Validity

Psychological Measurement Theory

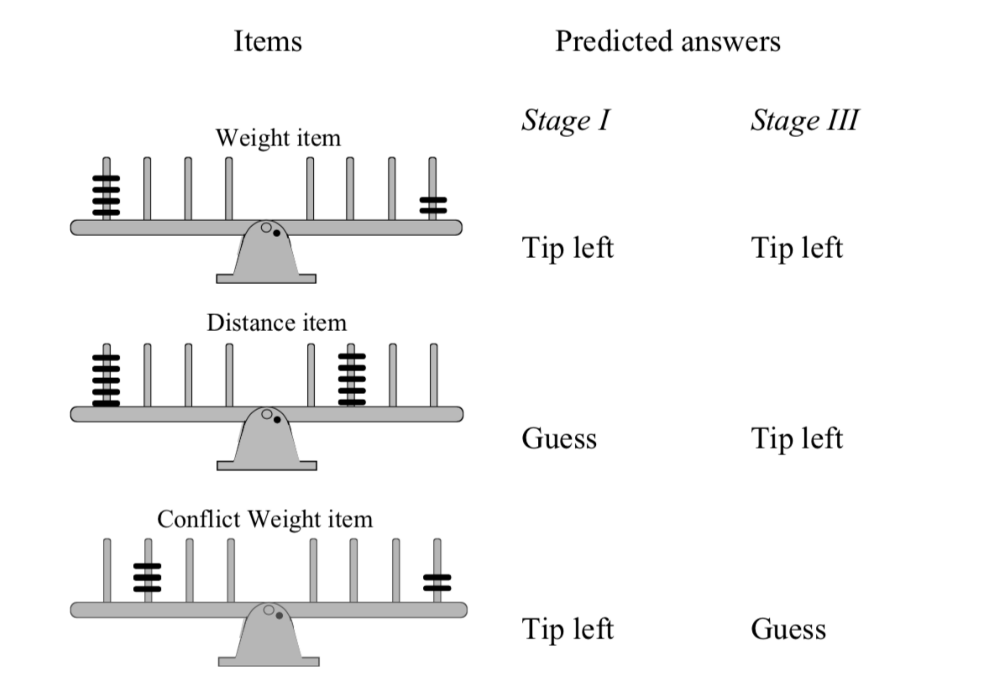
Validity in psychological measurement theory

"A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes."

[Borsboom et al., 2004, Borsboom, 2005, Borsboom and Mellenbergh, 2007]

- Measurement model explicates how the structure of theoretical attributes relates to the structure of observations
- Example: Measurement model for temperature stipulates how the level of mercury in a thermometer systematically relates to temperature

Example: Psychological Measurement of Developmental Stages



Items	Predicted answers: *Stage I*	Predicted answers: *Stage III*
Weight item	Tip left	Tip left
Distance item	Guess	Tip left
Conflict Weight item	Tip left	Guess

- Example: Psychological test of developmental stages by Jean Piaget [Inhelder and Piaget, 1958]
  - Different positions in the attribute (e.g. children of age 3-5 versus 10-12) lead to different test outcomes
  - Observed test outcomes can be used to infer position of children in one of four discrete stages of cognitive development

Example: Measurement in the train-dev-test Paradigm

Machine learning models as measurement models

- Example: Multiclass classification
  - Variation in attribute = variation of test pairs (x, y) where x are inputs, y gold standard outputs
  - Measurement outcome = model prediction $\hat{y}$ for input x
  - Causal relation = variation in feature values and labels correlates invariantly across environments (here: test (re-)splits)

    [Peters et al., 2016, Arjovsky et al., 2019]

- Is accurate prediction across test sets all we need to claim validity?
- No! Further criterion of **absence of circularity** from philosophy of science. [Balzer and Brendel, 2019]

Circularity in Machine Learning

Circular features

- **Indirect measurements:** Target label is determined by indirect measurement, but fundamental measurements needed to determine this indirect measurement are part of input feature representation.
- **Circularity:** Circular feature (= fundamental measurement) will lead to an exact reconstruction of the known deterministic function (= indirect measurement) by machine learning, but achieves nothing else.

