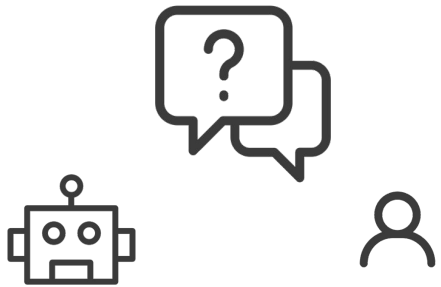
Evaluation I: Measurement Theory

EN. 601.792.01

Ziang Xiao
Department of Computer Science
Spring 2024

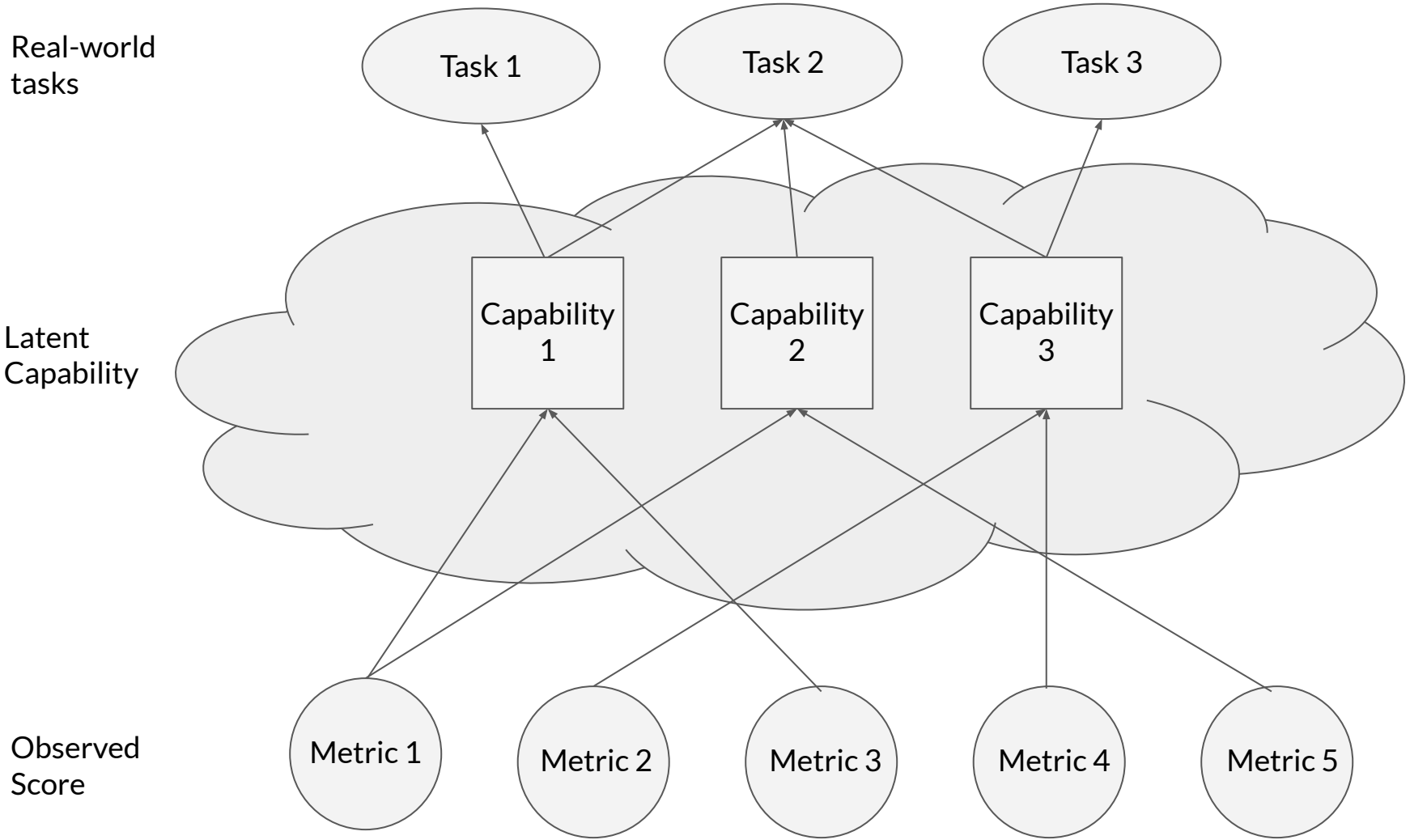


