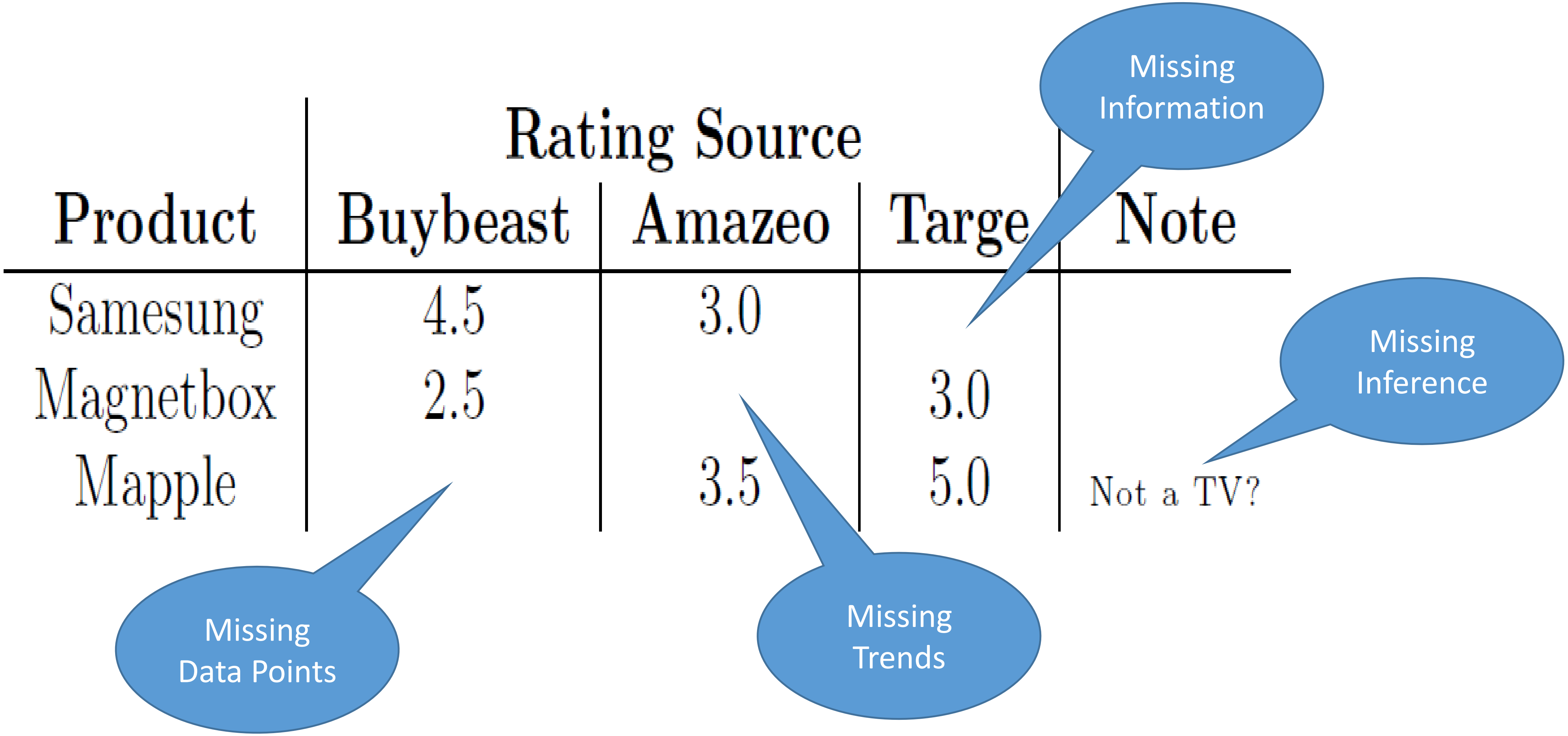
Communicating Data Quality in On-Demand Curation

Poonam Kumari*, Oliver Kennedy*, Said Achmiz

*University at Buffalo

Uncertain Data



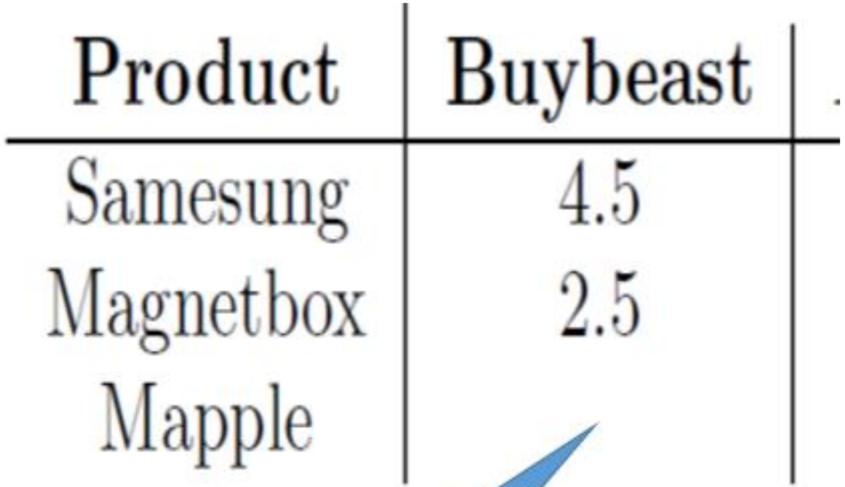
	Rating Source			
Product	Buybeast	Amazeo	Targe	Note
Samesung	4.5	3.0		
Magnetbox	2.5		3.0	
Mapple		3.5	5.0	Not a TV?

Missing Information

Missing Inference

Missing Data Points

Missing Trends



Product	Buybeast
Samesung	4.5
Magnetbox	2.5
Mapple	

Missing Data Points

What?

- ✓ Absence of Data in a column.
- ✓ Inability to get sufficient data points for analysis.

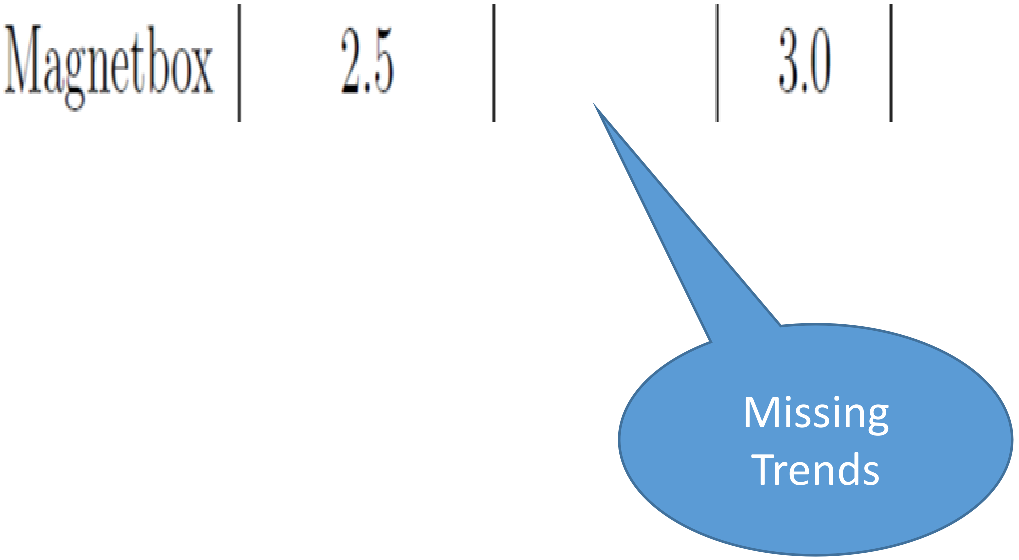
GUESS

How to Solve this?

- ✓ Guess based on available data.
- ✓ Fill with a default value.
- ✓ Ignore the absent data - Delete

What?

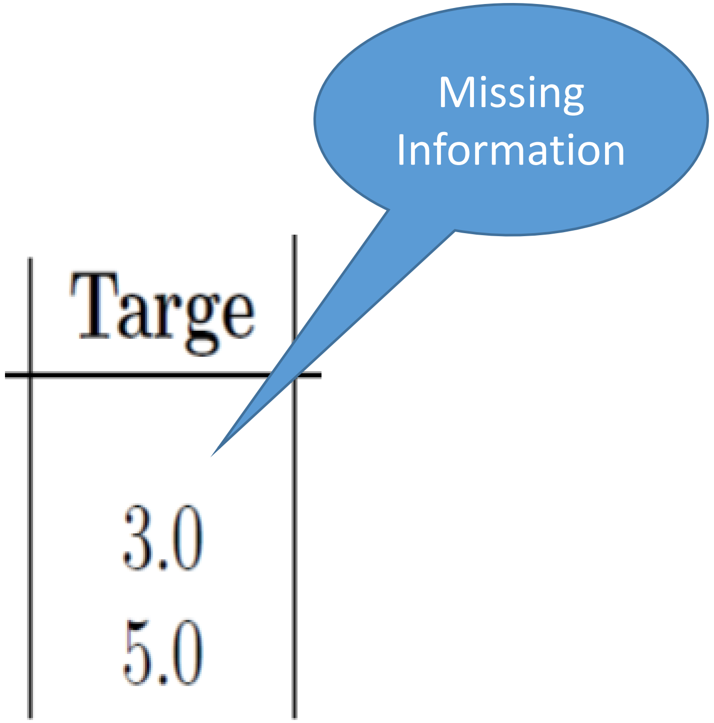
- ✓ Absence of Data in a column.
- ✓ Inability to determine the increase or decrease in values.



