Introduction to Qualitative Coding with nCoder

Amanda Siebert-Evenstone, Ph.D.
alevenstone@wisc.edu

Learning Analytics Masters Program
UNIVERSITY OF WISCONSIN–MADISON

Nelson Institute for Environmental Studies
UNIVERSITY OF WISCONSIN–MADISON

Jaeyoon Choi
jaeyoon.choi@wisc.edu

epistemic analytics
turning data into meaning

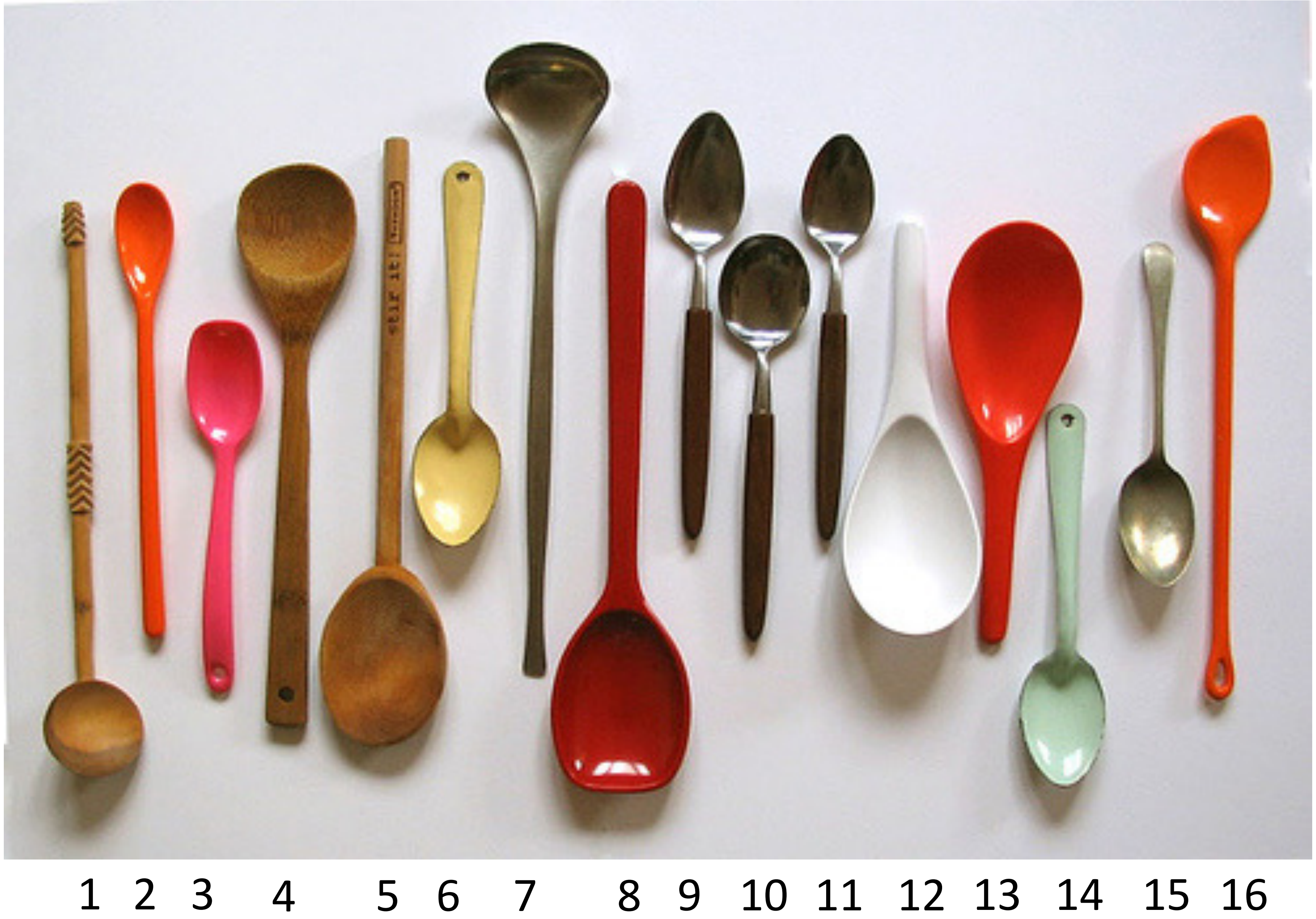
Brett Puetz
bpuetz@wisc.edu

School of Human Ecology
UNIVERSITY OF WISCONSIN–MADISON

Quantitative Research
Qualitative Research

Categories of Spoons

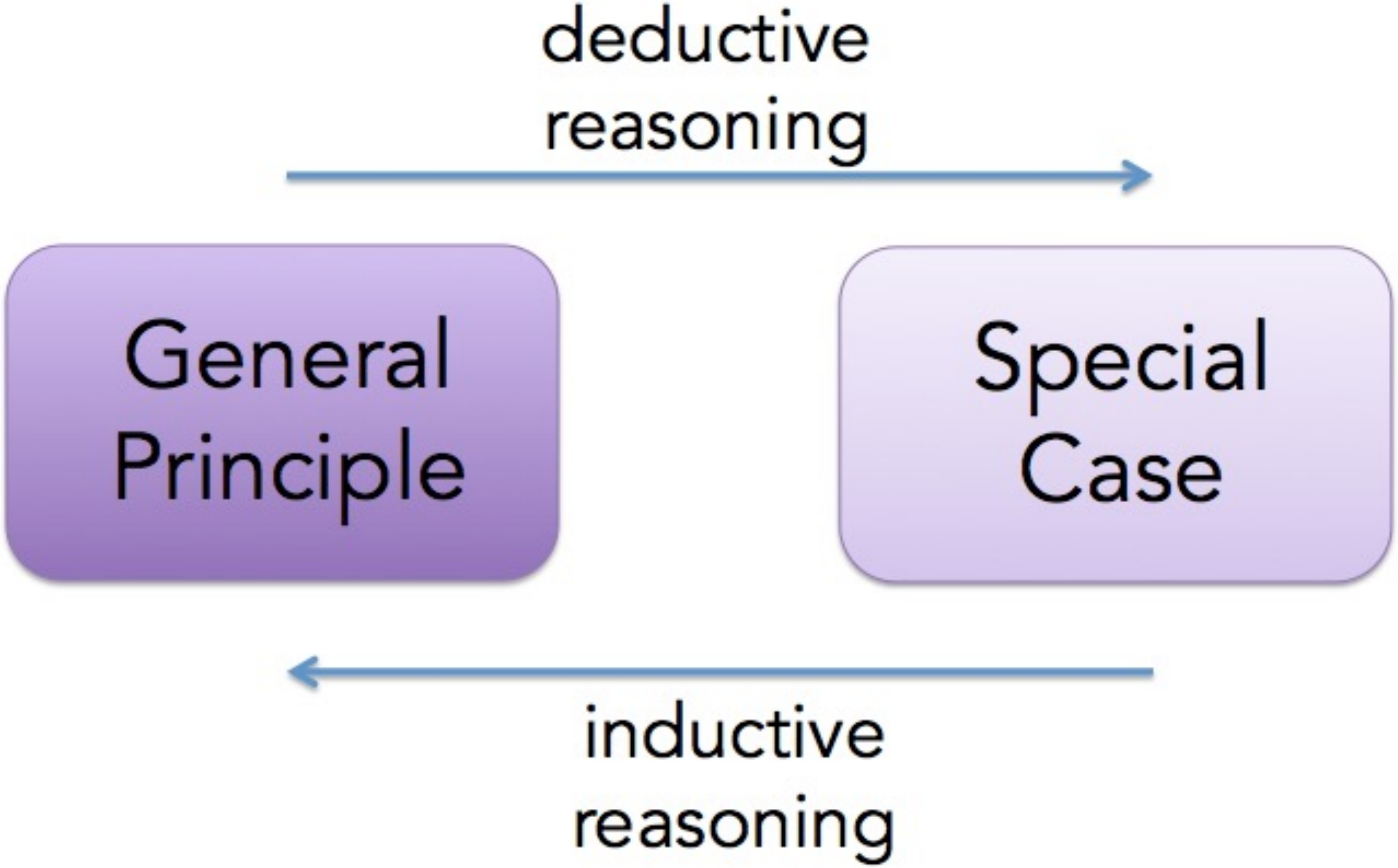
https://go.wisc.edu/0nxl70



Storytelling

Storytelling

What is our systematic explanation?



deductive reasoning

General Principle

Special Case

inductive reasoning

Top-Down

(aka a priori, theoretical, deductive)

Start with theory

Synonyms or Word Associations

Existing coding schemes

Top-Down

(aka a priori, theoretical, deductive)

Start with theory

Synonyms or Word Associations

Existing coding schemes

Bottom-Up

(aka Grounded theory, emergent coding, inductive)

N-grams

TFIDF

Topic Models

Word Counter or TextRazor

SIEBERT-EVENSTONE'S MAXIM

WHEN IN DOUBT, **READ YOUR DATA.**

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Science

Drawing

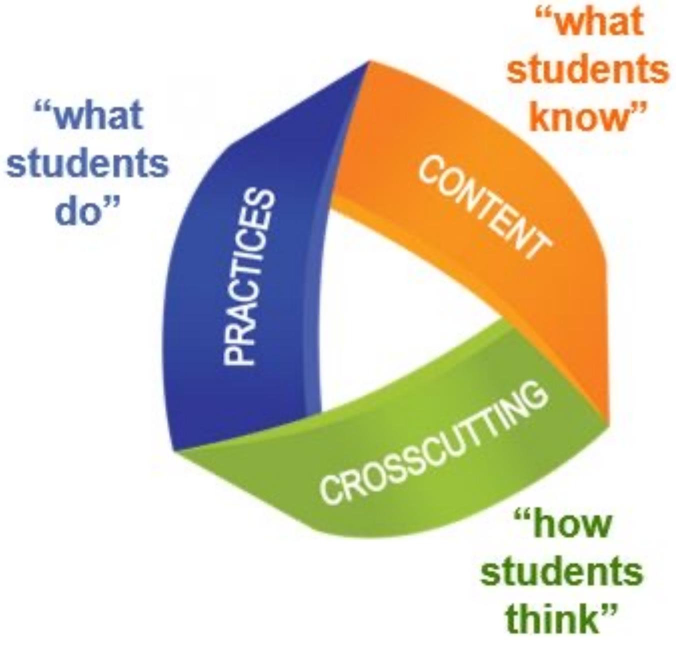
Modeling

Hypothesizing

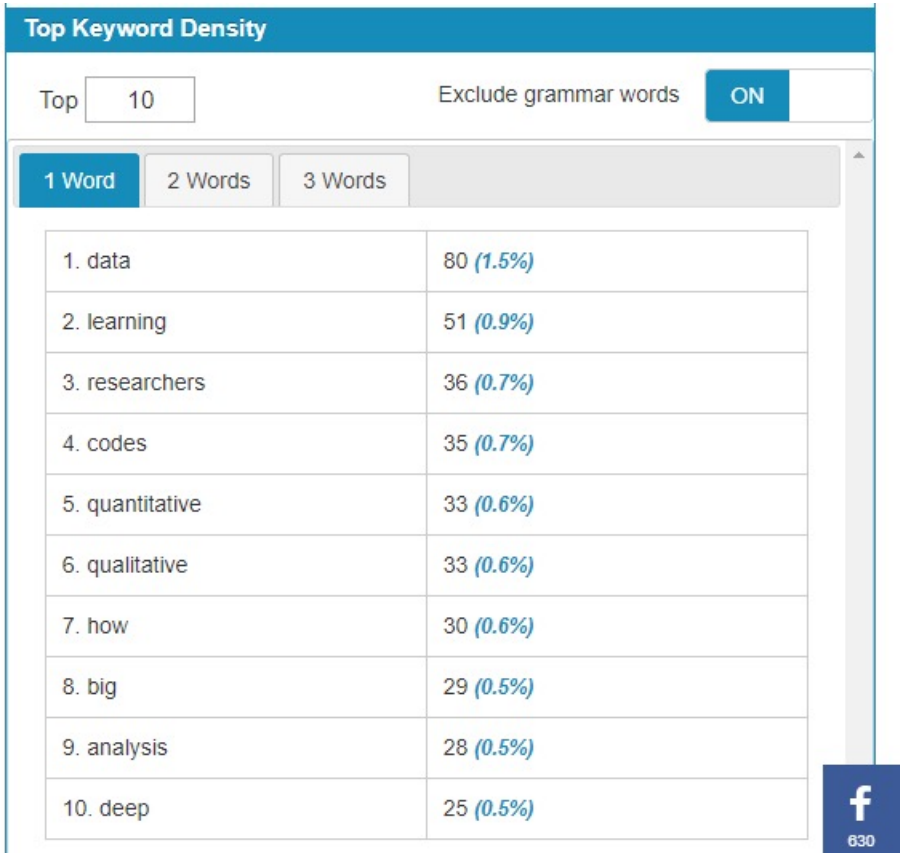
“imagine”
-in vivo code

Coding Process

1. Identify common theories or ideas about the topic
1. Read and get to know the data



Quoted text from Peter A'Hearn



Top Keyword Density

Top 10

Exclude grammar words ON

1. data	80 *(1.5%)*
2. learning	51 *(0.9%)*
3. researchers	36 *(0.7%)*
4. codes	35 *(0.7%)*
5. quantitative	33 *(0.6%)*
6. qualitative	33 *(0.6%)*
7. how	30 *(0.6%)*
8. big	29 *(0.5%)*
9. analysis	28 *(0.5%)*
10. deep	25 *(0.5%)*

Coding Process

1. Identify common theories or ideas about the topic
1. Read and get to know the data
2. Describe each line

How do you study / what is the best way for you to study?	Description
It depends on what I’m trying to learn	
Quizlet	
Flashcards, highlight notes	
It depends on the subject	
Have someone quiz me on the material	
I like using the Quizlet when I need note memorization and vocabulary. For more complex topics, I always do well reading from a textbook and taking handwritten notes as well as completing or working through practice problems.	

Coding Process

1. Identify common theories or ideas about the topic
1. Read and get to know the data
2. Describe each line
3. Identify ideas or codes (quizzing, context-based, processing info)

Coding Process

1. Identify common theories or ideas about the topic
1. Read and get to know the data
2. Describe each line
3. Identify ideas or codes
4. Building wordlists (we’ll get to this later)
5. Building categories of codes

What should I code?

- What’s interesting?
- Why is it interesting?
- Why am I interested in that? (Richards, 2009)

From Hatch 2002:

- Similarity (things happen the same way)
- Difference (they happen in predictably different ways)
- Frequency (they happen often or seldom)
- Sequence (they happen in a certain order)
- Correspondence (they happen in relation to other activities or events)
- Causation (one appears to cause another) (p.155)

(Amanda keeps a cheat sheet and multiple books of ideas to help inspire ideas)

Second (to nth) cycle of coding

- Recode the data because more accurate words and phrases were discovered for the original codes
- Merge together similar codes
- Separate codes that are too large
- Infrequent codes will be assessed for their utility (then kept or dropped)

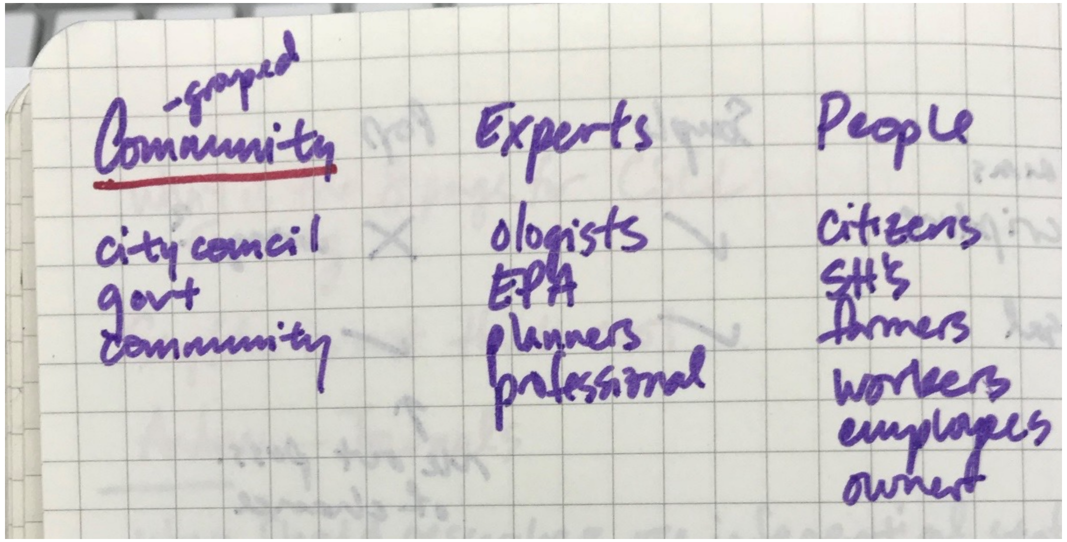
Memo

Memos are ways of summarizing where you are at during your analysis and potential interpretations you may have about your data.

