Adaptive Testing with Complex Specifications

Seung W. Choi & Sangdon Lim
University of Texas at Austin

Symposium: Advanced Technologies for Adaptive Testing, January 29, 2021

Introduction

- Most CAT applications in educational testing are highly constrained due in part to the use of complex test specifications for content coverage and the need for exposure control for test security.

- CAT with item sets, such as passage-based reading comprehension tests, can add another layer of complexity due to constraints imposed on both passages and items.

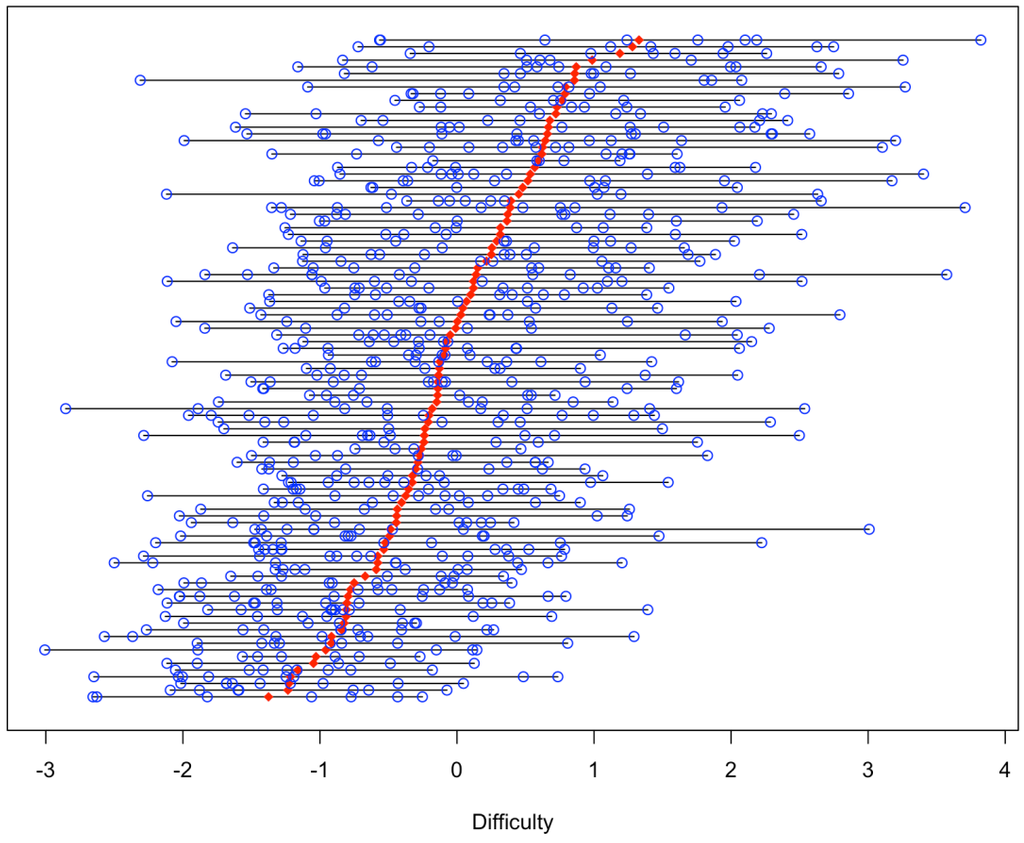
- Overexposing passages, compared to individual items, can have more adverse effects on validity and test security.

Introduction

- Items within passages can vary widely in difficulty to the extent that the variation within passages approaches that for the entire test.

- Common item development strategies may produce alternative items or enemy items that should not be administered together.

- These may call for within-passage adaptation in real time in lieu of pre-packaging multiple versions of item sets associated with a given passage.

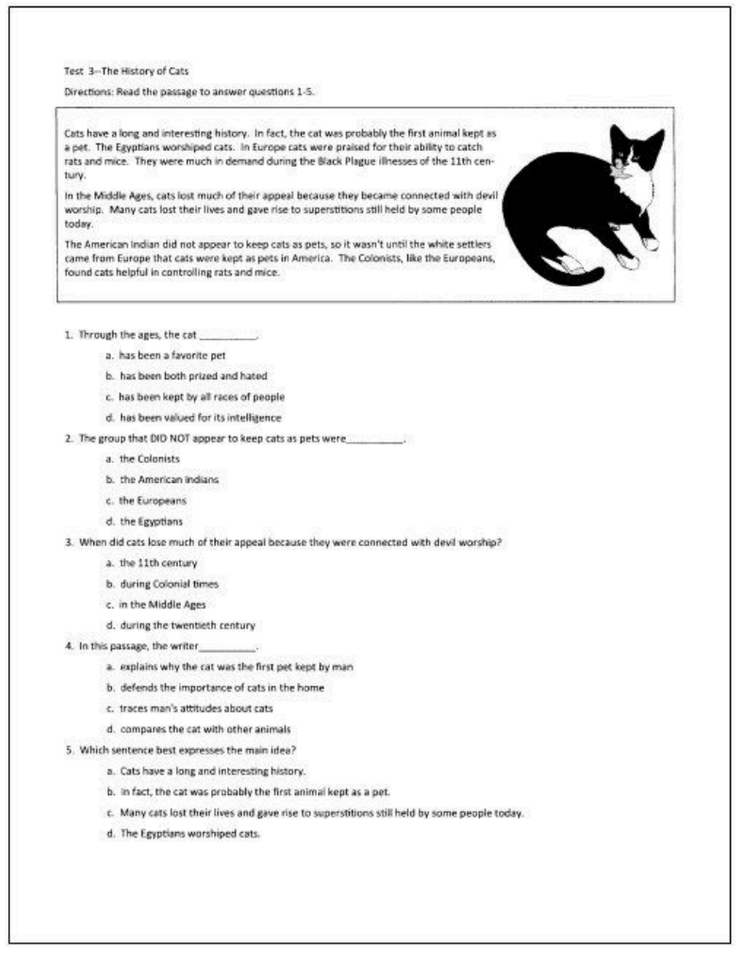


Introduction

- Conversely, some tests may require a provision for previewing (and reviewing) all items within passages.

- Typically in fixed tests, the presentation of items within passages follows a certain logical sequence (e.g., easy to hard).

- These constraints, however, may require passage-level adaptation.



Test 3--The History of Cats

Directions: Read the passage to answer questions 1-5.

Cats have a long and interesting history. In fact, the cat was probably the first animal kept as a pet. The Egyptians worshiped cats. In Europe cats were praised for their ability to catch rats and mice. They were much in demand during the Black Plague illnesses of the 11th century.

In the Middle Ages, cats lost much of their appeal because they became connected with devil worship. Many cats lost their lives and gave rise to superstitions still held by some people today.

The American Indian did not appear to keep cats as pets, so it wasn't until the white settlers came from Europe that cats were kept as pets in America. The Colonists, like the Europeans, found cats helpful in controlling rats and mice.

1. Through the ages, the cat _____.
   a. has been a favorite pet
   b. has been both prized and hated
   c. has been kept by all races of people
   d. has been valued for its intelligence
2. The group that DID NOT appear to keep cats as pets were_____.
   a. the Colonists
   b. the American Indians
   c. the Europeans
   d. the Egyptians
3. When did cats lose much of their appeal because they were connected with devil worship?
   a. the 11th century
   b. during Colonial times
   c. in the Middle Ages
   d. during the twentieth century
4. In this passage, the writer_____.
   a. explains why the cat was the first pet kept by man
   b. defends the importance of cats in the home
   c. traces man's attitudes about cats
   d. compares the cat with other animals
5. Which sentence best expresses the main idea?
   a. Cats have a long and interesting history.
   b. In fact, the cat was probably the first animal kept as a pet.
   c. Many cats lost their lives and gave rise to superstitions still held by some people today.
   d. The Egyptians worshiped cats.

Introduction

- The problem can become further complicated when the test consists of both discrete and set-based items, albeit this type of configuration is common in educational testing.

- Test specifications in that scenario should stipulate the desired attributes of the test and the adaptivity at all levels: stimuli, item sets, and discrete items.

- The CAT algorithm then selects an optimal set of passages and associated items subject to the test specifications and the desired adaptivity.

Introduction

- In this presentation, we will demonstrate how the universal shadow-test assembler framework (van der Linden & Diao, 2014) implemented in TestDesign* (Choi, Lim, & van der Linden, in preparation) can effectively handle complex test specifications in the context of a reading assessment comprised of both item sets and discrete items.

*https://cran.r-project.org/web/packages/TestDesign/index.html

Exposure Control

- The maximum-information item-selection criterion, which increases test efficiency in CAT, can also cause overexposure of a small proportion of items while underexposing the rest.
- The shadow-test approach addresses the problem by adding random item-eligibility constraints to the test-assembly model so that items with higher exposure rates have higher probabilities of being temporarily ineligible for the examinees.

