Modeling Disability Progression in Multiple Sclerosis Using Item Response Theory

Younes Boulaguiem, PhD
Postdoctoral Researcher in Statistics

University of Geneva – Faculty of Medicine
Clinical Research Center, HUG

Addressing inconsistencies of the EDSS by modeling disability progression through IRT.

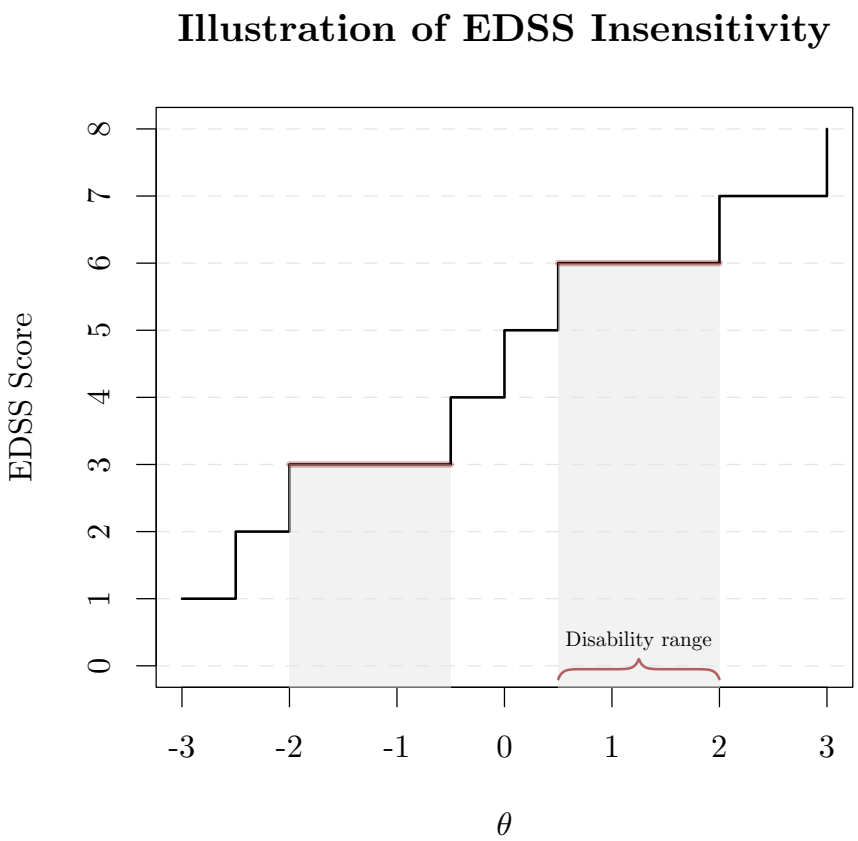
Delivered April 2025

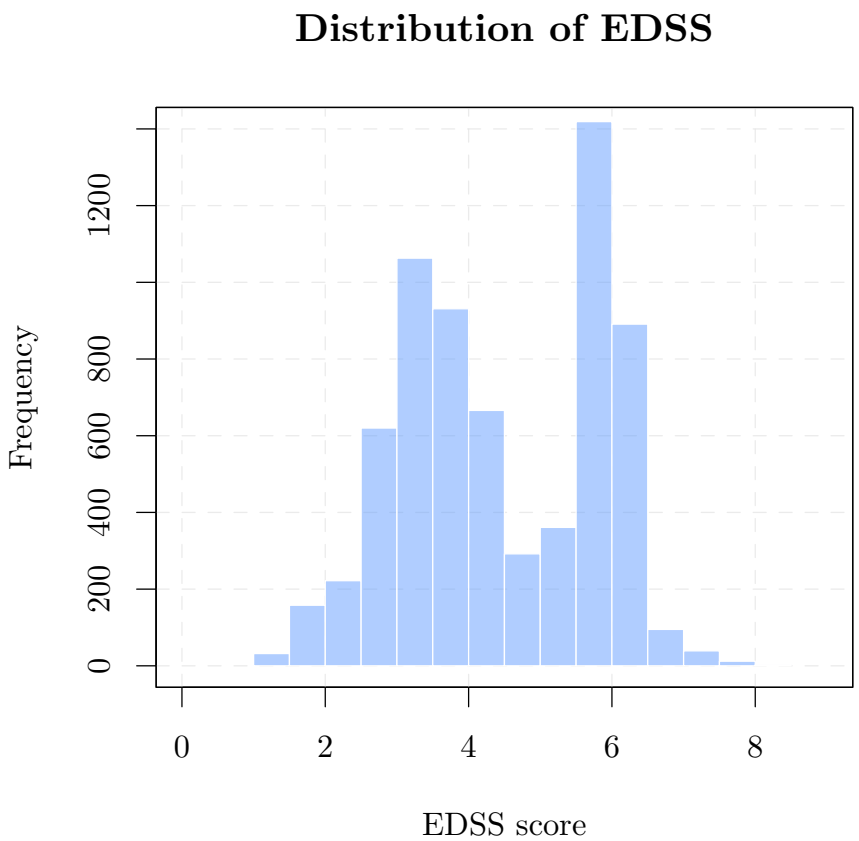
Expanded Disability Score

Drawbacks of the Expanded Disability Status Score:

- **Ordinal** scale, **bimodal** distribution, and **insensitive** to **congnition** and **upper limb** dysfunction.
  - $\rightarrow$ May not detect **small** but **meaningful** changes in disability.
- **Poor inter-rater** reliability (in theory).
  - $\rightarrow$ Likely due to **ambiguous score system** (loopholes) and **discrepancy** between **attributed** score and **patient experience**.
- **Heavily weighted** toward **ambulation**, especially for scores > 3.5.