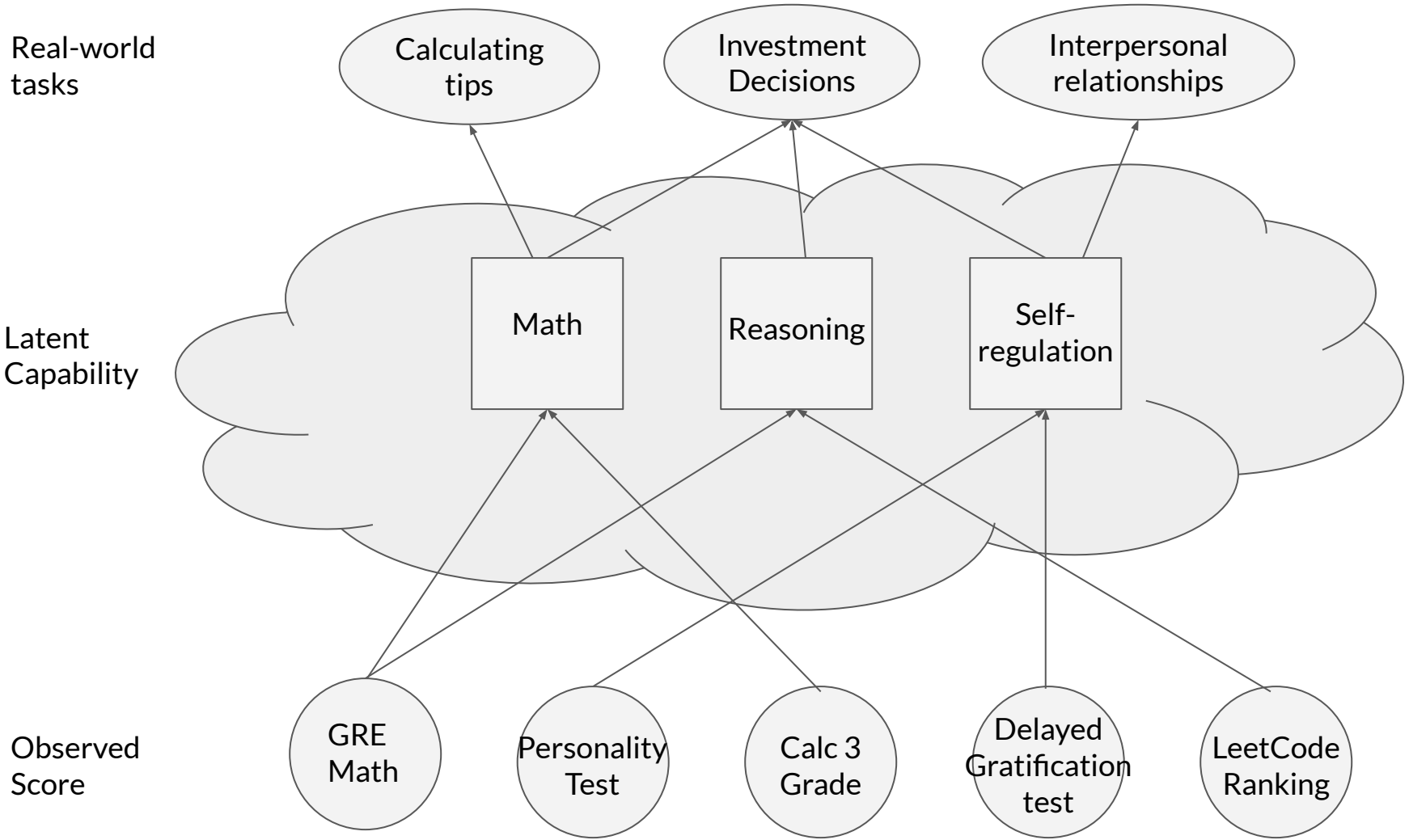
What is measurement theory?

Measurement theory

Measurement theory is the foundation for **developing**, **evaluating**, **interpreting** educational and psychological measurement.