Exposure Control

- Eligibility $P(E_i)$ is controlled by continuously monitoring exposure rates $P(A_i)$ against a target r_{max} as follows:

$$P(E_i) \leftarrow \min\left[\frac{r_{\max}}{P(A_i)}P(E_i), \quad 1\right]$$

- Overexposed items when tagged “ineligible” can be temporarily removed by imposing constraints or deterred by modifying the objective function (van der Linden & Choi, 2019).

Exposure Control

- Eligibility can be controlled for all test takers considered together or in separate segments conditioning on ability (θ):

$$P(E_i \mid \theta) \leftarrow \min\left[\frac{r_{\max}}{P(A_i \mid \theta)} P(E_i \mid \theta), \quad 1\right]$$

- See van der Linden & Choi (2019) for recent advances in the methodology.

Maximum Information Selection

- The most common item selection method in CAT is the maximum information criterion:

$$\arg\max_{i \in V} I_i(\theta)$$

- As the maximum-information item-selection criterion is essentially greedy, a relatively small portion of items can be heavily exposed while the rest of the items remain inactive.

- A more controlled (and less greedy) selection algorithm might help alleviate the overexposure issue.

Target Information Selection

- That is, instead of maximizing information, the target information selection method aims to hit a predetermined target using a goal programming approach (van der Linden, 2005).

$$\begin{aligned} &\text{minimize } y \\ &\text{subject to} \\ &\sum_i I_i(\hat{\theta}) x_i \leq T + y \\ &\sum_i I_i(\hat{\theta}) x_i \geq T - y \\ &y \geq 0 \end{aligned}$$

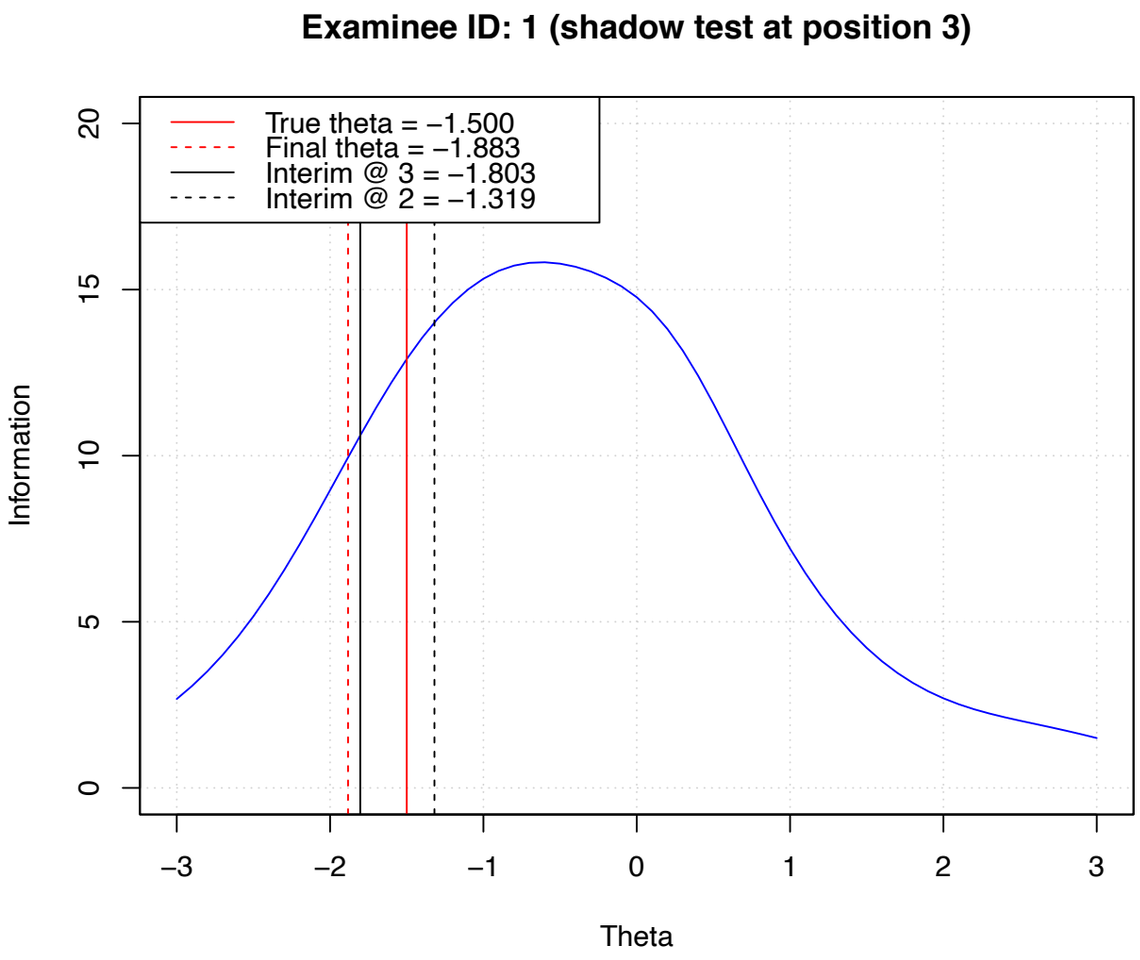
- It aims to minimize the deviance y between the current and a target information level T.
- This selection method has not been used in adaptive test assembly.

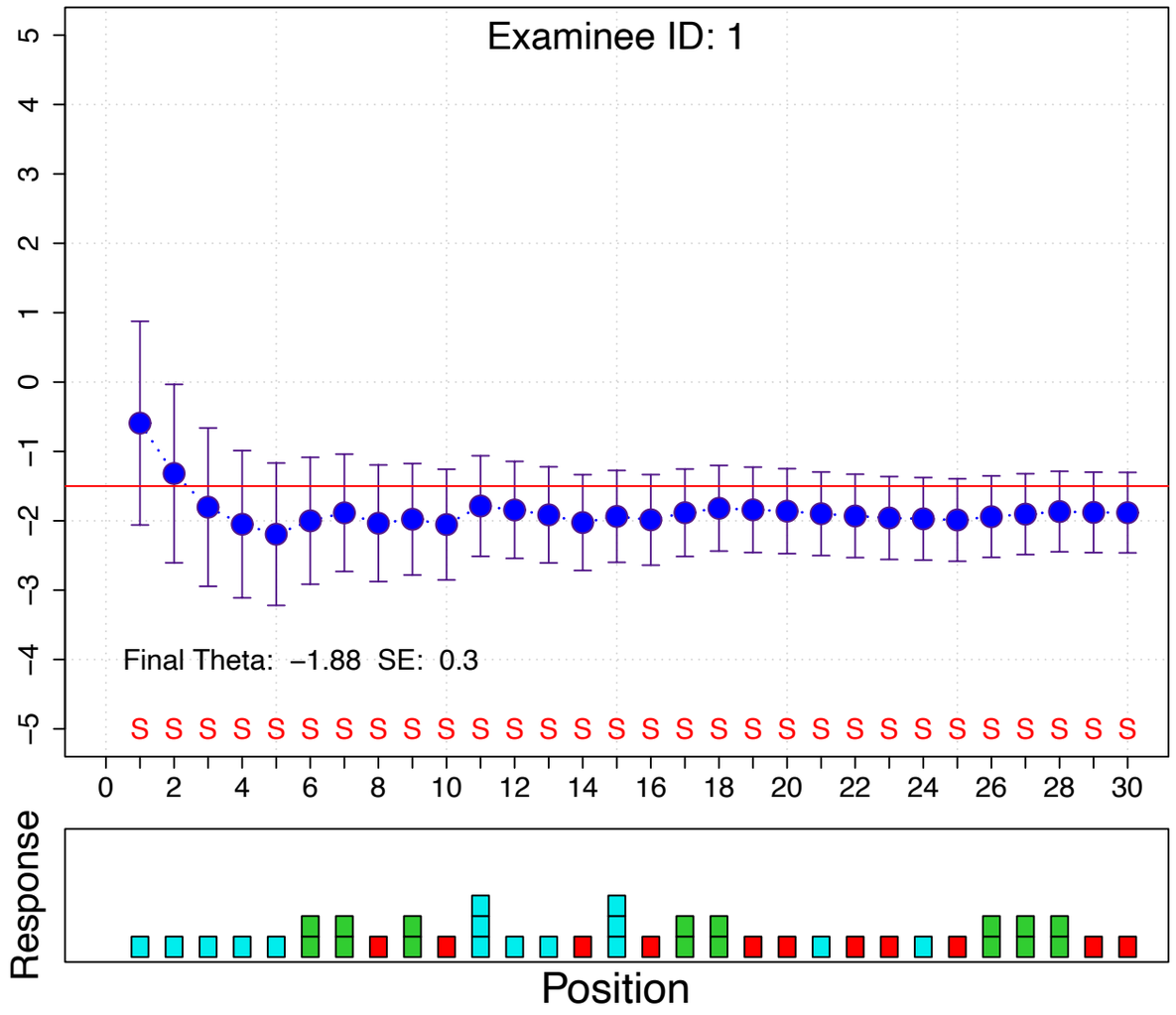
Adaptive Assembly using Target Information

