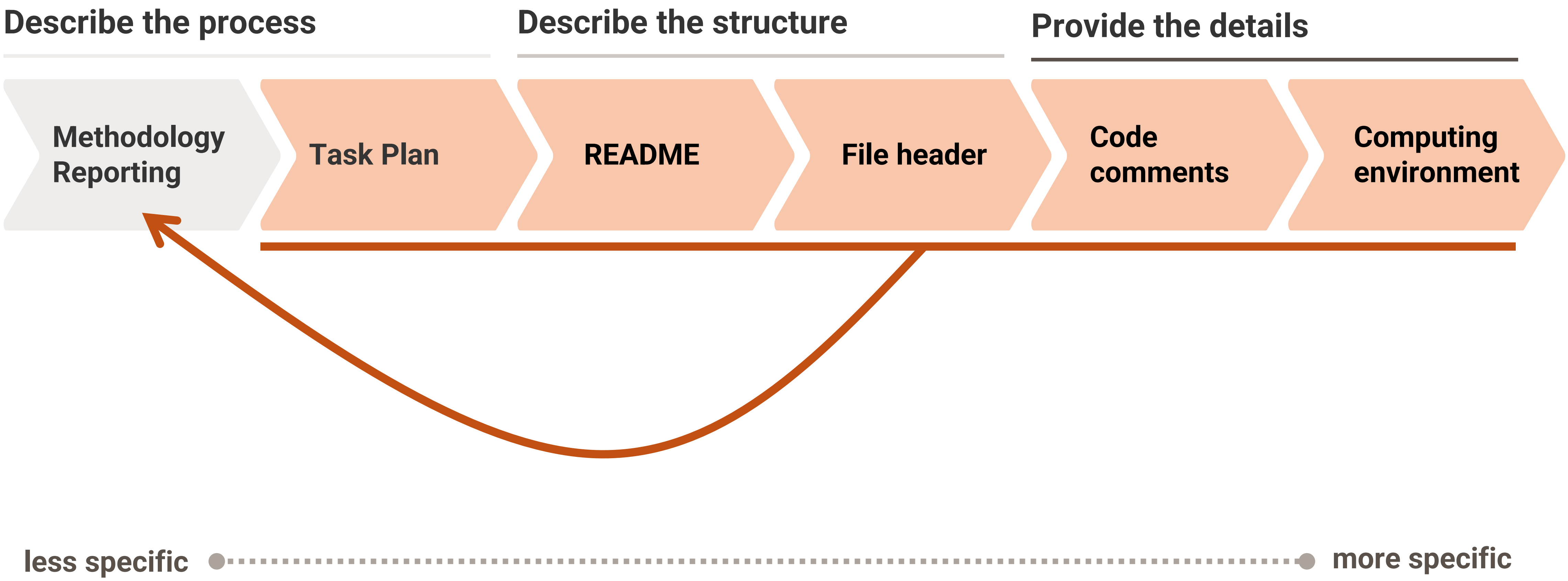
- Reuse of process across years
- Ability to make changes across years when needed
- File handoffs between distinct tasks
- Creation of distinct files: PUF and RUF
- Re-runs of some steps when upstream changes are needed

The resulting reproducibility framework helps us identify and prioritize focus areas for implementation on the project.

Developing consistent documentation across processes

- Task Plans laying out steps and expectations for each task
- READMEs describing file and folder structures
- File headers describing contents of each file
- Code comments
- Computing environment

The resulting reproducibility framework helps us identify and prioritize focus areas for implementation on the project.



Conclusions

NORC at the University of Chicago

Application of our framework and improving the documentation structure on EPOP aided the project in more quickly and effectively completing Y3 data processing.

- Shortened processing time
- QC items caught earlier in the process
- Improved methodological documentation process

Defining a framework and specific dimensions within that framework allows us to systematically review and improve reproducibility on projects.

- Identify areas for growth
- Identify areas of strength
- Allow for flexibility in how each requirement is met
  - Different statistical software
  - Different project needs & expectations

Documentation is important for ensuring all four principles of reproducibility are met.

Document the process

Document changes over time

Ensure reproducibility of analytic actions

Ensure consistency of documentation and process

Thank you.

Kiegan Rice
Senior Statistician
rice-kiegan@norc.org

Research You Can Trust™

NORC at the University of Chicago