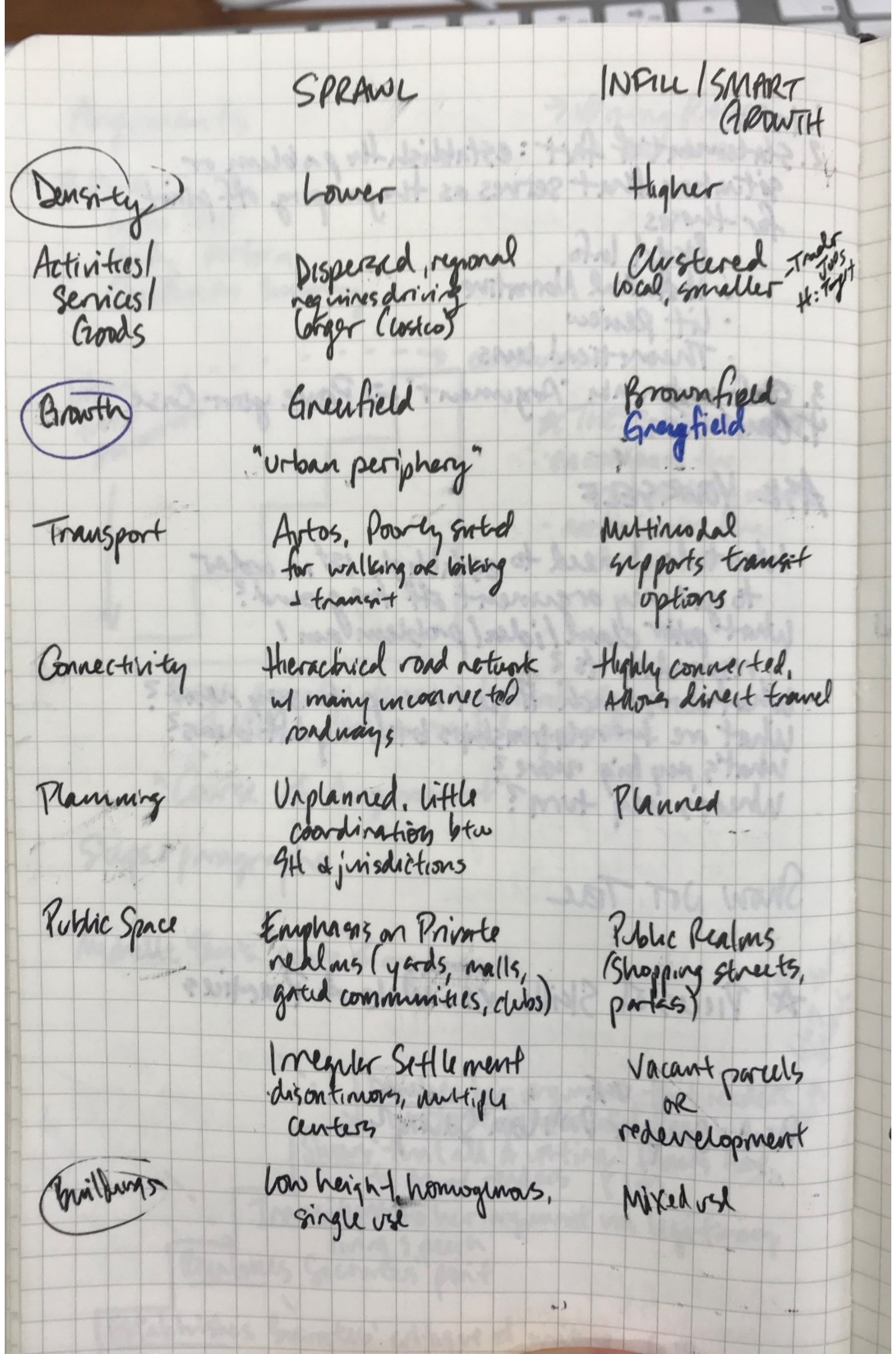
Memo

Memos are ways of summarizing where you are at during your analysis and potential interpretations you may have about your data.

- Codes, categories, and their relationships
- Initial thoughts on data analysis
- pulling together incidents that appear to have commonalities
- proposals for a specific new pattern code
- when the analyst does not have a clear concept in mind but is struggling to clarify one



-groped Community	Experts	People
city council	ologists	citizens
govt	EPA	SH's
community	planners	farmers
	professional	workers
		employees
		owner



	SPRAWL	INFILL / SMART GROWTH
Density	Lower	Higher
Activities / Services / Goods	Dispersed, regional requires driving larger (Costco)	Clustered local, smaller -Trader Joes H: Target
Growth	Greenfield "urban periphery"	Brownfield Greyfield
Transport	Autos, Poorly suited for walking or biking + transit	Multimodal supports transit options
Connectivity	Hierarchical road network w/ many unconnected roadways	Highly connected, Allows direct travel
Planning	Unplanned, little coordination btw SH & jurisdictions	Planned
Public Space	Emphasis on Private realms (yards, malls, gated communities, clubs)	Public Realms (Shopping streets, parks)
	Irregular Settlement discontinuous, multiple centers	Vacant parcels or redevelopment
Buildings	Low height, homogenous, single use	Mixed use

Role of Researcher

YOU are the data collection and analysis instrument

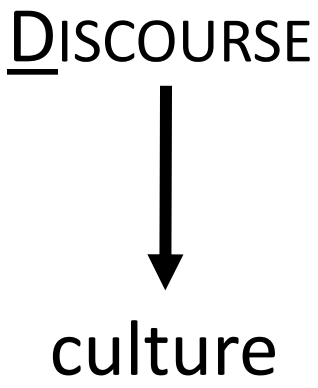
- You take notes and decide what topics to record
- What questions do you ask or not ask?
- What do you deem important?
- What are your implicit/explicit theories?
- What do you value?

Identities
& Cultures
Experience
& Context

"A <u>D</u>ISCOURSE is a socially accepted association among ways of using language, of thinking, feeling, believing, valuing, and of acting that can be used to identify oneself as a member of a socially meaningful group… or to signal (that one is playing) a socially meaningful role."

- Jim Gee

Learning is a process of Enculturation

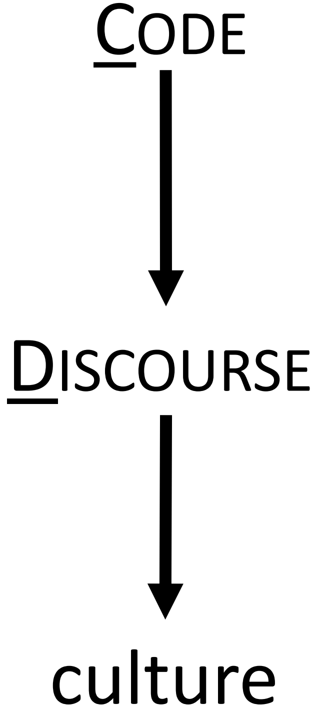


DISCOURSE

culture

CODES

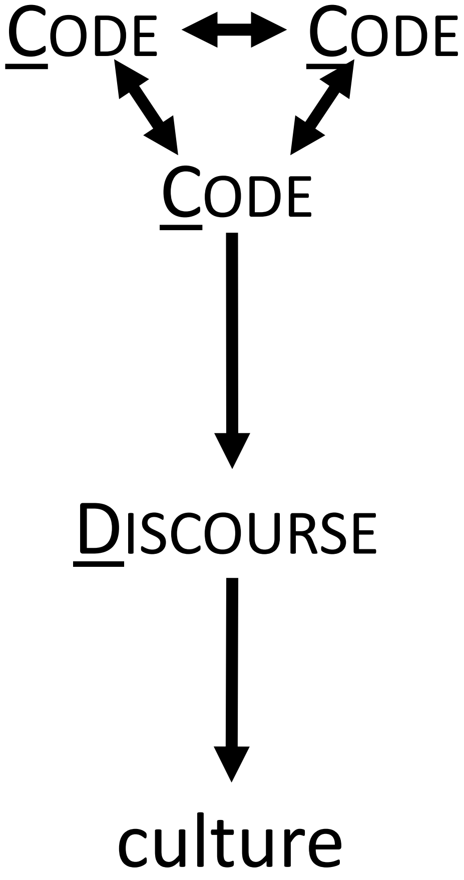
Culturally-relevant and meaningful aspects of a DISCOURSE



CODE

DISCOURSE

culture



CODE

CODE

CODE

DISCOURSE

culture

CODES

Culturally-relevant and meaningful aspects of a DISCOURSE

code → CODE

codes

Things that count as evidence or *warrants* for CODES

Codebooks

Codebooks

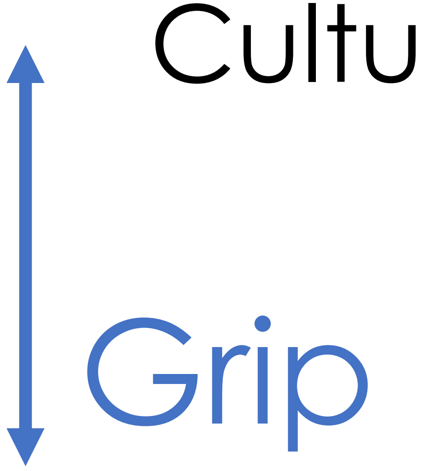
Names	Definition	Examples
Performance Metrics	Discussion of one or more criteria for device functionality: agility, payload, cost, recharge interval, and/or safety.	***I thought that safety near the maximum was not very good (close to 225 - one had 218 RPN), but other than that I was fine with the safety as long as it was around 200 or lower.***

Codebooks

Names	Definition	Examples
Performance Metrics	Discussion of one or more criteria for device functionality: agility, payload, cost, recharge interval, and/or safety.	***I thought that safety near the maximum was not very good (close to 225 - one had 218 RPN), but other than that I was fine with the safety as long as it was around 200 or lower.***
CODES	CODES	codes