Magnetbox	2.5		3.0

How to Solve this?

- ✓ Guess based on available data.
- ✓ Fill with a statistically derived value.



Missing Information

Targe
3.0
5.0

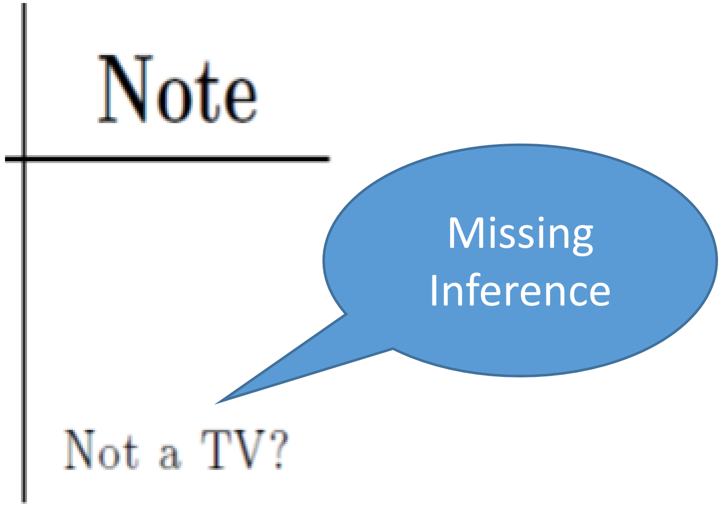
What?

- ✓ Absence of Data in a column.
- ✓ Inability to derive the range of values for a column.
- ✓ Inability to determine the if the data is continuous or discrete.

GUESS

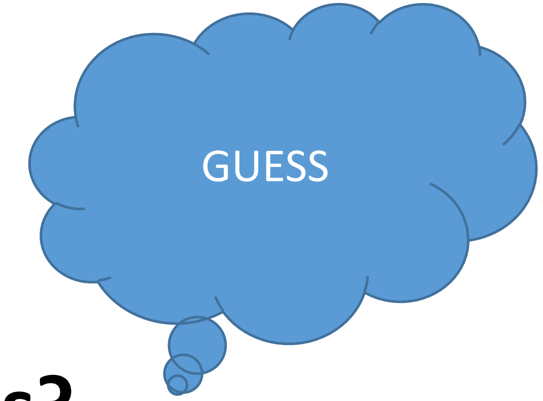
How to Solve this?

- ✓ Guess based on available data.
- ✓ Fill with a value based on the row and column data type.


Note
Missing Inference
Not a TV?

What?

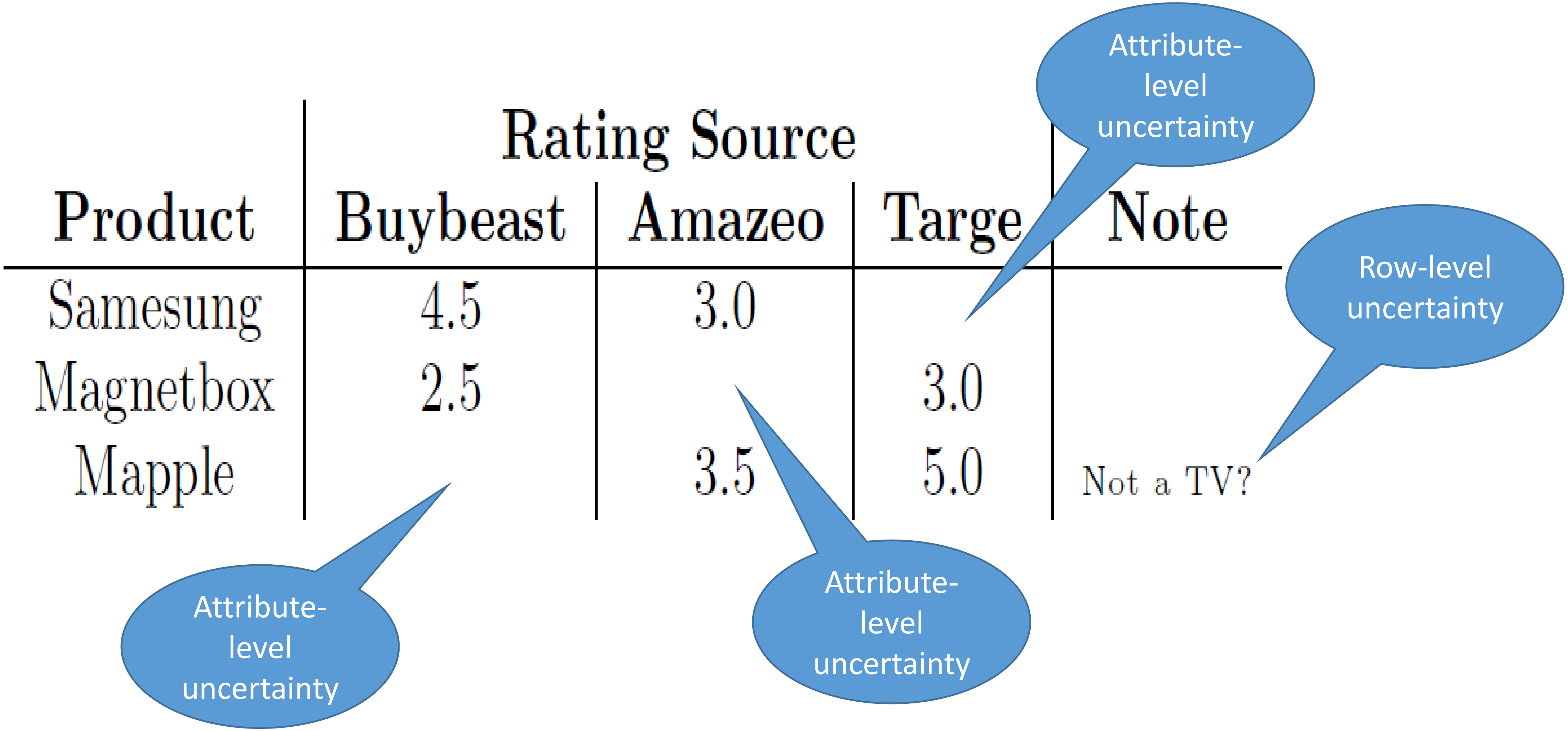
- ✓ Absence of Data in a column.
- ✓ Inability to make any conclusion about the data.



How to Solve this?

- ✓ Manually enter the conclusion.
- ✓ Use Machine Learning Models.

Types of uncertainty



	Rating Source			
Product	Buybeast	Amazeo	Targe	Note
Samesung	4.5	3.0		
Magnetbox	2.5		3.0	
Mapple		3.5	5.0	Not a TV?

Attribute-level uncertainty
Row-level uncertainty
Attribute-level uncertainty
Attribute-level uncertainty

Classical Databases vs On Demand Data Curation

Classical Databases

- Erroneous data cannot be queried.

On Demand Data Curation

- Paygo, KATARA and Mimir
- Defer curation effort until necessary
- **Guesses or approximations as answers**
- **Quality and scope of guesses vary**
- **Communicate information to end user**

On Demand Data Curation

- Initial data is of low quality
- Queries liable to produce incomplete/incorrect result.
- Mitigate unreliability by providing a form of lineage
- Query tagging with quality metrics.
- ODC efforts are specialized form of probabilistic DB.
- Answers in form of certain data or probability distribution.

ODC represent low quality data in the form of probability distribution which is not understood by average database user.

What needs to be changed?

- Although we can fix the data using ODC tools.
- No matter how we fix it we are not guaranteed automated system would do a good job.
- Users uncomfortable with probability distribution.

Can we represent uncertain data in simpler form (user friendly)?

Long Term Goals

- The quality and scope of guesses may vary.
- How to communicate this information to an end-user.
- Set of UI design guidelines.
- Best practices for conveying uncertainty.