The Circularity Problem

Why is circularity a problem?

- Machine learning models trained on data including circular features will yield **nearly perfect predictions** on input data including the defining measurements, but they **cannot be transferred to unseen data** where the defining features are not or only incompletely available.
- Circular features in machine learning models will **nullify the contribution of all features except those defining the target**, thus such models will **not learn new predictive patterns** that involve features other than the known defining measurements.

Circularity Test

Goal: A Circularity Test for Black-Box Models

- Assume we know the functional definition of the target, but not the training data of the machine learning model
- Our data are model predictions on test data $T = \{(\mathrm{x}^m, \hat{y}^m)\}_{m=1}^{M}$
- Detect whether black-box model used circular features and remove them from dataset.

GAMs, Deviance, and Nullification

- **GAMs:** Expressive and yet interpretable model class [Wood, 2017]
  - Decompose complex function into sum of non-linear *feature shapes* $f_k(x_k)$, e.g., regression splines [Hastie and Tibshirani, 1990]

$$Y^n = \sum_{k=1}^{p} f_k(x_k^n) + \sum_{i \neq j} f_{ij}(x_i^n, x_j^n) + \epsilon^n, \text{ where } \epsilon^n \sim \mathcal{N}(0, \sigma^2).$$

- **Deviance:** $D^2(\mu) \in [0,1]$ measures proportion of log-likelihood of model μ out of maximal data fit.
- **Nullification:** Identifies circular features by non-null feature shapes, based on identifiability and consistency of maximum likelihood estimators for GAMs.

Statistical Test for Circularity

- **Dataset** $T = \{(\mathrm{x}^m, \hat{y}^m)\}_{m=1}^M$ where $\mathrm{x}^m \in \mathbb{R}^p$,
- **Candidate circular features** $C \subseteq \mathcal{P}(\{1, \ldots, p\})$,
- **Models** $\mathcal{M} := \{\mu_c \colon c \in C\}$ obtained by fitting a GAM based on feature set c to data T.

Two-step test to detect circular features c^*

1. **Deviance:** $c^* = \mathrm{argmax}_{c \in C}\, D^2(\mu_c)$ where $D^2(\mu_{c^*})$ is close to 1, and in case the maximizer is not unique, the maximizer is chosen whose associated GAM μ_{c^*} has the smallest degrees of freedom.
2. **Nullification:** The feature shape of every feature $x_j : j \in \{1, \ldots, p\} \setminus c^*$ added to the GAM μ_{c^*} is nullified in the resulting model.

Example: Circularity in Patent Prior-Art Search

Condition	Relevance Score
no citation	0
inventor citation	1
examiner citation	2
family patent	3

- Data: Construction of gold standard relevance judgements from citations of patents in other patents. [Graf and Azzopardi, 2008]
- Model: Non-linear combination of features by MLP, trained for logistic regression on binarized relevance ranks (level 1 for citations, 0 else).
- Teacher MLP trained on 1,500 patent queries, resulting in 318,375 observations of query-document pairs.

Example: Circularity in Patent Prior-Art Search

- **What if patent citations are included as features in ranking model** (e.g., KISS principle of [Magdy and Jones, 2010])?