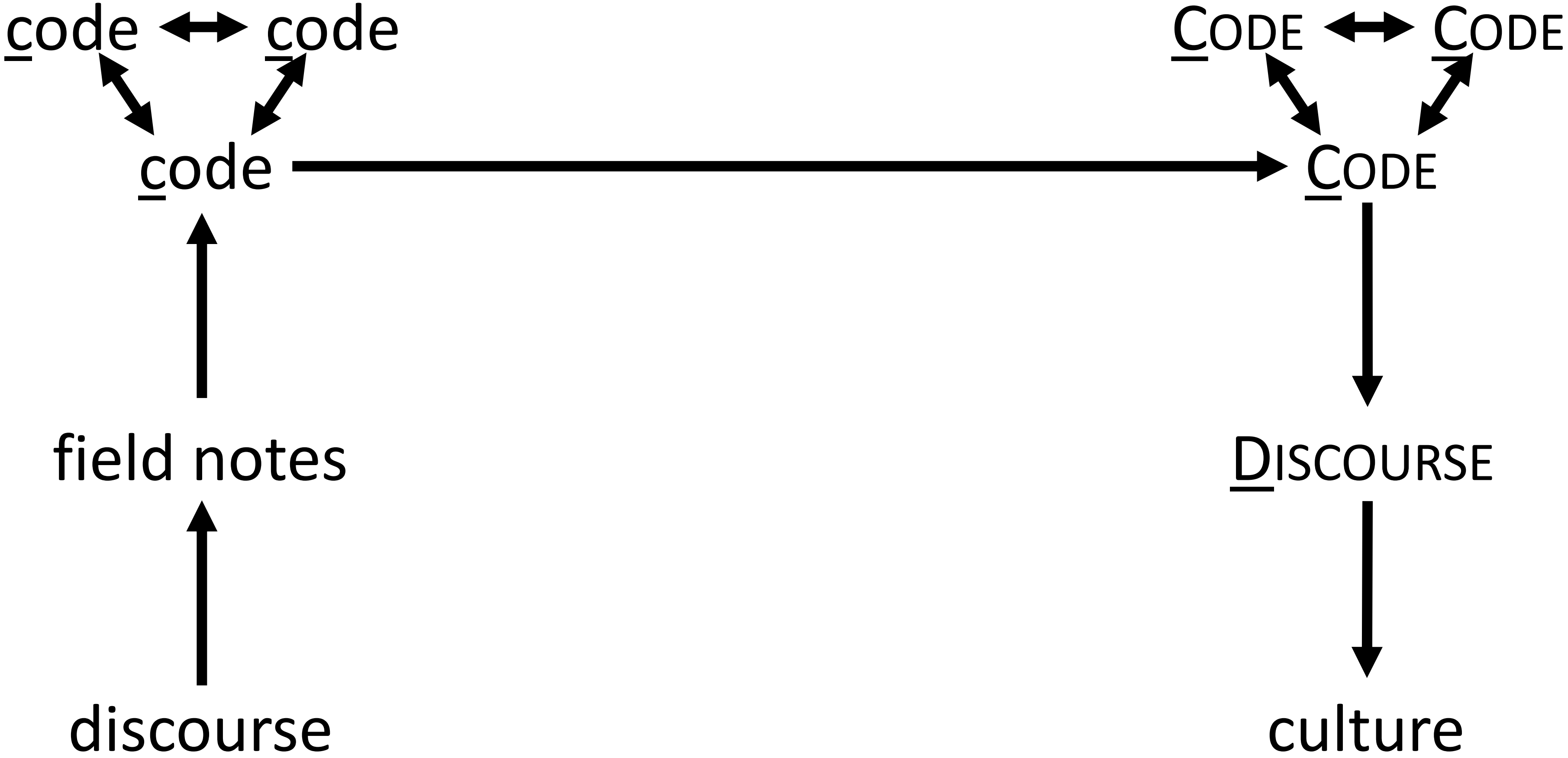
CODES

Culturally-relevant and meaningful aspects of a DISCOURSE



codes

Things that count as evidence or *warrants* for CODES



code ↔ code

code

field notes

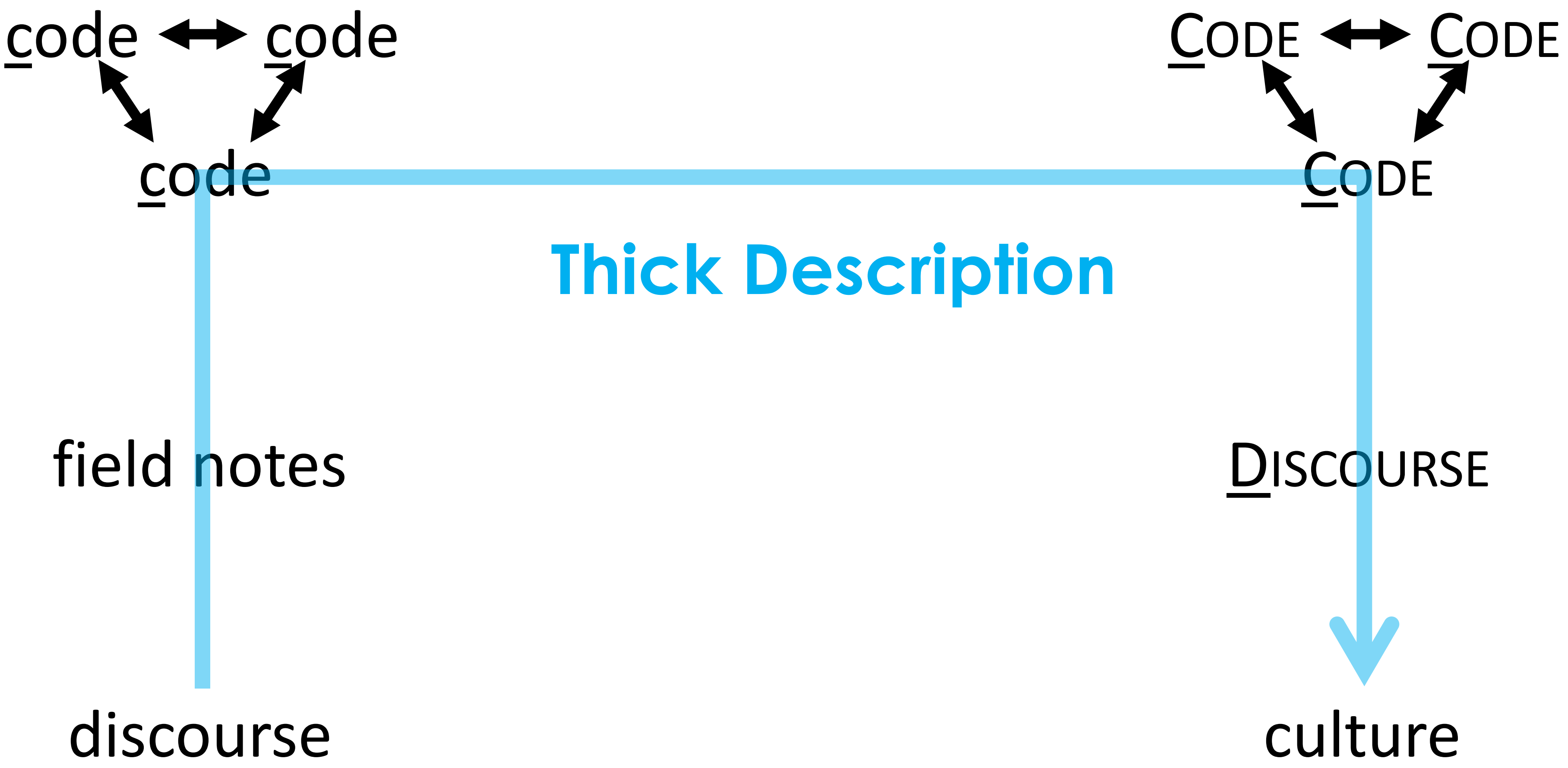
discourse

CODE ↔ CODE

CODE

DISCOURSE

culture



code ↔ code

code

field notes

discourse

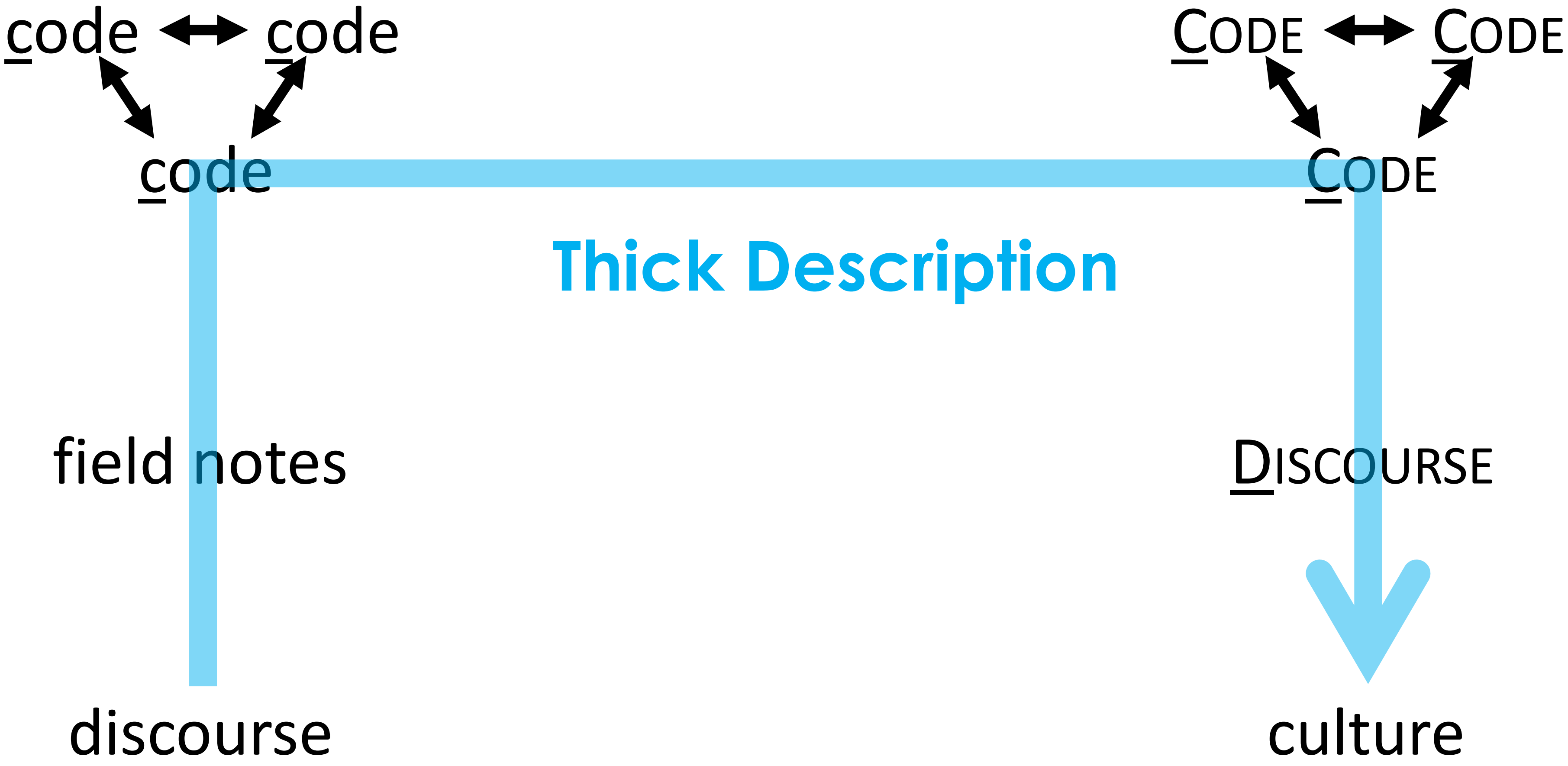
Thick Description

CODE ↔ CODE

CODE

DISCOURSE

culture



code ↔ code

code

field notes

discourse

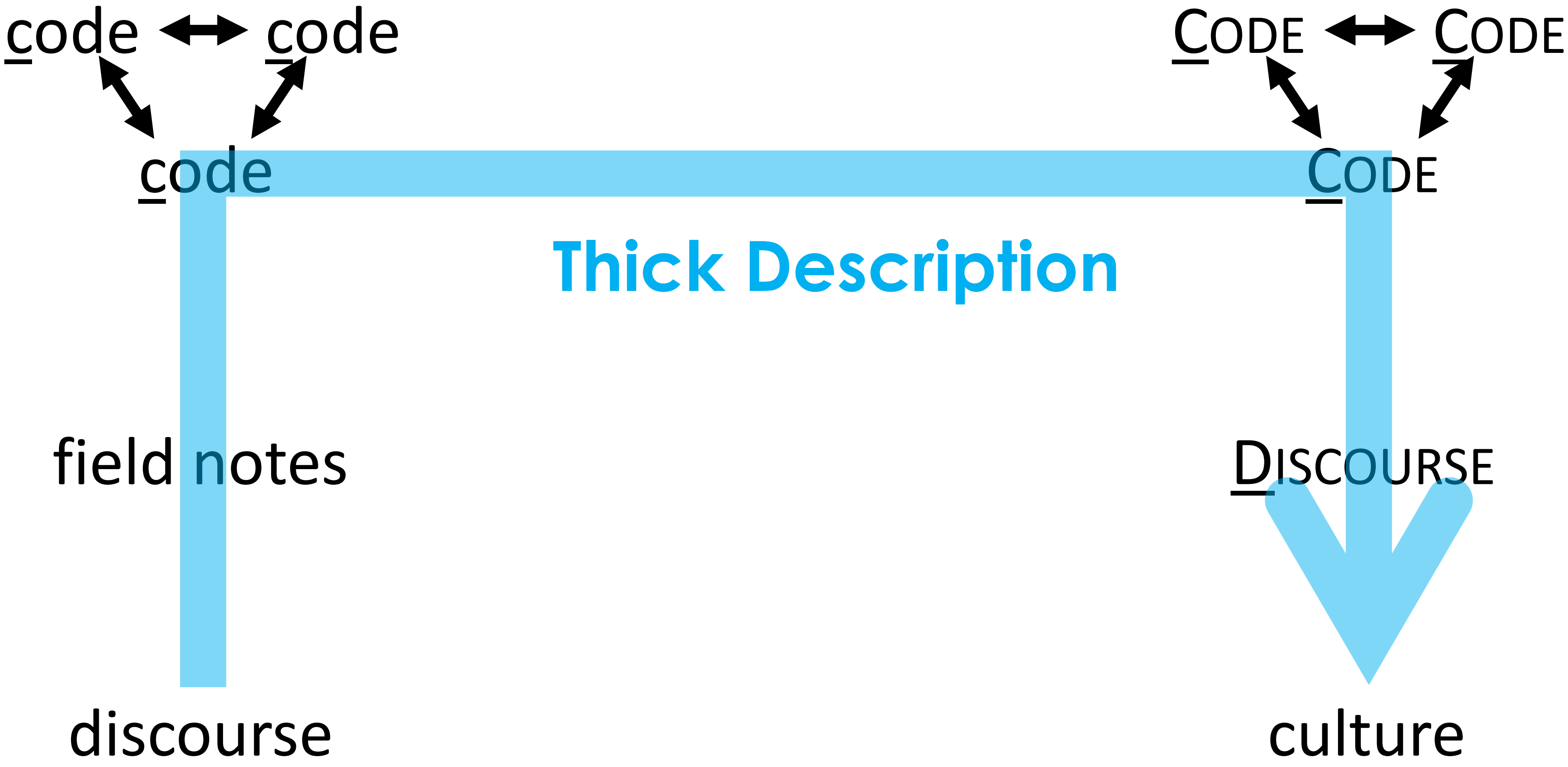
Thick Description

CODE ↔ CODE

CODE

DISCOURSE

culture



code ↔ code

code

Thick Description

field notes

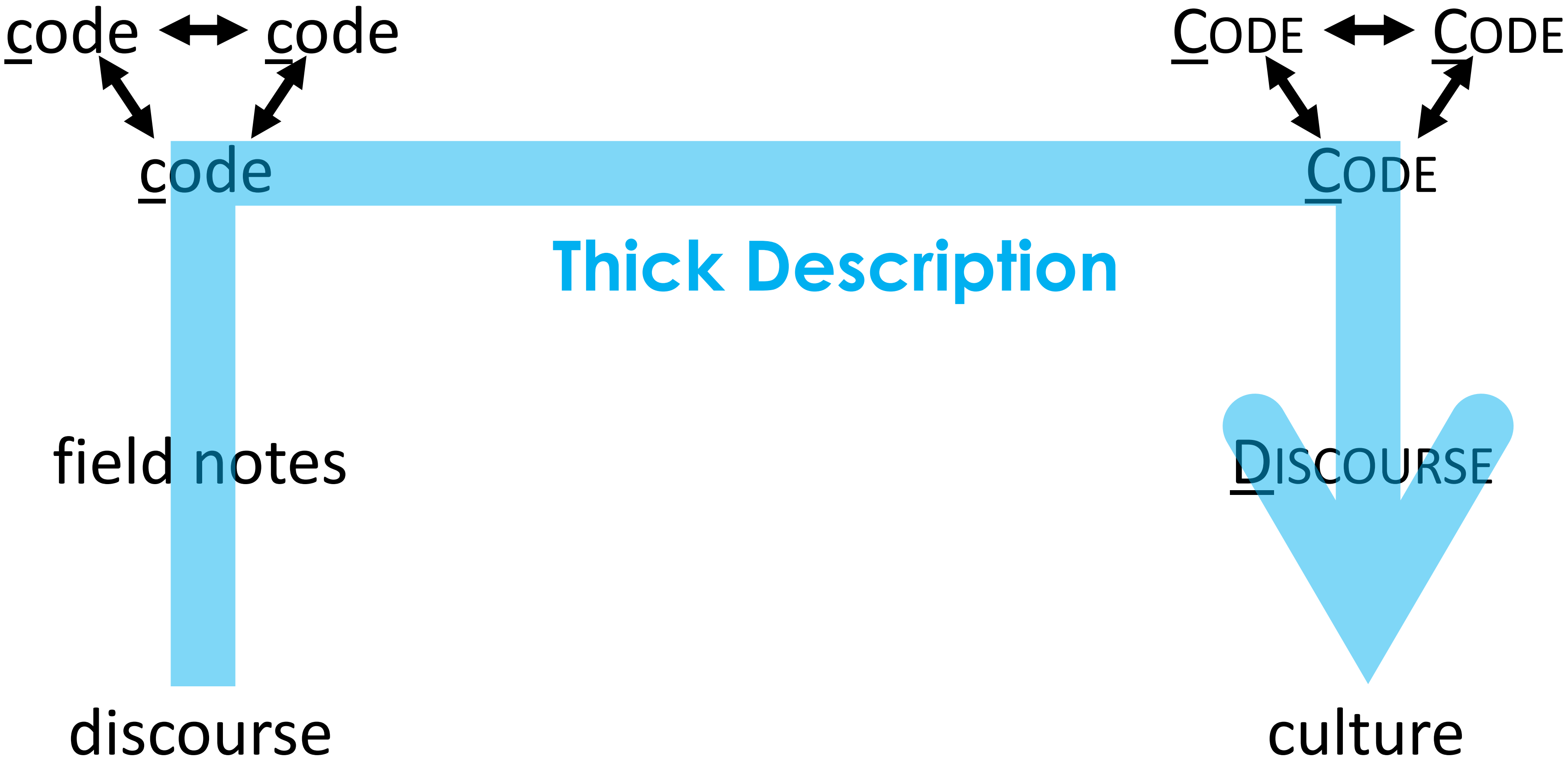
discourse

CODE ↔ CODE

CODE

DISCOURSE

culture



code ↔ code

code

field notes

discourse

Thick Description

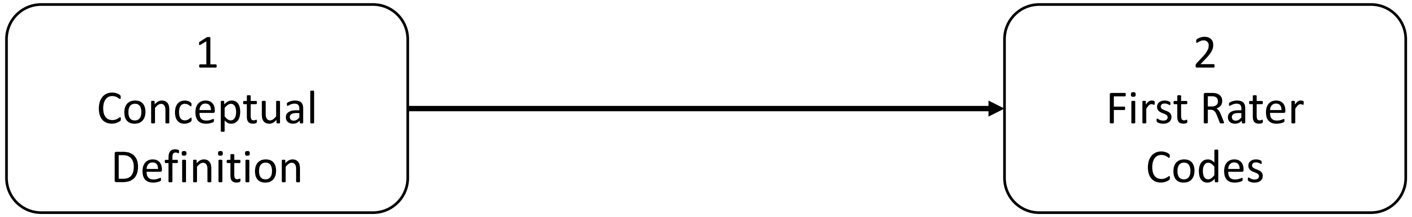
CODE ↔ CODE

CODE

DISCOURSE

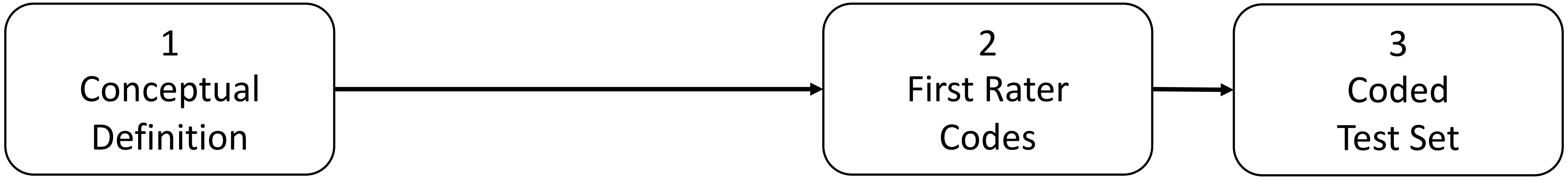
culture

1
Conceptual
Definition



1
Conceptual
Definition