- Traditional shadow test assembly
    - attempts to maximize information at the outset
    - uses too much information in early stages
- Target information selection
    - attempts to hit a (more modest) target
    - may preserve highly informative items until needed

Adaptive Assembly using Target Information

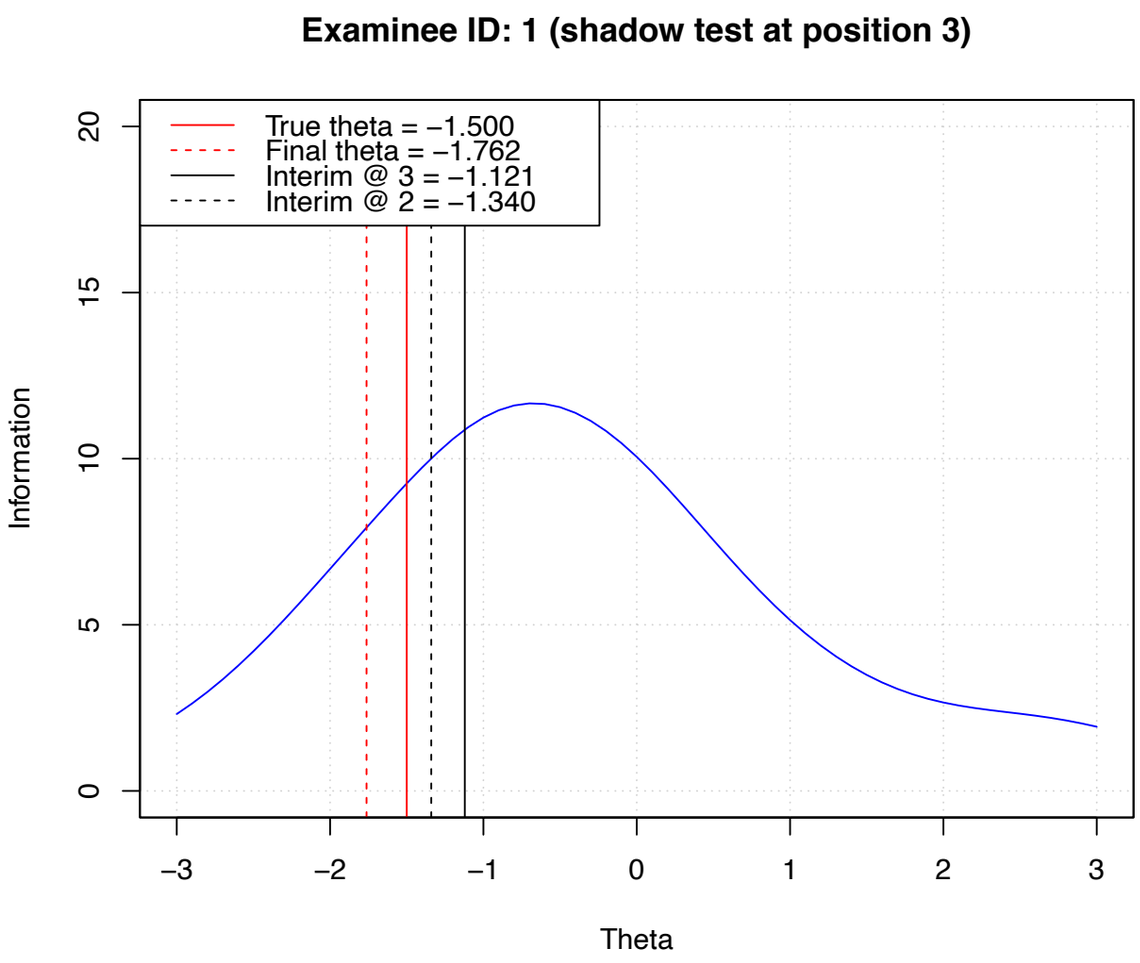
Maximum Information Selection

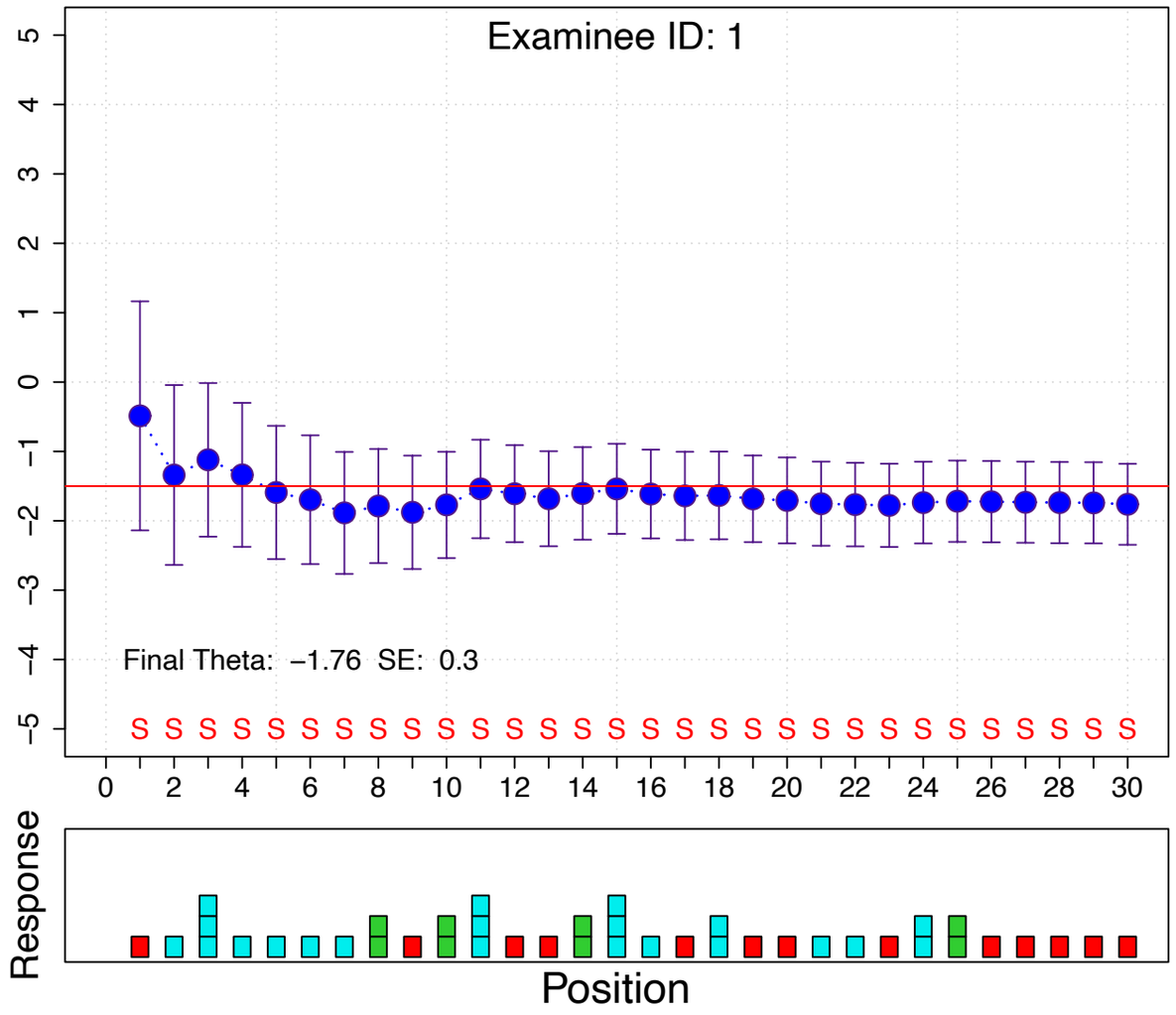




Adaptive Assembly using Target Information

Target Information Selection

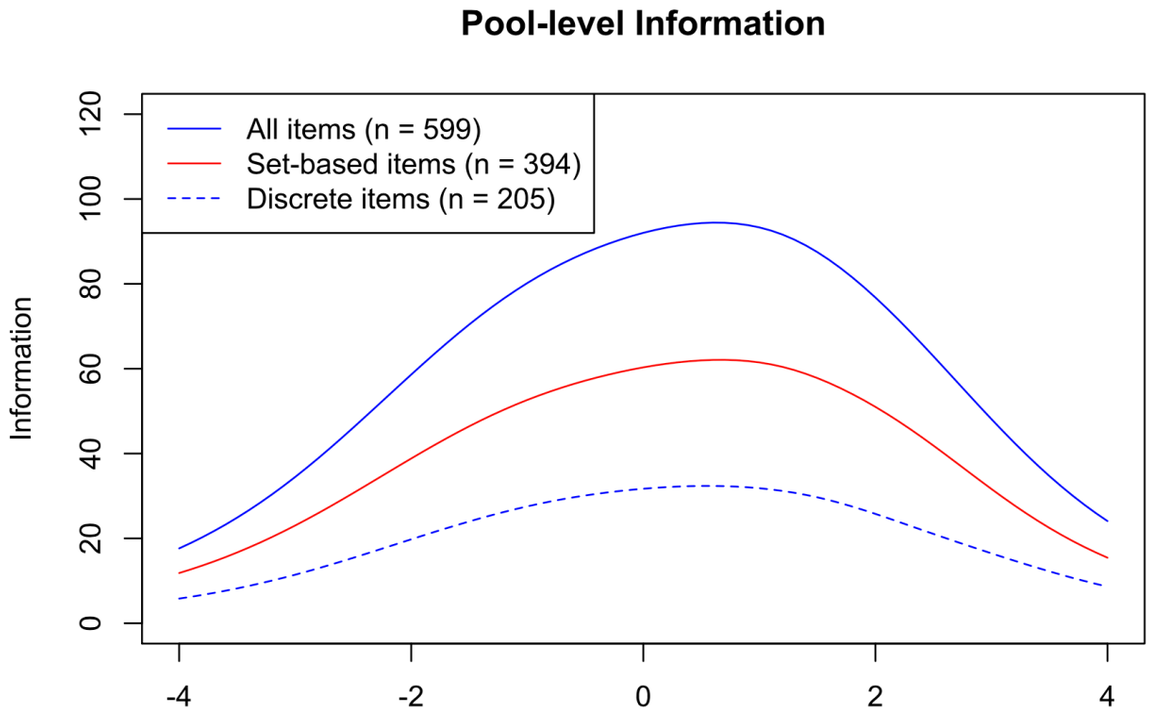




Illustration

Item Pool

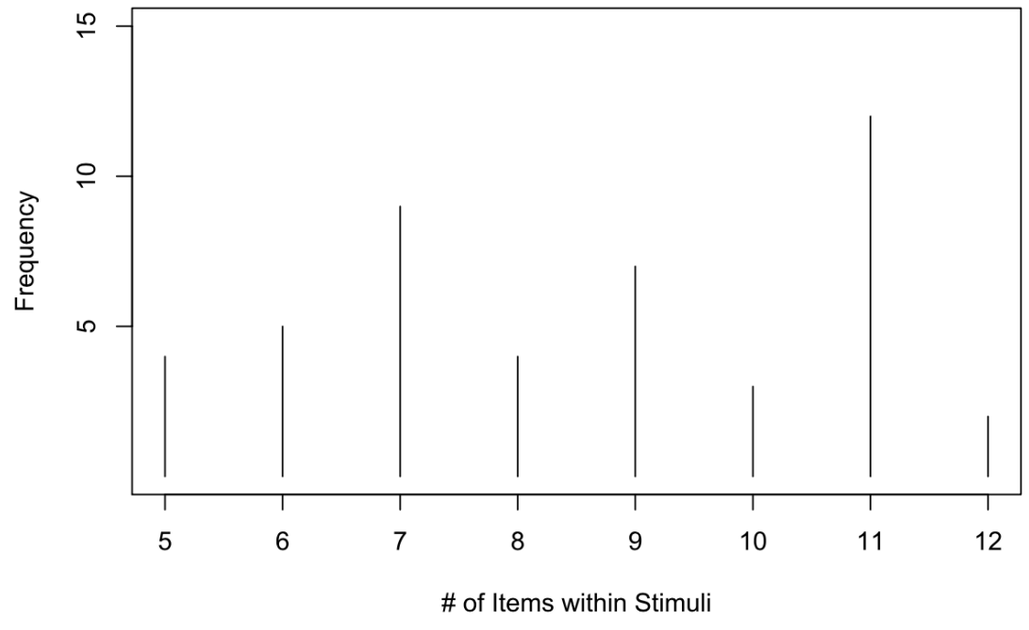
- 599 items total
  - 394 set-based items
  - 205 discrete items
  - 46 passages

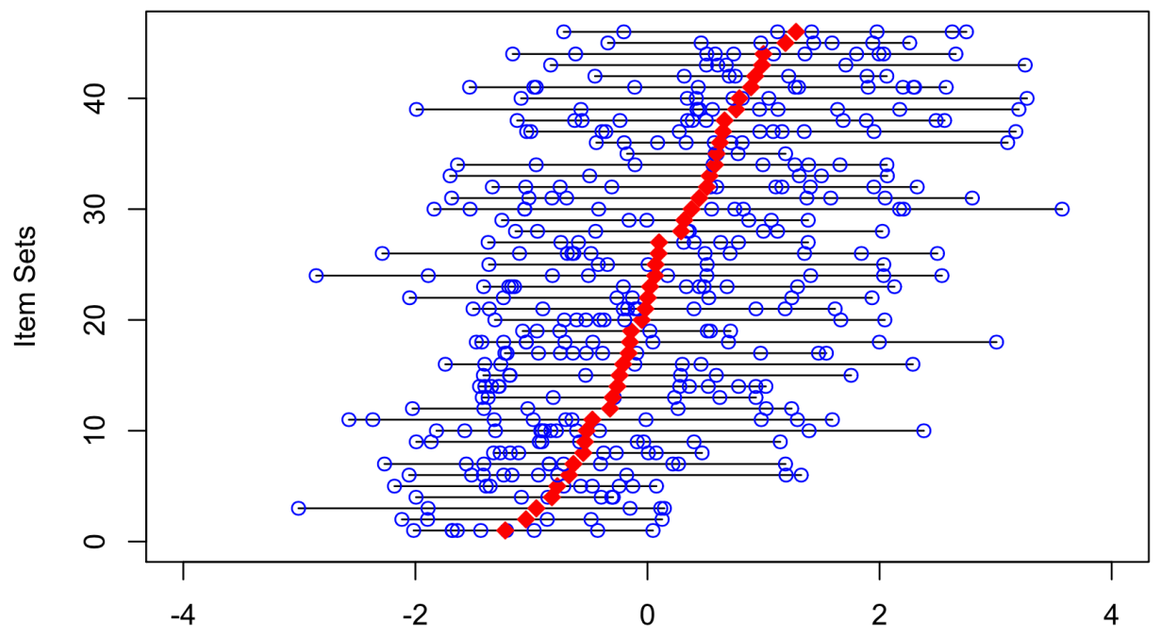


Illustration

Item Pool

- 599 items total
  - 394 set-based items
  - 205 discrete items
  - 46 passages





Illustration

Test Specification

- 30-item test
- 20 set-based items
- 10 discrete items
- 4 passages
- 5 items per passage

CONSTRAINT_ID	TYPE	WHAT	CONDITION	LB	UB	ONOFF	COUNT	ST_COUNT
C1	NUMBER	ITEM		30	30		599	NA
C2	NUMBER	ITEM	IS_DISC == FALSE	20	20		394	NA
C3	NUMBER	ITEM	IS_DISC == TRUE	10	10		205	NA
C4	NUMBER	STIMULUS		4	4		NA	46
C5	NUMBER	ITEM	Per Stimulus	5	5		NA	NA
C6	NUMBER	STIMULUS	CONTENTCAT == 1	2	2		NA	25
C7	NUMBER	STIMULUS	CONTENTCAT == 2	2	2		NA	21
C8	NUMBER	ITEM	IS_DISC == FALSE & CONTENTCAT == 1	10	10		203	NA
C9	NUMBER	ITEM	IS_DISC == FALSE & CONTENTCAT == 2	10	10		191	NA
C10	NUMBER	ITEM	IS_DISC == TRUE & CONTENTCAT == 1	5	5		78	NA
C11	NUMBER	ITEM	IS_DISC == TRUE & CONTENTCAT == 2	5	5		127	NA
C12	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 == 2 & ITEMTYPE == 'MC'	1	2		23	NA