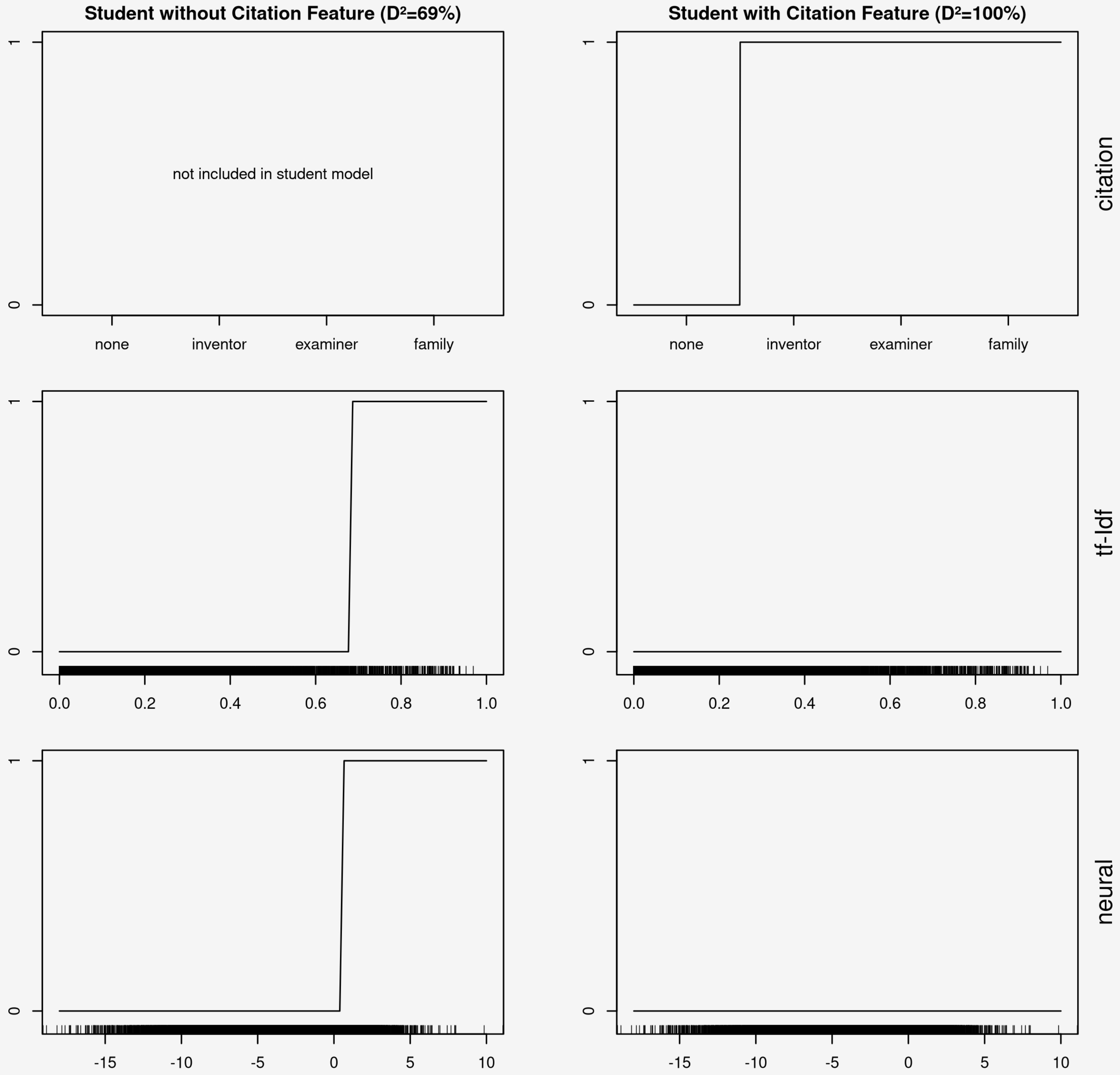
	Feature	Meaning	Range
(1)	`neural`	similarity score learned by neural network	$\mathbb{R}$
(2)	`tf-Idf`	cosine similarity of tf-Idf scores	$\mathbb{R}$
(3)	`inventor`	indicator for inventor citation	$\{0,1\}$
(4)	`examiner`	indicator for examiner citation	$\{0,1\}$
(5)	`family`	indicator for family patent	$\{0,1\}$

Deviance of Student GAMs

Rank	Included Features	D^2	Complexity
1	`{inventor, examiner, family}`	100%	5
2	`{inventor, examiner, family, neural}`	100%	6.33
3	`{inventor, examiner, family, tf-Idf}`	100%	7.95
4	`{inventor, examiner, family, neural, tf-Idf}`	100%	11.1
5	`{examiner, family, neural, tf-Idf}`	95%	22

- Top five models visited during circularity search for IR training data.