Expanded Disability Score II



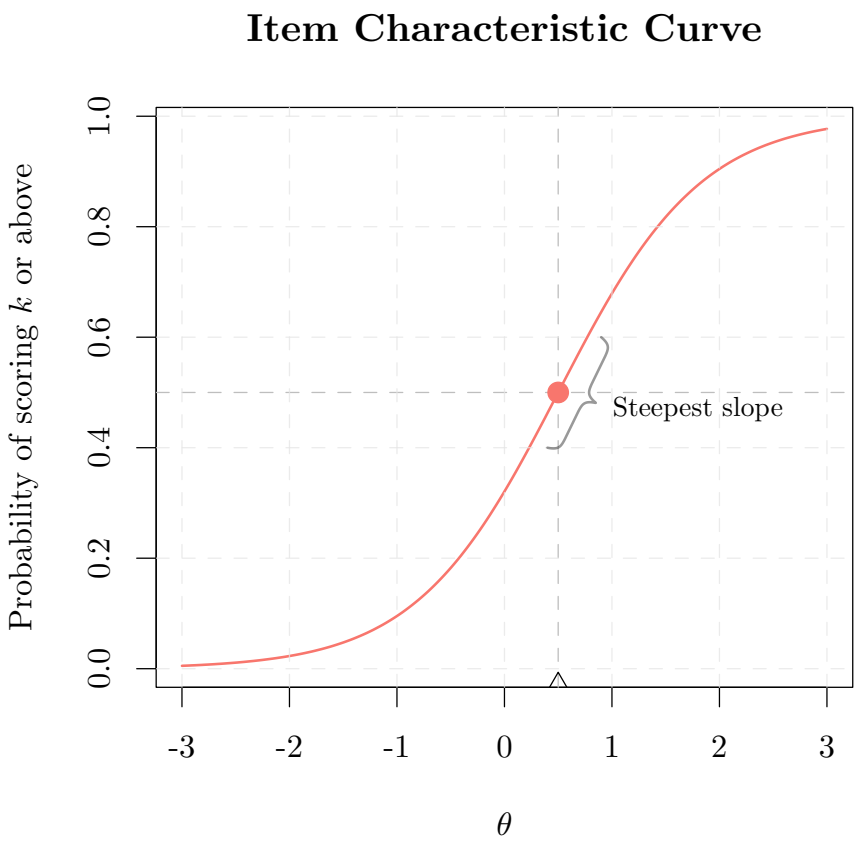


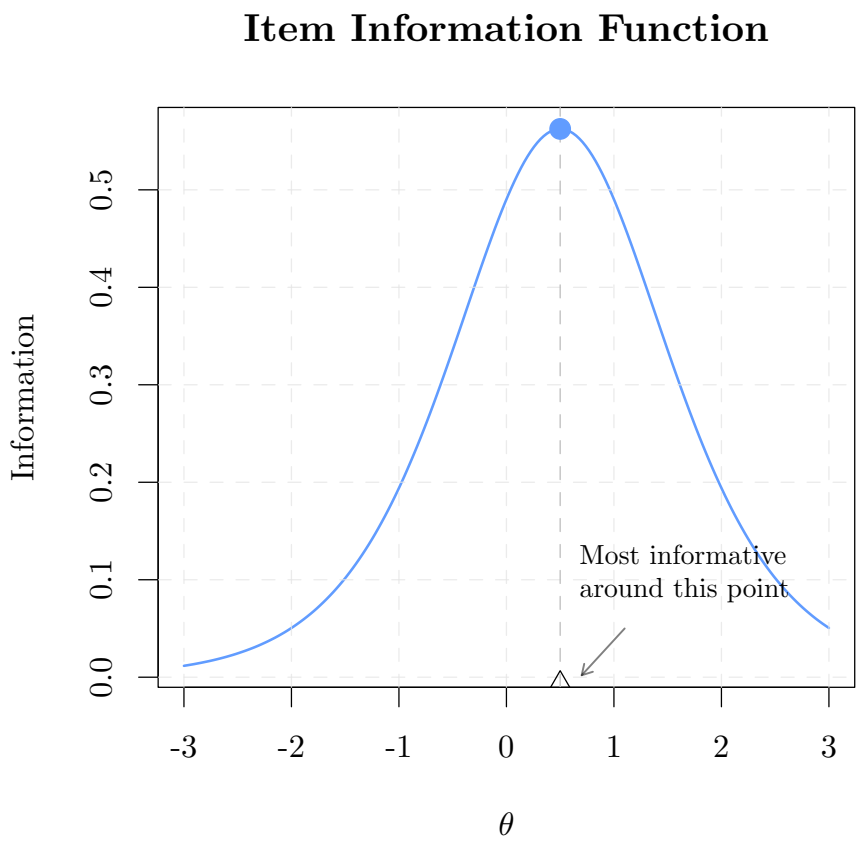
Item Response Theory as an Alternative

Why use Item Response Theory (IRT)?

- **Directly infers** disability on a **continuous scale**, rather than a summary score.
  - → Mitigates the stepwise shape and therefore **preserves granular information** that is lost in the composite scales.
- **Flexible** and captures information at the **item level**.
  - → Evaluates the informativeness of each item **individually**, and allows for the **flexible inclusion/exclusion** of (additional) items — unlike the EDSS, which is restricted to a fixed set of Functional System scores.