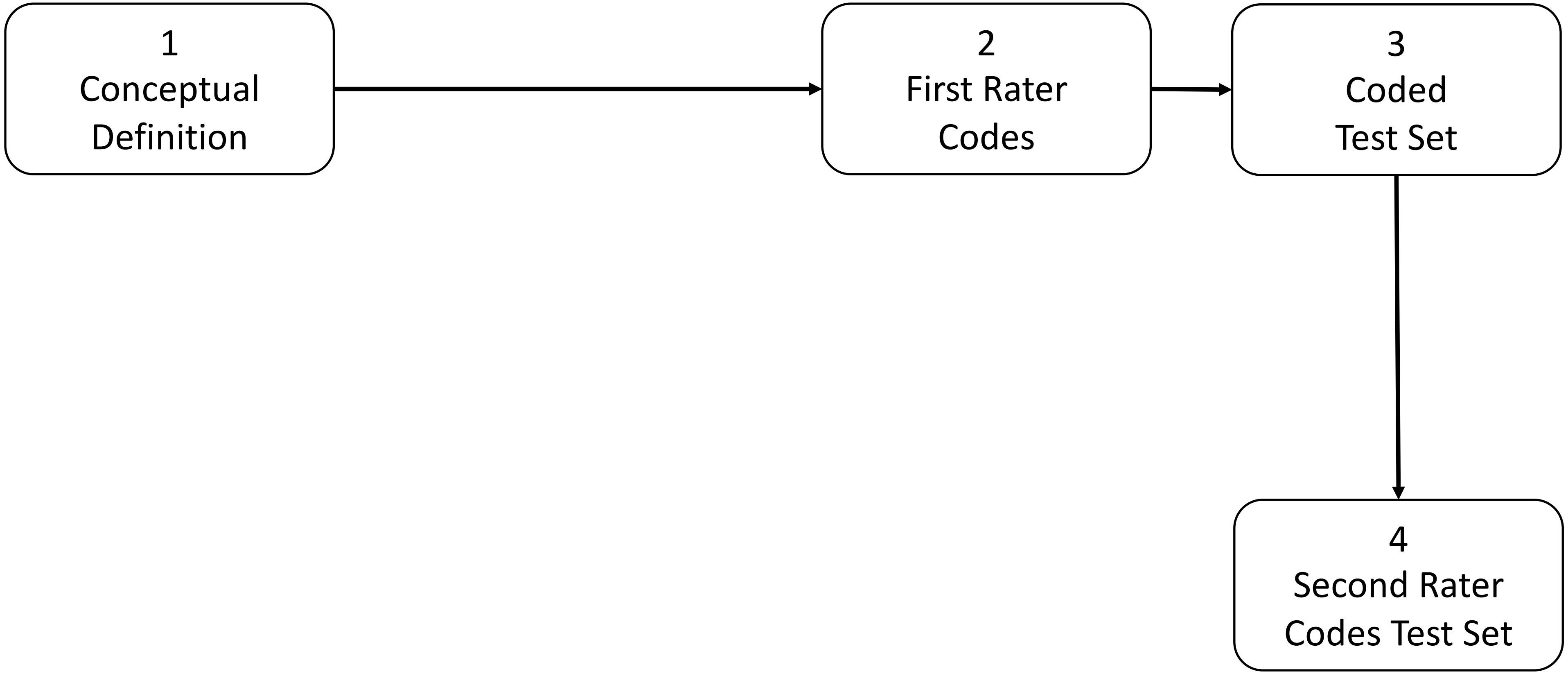
2
First Rater
Codes



1
Conceptual
Definition

2
First Rater
Codes

3
Coded
Test Set

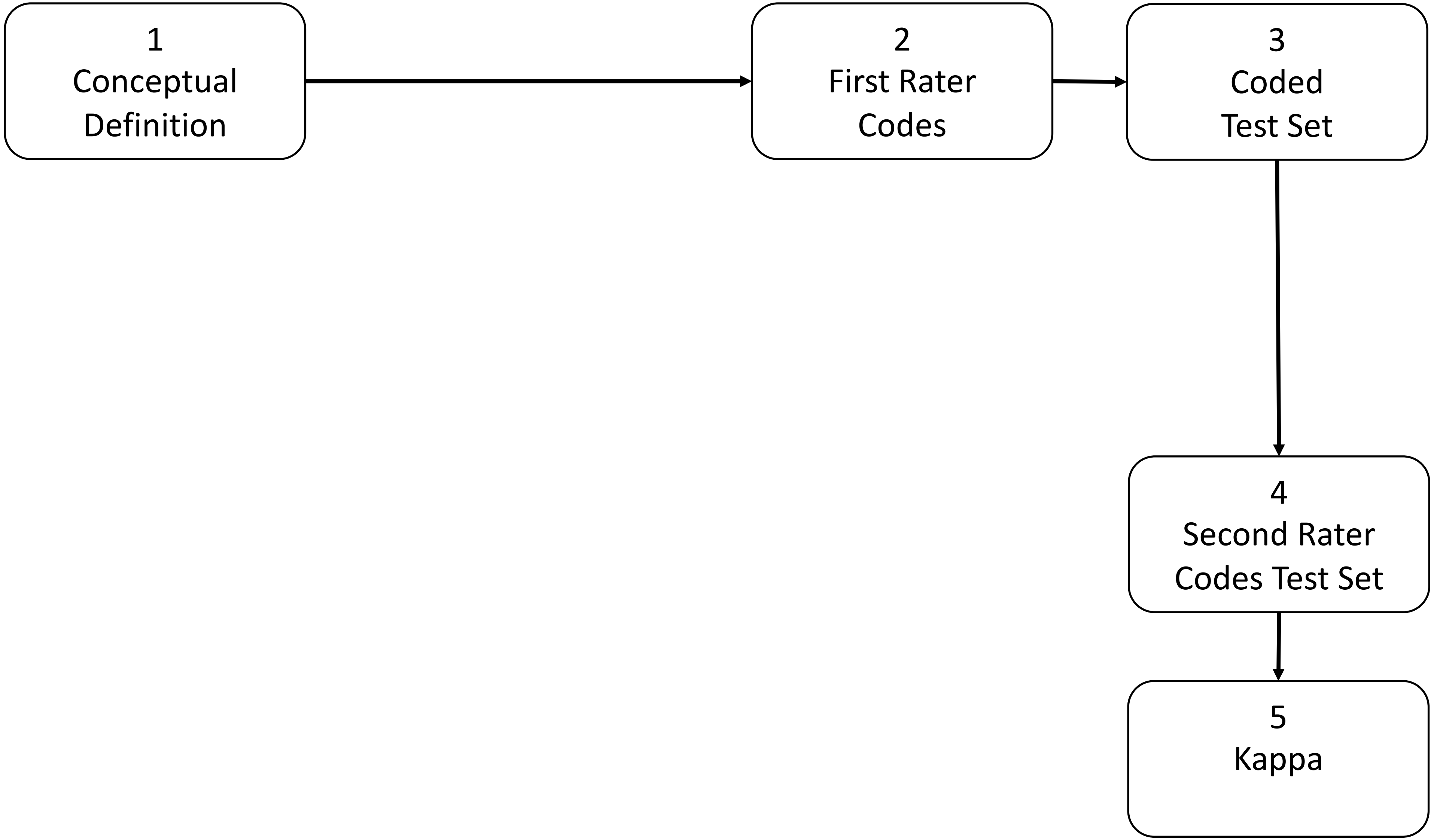


1
Conceptual Definition

2
First Rater Codes

3
Coded Test Set

4
Second Rater Codes Test Set



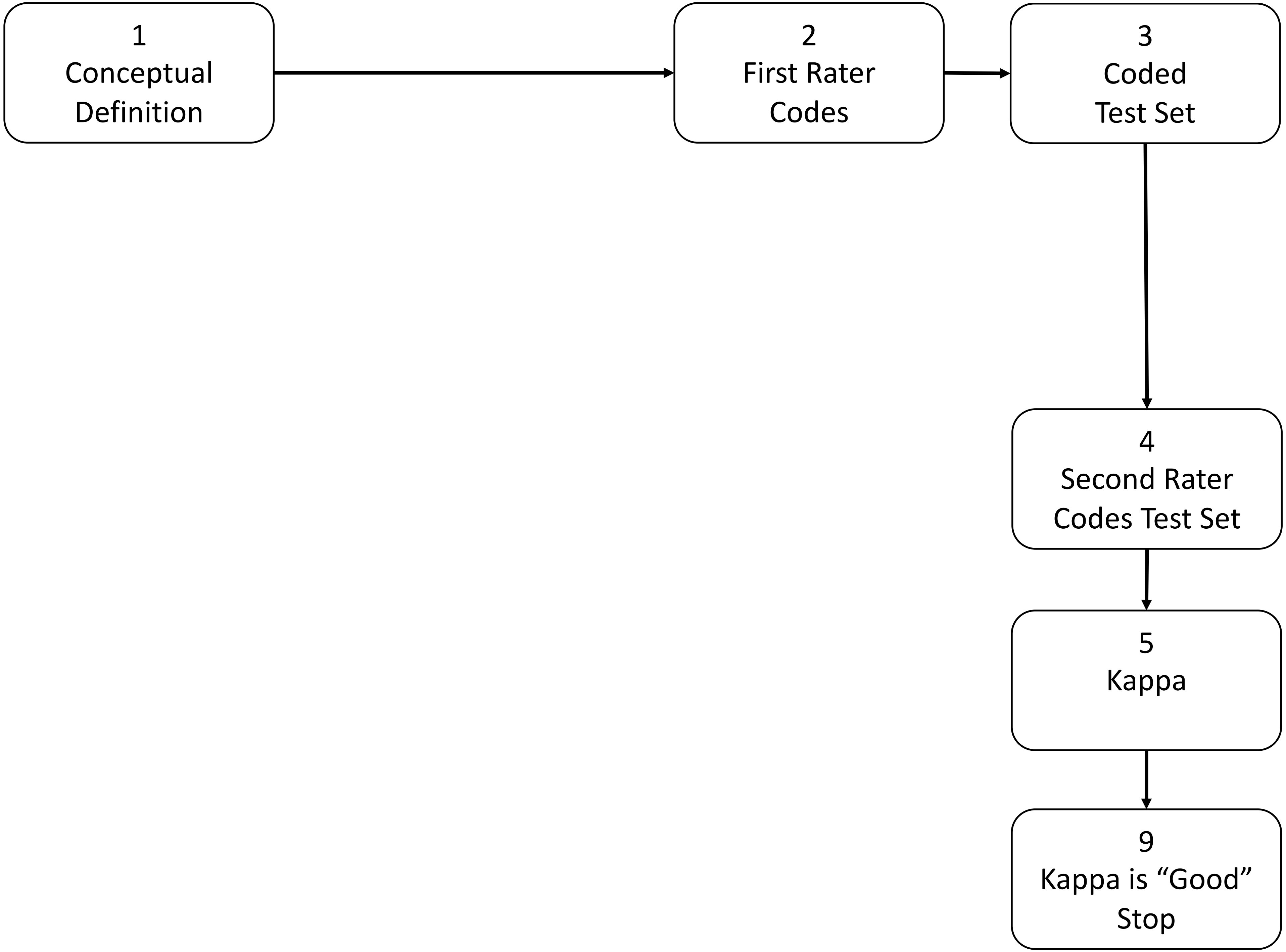
1
Conceptual
Definition

2
First Rater
Codes

3
Coded
Test Set

4
Second Rater
Codes Test Set

5
Kappa



1
Conceptual
Definition

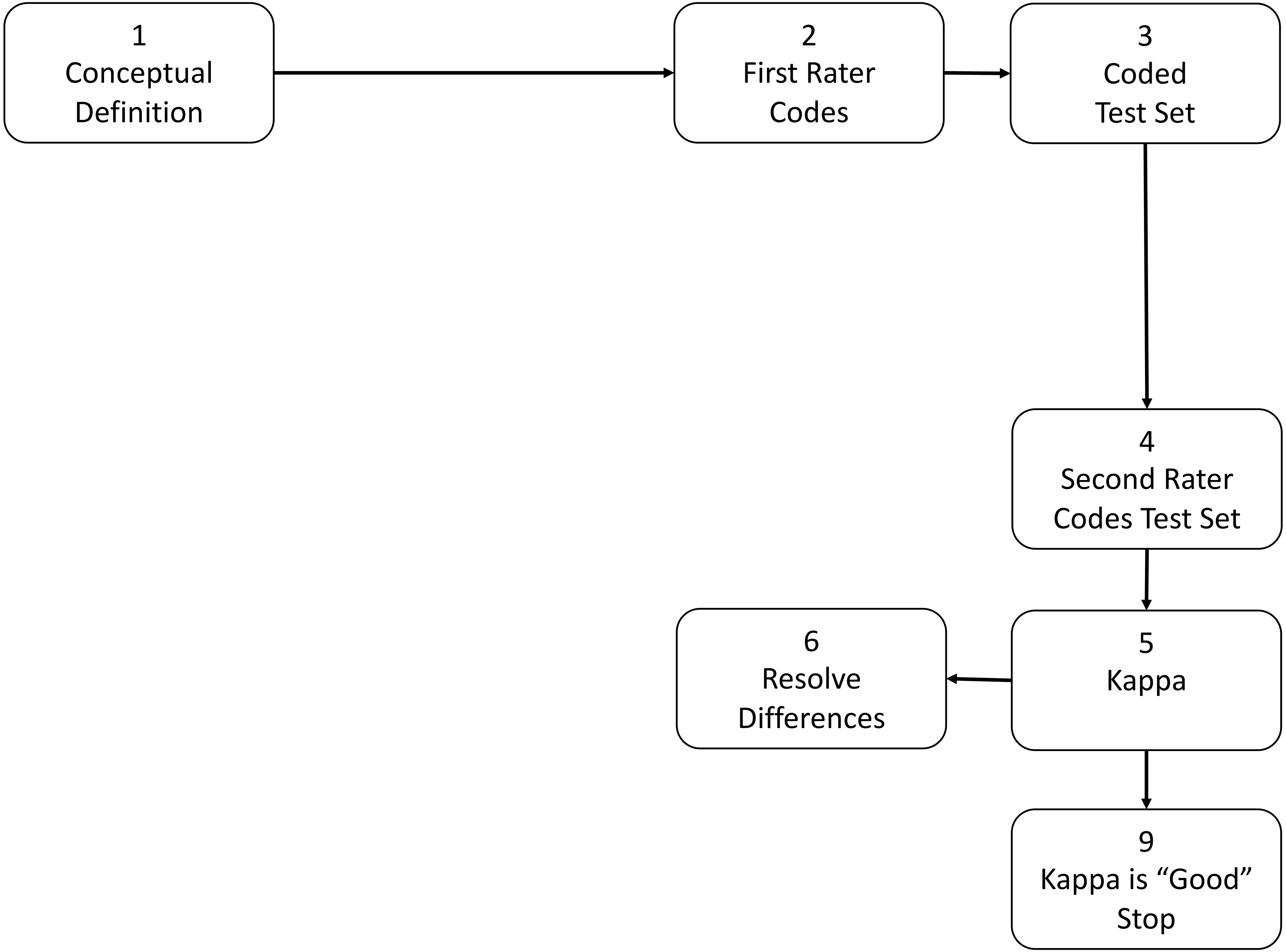
2
First Rater
Codes

3
Coded
Test Set

4
Second Rater
Codes Test Set

5
Kappa

9
Kappa is “Good”
Stop



1
Conceptual
Definition

2
First Rater
Codes

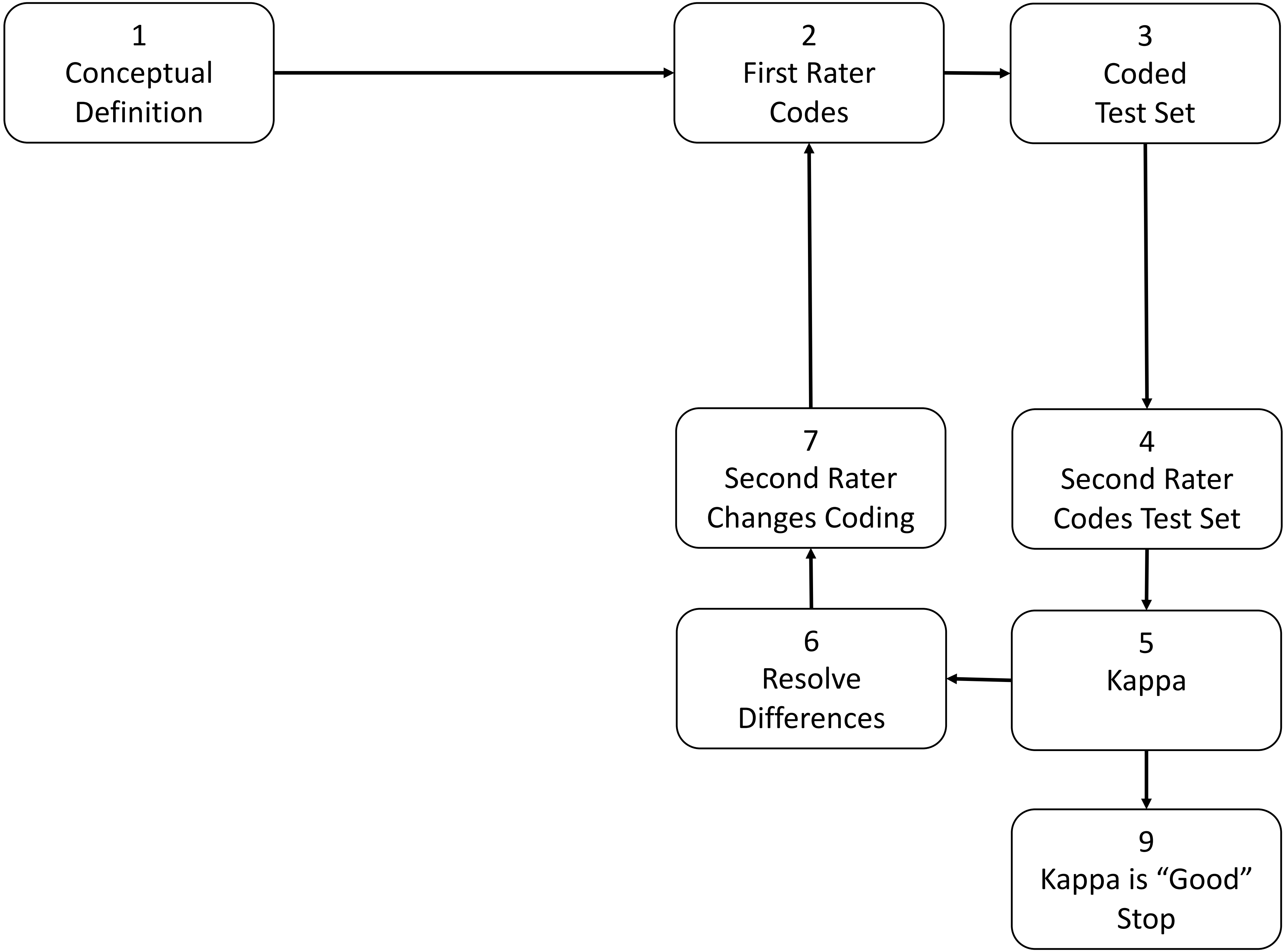
3
Coded
Test Set

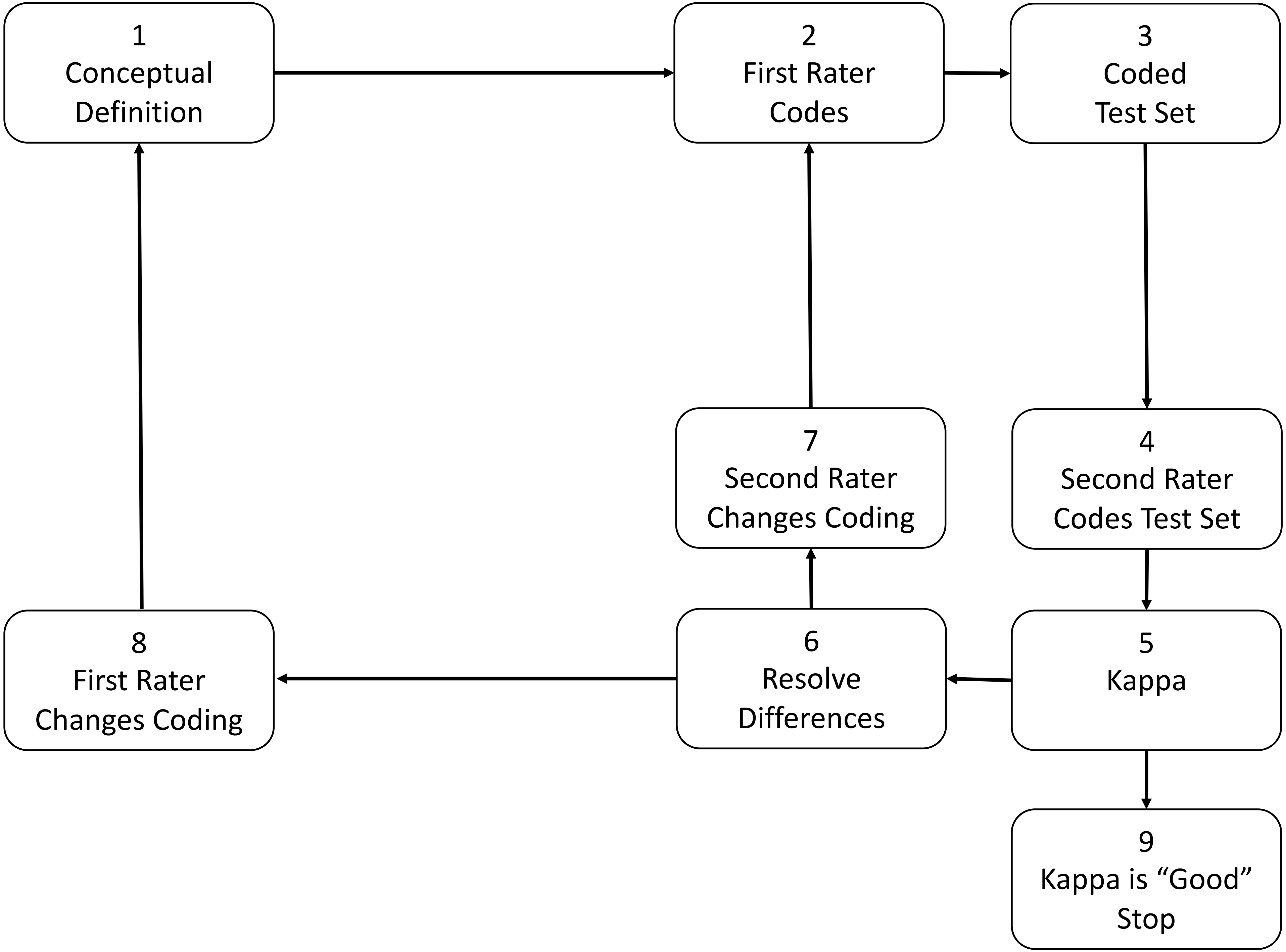
4
Second Rater
Codes Test Set

5
Kappa

6
Resolve
Differences

9
Kappa is “Good”
Stop


1
Conceptual
Definition
2
First Rater
Codes
3
Coded
Test Set
4
Second Rater
Codes Test Set
5
Kappa
6
Resolve
Differences
7
Second Rater
Changes Coding
9
Kappa is “Good”
Stop