Nullification in Student GAMs



Example: Circularity in Medical Data Science

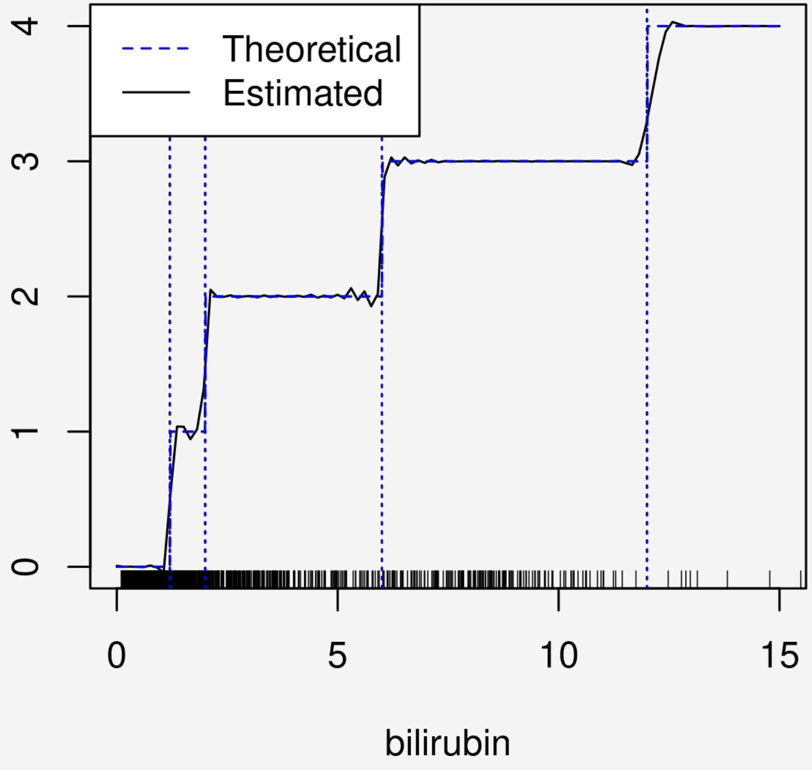
Circularity in SOFA/Sepsis Score Prediction

- Sepsis-3 consensus definition defines sepsis as a change in total SOFA (sequential organ failure assessment) score of at least 2 points consequent to an infection. [Singer et al., 2016, Seymour et al., 2016]
- SOFA scoring system itself is defined for 6 organ systems whose scores are defined by thresholds on measurable physiological quantities like heart rate, creatinin, bilirubine, urine output etc.

  [Vincent et al., 1996]
- Recent overview examined 22 studies on machine learning for (early) prediction of sepsis, with the exception of one, all studies define ground-truth sepsis labels using the deterministic rules of the consensus definition like Sepsis-3. [Moor et al., 2021]