Goals of user study

- Preliminary user study
- Evaluate cognitive burden and expressiveness
- Focus on attribute level uncertainty
- Four ways of representing uncertainty

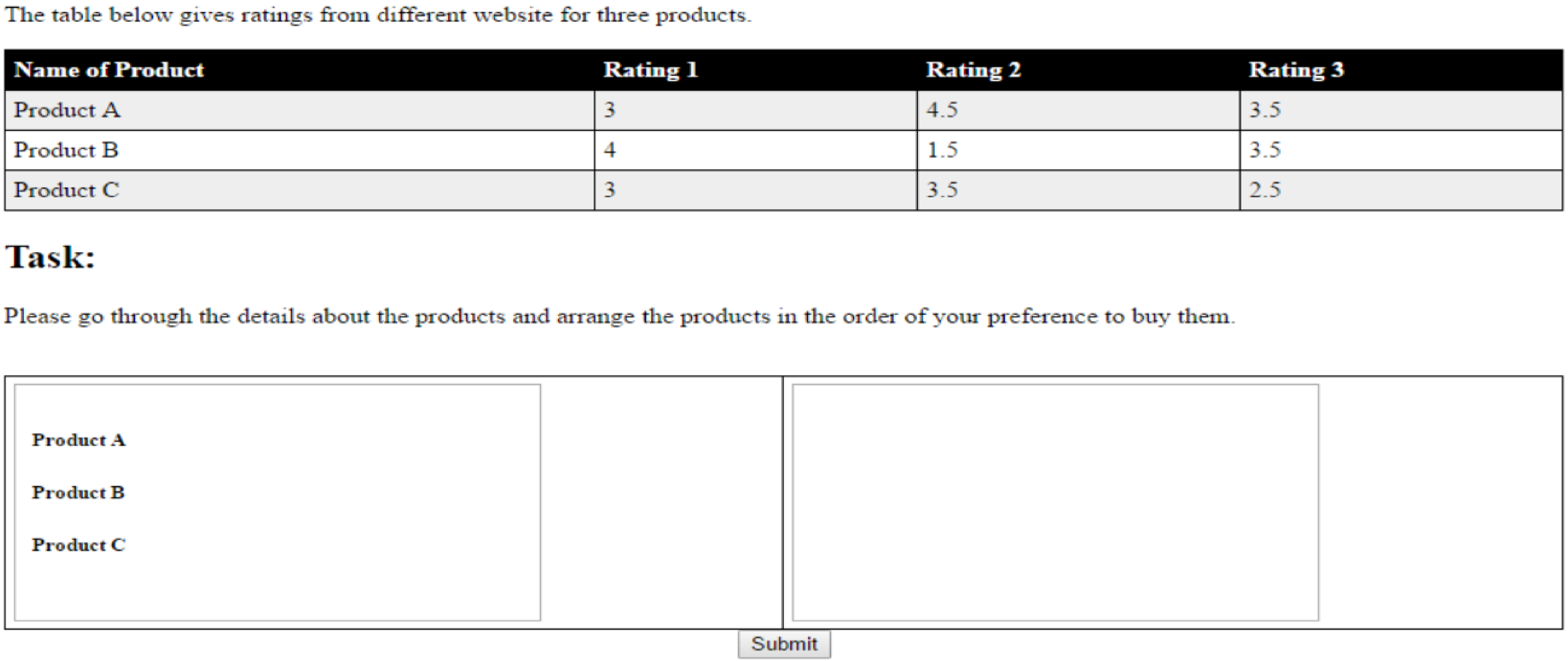
Primary Questions for the User Study

- Is the representation effective at communicating uncertainty?
- What is the cognitive burden of interpreting the representation?
- 14 participants, predominantly from CSE department at UB.

Experimental Setup

- Ranking task
- Web form with 3*3 matrix
- 3 products with ratings from 3 different website
- Participants presented with same set of information
- Multiple rounds with each round of five trials

User Interface



The table below gives ratings from different website for three products.

Name of Product	Rating 1	Rating 2	Rating 3
Product A	3	4.5	3.5
Product B	4	1.5	3.5
Product C	3	3.5	2.5

Task:

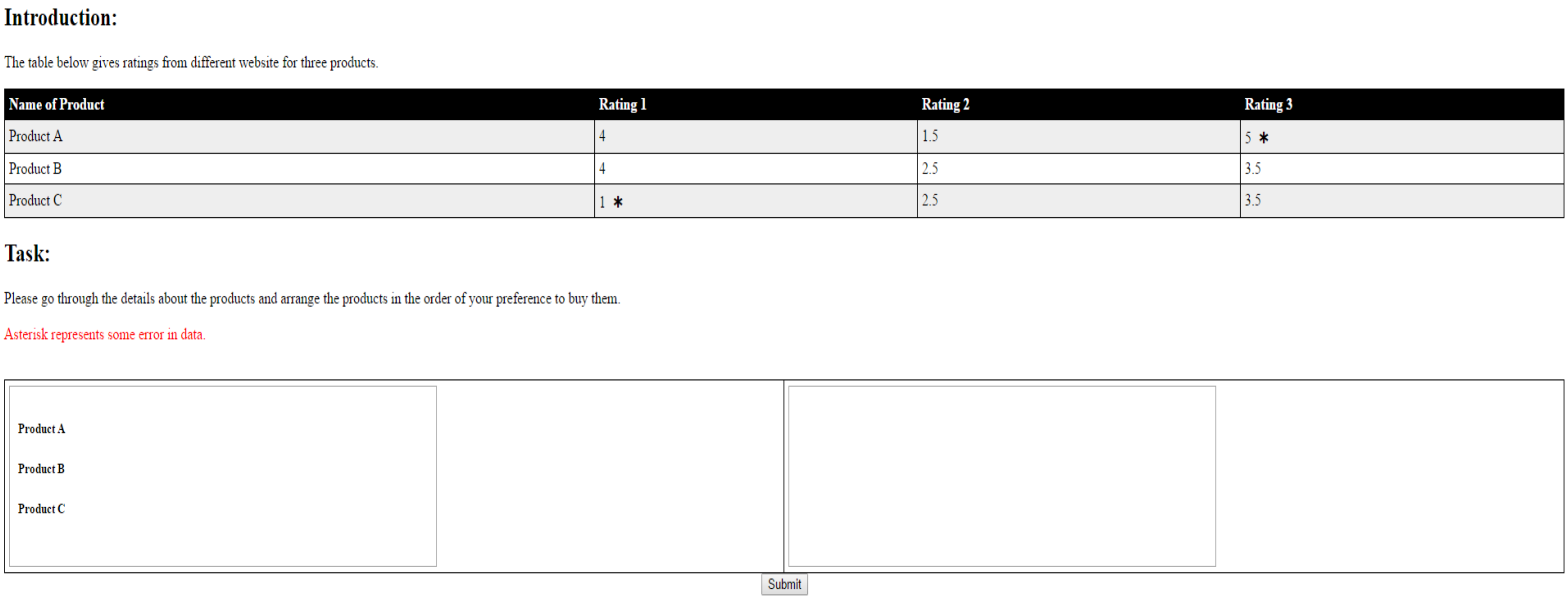
Please go through the details about the products and arrange the products in the order of your preference to buy them.

Product A

Product B

Product C

Asterisk



Introduction:

The table below gives ratings from different website for three products.