C13	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 == 4 & ITEMTYPE == 'MC'	1	2		22	NA
C14	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 == 2 & ITEMTYPE == 'MC'	1	1		11	NA
C15	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 == 4 & ITEMTYPE == 'MC'	1	1		9	NA
C16	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 %in% c(1, 3, 5, 6, 7) & ITEMTYPE == 'MC'	3	6		132	NA
C17	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 %in% c(1, 3, 5, 6, 7) & ITEMTYPE == 'MC'	2	3		50	NA
C18	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 %in% c(2, 4) & ITEMTYPE == 'CR'	2	2		26	NA
C19	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 %in% c(2, 4) & ITEMTYPE == 'CR'	1	1		8	NA
C20	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 == 9 & ITEMTYPE == 'MC'	1	2		25	NA
C21	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 == 11 & ITEMTYPE == 'MC'	1	2		26	NA
C22	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 == 9 & ITEMTYPE == 'MC'	1	1		15	NA
C23	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 == 11 & ITEMTYPE == 'MC'	1	1		12	NA
C24	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 %in% c(8, 10, 12, 13, 14) & ITEMTYPE == 'MC'	3	6		124	NA
C25	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 %in% c(8, 10, 12, 13, 14) & ITEMTYPE == 'MC'	2	3		85	NA
C26	NUMBER	ITEM	IS_DISC == FALSE & TESTINGLEVEL2 %in% c(9, 11) & ITEMTYPE == 'CR'	2	2		16	NA
C27	NUMBER	ITEM	IS_DISC == TRUE & TESTINGLEVEL2 %in% c(9, 11) & ITEMTYPE == 'CR'	1	1		15	NA
C28	NUMBER	ITEM	IS_DISC == FALSE & DOKLEVELCODE == 2	10	20		172	NA
C29	NUMBER	ITEM	IS_DISC == TRUE & DOKLEVELCODE == 2	5	10		90	NA