Item Response Theory as an Alternative II



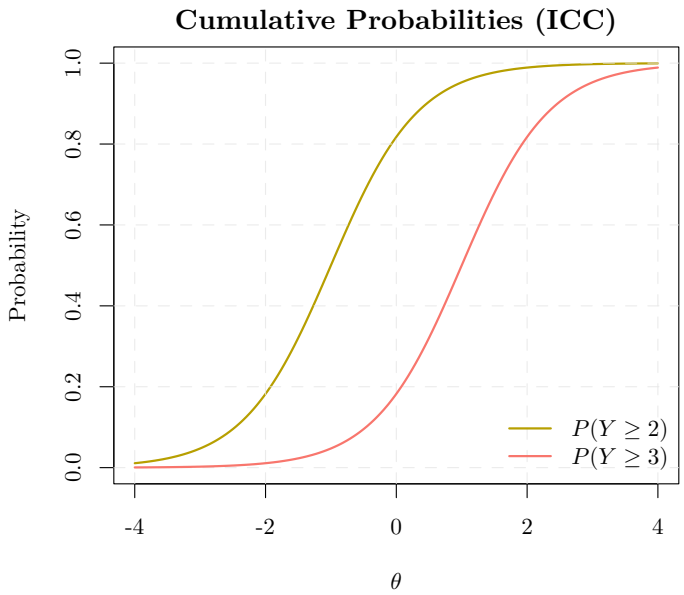


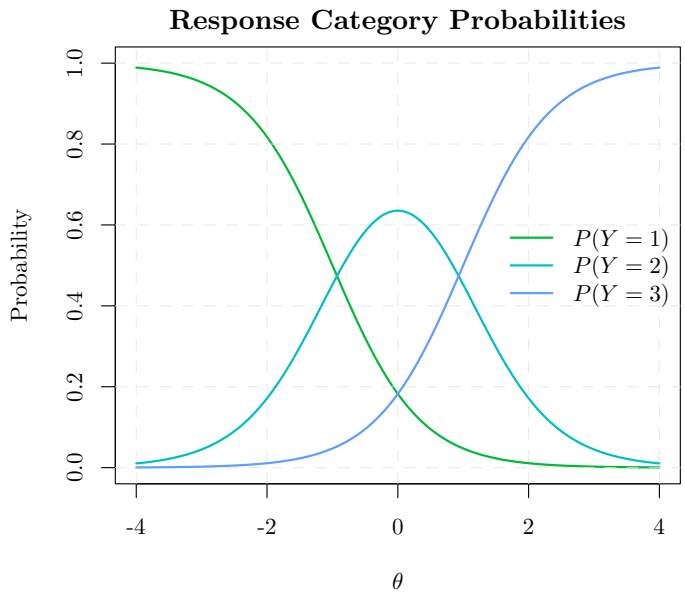
Mechanisms of the IRT model

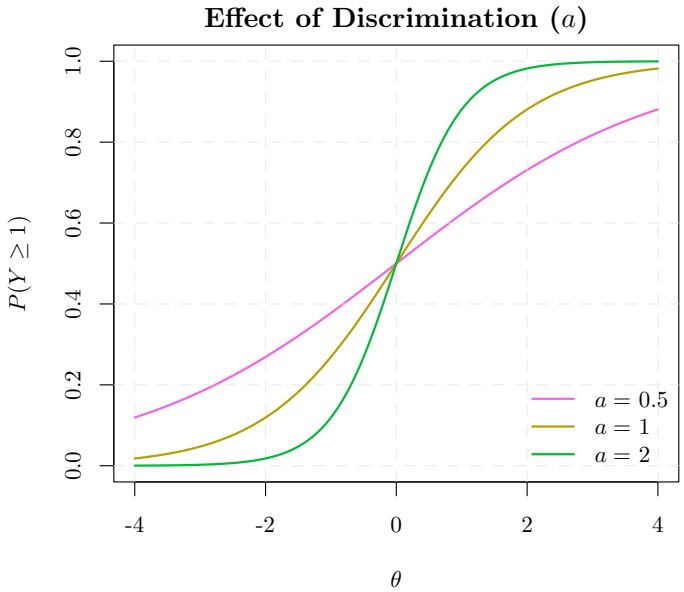
How does IRT work?

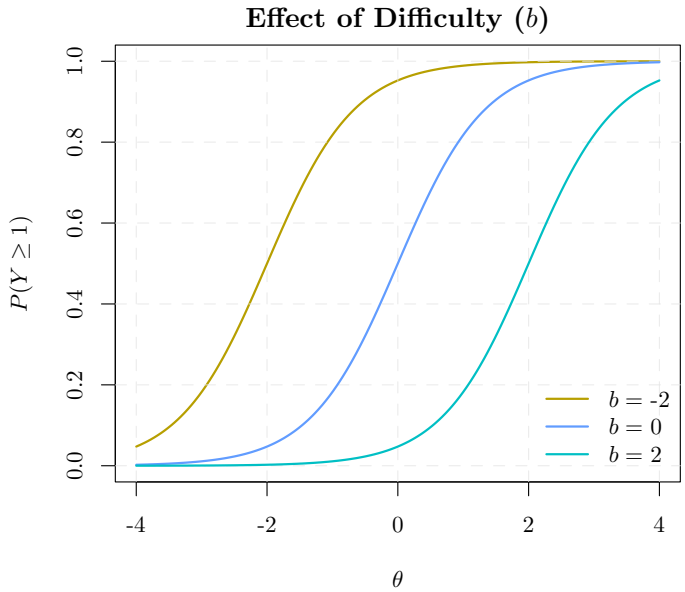
- IRT models the **probability of scoring at or above a certain level** on an item, as a function of disability.
- The model is represented by a **parametric mathematical function**, usually between 1 and 3, which defines the overall **shape** and **behavior** of the curve.
- In the context of MS, we consider the two-parameter model: a **difficulty parameter** (location on the ability scale), and a **discrimination parameter** (slope/steepness of the curve).
  - $\rightarrow$ The **higher the difficulty** for a given level, the **more disabled** one must be to score on that level.
  - $\rightarrow$ The **higher the discrimination** parameter, the steeper the curve (i.e., probabilities changing rapidly across θ), and the **easier it is to discriminate** between individuals.
- The model searches for the **best-fitting curves** for each item, those that **maximize the likelihood** of observing the response patterns in our data.