1
Conceptual Definition

2
First Rater Codes

3
Coded Test Set

4
Second Rater Codes Test Set

5
Kappa

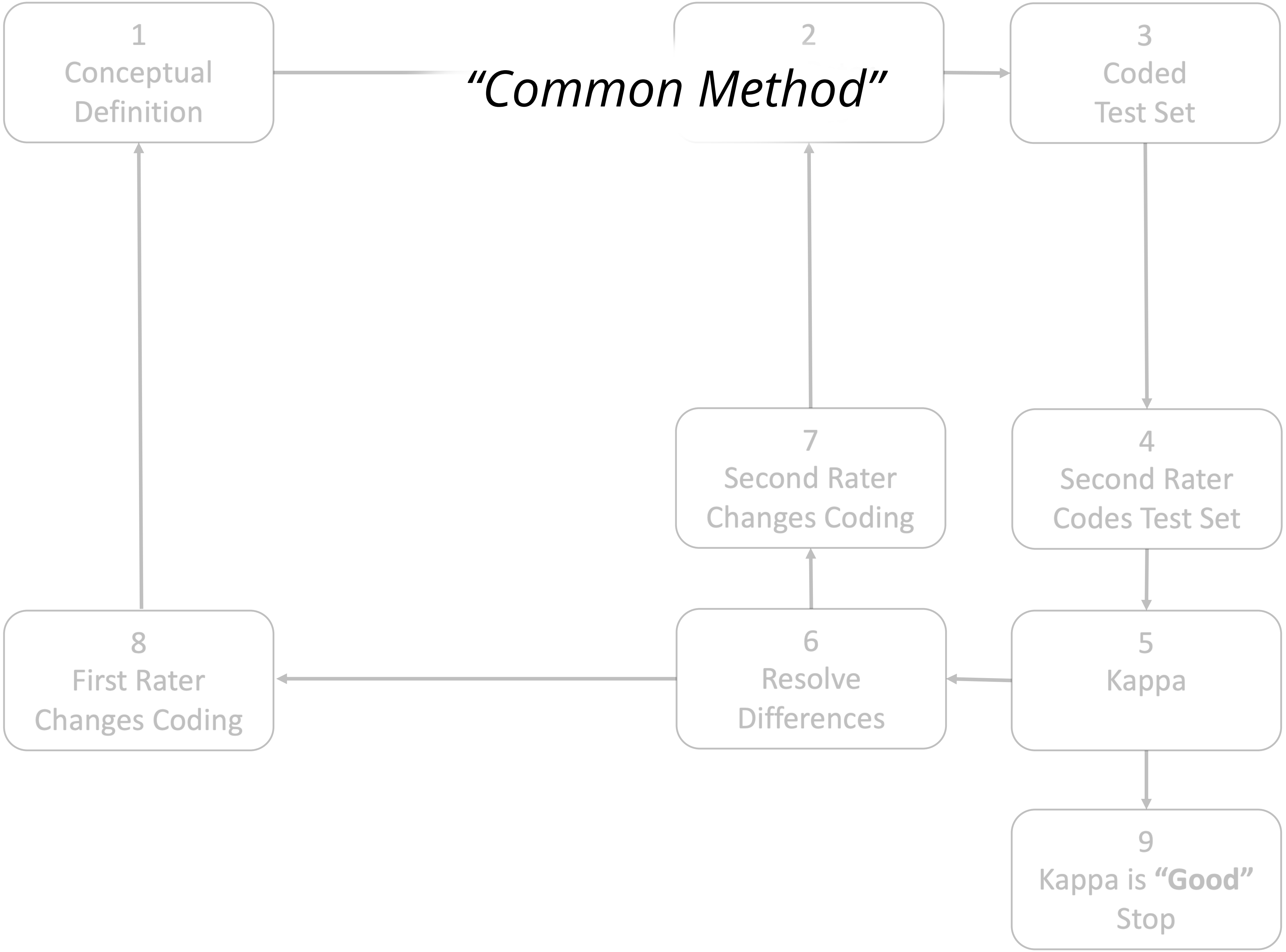
6
Resolve Differences

7
Second Rater Changes Coding

8
First Rater Changes Coding

9
Kappa is “Good” Stop

“Common Method”



1
Conceptual
Definition

2

3
Coded
Test Set

4
Second Rater
Codes Test Set

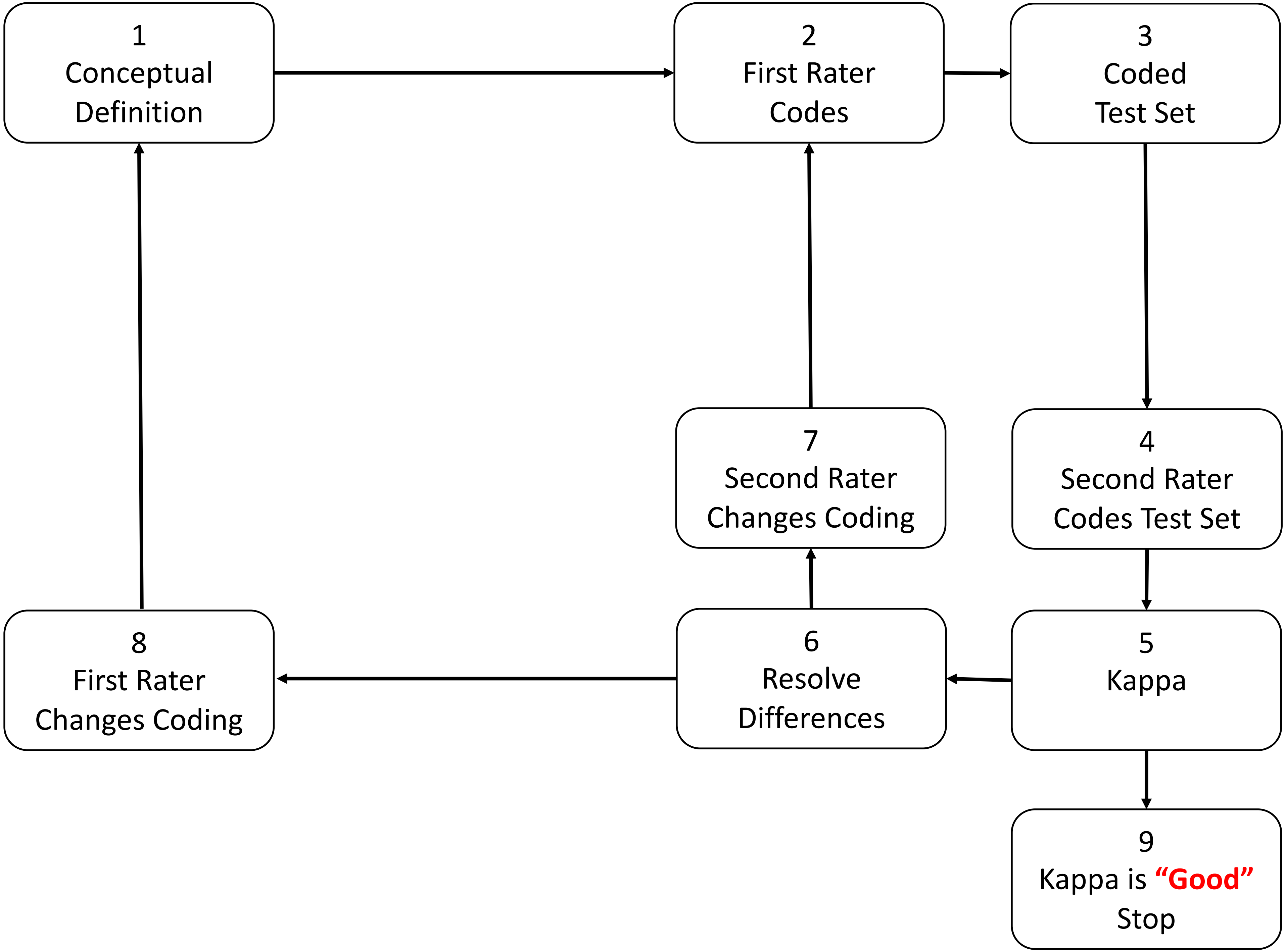
5
Kappa

6
Resolve
Differences

7
Second Rater
Changes Coding

8
First Rater
Changes Coding

9
Kappa is **“Good”**
Stop



1
Conceptual
Definition

2
First Rater
Codes

3
Coded
Test Set

4
Second Rater
Codes Test Set

5
Kappa

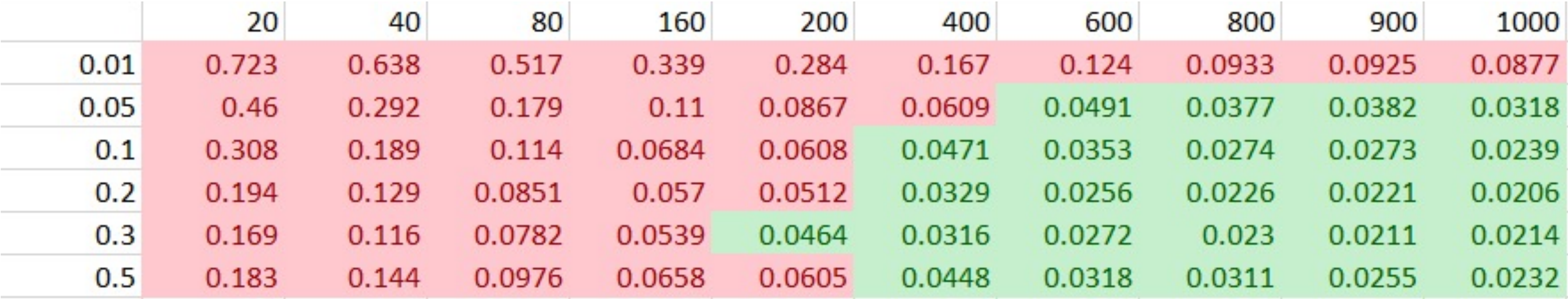
6
Resolve
Differences

7
Second Rater
Changes Coding

8
First Rater
Changes Coding

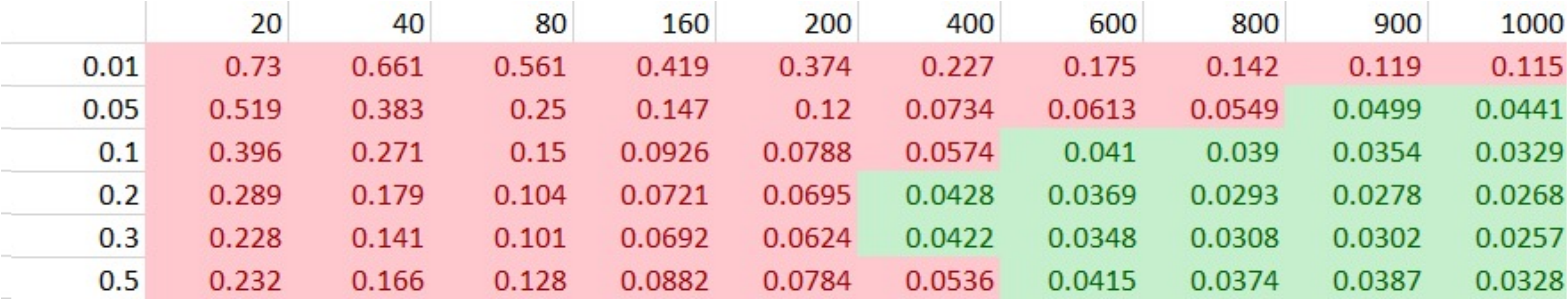
9
Kappa is **“Good”**
Stop

Percent positive agreement (>70%)



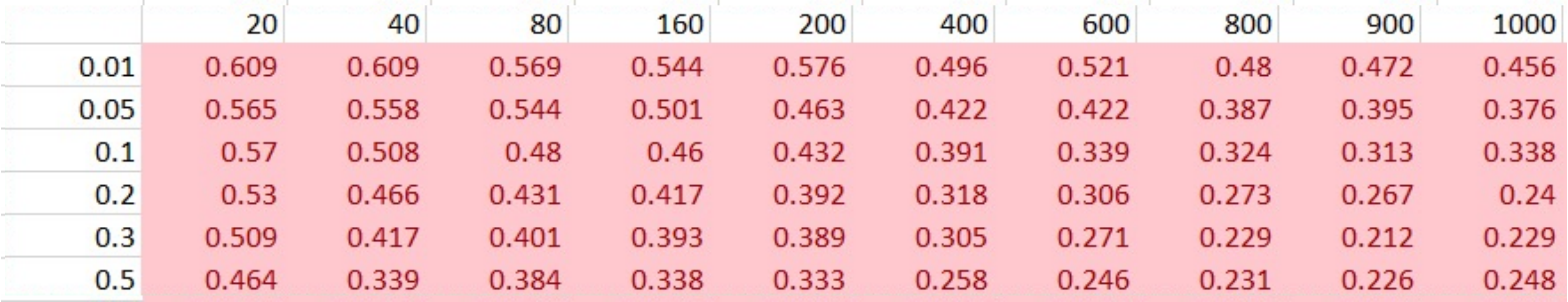
	20	40	80	160	200	400	600	800	900	1000
0.01	0.723	0.638	0.517	0.339	0.284	0.167	0.124	0.0933	0.0925	0.0877
0.05	0.46	0.292	0.179	0.11	0.0867	0.0609	0.0491	0.0377	0.0382	0.0318
0.1	0.308	0.189	0.114	0.0684	0.0608	0.0471	0.0353	0.0274	0.0273	0.0239
0.2	0.194	0.129	0.0851	0.057	0.0512	0.0329	0.0256	0.0226	0.0221	0.0206
0.3	0.169	0.116	0.0782	0.0539	0.0464	0.0316	0.0272	0.023	0.0211	0.0214
0.5	0.183	0.144	0.0976	0.0658	0.0605	0.0448	0.0318	0.0311	0.0255	0.0232

Recall (>0.65)



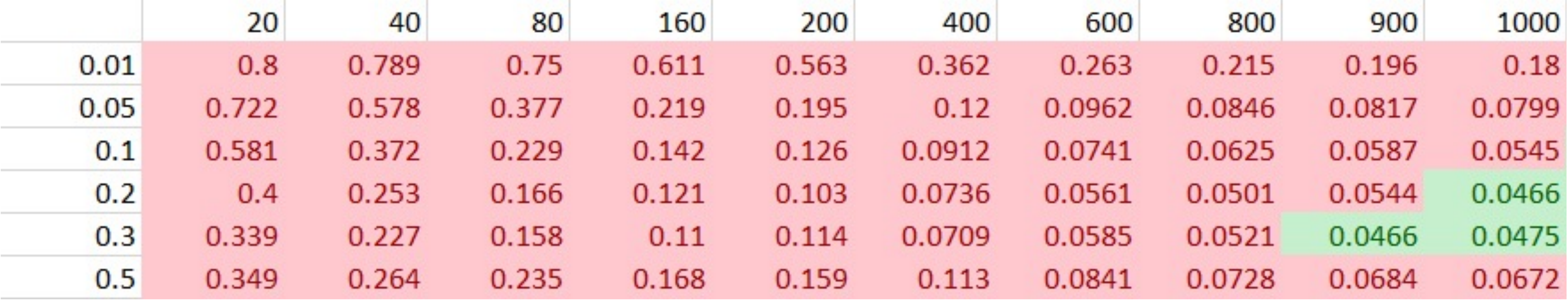
	20	40	80	160	200	400	600	800	900	1000
0.01	0.73	0.661	0.561	0.419	0.374	0.227	0.175	0.142	0.119	0.115
0.05	0.519	0.383	0.25	0.147	0.12	0.0734	0.0613	0.0549	0.0499	0.0441
0.1	0.396	0.271	0.15	0.0926	0.0788	0.0574	0.041	0.039	0.0354	0.0329
0.2	0.289	0.179	0.104	0.0721	0.0695	0.0428	0.0369	0.0293	0.0278	0.0268
0.3	0.228	0.141	0.101	0.0692	0.0624	0.0422	0.0348	0.0308	0.0302	0.0257
0.5	0.232	0.166	0.128	0.0882	0.0784	0.0536	0.0415	0.0374	0.0387	0.0328

Precision (>0.65)