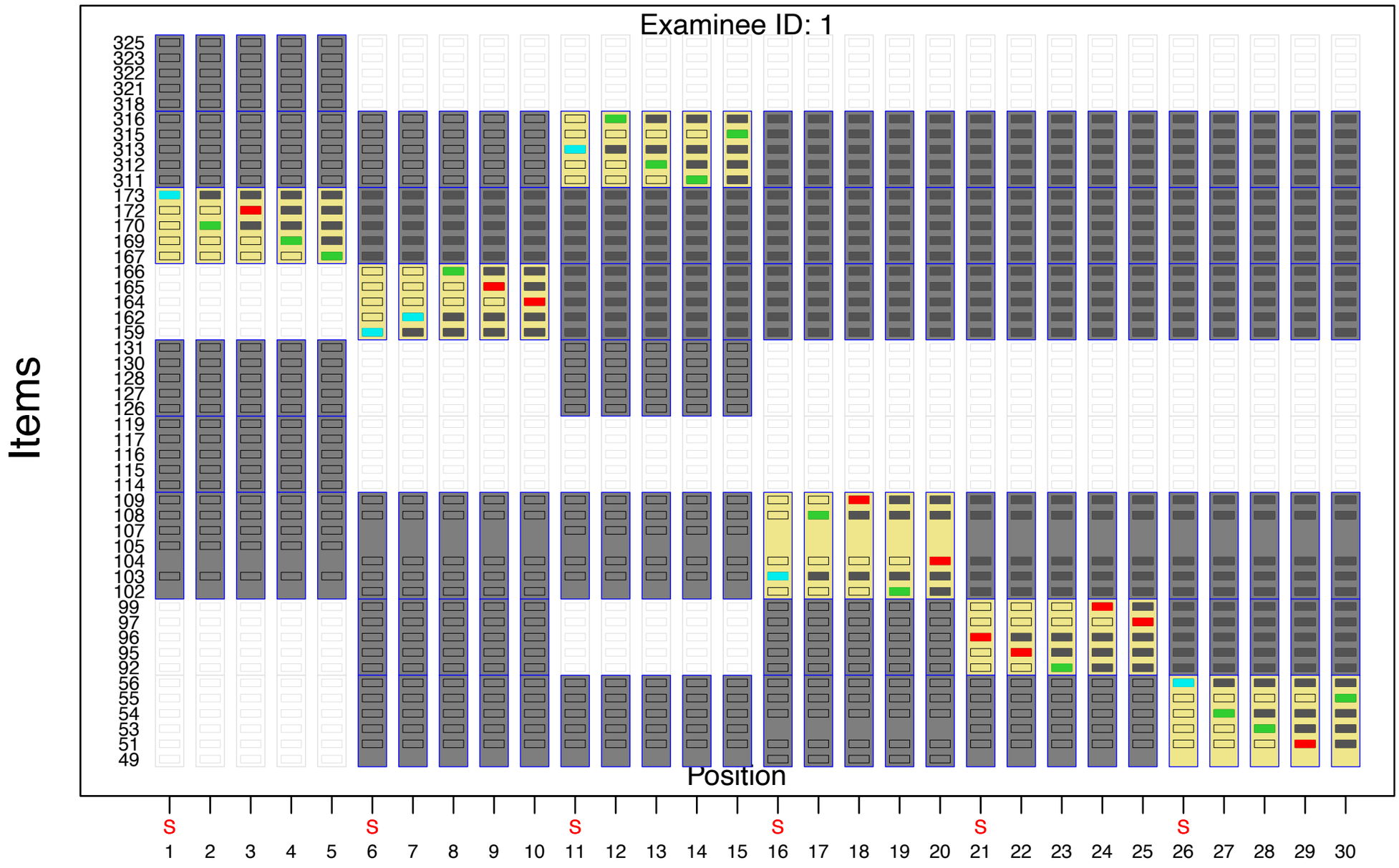
Illustration

Simulation Design

- Placement of set-based and discrete items:
    - Fully set-based without discrete items (6 passages x 5 items)
    - Set-based first, followed by discrete
    - Discrete first, followed by set-based
    - Freely interspersed (no constraints on the placement)
- Exposure control:
    - Eligibility control (via Big-M) with $r_{max} = 0.25$
    - No exposure control
- Item selection criterion:
    - Maximum information selection
    - Target information selection ($T = 8$)