Mechanisms of the IRT model II









IRT's Key Assumptions

What does IRT need to work well?

- **Unidimensionality**: responses are driven by a single latent trait.
  - → MS disability needs to manifest **simultaneously** across all items.
- **Local independence**: item responses are independent given the latent trait.
  - → No hidden dependencies between items once disability is accounted for, i.e., two individuals with the same disability will score similarly.
- **Monotonicity**: the probability of higher item scores increases with higher latent trait values.
  - → Individuals with greater disability should be more likely to score higher on each item.
- When these assumptions hold, the model is **conceptually sound**.
  - → Any remaining variability is **statistical**, **not structural**.
  - → This contrasts with EDSS, where the **limitations are built into the scale** itself.

IRT's assumptions in the context of MS

Do these assumptions hold with FS scores?

- **Independence of FS scores**: Each FS score captures a specific functional domain (e.g., pyramidal = motor strength, cerebral = cognition, etc.), and increases in disability can be **localized to one domain** with **no parallel increase** in other FS scores:

  → A patient can have severe visual impairment but no motor issues (e.g., lesions affecting vision only), or severe bladder dysfunction but normal ambulation and cognition.

  So: **Disability increase $\neq$ uniform FS worsening**

Violation of unidimensionality?

- Unidimensional IRT assumes **all items reflect a single underlying trait**, but FS scores relect multiple traits, **possibly correlated**, but **not reducible to one**.

  → Fitting a unidimensional graded response model would likely give **subopitmal** parameter estimates.

IRT's assumptions in the context of MS: An Analogy

- Assessing overall disability in MS using FS scores is like trying to evaluate how good a student is overall by giving them a set of tests from various unrelated subjects — for example, literature, chemistry, history, geography, arts, and math tests.
- While high-performing students often do well across multiple domains, these scores primarily reflect distinct abilities. Therefore, using them to infer a single general trait introduces noise and reduces measurement precision.
- Even more caution needs to be shown when dealing with MS as disability seldom uniformly manifests across items (only the same brain region is increasingly attacked)...