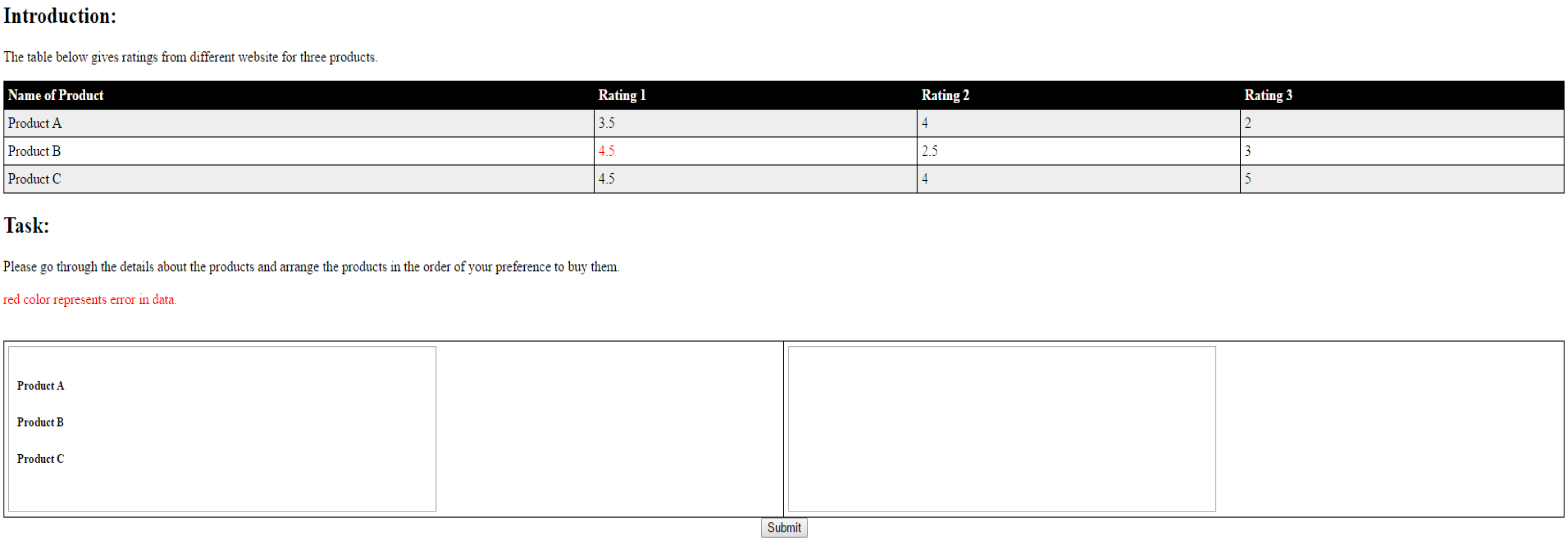
Name of Product	Rating 1	Rating 2	Rating 3
Product A	4	1.5	5 *
Product B	4	2.5	3.5
Product C	1 *	2.5	3.5

Task:

Please go through the details about the products and arrange the products in the order of your preference to buy them.

Asterisk represents some error in data.

Colored Text



Introduction:

The table below gives ratings from different website for three products.

Name of Product	Rating 1	Rating 2	Rating 3
Product A	3.5	4	2
Product B	4.5	2.5	3
Product C	4.5	4	5

Task:

Please go through the details about the products and arrange the products in the order of your preference to buy them.

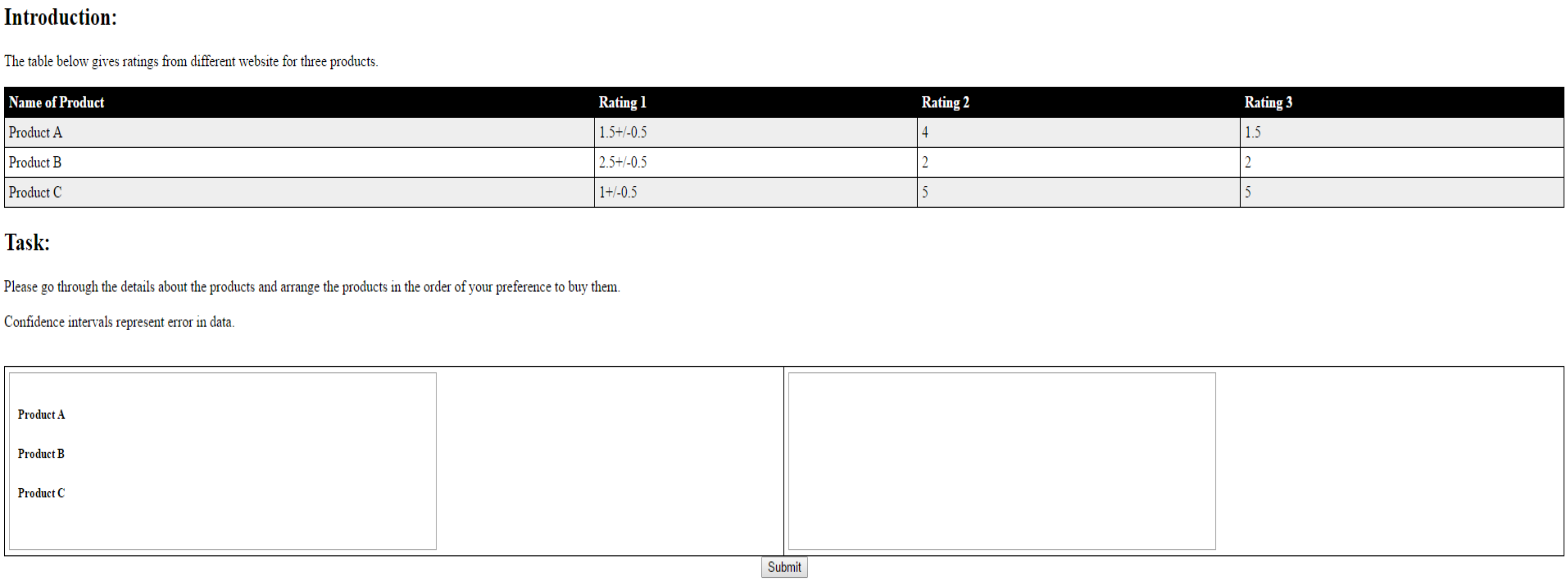
red color represents error in data.

Product A

Product B

Product C

Confidence Interval



Introduction:

The table below gives ratings from different website for three products.

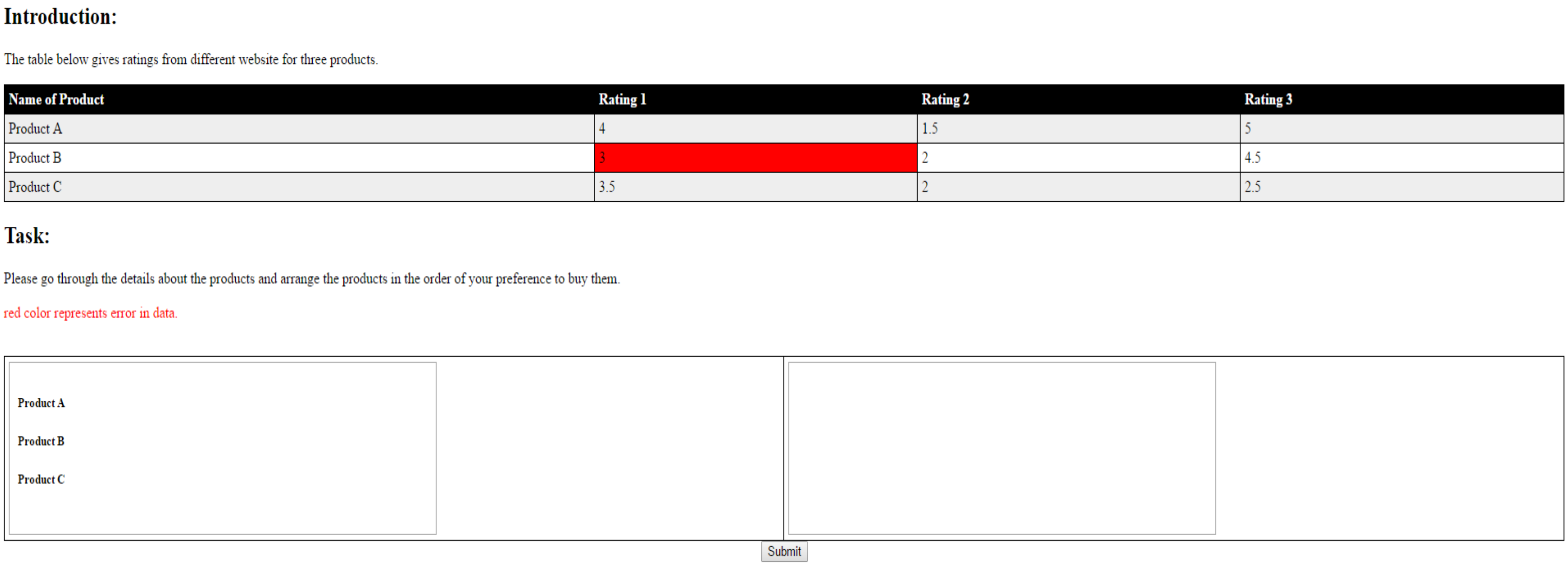
Name of Product	Rating 1	Rating 2	Rating 3
Product A	1.5+/-0.5	4	1.5
Product B	2.5+/-0.5	2	2
Product C	1+/-0.5	5	5

Task:

Please go through the details about the products and arrange the products in the order of your preference to buy them.

Confidence intervals represent error in data.

Color Box



Introduction:

The table below gives ratings from different website for three products.

Name of Product	Rating 1	Rating 2	Rating 3
Product A	4	1.5	5
Product B	3	2	4.5
Product C	3.5	2	2.5

Task:

Please go through the details about the products and arrange the products in the order of your preference to buy them.

red color represents error in data.