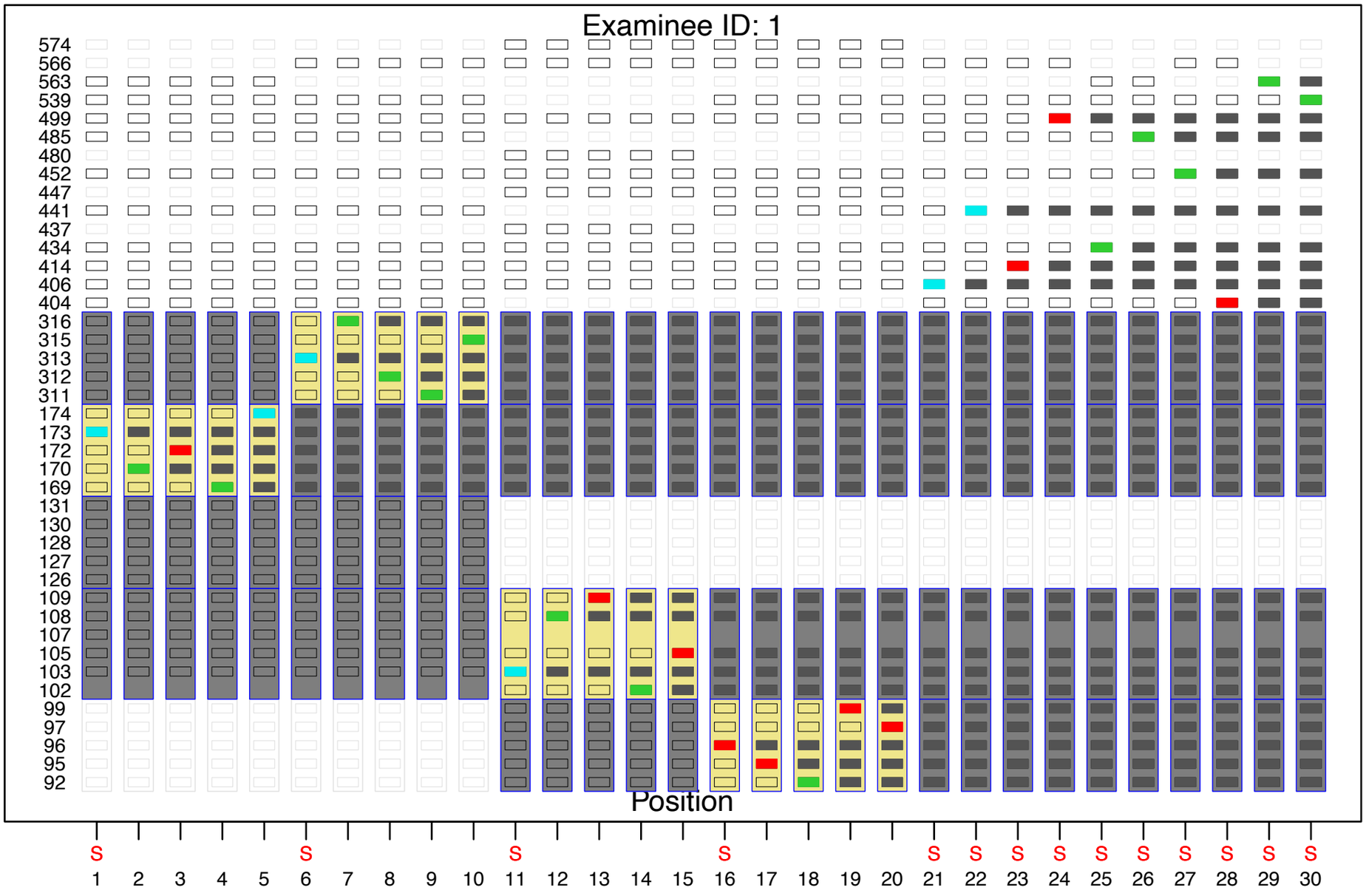
Illustration

Fully set-based



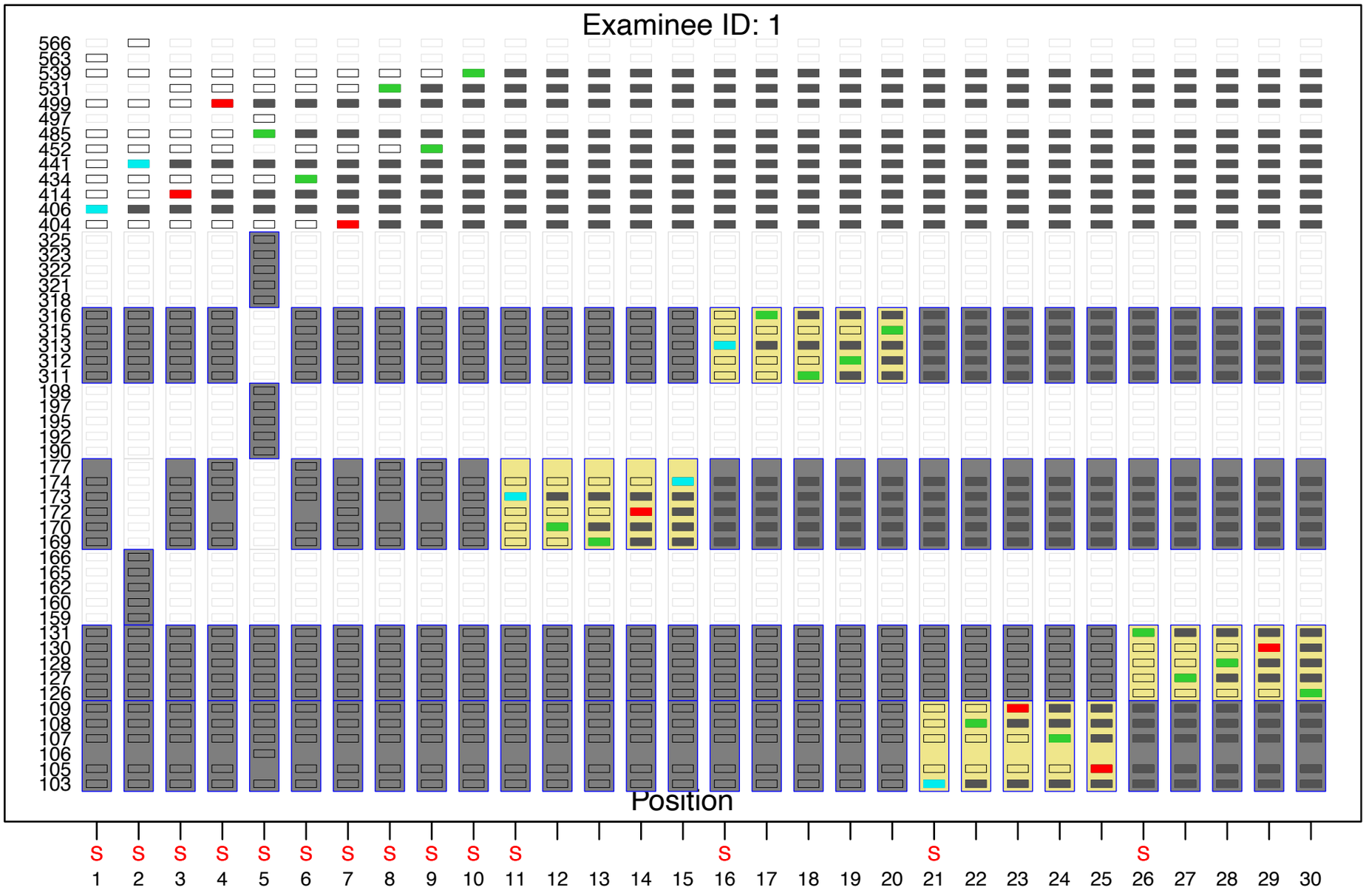
Illustration

Set-based items first



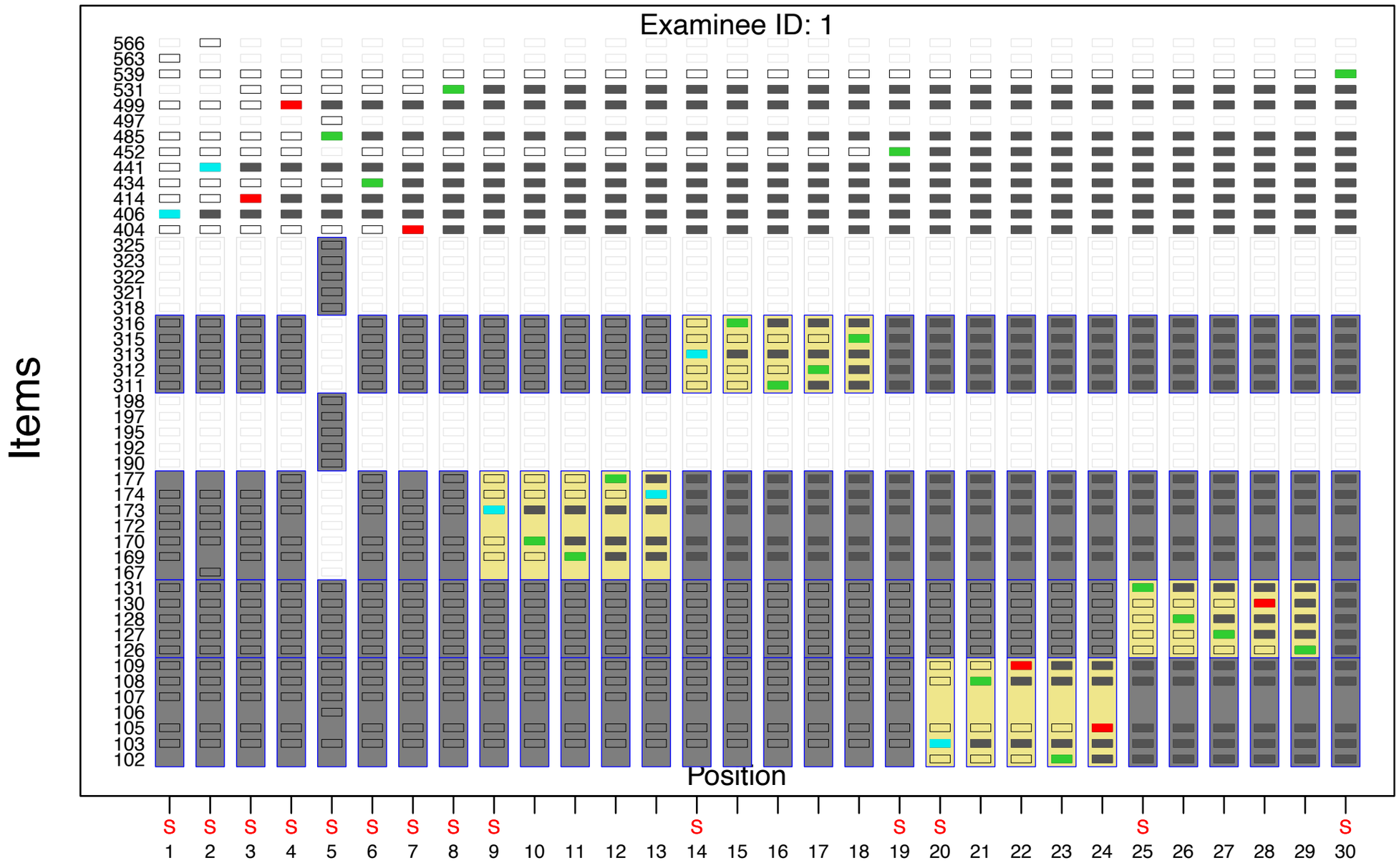
Illustration

Discrete items first



Illustration

Interleaved



Illustration

Study 1

- Unconditional exposure control
  - number of θ segment set to 1
- True θs ($n = 1000$) drawn from $N(0,1)$
- Evaluation criteria:
  - RMSE of θ
  - Exposure rates

Illustration

Study 2

- Conditional exposure control:
  - 5 θ segments: [-Inf, -1.5), [-1.5, 0.5), [-0.5, 0.5), …, [1.5, Inf]
- 1000 θs each at $-2(1)2$ for a total $N = 5000$
- Evaluation criteria:
  - RMSE of θ
  - Exposure rates