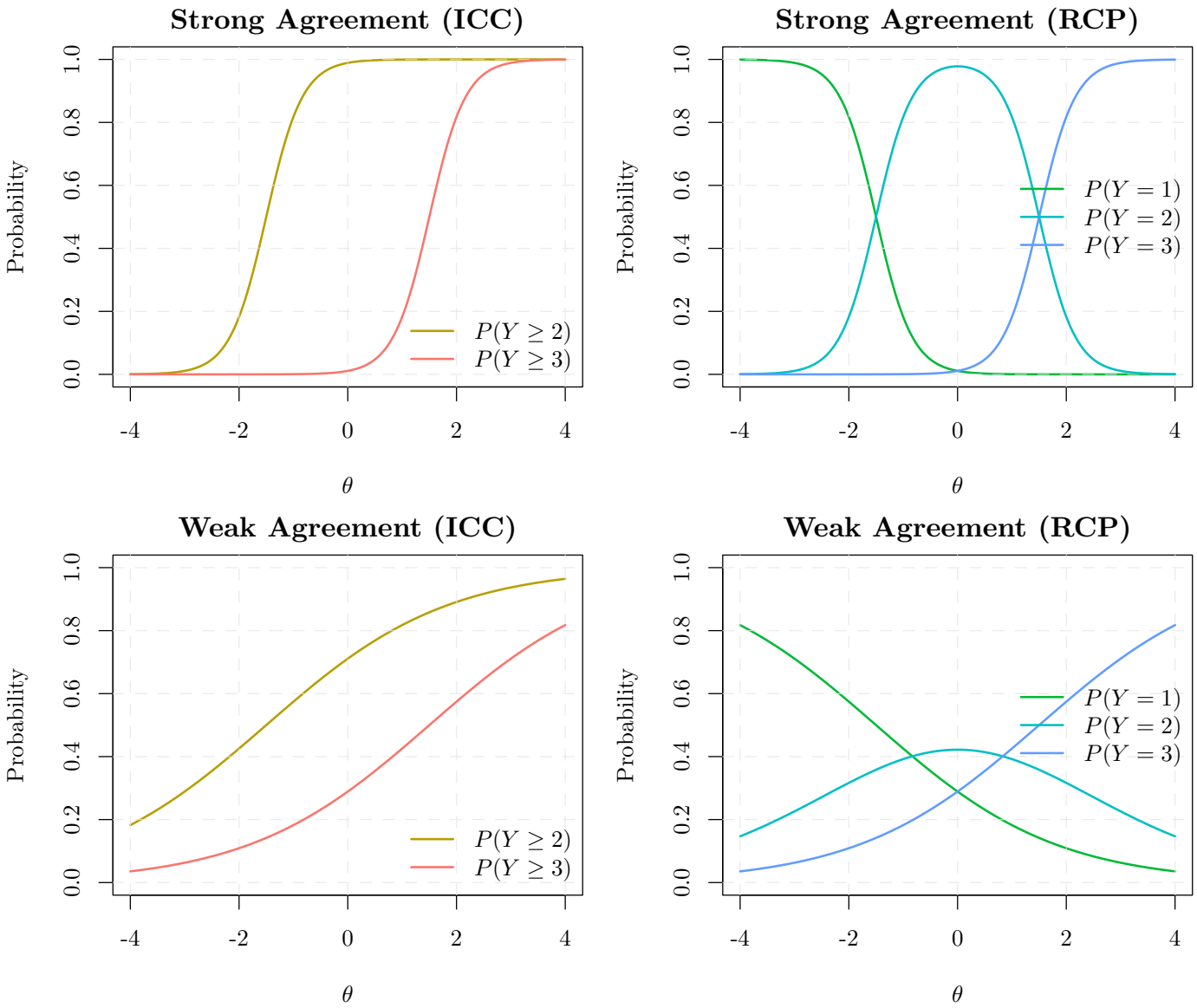
Effect on Parameter Estimation

What effect does this have on parameter estimation?

- Back to measuring a student's **general academic performance**, but we use both a **Math** and a **History** test.
- The student scores **very high in Math**, but **very low in History**.
- Math says: *"This student is strong!"* $\rightarrow$ votes for **high ability**.
  History says: *"This student is weak!"* $\rightarrow$ votes for **low ability**.
- The IRT model assumes a **single latent trait**, so forces them to **compromise on a middle ability**.
- But if both items keep **narrow, precise probability distributions for their response curves**, they will **misfit** the observed pattern:
  - $\rightarrow$ History can't explain a low score from a high(er)-ability person.
  - $\rightarrow$ Math can't explain a high score from a low(er)-ability person.
- To resolve this, both are forced to **flatten their response curves**, which reduces precision.
- This illustrates the **cost of violating unidimensionality**:

  Sacrifice precision to resolve conceptual disagreement

Illustration of Disagreement



Solutions

This actually can be not as much of a big deal...

- as long as it still **does better** than EDSS!
- But we need to trust the model under these circumstances.
  $\Rightarrow$ **Type-I error is key**.
- The problem is that the **simulations do not reflect reality**.
  On average it does, but **not for CDP**. So:

Leverage multidimensional IRT...

- to **at least simulate** the data where each FS score loads on a separate latent variable...
- but we would at a certain point **need to tune the item parameters** to make the simulated data look like the trial data.

References

Fred D Lublin, Stephen C Reingold, Jeffrey A Cohen, Gary R Cutter, et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, 83(3):278–286, 2014.

Aleksandra M Novakovic, Elke HJ Krekels, Alessandro Munafo, Sebastian Ueckert, and Mats O Karlsson. Application of item response theory to modeling of expanded disability status scale in multiple sclerosis. *The AAPS Journal*, 19(1):172–179, 2017.

Chris H Polman, Stephen C Reingold, Brenda Banwell, Michel Clanet, Jeffrey A Cohen, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria. *Annals of Neurology*, 69(2):292–302, 2011.

Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2):100–114, 1969.

Sebastian Ueckert. Modeling composite assessment data using item response theory. *CPT: Pharmacometrics & Systems Pharmacology*, 7(3):205–218, 2018.