Product A

Product B

Product C

Best Of 3

- Ratings were random with a bias towards a predictable ordering.
- Ratings generated using rejection sampling
- A had to have one extremely favorable rating compared to B (1 point higher)
- One slightly more favorable rating (0, 0.5, or 1 point higher)
- one slightly less favorable rating (0, 0.5, or 1 point lower).

With the preselected pattern we can detect change in user behavior.

Introducing Uncertainty

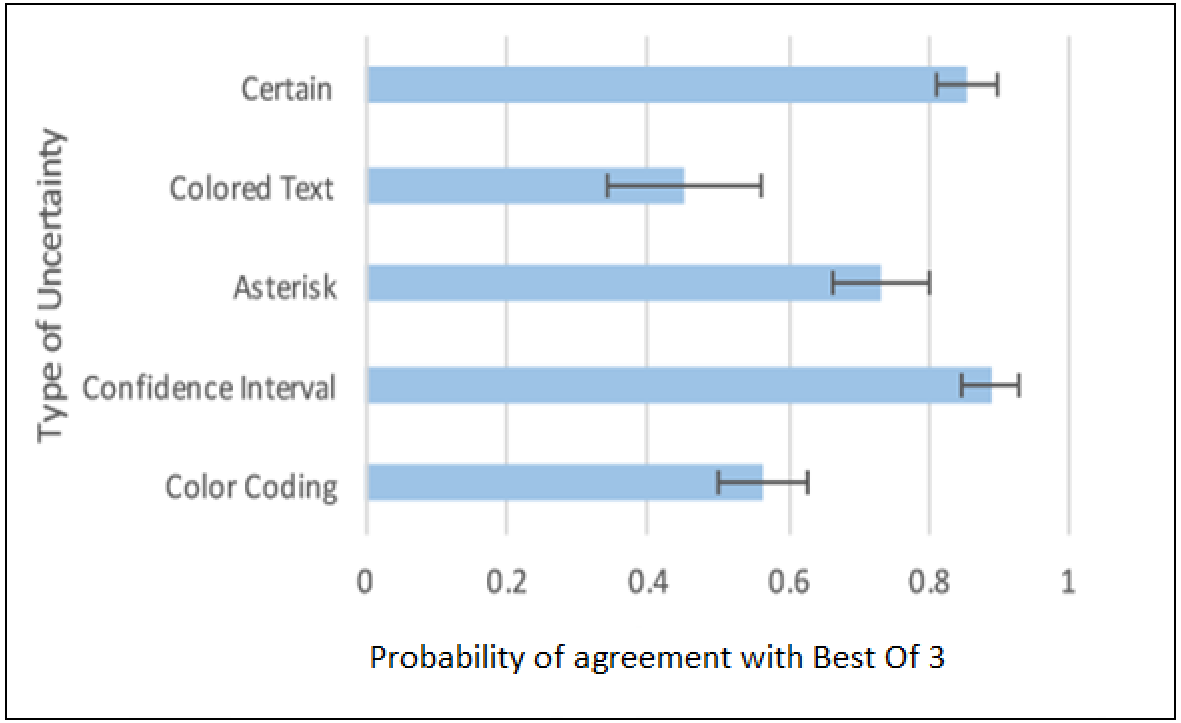
- In uncertain trials, base data generation followed Best Of 3
- Between 2 and 4 randomly chosen values were labeled as uncertain.
- Difference in user behavior due to certain values being uncertain

Rating generation process remained the same but the user were told that certain values were uncertain.

Effectiveness

Compliance with best of 3 changes with uncertainty type

Colored text and color coding altered participant behavior.



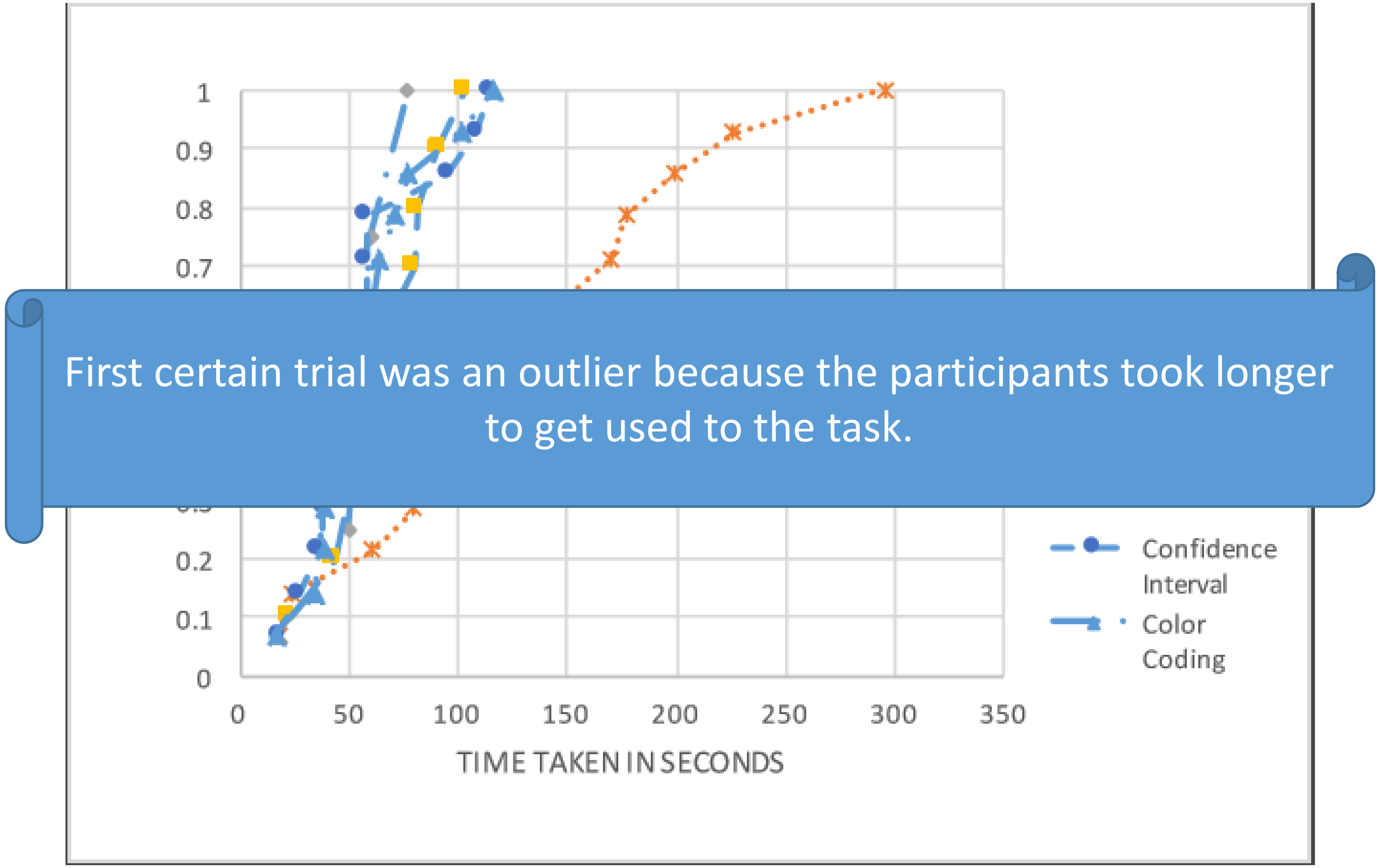
Experiment tests for changes is user behavior.

Certain and confidence interval show a consistent agreement with Best of 3.

Efficacy

First round, participants initially encounter the task and representation

Time taken per representation was relatively consistent across all forms of uncertainty