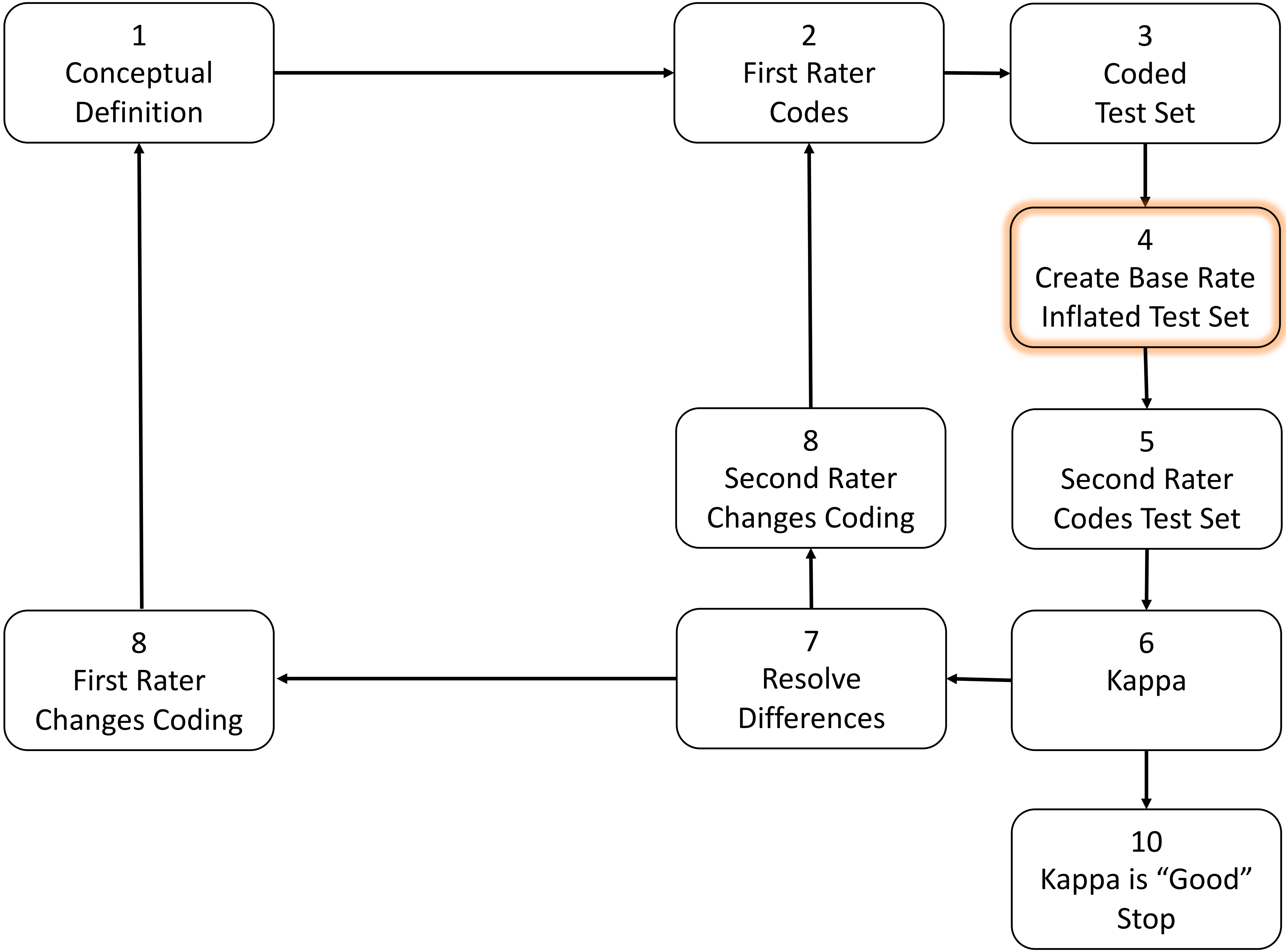


	20	40	80	160	200	400	600	800	900	1000
0.01	0.609	0.609	0.569	0.544	0.576	0.496	0.521	0.48	0.472	0.456
0.05	0.565	0.558	0.544	0.501	0.463	0.422	0.422	0.387	0.395	0.376
0.1	0.57	0.508	0.48	0.46	0.432	0.391	0.339	0.324	0.313	0.338
0.2	0.53	0.466	0.431	0.417	0.392	0.318	0.306	0.273	0.267	0.24
0.3	0.509	0.417	0.401	0.393	0.389	0.305	0.271	0.229	0.212	0.229
0.5	0.464	0.339	0.384	0.338	0.333	0.258	0.246	0.231	0.226	0.248

F statistic (>0.65)



	20	40	80	160	200	400	600	800	900	1000
0.01	0.8	0.789	0.75	0.611	0.563	0.362	0.263	0.215	0.196	0.18
0.05	0.722	0.578	0.377	0.219	0.195	0.12	0.0962	0.0846	0.0817	0.0799
0.1	0.581	0.372	0.229	0.142	0.126	0.0912	0.0741	0.0625	0.0587	0.0545
0.2	0.4	0.253	0.166	0.121	0.103	0.0736	0.0561	0.0501	0.0544	0.0466
0.3	0.339	0.227	0.158	0.11	0.114	0.0709	0.0585	0.0521	0.0466	0.0475
0.5	0.349	0.264	0.235	0.168	0.159	0.113	0.0841	0.0728	0.0684	0.0672



1
Conceptual
Definition

2
First Rater
Codes

3
Coded
Test Set

4
Create Base Rate
Inflated Test Set

5
Second Rater
Codes Test Set

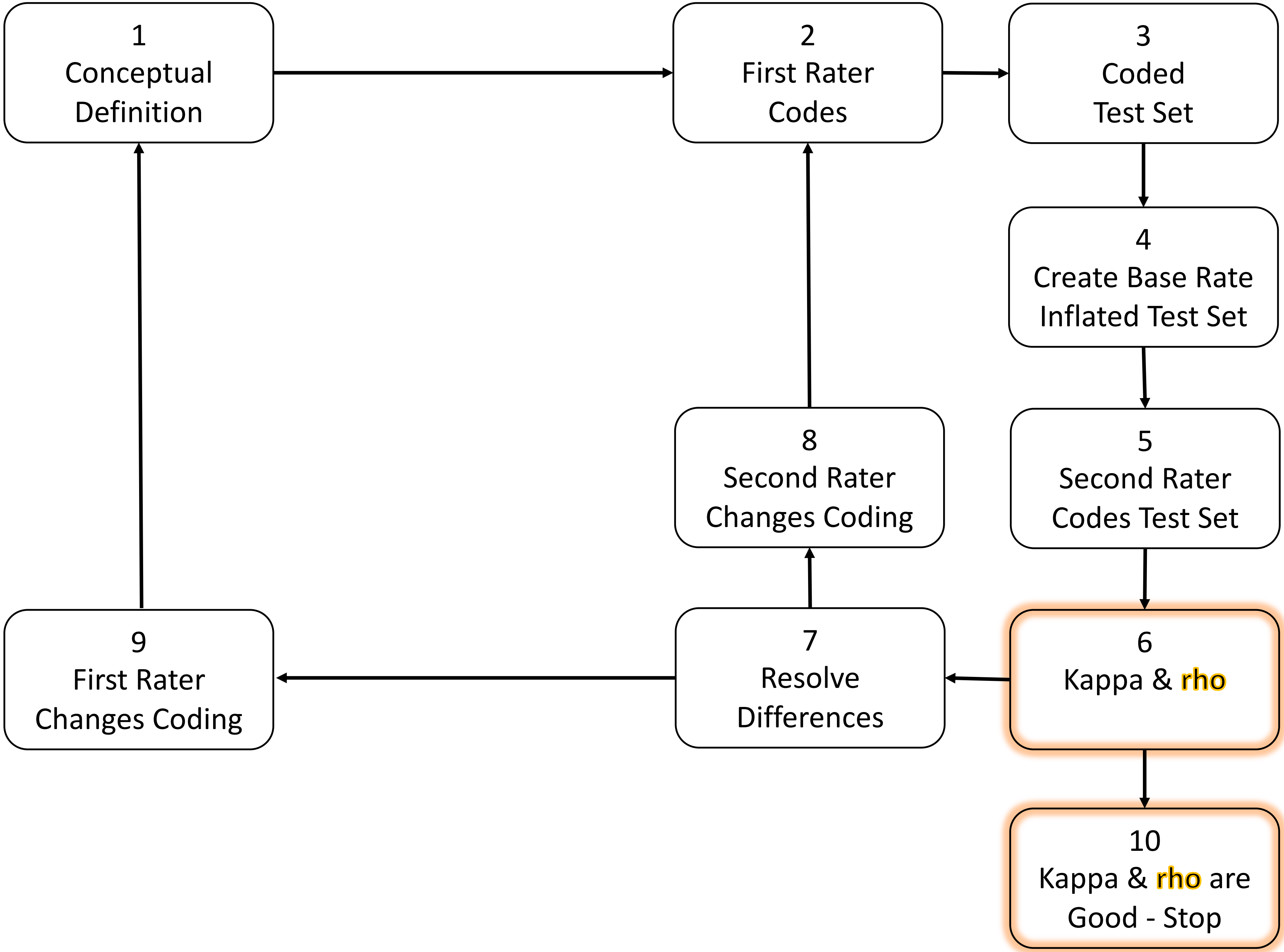
6
Kappa

7
Resolve
Differences

8
Second Rater
Changes Coding

8
First Rater
Changes Coding

10
Kappa is "Good"
Stop



1
Conceptual
Definition

2
First Rater
Codes

3
Coded
Test Set

4
Create Base Rate
Inflated Test Set

5
Second Rater
Codes Test Set

6
Kappa & rho

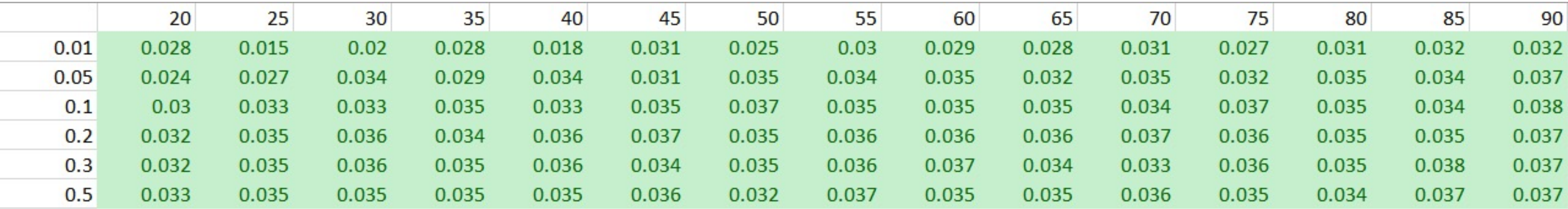
7
Resolve
Differences

8
Second Rater
Changes Coding

9
First Rater
Changes Coding

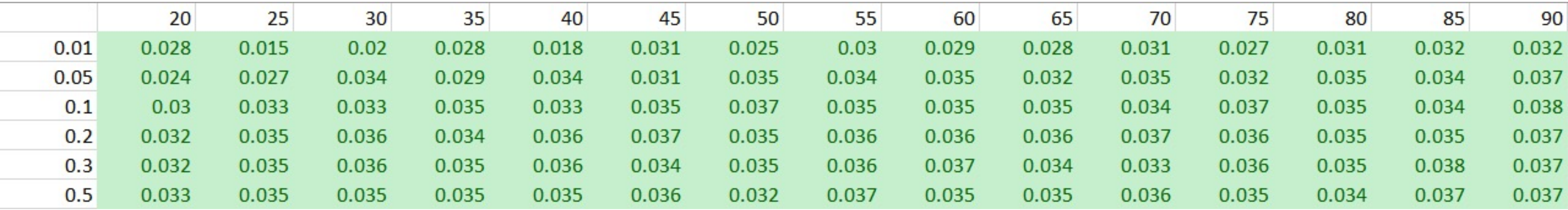
10
Kappa & rho are
Good - Stop

Type I error rate rho, using kappa (threshold = 0.65) base rate inflation



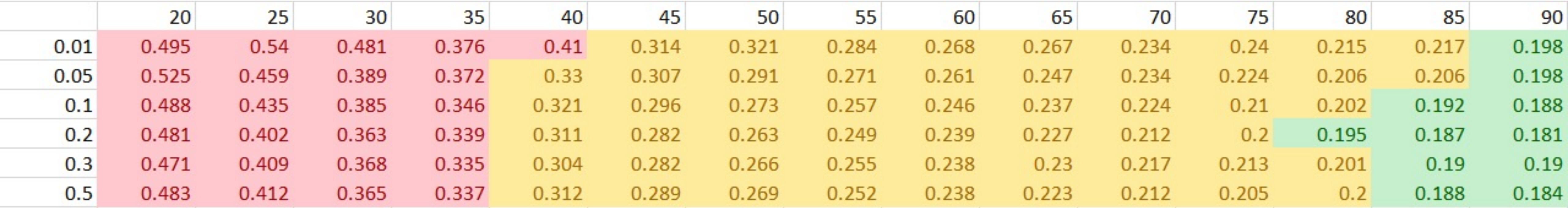
	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
0.01	0.028	0.015	0.02	0.028	0.018	0.031	0.025	0.03	0.029	0.028	0.031	0.027	0.031	0.032	0.032
0.05	0.024	0.027	0.034	0.029	0.034	0.031	0.035	0.034	0.035	0.032	0.035	0.032	0.035	0.034	0.037
0.1	0.03	0.033	0.033	0.035	0.033	0.035	0.037	0.035	0.035	0.035	0.034	0.037	0.035	0.034	0.038
0.2	0.032	0.035	0.036	0.034	0.036	0.037	0.035	0.036	0.036	0.036	0.037	0.036	0.035	0.035	0.037
0.3	0.032	0.035	0.036	0.035	0.036	0.034	0.035	0.036	0.037	0.034	0.033	0.036	0.035	0.038	0.037
0.5	0.033	0.035	0.035	0.035	0.035	0.036	0.032	0.037	0.035	0.035	0.036	0.035	0.034	0.037	0.037

Type I error rate rho, using kappa (threshold = 0.65) base rate inflation



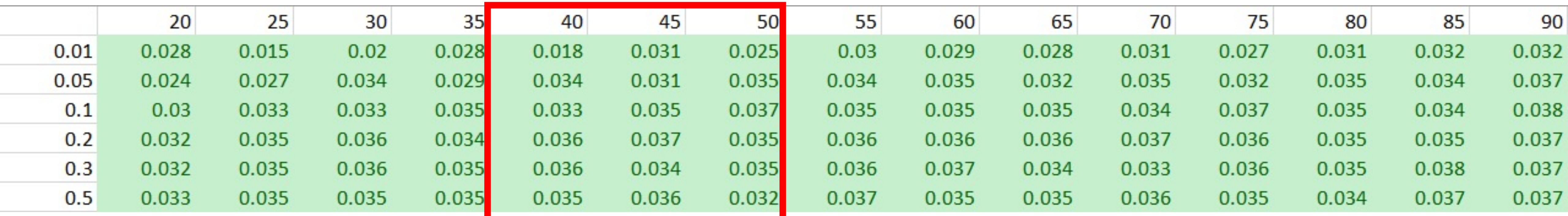
	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
0.01	0.028	0.015	0.02	0.028	0.018	0.031	0.025	0.03	0.029	0.028	0.031	0.027	0.031	0.032	0.032
0.05	0.024	0.027	0.034	0.029	0.034	0.031	0.035	0.034	0.035	0.032	0.035	0.032	0.035	0.034	0.037
0.1	0.03	0.033	0.033	0.035	0.033	0.035	0.037	0.035	0.035	0.035	0.034	0.037	0.035	0.034	0.038
0.2	0.032	0.035	0.036	0.034	0.036	0.037	0.035	0.036	0.036	0.036	0.037	0.036	0.035	0.035	0.037
0.3	0.032	0.035	0.036	0.035	0.036	0.034	0.035	0.036	0.037	0.034	0.033	0.036	0.035	0.038	0.037
0.5	0.033	0.035	0.035	0.035	0.035	0.036	0.032	0.037	0.035	0.035	0.036	0.035	0.034	0.037	0.037

Type II error rate rho, using kappa (threshold = 0.65) base rate inflation



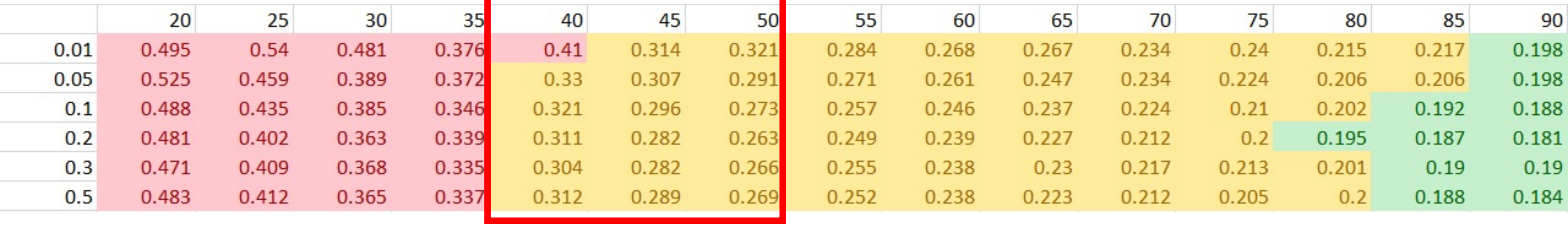
	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
0.01	0.495	0.54	0.481	0.376	0.41	0.314	0.321	0.284	0.268	0.267	0.234	0.24	0.215	0.217	0.198
0.05	0.525	0.459	0.389	0.372	0.33	0.307	0.291	0.271	0.261	0.247	0.234	0.224	0.206	0.206	0.198
0.1	0.488	0.435	0.385	0.346	0.321	0.296	0.273	0.257	0.246	0.237	0.224	0.21	0.202	0.192	0.188
0.2	0.481	0.402	0.363	0.339	0.311	0.282	0.263	0.249	0.239	0.227	0.212	0.2	0.195	0.187	0.181
0.3	0.471	0.409	0.368	0.335	0.304	0.282	0.266	0.255	0.238	0.23	0.217	0.213	0.201	0.19	0.19
0.5	0.483	0.412	0.365	0.337	0.312	0.289	0.269	0.252	0.238	0.223	0.212	0.205	0.2	0.188	0.184