Results

Study 1

- No meaningful difference between item/set order conditions
- Slightly worse RMSE for exposure control
- Slightly worse RMSE for target information
- Slightly better exposure control for target information

Results

Study 1

Weave Order	Exposure Control	Information Type	Correlation	RMSE
D > S	None	Maximize	0.955	0.308
S > D			0.954	0.311
Interspersed			0.954	0.311
Set-based			0.945	0.339
D > S	Big-M	Maximize	0.947	0.332
S > D			0.946	0.335
Interspersed			0.945	0.337
Set-based			0.939	0.357
D > S	None	Target = 8	0.947	0.332
S > D			0.946	0.336
Interspersed			0.944	0.341
Set-based			0.945	0.340
D > S	Big-M	Target = 8	0.942	0.348
S > D			0.939	0.356
Interspersed			0.943	0.345
Set-based			0.938	0.361

Results

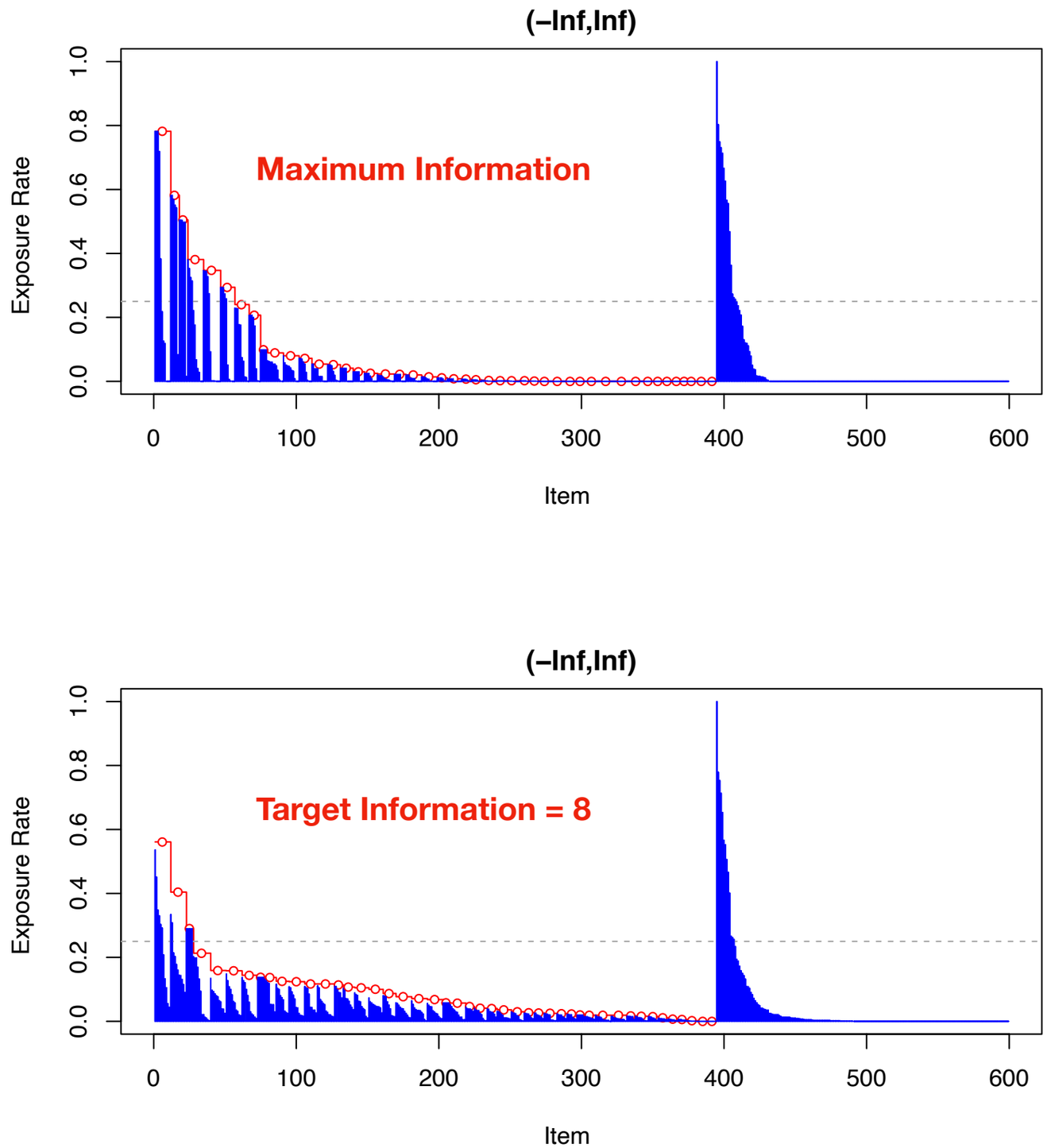
Study 1

Weave Order	Exposure Control	Information Type	Correlation	RMSE
D > S	None	Maximize	0.955	0.308
S > D			0.954	0.311
Interspersed			0.954	0.311
Set-based			0.945	0.339
D > S	Big-M	Maximize	0.947	0.332
S > D			0.946	0.335
Interspersed			0.945	0.337
Set-based			0.939	0.357
D > S	None	Target = 6	0.929	0.383
S > D			0.933	0.373
Interspersed			0.930	0.379
Set-based			0.936	0.364
D > S	Big-M	Target = 6	0.930	0.380
S > D			0.927	0.388
Interspersed			0.924	0.395
Set-based			0.935	0.368