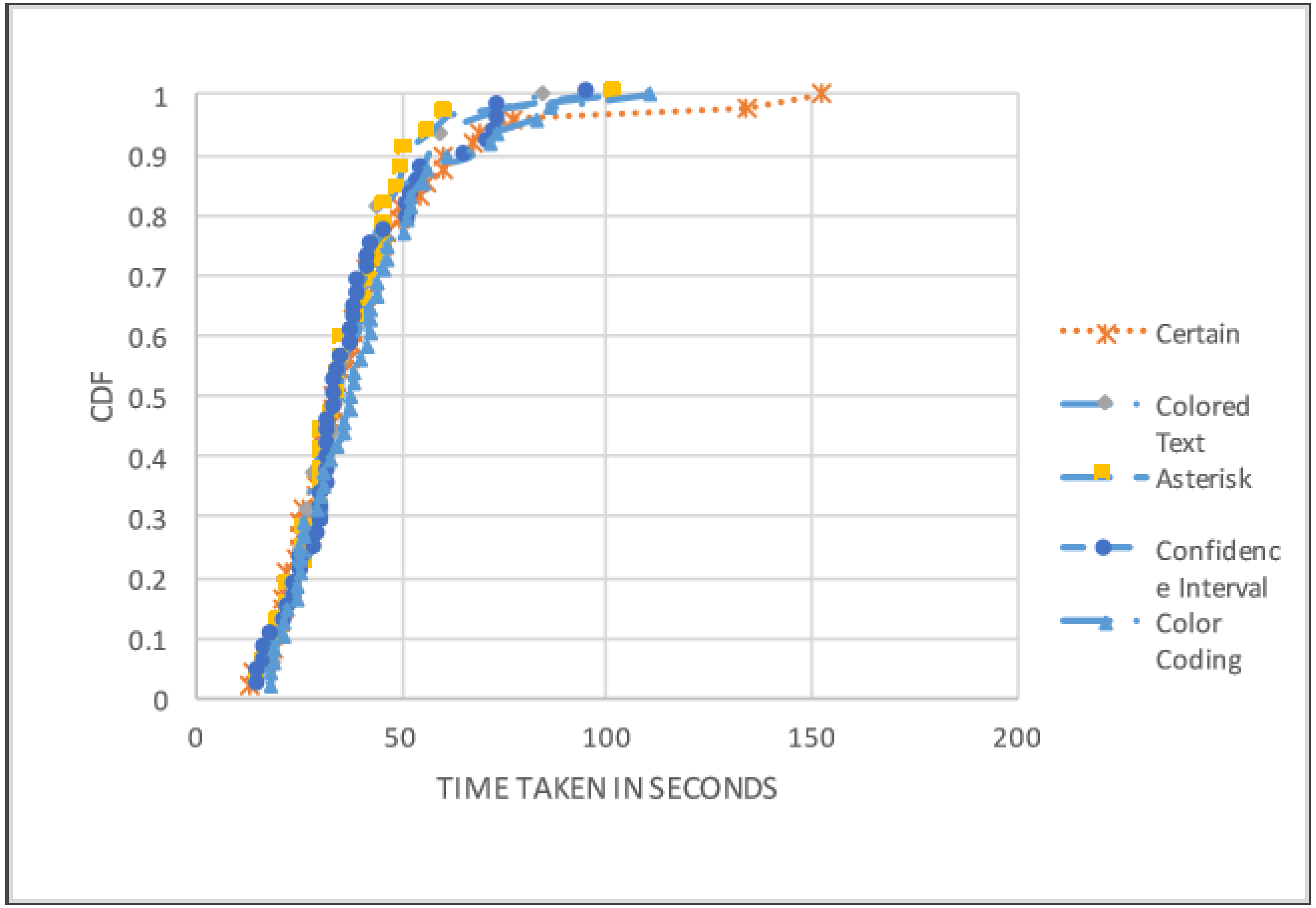


First certain trial was an outlier because the participants took longer to get used to the task.

Participants spent significantly more time familiarizing themselves with the overall ranking

Slower trial was deterministic

Efficacy



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CDF
0
50
100
150
200
TIME TAKEN IN SECONDS
Certain
Colored Text
Asterisk
Confidence Interval
Color Coding

Results

- Is the representation effective at communicating uncertainty?
  - At least three distinct behavioral responses to uncertainty in the data were identified
  - suggesting differences in the efficacy of each representation.
- What is the cognitive burden of interpreting the representation?
  - All uncertainty representations required a similar amount of decision time
  - Impose similar cognitive burdens in the population under study
- Participants were predominantly from CS department.
- Participants conveyed a strong negative emotional reaction to the color coding representation.
- Several participants suggested feelings of comfort associated with the additional information that the confidence interval supplied.

Future Work

- Focus of this study was attribute level uncertainty
- Explore other types of uncertainty in relational data (row-level and open-world)
- qualitative feedback such as explanations
- giving the user mechanisms to dynamically control the level and complexity of uncertainty representation being shown
- incorporating our findings into the Mimir on-demand curation system

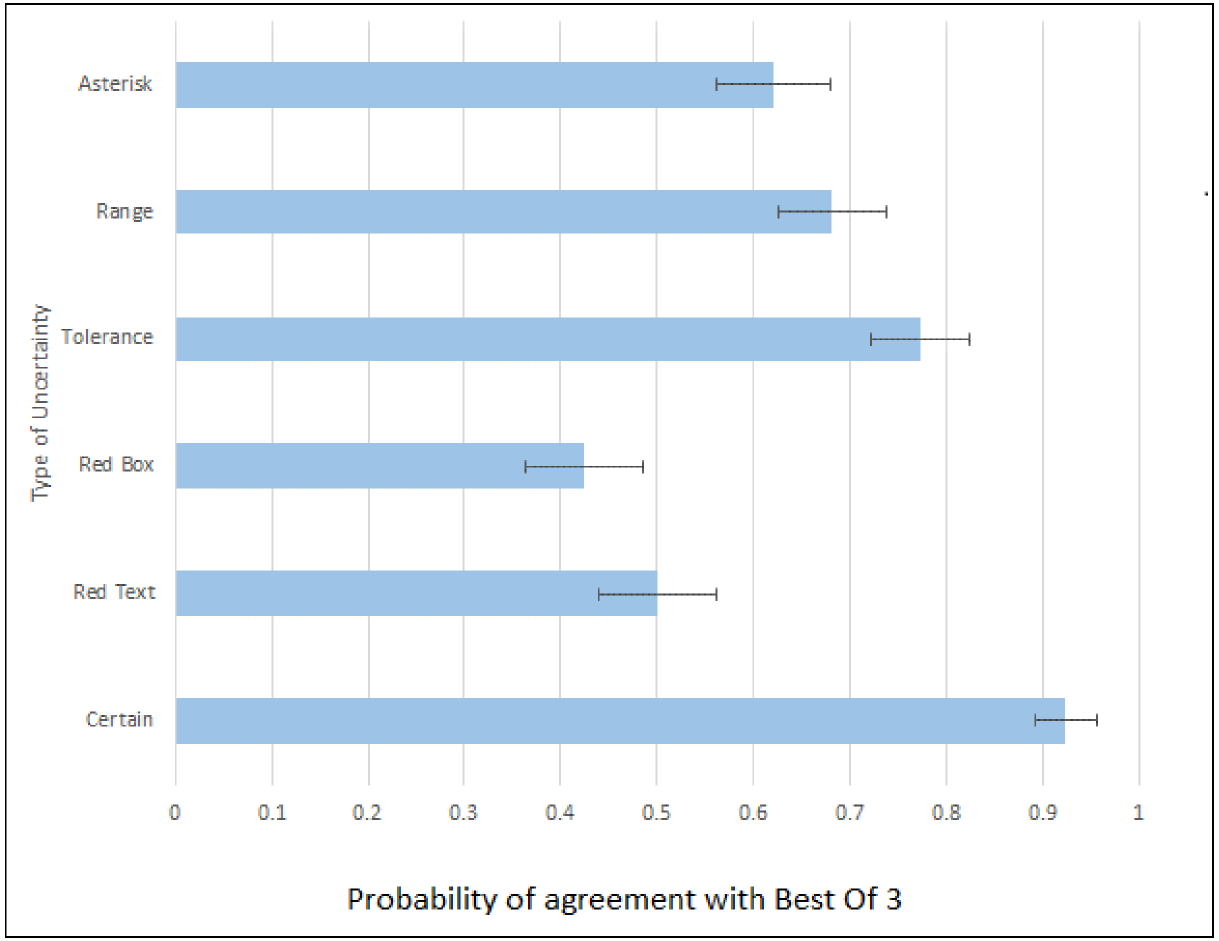
Changes made to Study 2

- Added two more representations of uncertainty.
- 3 rounds per user with 6 set of uncertainties.
- A mix of CS and non CS students were recruited as participants.
- Correlation between user characteristics and results analyzed.