Type I error rate rho, using kappa (threshold = 0.65) base rate inflation



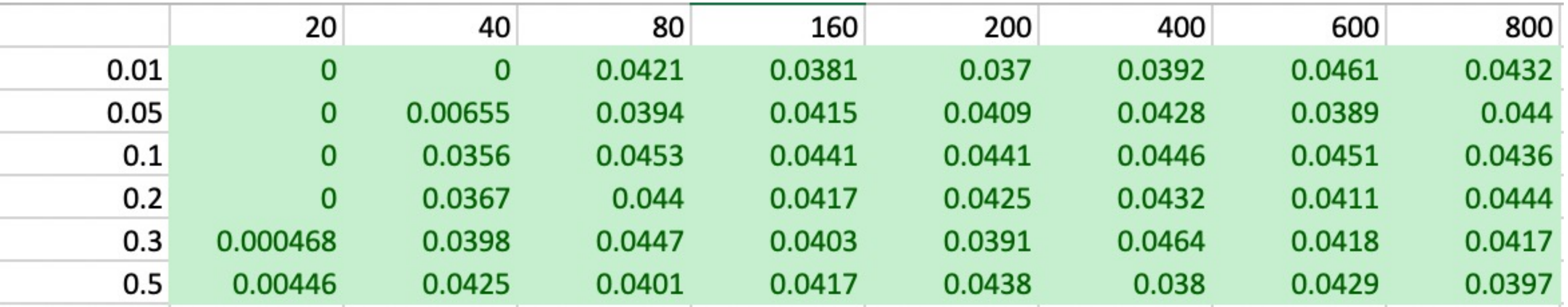
	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
0.01	0.028	0.015	0.02	0.028	0.018	0.031	0.025	0.03	0.029	0.028	0.031	0.027	0.031	0.032	0.032
0.05	0.024	0.027	0.034	0.029	0.034	0.031	0.035	0.034	0.035	0.032	0.035	0.032	0.035	0.034	0.037
0.1	0.03	0.033	0.033	0.035	0.033	0.035	0.037	0.035	0.035	0.035	0.034	0.037	0.035	0.034	0.038
0.2	0.032	0.035	0.036	0.034	0.036	0.037	0.035	0.036	0.036	0.036	0.037	0.036	0.035	0.035	0.037
0.3	0.032	0.035	0.036	0.035	0.036	0.034	0.035	0.036	0.037	0.034	0.033	0.036	0.035	0.038	0.037
0.5	0.033	0.035	0.035	0.035	0.035	0.036	0.032	0.037	0.035	0.035	0.036	0.035	0.034	0.037	0.037

Type II error rate rho, using kappa (threshold = 0.65) base rate inflation



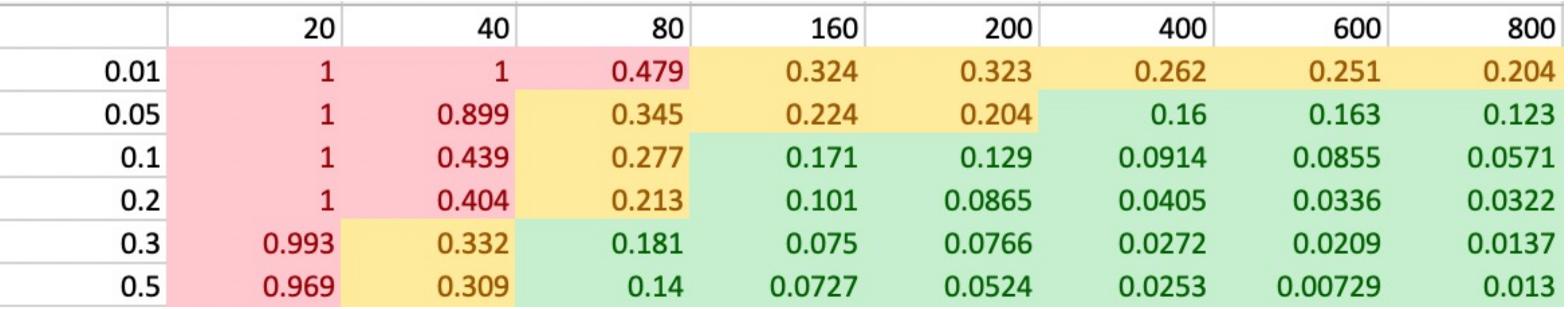
	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
0.01	0.495	0.54	0.481	0.376	0.41	0.314	0.321	0.284	0.268	0.267	0.234	0.24	0.215	0.217	0.198
0.05	0.525	0.459	0.389	0.372	0.33	0.307	0.291	0.271	0.261	0.247	0.234	0.224	0.206	0.206	0.198
0.1	0.488	0.435	0.385	0.346	0.321	0.296	0.273	0.257	0.246	0.237	0.224	0.21	0.202	0.192	0.188
0.2	0.481	0.402	0.363	0.339	0.311	0.282	0.263	0.249	0.239	0.227	0.212	0.2	0.195	0.187	0.181
0.3	0.471	0.409	0.368	0.335	0.304	0.282	0.266	0.255	0.238	0.23	0.217	0.213	0.201	0.19	0.19
0.5	0.483	0.412	0.365	0.337	0.312	0.289	0.269	0.252	0.238	0.223	0.212	0.205	0.2	0.188	0.184

Type I error rate rho, using kappa (threshold = 0.9) base rate inflation



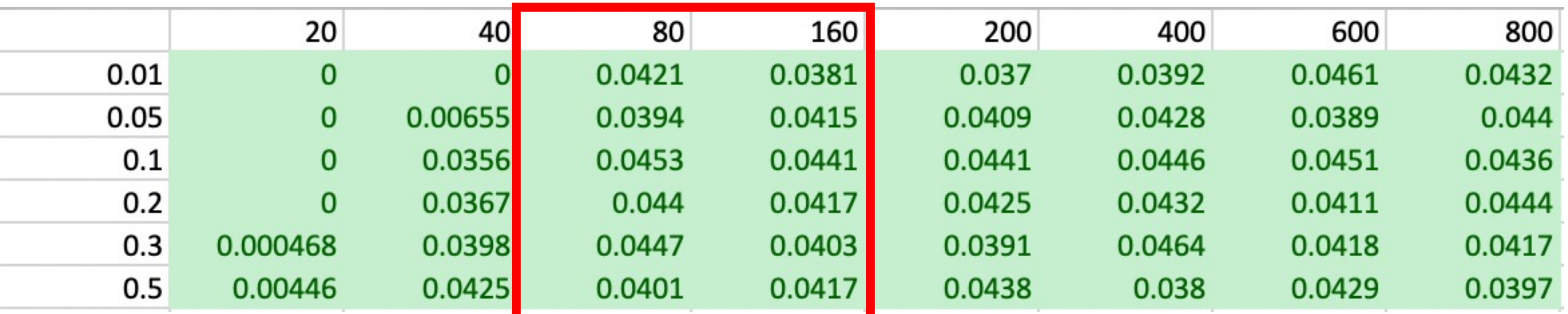
	20	40	80	160	200	400	600	800
0.01	0	0	0.0421	0.0381	0.037	0.0392	0.0461	0.0432
0.05	0	0.00655	0.0394	0.0415	0.0409	0.0428	0.0389	0.044
0.1	0	0.0356	0.0453	0.0441	0.0441	0.0446	0.0451	0.0436
0.2	0	0.0367	0.044	0.0417	0.0425	0.0432	0.0411	0.0444
0.3	0.000468	0.0398	0.0447	0.0403	0.0391	0.0464	0.0418	0.0417
0.5	0.00446	0.0425	0.0401	0.0417	0.0438	0.038	0.0429	0.0397

Type II error rate rho, using kappa (threshold = 0.9) base rate inflation



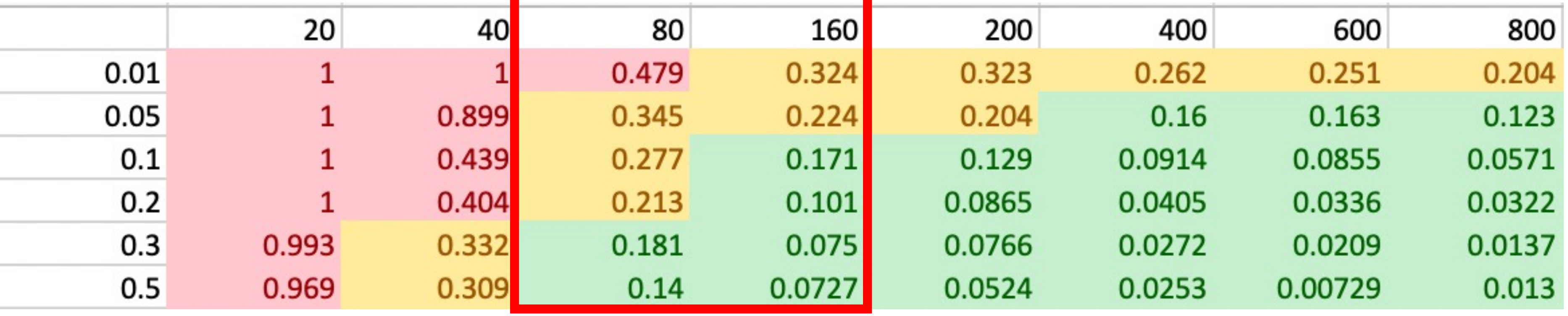
	20	40	80	160	200	400	600	800
0.01	1	1	0.479	0.324	0.323	0.262	0.251	0.204
0.05	1	0.899	0.345	0.224	0.204	0.16	0.163	0.123
0.1	1	0.439	0.277	0.171	0.129	0.0914	0.0855	0.0571
0.2	1	0.404	0.213	0.101	0.0865	0.0405	0.0336	0.0322
0.3	0.993	0.332	0.181	0.075	0.0766	0.0272	0.0209	0.0137
0.5	0.969	0.309	0.14	0.0727	0.0524	0.0253	0.00729	0.013

Type I error rate rho, using kappa (threshold = 0.9) base rate inflation

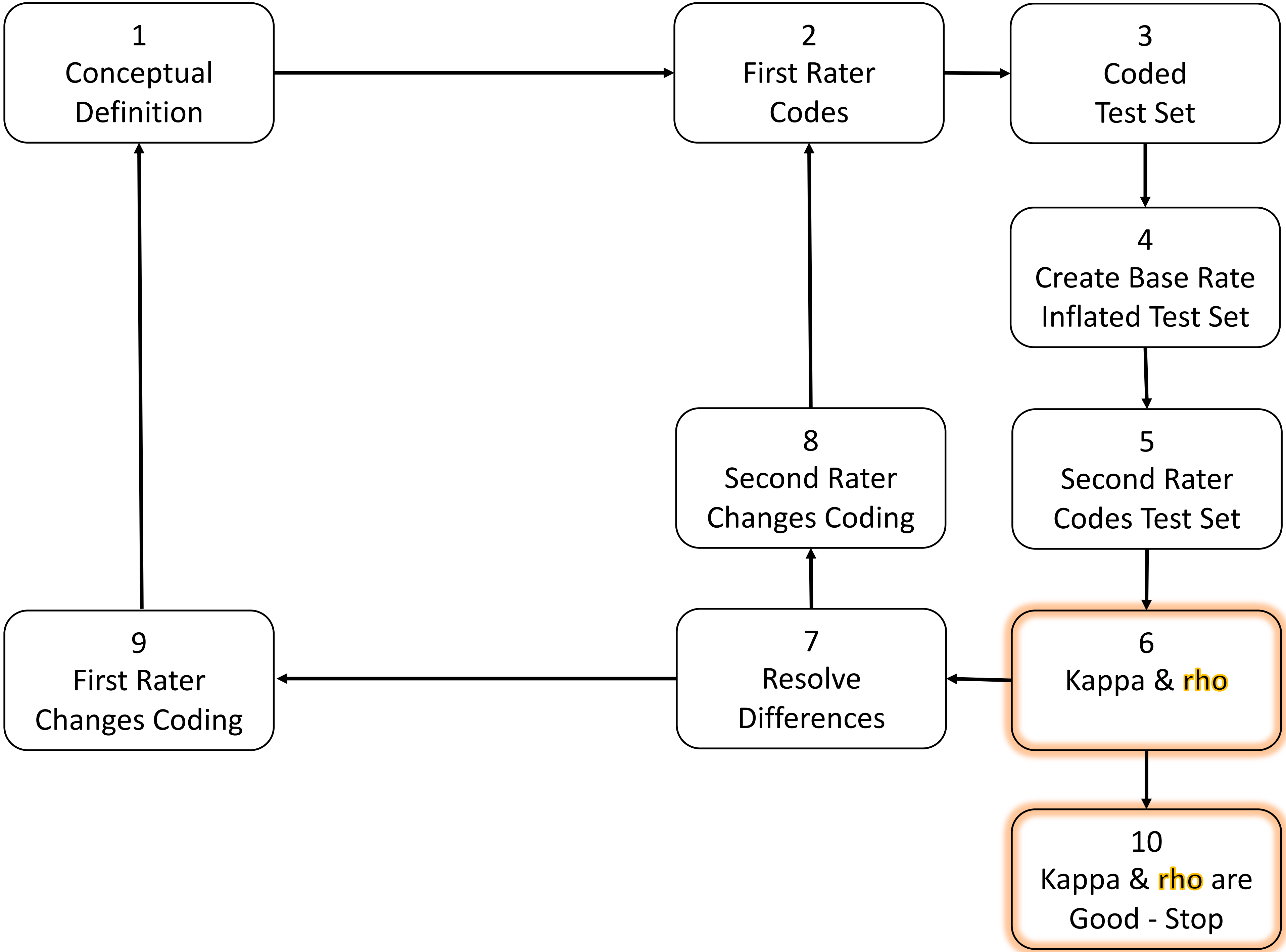


	20	40	80	160	200	400	600	800
0.01	0	0	0.0421	0.0381	0.037	0.0392	0.0461	0.0432
0.05	0	0.00655	0.0394	0.0415	0.0409	0.0428	0.0389	0.044
0.1	0	0.0356	0.0453	0.0441	0.0441	0.0446	0.0451	0.0436
0.2	0	0.0367	0.044	0.0417	0.0425	0.0432	0.0411	0.0444
0.3	0.000468	0.0398	0.0447	0.0403	0.0391	0.0464	0.0418	0.0417
0.5	0.00446	0.0425	0.0401	0.0417	0.0438	0.038	0.0429	0.0397

Type II error rate rho, using kappa (threshold = 0.9) base rate inflation



	20	40	80	160	200	400	600	800
0.01	1	1	0.479	0.324	0.323	0.262	0.251	0.204
0.05	1	0.899	0.345	0.224	0.204	0.16	0.163	0.123
0.1	1	0.439	0.277	0.171	0.129	0.0914	0.0855	0.0571
0.2	1	0.404	0.213	0.101	0.0865	0.0405	0.0336	0.0322
0.3	0.993	0.332	0.181	0.075	0.0766	0.0272	0.0209	0.0137
0.5	0.969	0.309	0.14	0.0727	0.0524	0.0253	0.00729	0.013