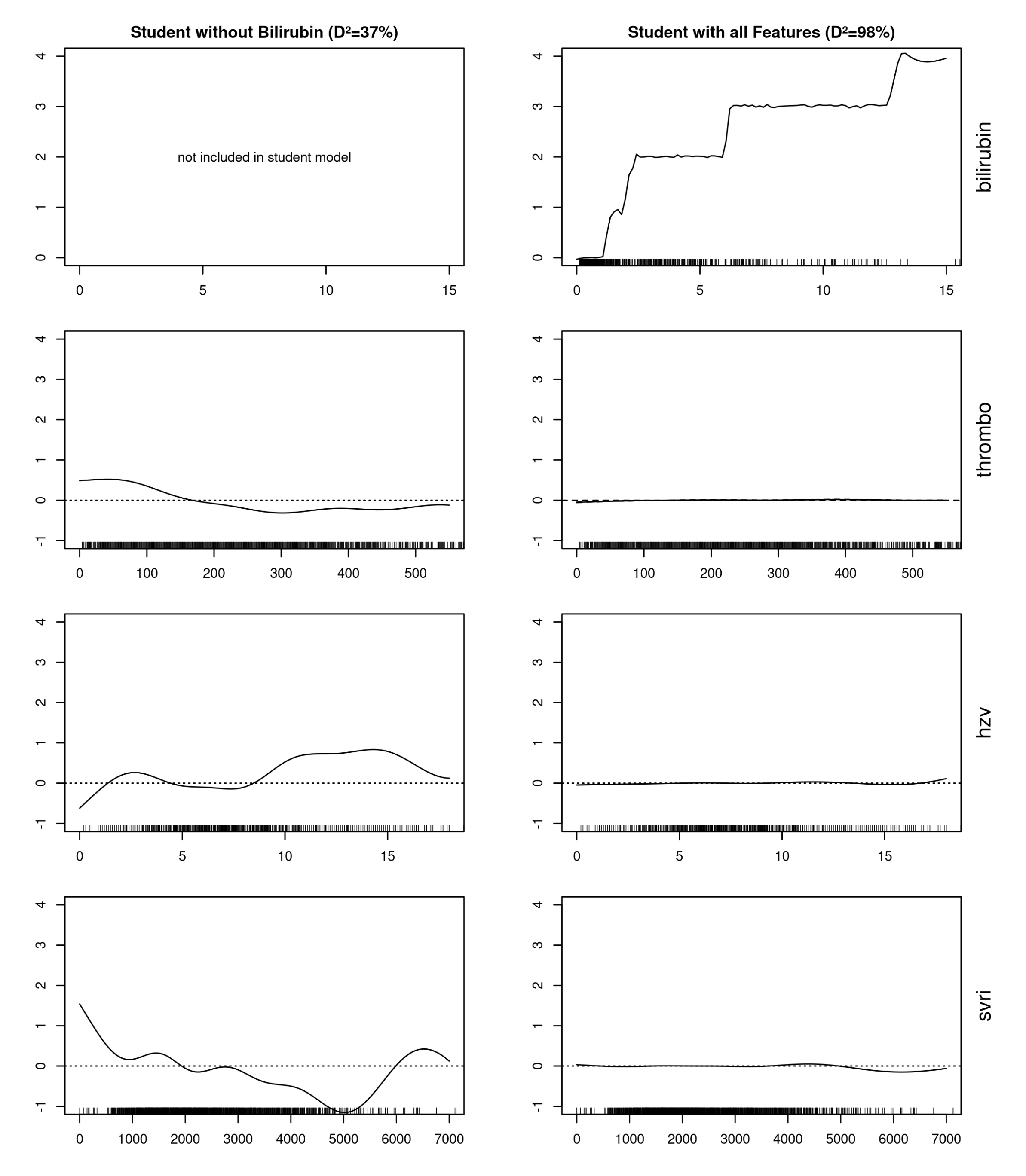
Circularity in Liver SOFA Datasets

Condition	Liver SOFA Score
$0 < \text{bilirubin} \leq 1.2$	0
$1.2 < \text{bilirubin} \leq 1.9$	1
$1.9 < \text{bilirubin} \leq 5.9$	2
$5.9 < \text{bilirubin} \leq 11.9$	3
$\text{bilirubin} > 11.9$	4



- Definition and reconstruction of liver SOFA score.

Circularity in Liver SOFA Prediction



Circularity in Medical Data Science

- How likely is it that other datasets and machine learning models exhibit a yet undetected circularity problem?
    - **Critical candidates** are machine learning applications in measurement-based sciences like medicine that define the objects of their research, e.g., diseases, by **rigid measurement procedures** (e.g., on physiological features, images, or text data).
    - **Circularity problems extend to a longitudinal design for prognosis** where feature measurements at a current point in time are used to forecast disease status at future points in time.
        - Circularity will be introduced by the auto-correlation of the time series, especially if data imputation methods like last-value carried forward are used in dataset creation.

Summary: Circularity

- **Circularity inhibits machine learning at its core**
  - If circular features are included in data/models, **nothing else but a reconstruction of the known functional definition of the target will be learned.**
  - If circular features are not or only incompletely available, **reproducibility is lost** in any case.
  - **No insights into new predictive patterns**, no transfer to data labeled in other ways.
- Remedy: Detect and remove circular features in data/models!