Effectiveness

Certain trial still shows consistent agreement with Best Of 3

Red Box and Red Text were the most effective in altering participant behavior



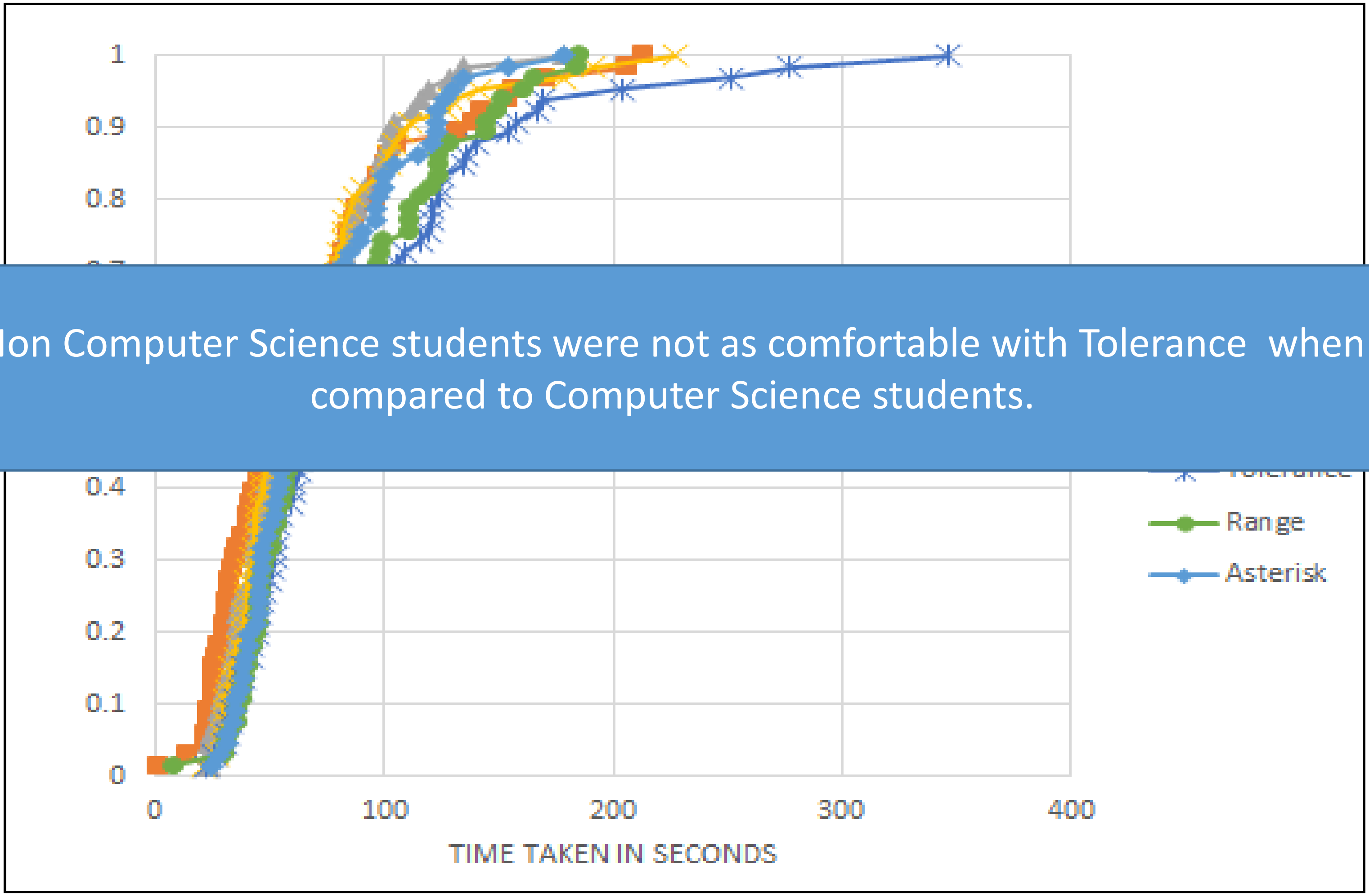
A dip was seen in agreement with Best Of 3 for Tolerance

Participants requested more information in asterisk trial

Efficacy

Participants were introduced to the task before the trial

Time taken per representation was relatively consistent across all forms of uncertainty



Non Computer Science students were not as comfortable with Tolerance when compared to Computer Science students.

Participants still spent more time on deterministic trial

Slower trial was Tolerance

Low quality data can be represented as a certain value by making a guess

We can represent low quality data in simpler form (user friendly)

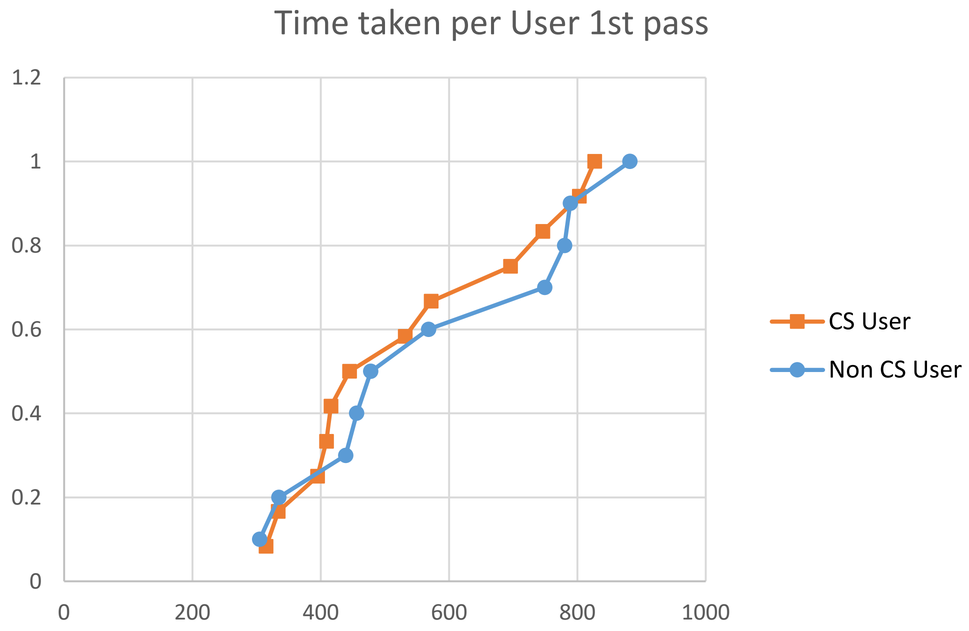
Questions?

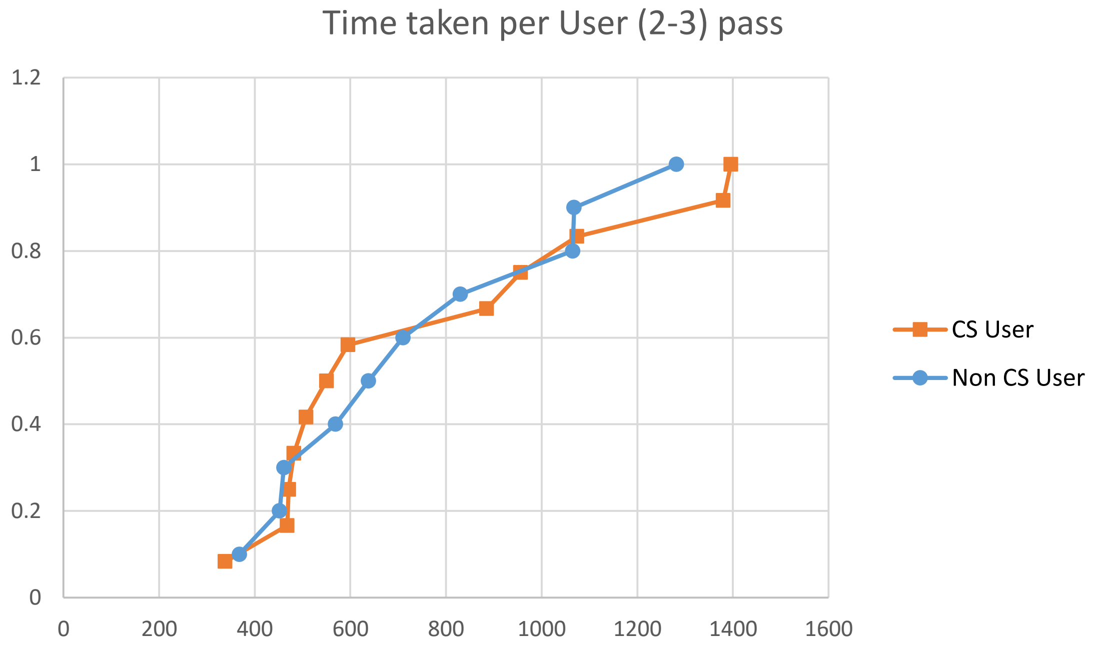
ODC tools make these guesses and represent low quality data as probability distribution

Time taken per representation was relatively consistent across all forms of uncertainty