1
Conceptual Definition

2
First Rater Codes

3
Coded Test Set

4
Create Base Rate Inflated Test Set

5
Second Rater Codes Test Set

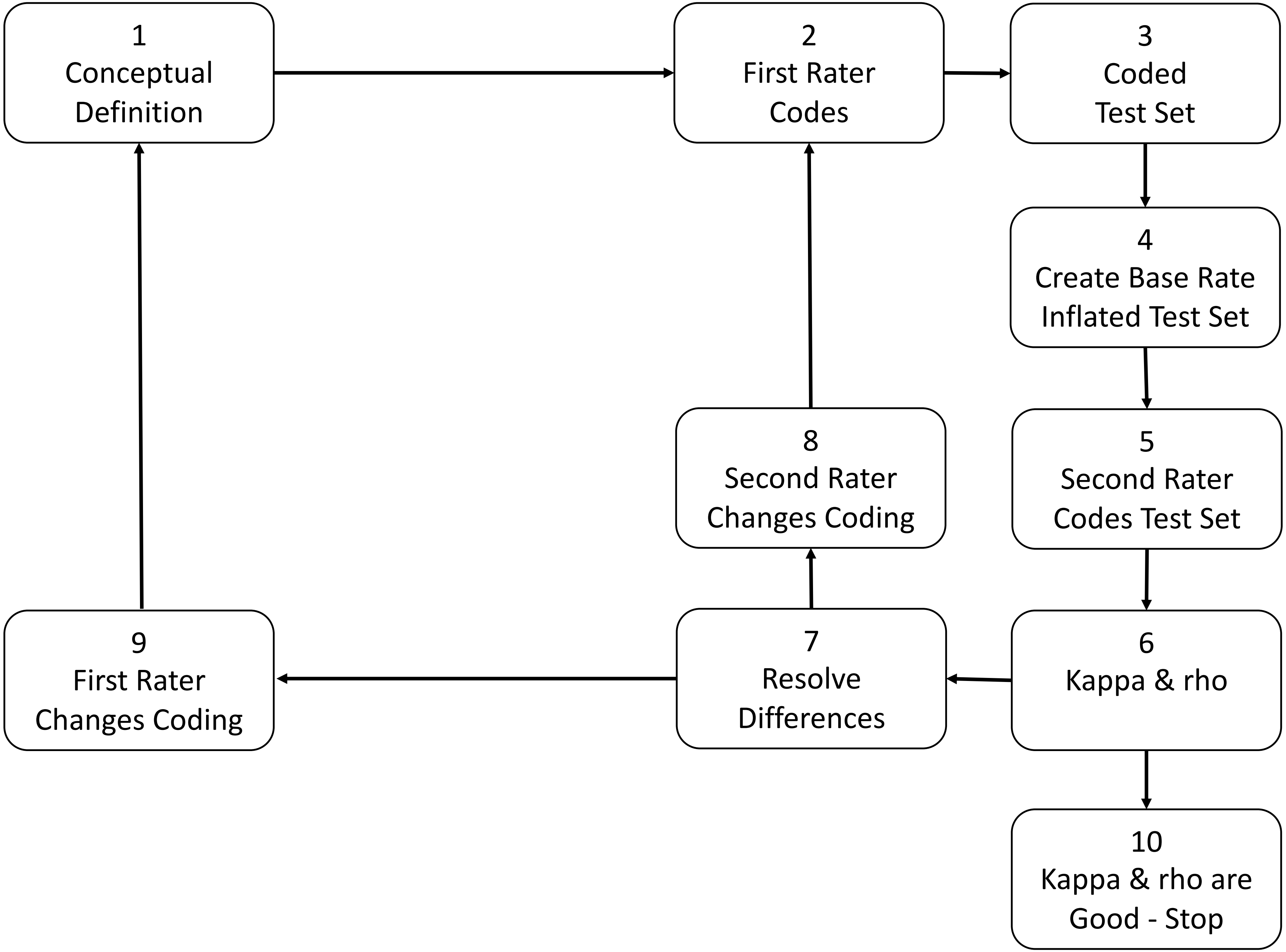
6
Kappa & rho

7
Resolve Differences

8
Second Rater Changes Coding

9
First Rater Changes Coding

10
Kappa & rho are Good - Stop



1
Conceptual Definition

2
First Rater Codes

3
Coded Test Set

4
Create Base Rate Inflated Test Set

5
Second Rater Codes Test Set

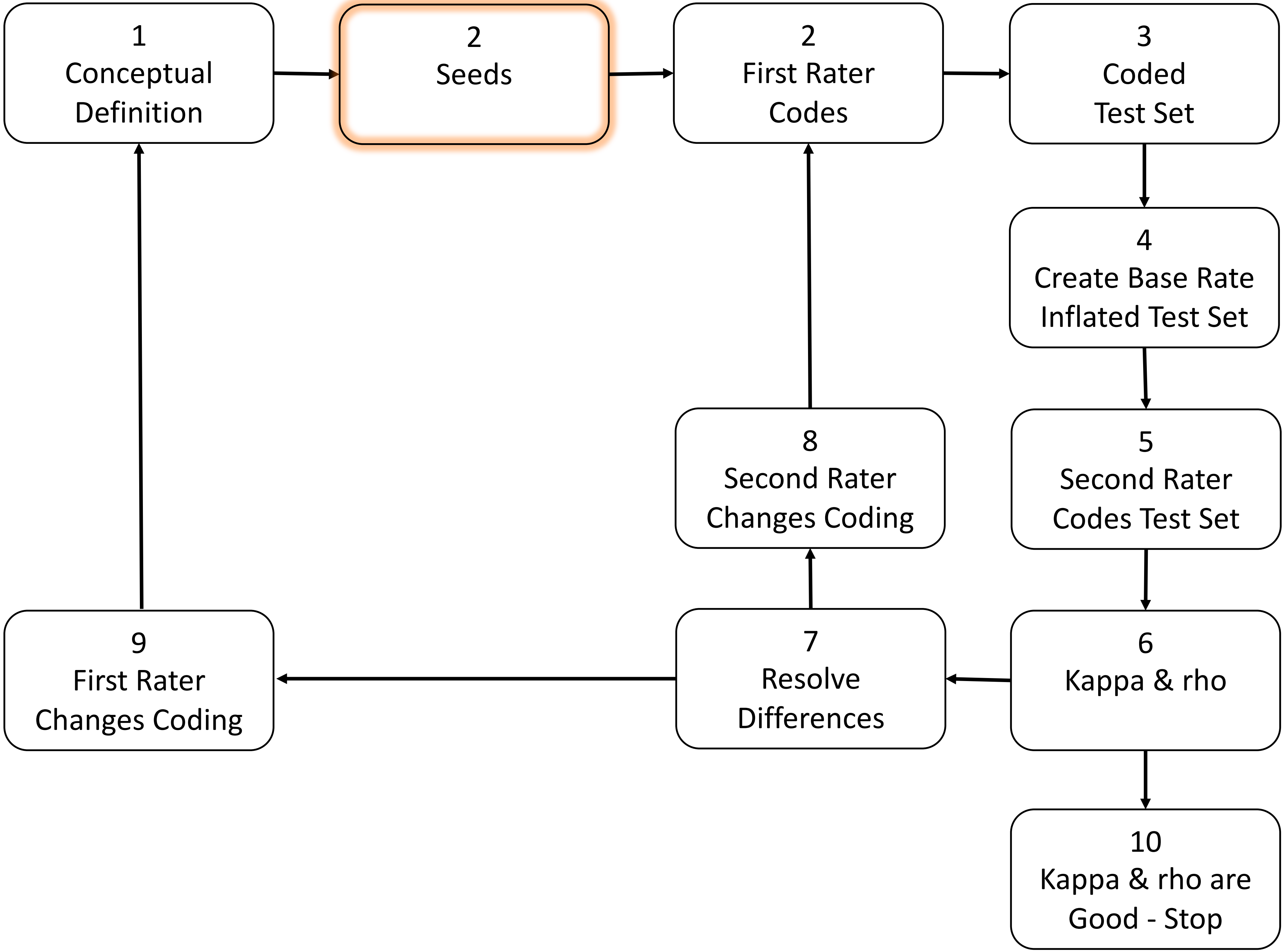
6
Kappa & rho

7
Resolve Differences

8
Second Rater Changes Coding

9
First Rater Changes Coding

10
Kappa & rho are Good - Stop



1 Conceptual Definition

2 Seeds

2 First Rater Codes

3 Coded Test Set

4 Create Base Rate Inflated Test Set

5 Second Rater Codes Test Set

6 Kappa & rho

7 Resolve Differences

8 Second Rater Changes Coding

9 First Rater Changes Coding

10 Kappa & rho are Good - Stop

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Imagine that you have a special instrument that allows you to see what makes up odor.
The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.
Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Segmentation: how you divide your data into meaningful parts

Imagine that you have a special instrument that allows you to see what makes up odor.
The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.
Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Segmentation: how you divide your data into meaningful parts

How have you segmented your data?

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Modeling:

Imagine that you have a special instrument that allows you to see what makes up odor.
The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.
Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Modeling:

Draw
Represent
Circle

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Modeling:

Draw
Represent
Circle
Model

Imagine that you have a special instrument that allows you to see what makes up odor.

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.

Create a model of what you would see if you could focus on one tiny spot in the area between the jar and your nose.

What is this about?

Modeling:

Draw
Represent
Circle
Model

Imagine – create a mental representation? hypothesize? thought experiment?

THIS COMIC MADE POSSIBLE THANKS TO ADAM LINGELBACH

MRLOVENSTEIN.COM

Modeling:

The large circle in the drawing represents a spot that is magnified many times, so you can see it up close.
Let’s sit in a circle.

Circle

5-minute break

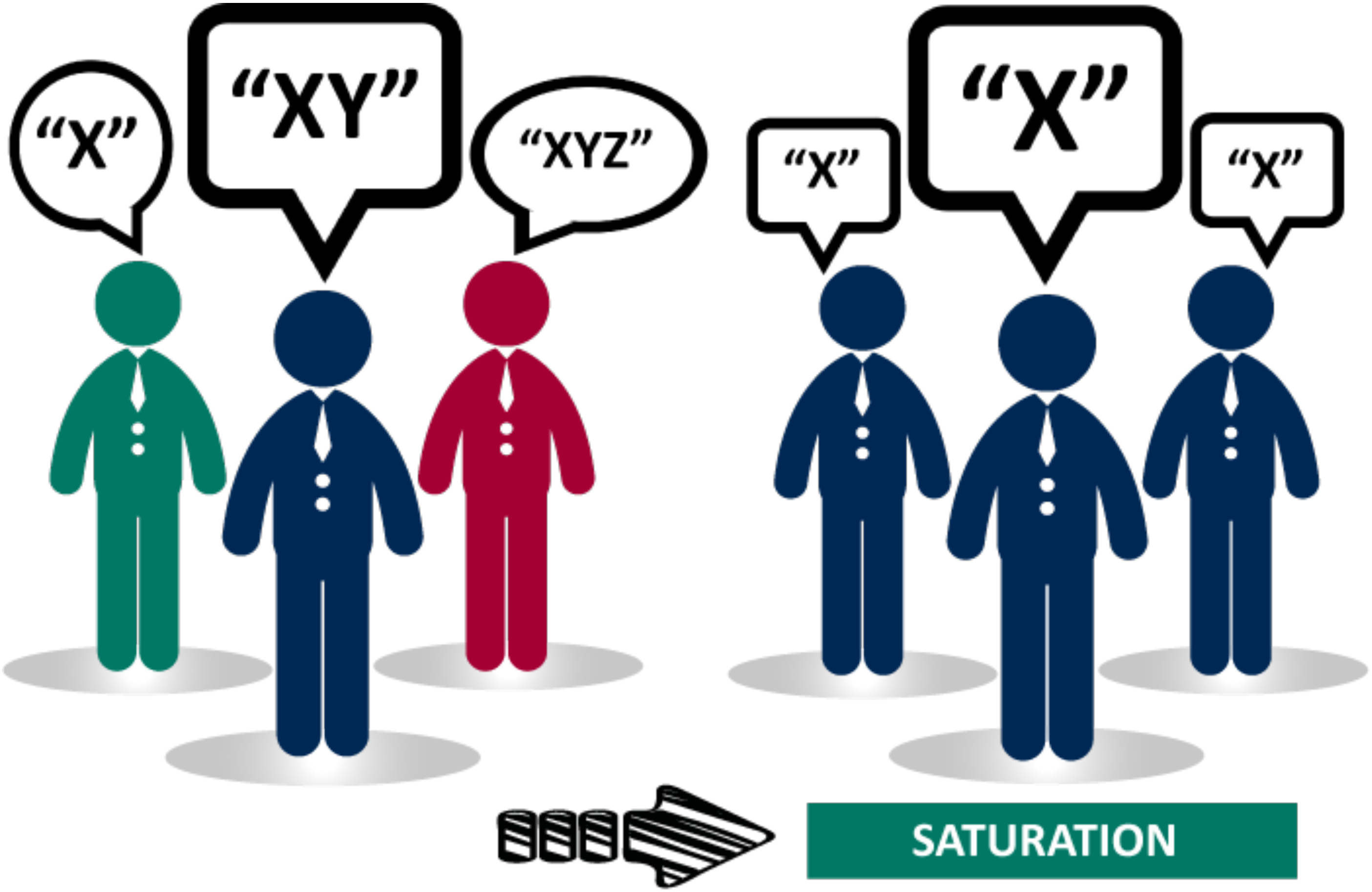
Hydration, snack, chatting, etc.

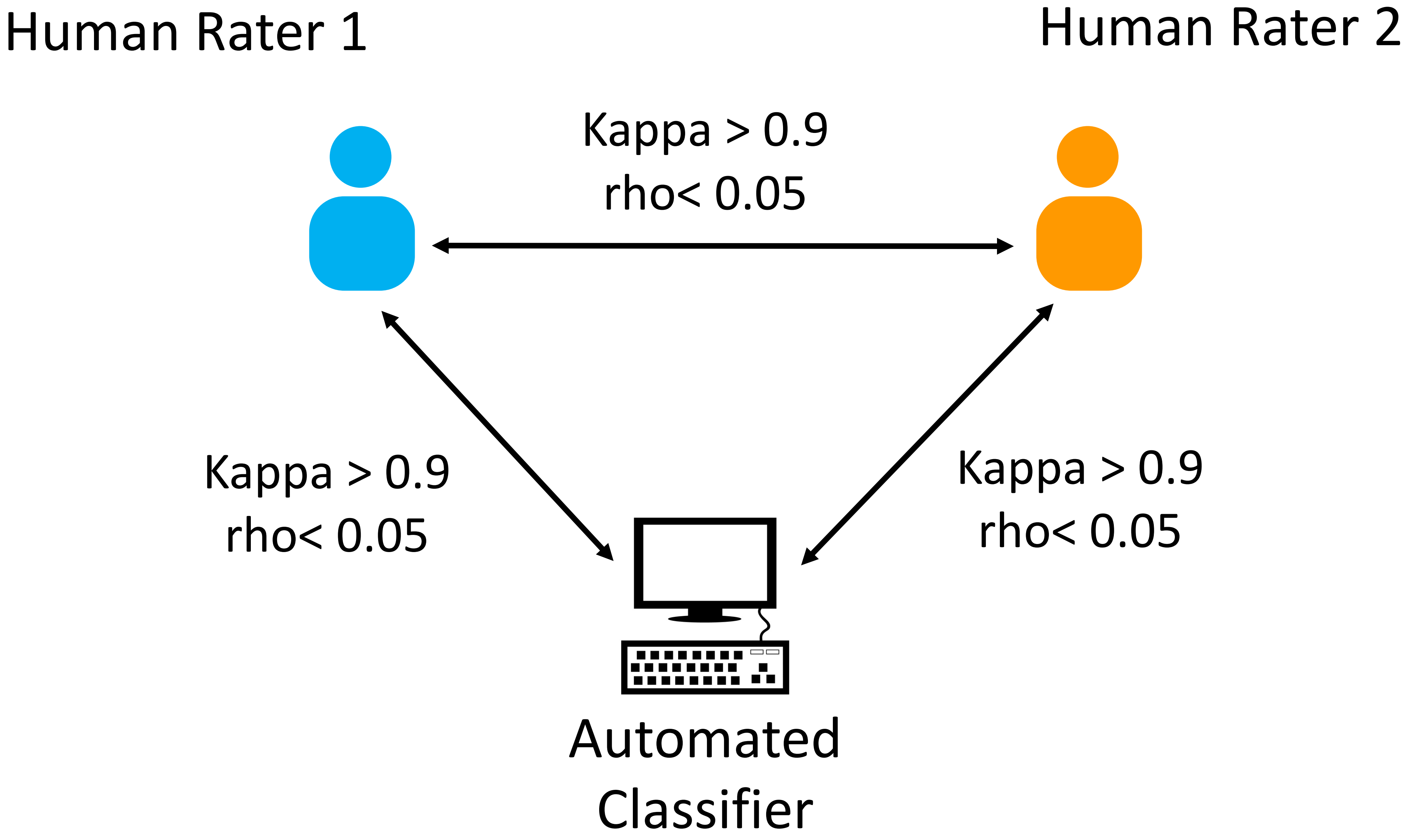
nCoder

https://app.n-coder.org/

https://go.wisc.edu/0nxl70

When have I coded enough?





Human Rater 1

Human Rater 2

Kappa > 0.9
rho< 0.05

Kappa > 0.9
rho< 0.05

Kappa > 0.9
rho< 0.05

Automated
Classifier

ICQE21

Workshop 2B: Introduction to nCoder

An Applied Example

Brett Puetz
University of Wisconsin – Madison

Introduction

Social Security Disability Insurance (SSDI)

- Exploration of individual’s discussions in applying for SSDI benefits
- Data consists of posts scraped from seven online forums
    - Spans years from 2004 to 2019
- How do conversations differ between those initially applying and those who have been denied and are appealing?
    - Specific interest in the central theme of medical evidence relative to different types of medical conditions

Data & Analysis

- Large amount of unstructured data (~150,000 posts)
- Approach to getting started with automated coding:
  - Unsupervised machine learning techniques
    - Topic modeling using LDA
    - Efficiently analyze large data sets for latent topics and associated words
  - Read the data set
    - Seems simple but useful for idiosyncratic nature of specific data
    - Helpful for grasping emic nature of discourse (ex. acronym usage)
  - Use related external resources as source material
    - Lots of material from the Social Security Administration such as Blue Book

Initial Results

- Achieved $\kappa > 0.90$ and $\rho < 0.05$ for all codes between human rater and nCoder
  - Needed between 100 and 700 lines of data per code
  - Full three-way validation for Denial/Appeals code

Code	Training Lines	Testing Lines	IRR – $\kappa(\rho)$
Denial/Appeals	460	100	0.97 (0.00)
Initial Application	300	100	0.97 (0.01)
Medical Evidence	700	100	0.91 (0.05)
Mental Health	90	100	0.97 (0.01)
Neurological Conditions	100	100	0.97 (0.00)
Pain	110	100	0.97 (0.00)

Code Example

<table>
<tr><th>Code</th><th>Definition</th><th>Examples</th><th>Classifiers</th><th>IRR</th></tr>
<tr><td>Denial and/or Appeals Process</td><td>References being denied at any stage of the Social Security Disability Insurance application process, however not the initial application itself; this may refer to the initial denial, the appeals process, or references to prior experience with being denied of the appeals process. For reference, the appeals process is a complex, multi-stage process administered by a reconsideration process; hearings by an administrative law judge, or ALJ; and the Social Security appeals council.</td><td><i>Likewise, without having the entire case record, including oral testimony, to review, it would be pure speculation to try to predict the ALJ's decision on a case as complex as has been presented.</i>
<i>I will be looking for someone else after the reconsideration phase because I know I will not be approved.</i></td><td>\balj
\bdeny
\bdeni\w+
\bappeal
\breconsideration
\bjudge
\bcouncil
\btestif
\bprocess.*?\bdecision
\bexplor.*?\boption
\badjudicator
\bhearing(?!(\w|\s)*?\b((voice)|(from)|(loss)))</td><td>0.97
(0.00)</td></tr>
</table>

Tips When Using nCoder

- Generally easier to code with smaller segmentation sizes (data permitting)
- Usually easier to split codes than combine them
- Worthwhile to get basic understanding of regular expression bestiary
  - Character classes - \w \d \s \b and Special characters - ^ $. * + ?
  - R – https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html
- Keep the regex patterns simple
  - Sometimes more powerful
    - \bteach.*\b matches teach, teaching, teacher, and teachable etc., but not reteach…
  - Easier to debug
- Easier to look for the presence rather than absence of patterns

nCoder

https://app.n-coder.org/

Amanda Siebert-Evenstone, Ph.D.
alevenstone@wisc.edu

Jaeyoon Choi
jaeyoon.choi@wisc.edu

Brett Puetz
bpuetz@wisc.edu