Test length doesn't matter, it's how you use the items that counts: An intelligent procedure for item selection in Item Response Theory

Ottavia M. Epifania[1,2], Pasquale Anselmi[3], Egidio Robusto[3]

[1] Psychology and Cognitive Science Department, University of Trento, Italy
[2] Psicostat, University of Padova, Italy
[3] Department of Philosophy, Sociology, Education, and Applied Psychology, University of Padova, Italy

Convegno ASA 2024, Contributed session:
Developing, administering and refining measurement instruments in Social Sciences

Item Response Theory (IRT) for the development of Short Test Form (STF):

Typical procedure: Manually inspecting the item characteristics to recreate the desired characteristics of a test

Automated (new) procedure: A priori definition of latent trait levels of interest on which the STF should be focusing the most

Item Response Theory (IRT) for the development of Short Test Form (STF):

Typical procedure: Manually inspecting the item characteristics to recreate the desired characteristics of a test

Automated (new) procedure: A priori definition of latent trait levels of interest on which the STF should be focusing the most

Item Response Theory (IRT) for the development of Short Test Form (STF):

Typical procedure: Manually inspecting the item characteristics to recreate the desired characteristics of a test

Issue
Not an automated procedure → depends on the subjectivity of the researcher

Automated (new) procedure: A priori definition of latent trait levels of interest on which the STF should be focusing the most

Issue
Punctual definition of the specific latent trait levels of interest influences the number of selected items

Item Response Theory (IRT) for the development of Short Test Form (STF):

Typical procedure: Manually inspecting the item characteristics to recreate the desired characteristics of a test

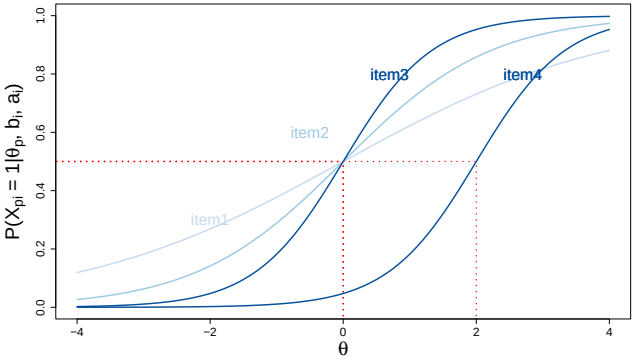
Issue
Not an automated procedure → depends on the subjectivity of the researcher

Automated (new) procedure: A priori definition of latent trait levels of interest on which the STF should be focusing the most

Issue
Punctual definition of the specific latent trait levels of interest influences the number of selected items

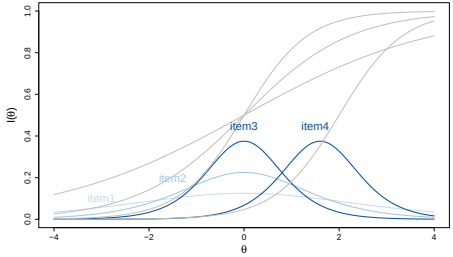
AIM
New automated procedure for item selection in IRT that only requires the definition of the desired characteristics of a test

$$P(x_{pi} = 1|\theta_p, b_i, a_i) = \frac{\exp[a_i(\theta_p - b_i)]}{1+\exp[a_i(\theta_p - b_i)]}$$

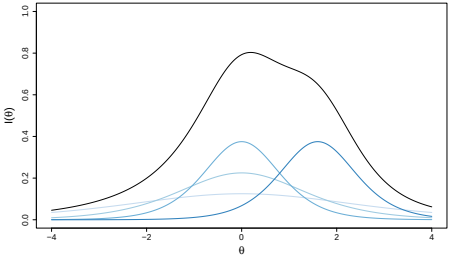


θ_p: Latent trait level of person p
b_i: Location of item i on θ
a_i: Discrimination ability of item i

Item Information Function (IIF):
$I_i(\theta) = a_i^2 P_i(\theta, b_i, a_i)[1 - P_i(\theta, b_i, a_i)]$



Test Information Function (TIF):
$I(\theta) = \sum_{i=1}^{N} I_i(\theta)$



Set up:

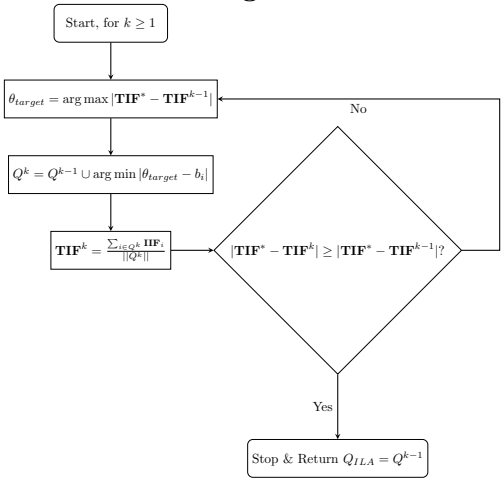
N: number of items included in the item bank

Q^k: Set of item indexes selected for inclusion in the STF up to iteration k ($Q^0 = \emptyset$)

$\mathbf{TIF}^*$: TIF target

$\mathbf{TIF}^0 = (0, 0, \ldots, 0)$

ILA Algorithm:



For each $Q_m \subset Q$ with $Q_m \neq \emptyset$, calculate:

1. $\mathbf{TIF}^{Q_m} = \dfrac{\sum_{i \in Q_m} IIF_i}{||Q_m||}$
2. $\overline{\Delta}_{\mathbf{TIF}^{Q_m}} = mean(|\mathbf{TIF}^{*} - \mathbf{TIF}^{Q_m}|)$

$Q_{BFP} = \arg\min_{\emptyset \neq Q_m \subset Q} \overline{\Delta}_{\mathbf{TIF}^{Q_m}}$

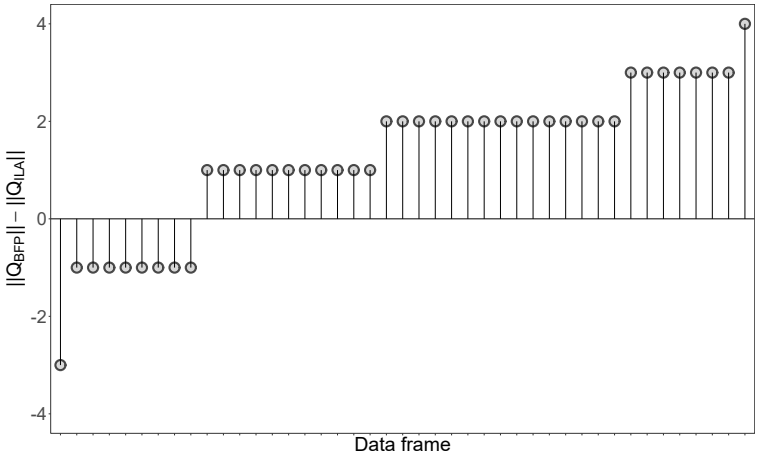
100 data frames:

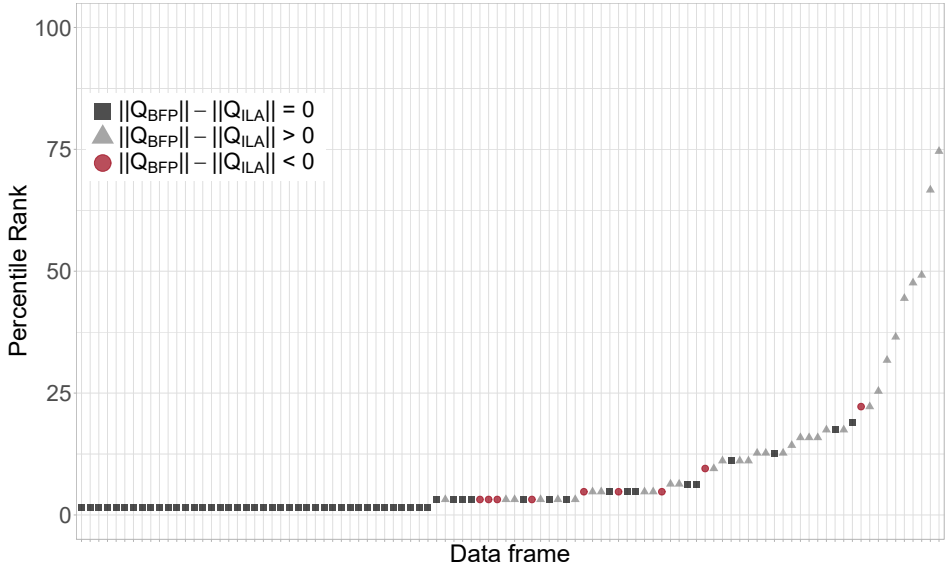
1. Generate an item bank B of $N = 6$ items:
   - Difficulty parameters: $\mathcal{U}(-3, 3)$
   - Discrimination parameters: $\mathcal{U}(.90, 2.0)$
2. Random item selections of lengths l from B ($M_l = 3.34 \pm 1.13$) + modification parameters $\mathcal{U}(-0.20, 0.20) \rightarrow \mathbf{TIF}^*$
3. Considering $\mathbf{TIF}^*$ at Step 2 and item parameters at Step 1:
   - ILA $\rightarrow$ *Forwardly searches*
   - BFP $\rightarrow$ *Systematically tests*

Comparison:

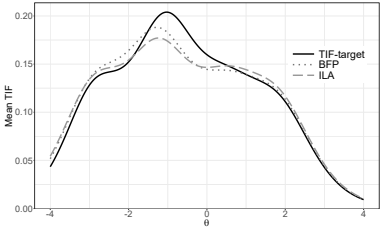
- $||Q_{\text{BFP}}|| - ||Q_{\text{ILA}}||$
- Percentile rank of the distance $\mathbf{TIF}_{\text{BFP}} - \mathbf{TIF}_{\text{ILA}}$

$||Q_{\text{BFP}}|| - ||Q_{\text{ILA}}|| = 0$ in 57% of cases

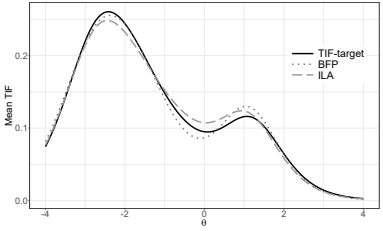



100
75
50
25
0
Percentile Rank
Data frame
||Q_BFP|| − ||Q_ILA|| = 0
||Q_BFP|| − ||Q_ILA|| > 0
||Q_BFP|| − ||Q_ILA|| < 0

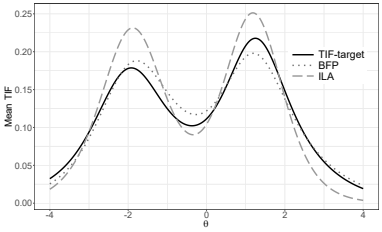
$||Q_{\text{BFP}}|| = ||Q_{\text{ILA}}||$, $Q_{\text{BFP}} \neq Q_{\text{ILA}}$, $RP = 3.17$



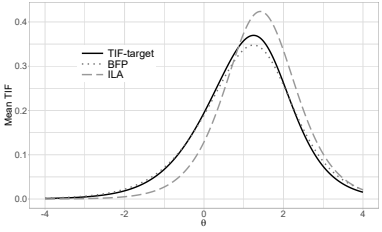
$||Q_{\text{BFP}}|| < ||Q_{\text{ILA}}||$, $RP = 3.17$



$||Q_{\text{BFP}}|| > ||Q_{\text{ILA}}||$, $RP = 4.76$



$||Q_{\text{BFP}}|| > ||Q_{\text{ILA}}||$, $RP = 12.70$



Pros of ILA

- It selects items that are able to recreate the desired characteristics of a test (usually)
- It is computationally "Light"

Cons of ILA

- It grounds its selection on a single θ_{target} at a time $\rightarrow$ it might select items minimizing the distance on that target but that are not very useful for the test
- It only forwardly searches an item $\rightarrow$ once it is in, it can't get out
- It does not account for the discrimination parameters of the items