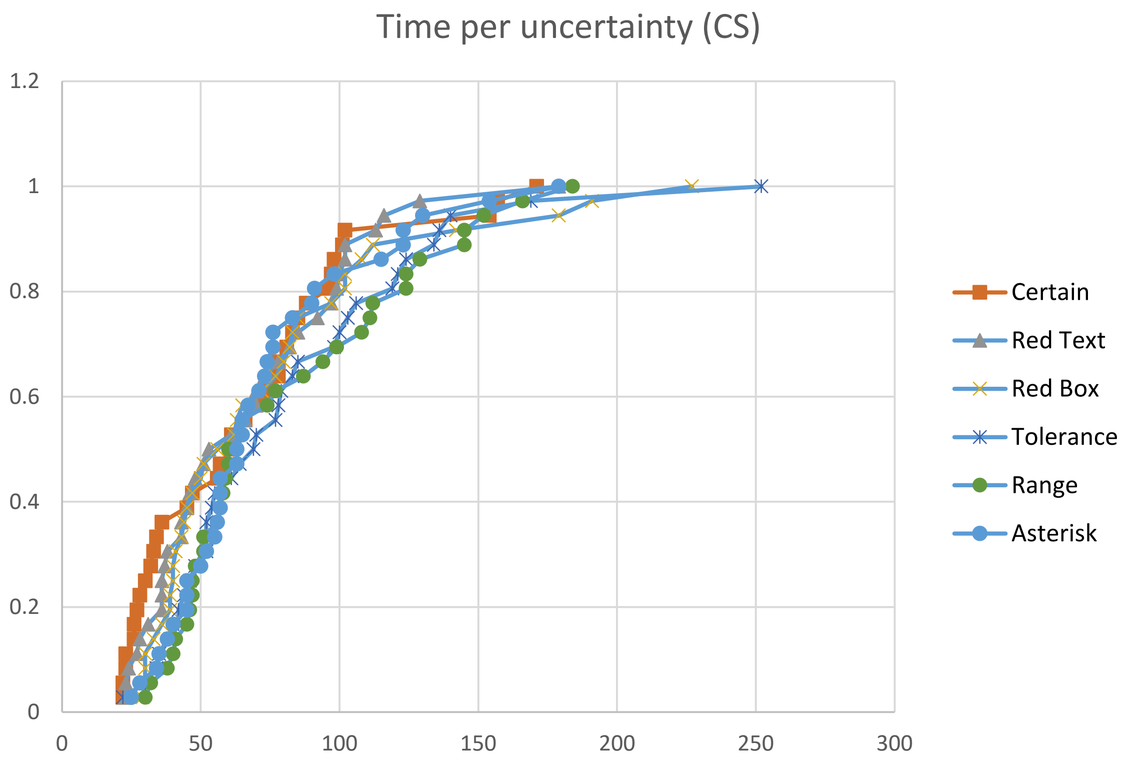
Thank You

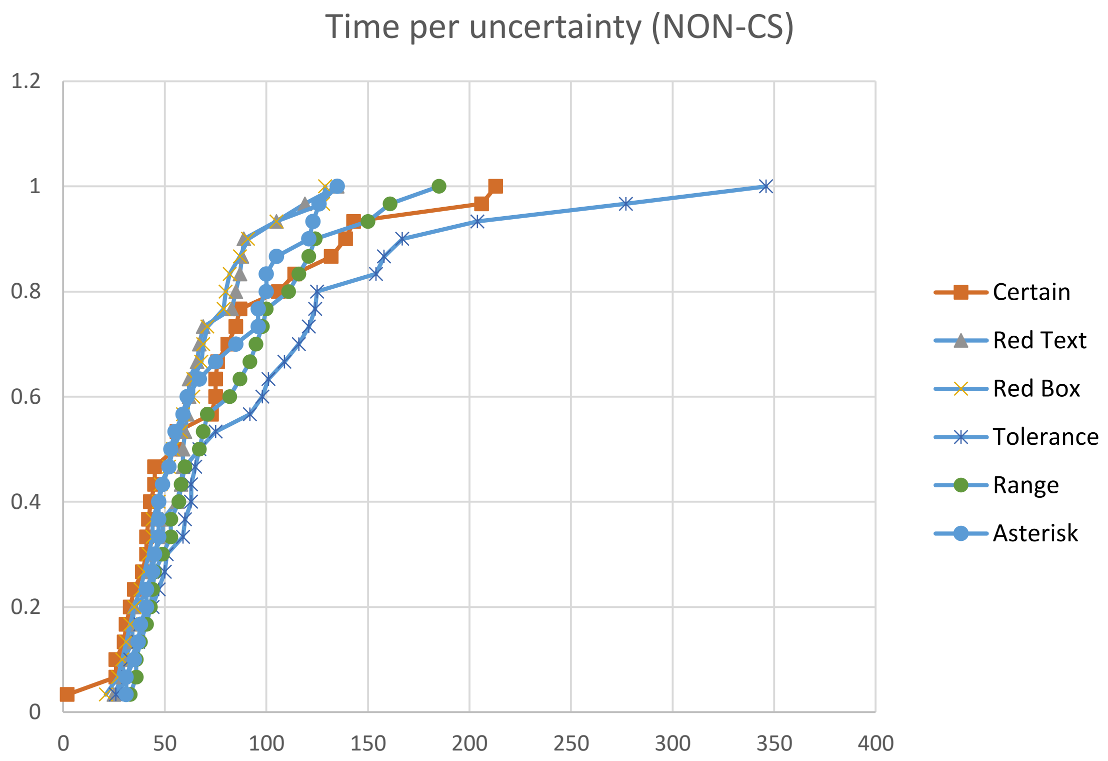
Biases between CS and Non-CS students



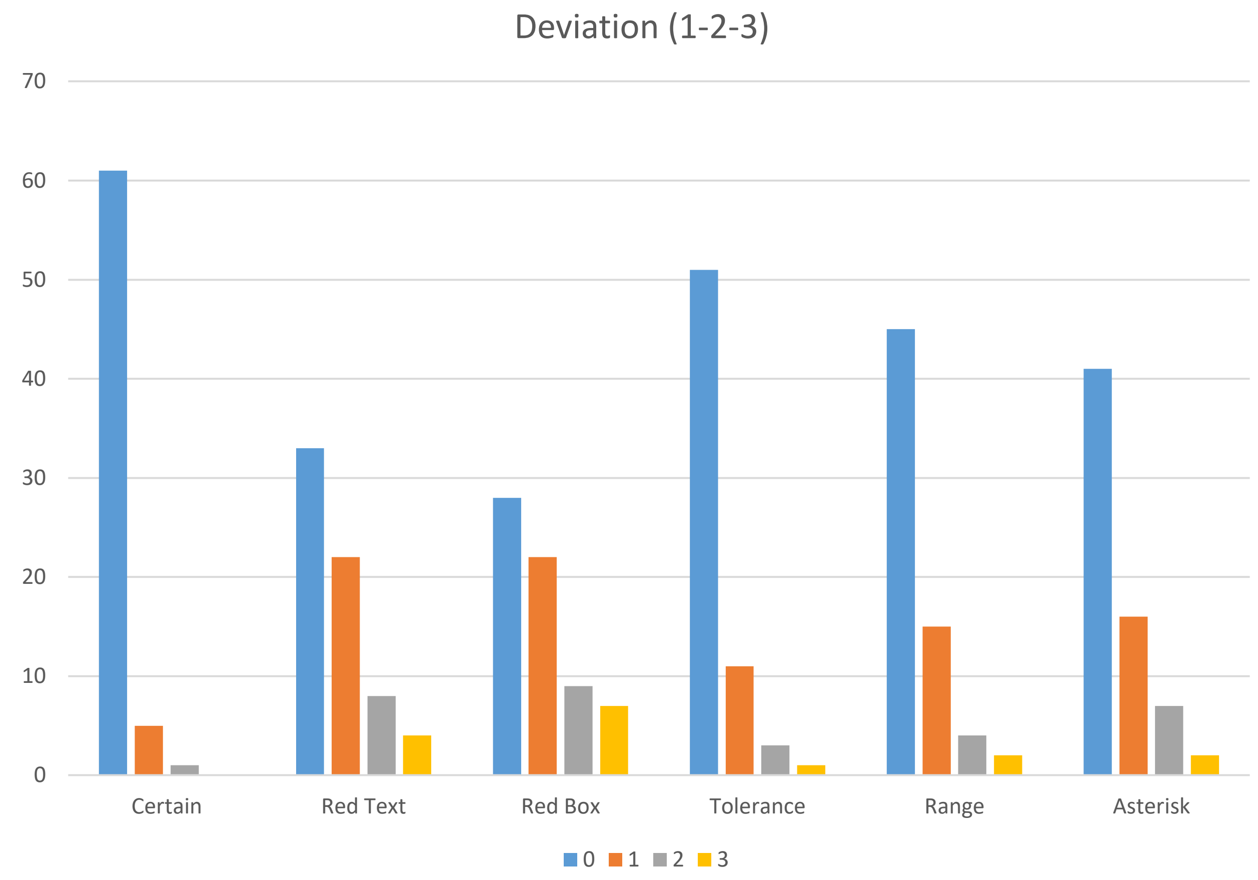


Time taken per uncertainty





Different Regularities


Deviation (1-2-3)
Certain
Red Text
Red Box
Tolerance
Range
Asterisk