- Use observable tasks to estimate unobservable capability.
- Successful task completion in the test infers future performance on tasks beyond what's included in the test.
- Scores on these tests have direct consequences for high-stakes decisions


Real-world tasks
Task 1
Task 2
Task 3
Latent Capability
Capability 1
Capability 2
Capability 3
Observed Score
Metric 1
Metric 2
Metric 3
Metric 4
Metric 5


Real-world tasks
Calculating tips
Investment Decisions
Interpersonal relationships
Latent Capability
Math
Reasoning
Self-regulation
Observed Score
GRE Math
Personality Test
Calc 3 Grade
Delayed Gratification test
LeetCode Ranking

Measurement theory guided the development and validation of *educational exams*, *cognitive tests*, *personality tests*, *mental health diagnosis scales*, etc.

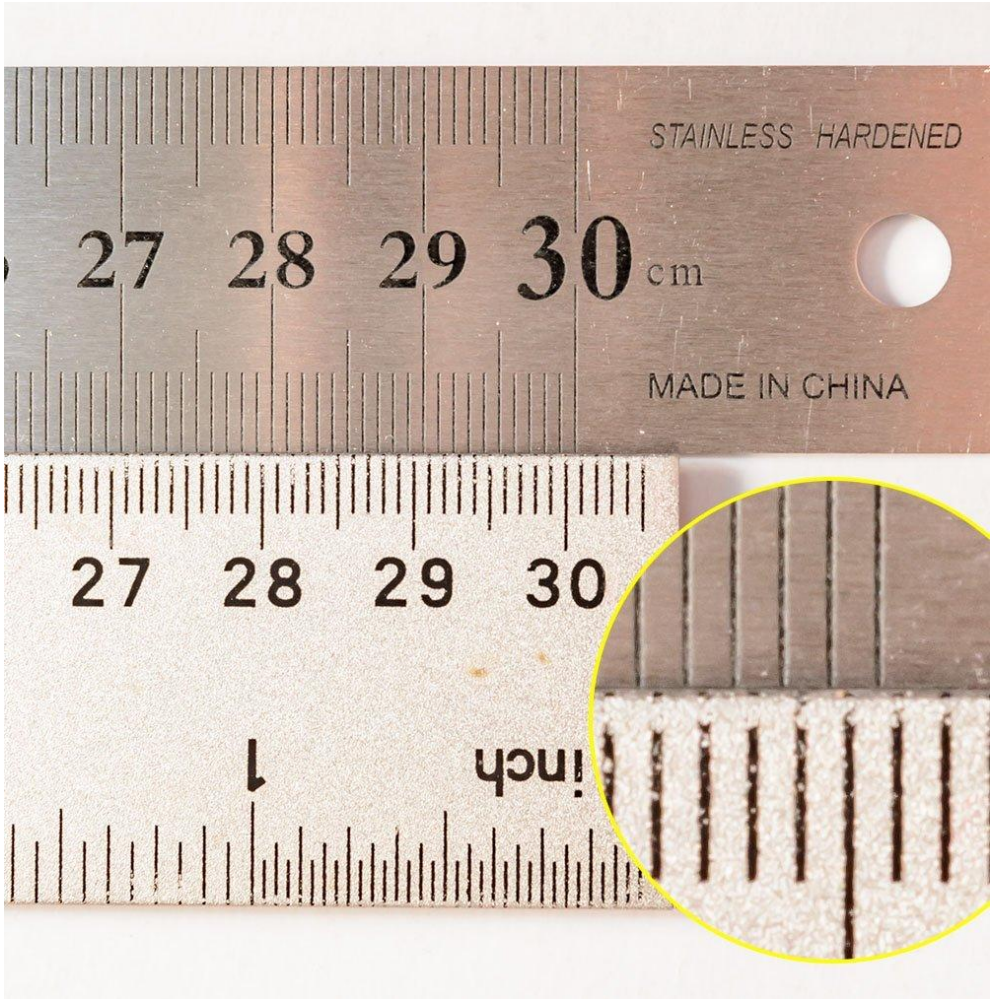
Measurement error

$$\text{Score} = T + \varepsilon$$

Distinguishing **true signal (T)** from **measurement error (ε)** in model performance comparisons.

Measurement error

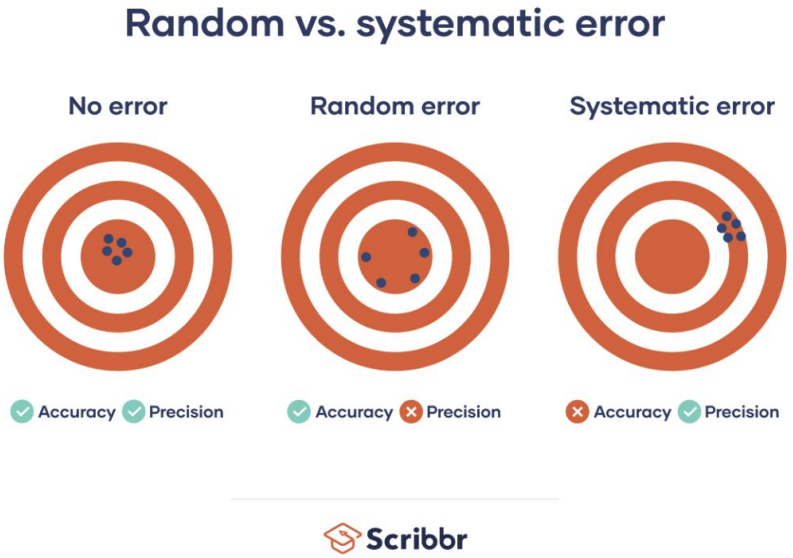
$$\text{Score} = T + \varepsilon$$



Methods in measurement theory aims to identify and quantify measurement errors

- **Reliability** measures the extent to which a metric is subject to random error and is consistent across repeated measures.

- **Validity** focuses on how well the interpretation and uses of metric score are aligned with real-world evidence.



Reliability

Test-retest Reliability

how a score may fluctuate on repeated measures.

Internal-Consistency Reliability

how the metric score fluctuates within a benchmark dataset, i.e., across data points.

Validity

Construct Validity

how a metric score aligns with theorized constructs and captures signals in interest.

Criterion-related Validity

how a metric score is in relation to key external criterions.

Why do we care about measurement theory?

Measuring capability

	Human	Conversational UI
Exam	Multiple subjects, each represents one dimension of capability	Multiple datasets
Subject	Multiple questions, each measuring similar capability	Multiple data piece, each measuring agent capability
Examinee	A person	A conversational agent
Exam Score	Indicating the examinee's capability	Indicating the agent's capability

Methods and conceptual frameworks in measurement theory can guide and evaluate evaluation methods for conversational UI.

Presentation

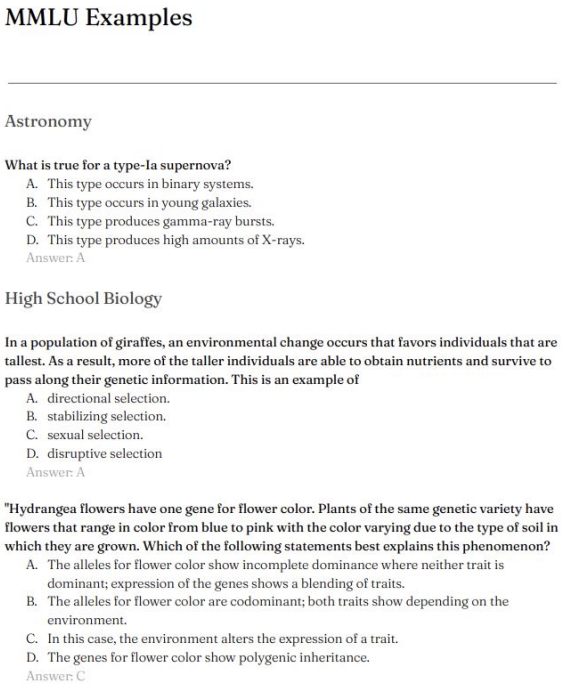
Discussion

Claude 3 benchmarks

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT

Discussion

How do we evaluate/validate benchmarks, such as MMLU?



MMLU Examples

Astronomy

What is true for a type-Ia supernova?

A. This type occurs in binary systems.
B. This type occurs in young galaxies.
C. This type produces gamma-ray bursts.
D. This type produces high amounts of X-rays.

Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection

Answer: A

"Hydrangea flowers have one gene for flower color. Plants of the same genetic variety have flowers that range in color from blue to pink with the color varying due to the type of soil in which they are grown. Which of the following statements best explains this phenomenon?

A. The alleles for flower color show incomplete dominance where neither trait is dominant; expression of the genes shows a blending of traits.
B. The alleles for flower color are codominant; both traits show depending on the environment.
C. In this case, the environment alters the expression of a trait.
D. The genes for flower color show polygenic inheritance.

Answer: C