Results

Study 1

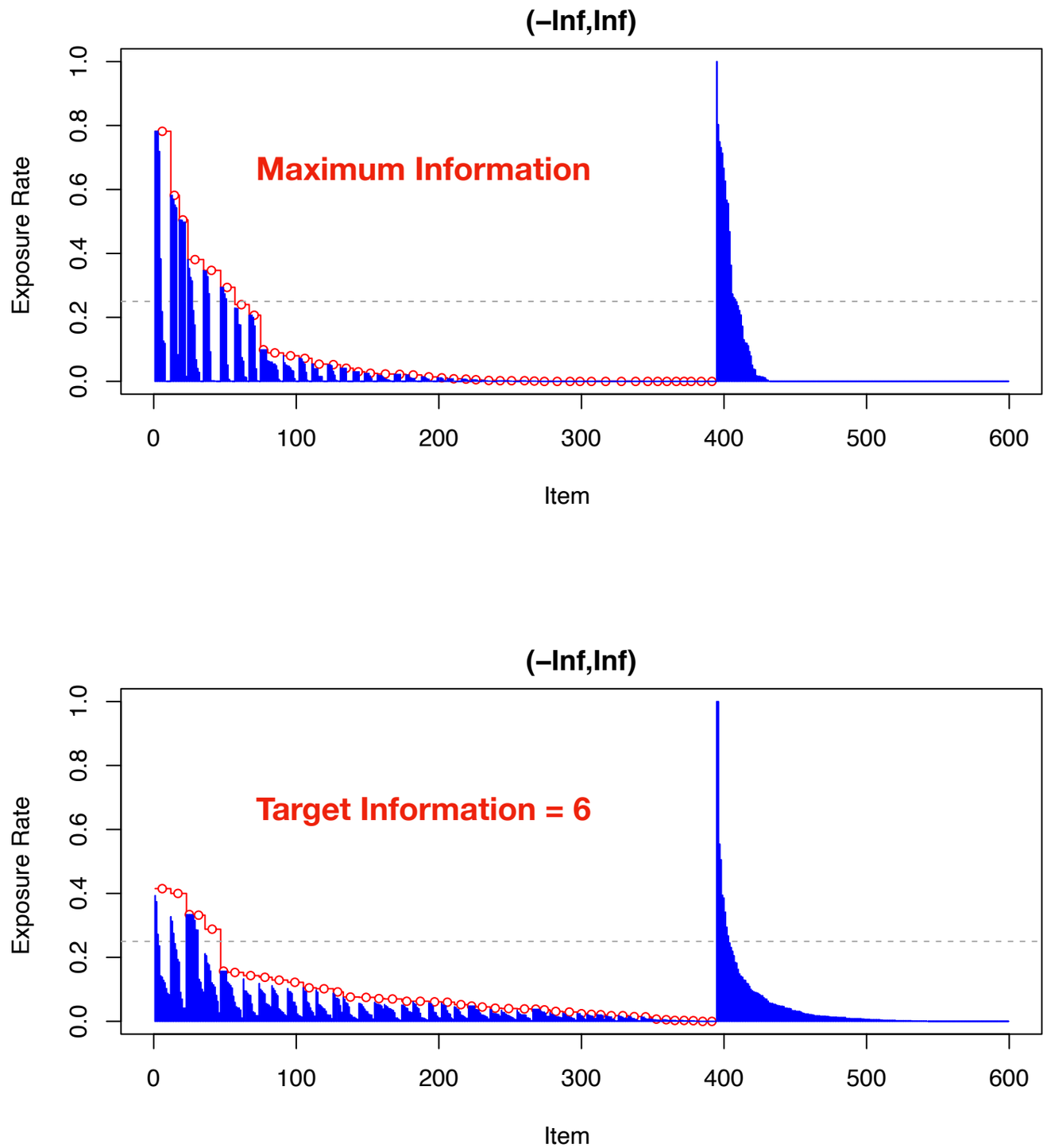
- Discrete -> Set
- No exposure control



Results

Study 1

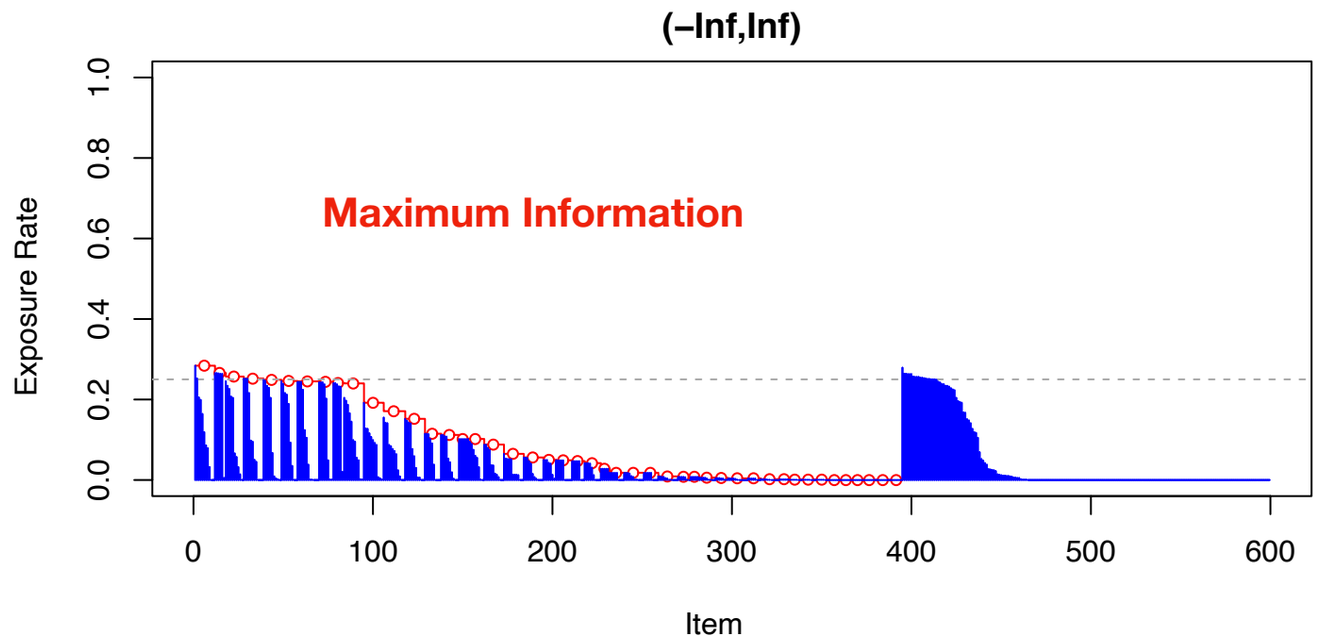
- Discrete -> Set
- No exposure control

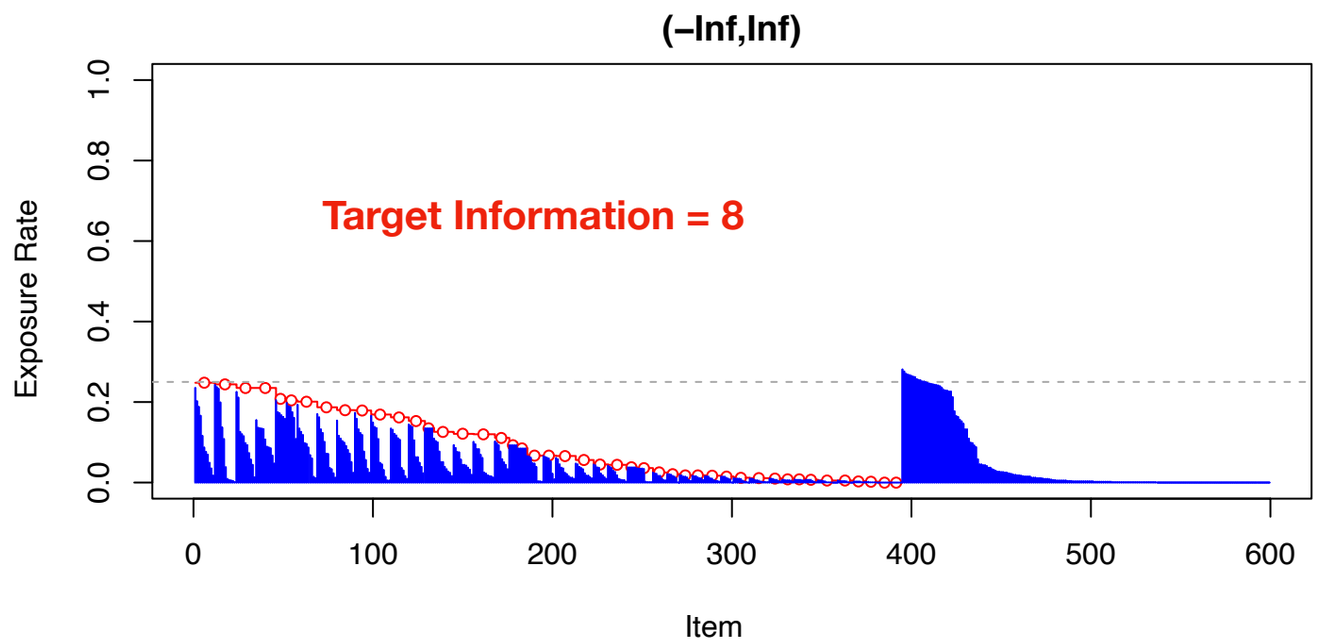


Results

Study 1

- Discrete -> Set
- Eligibility control

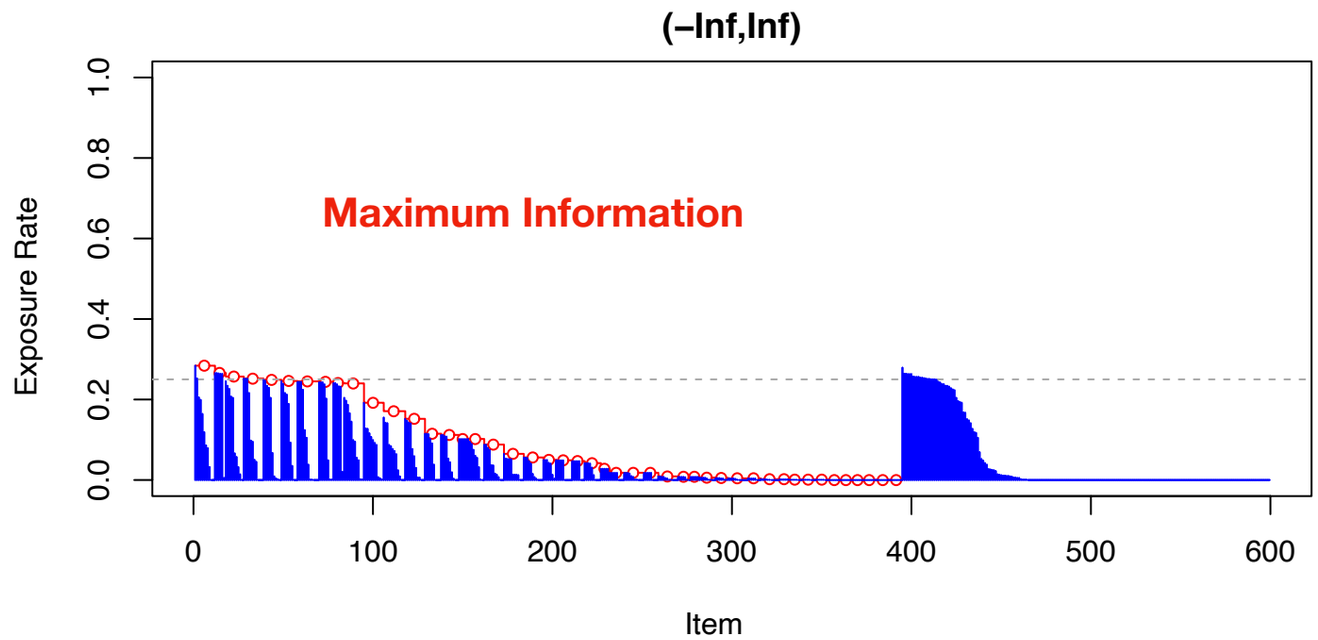


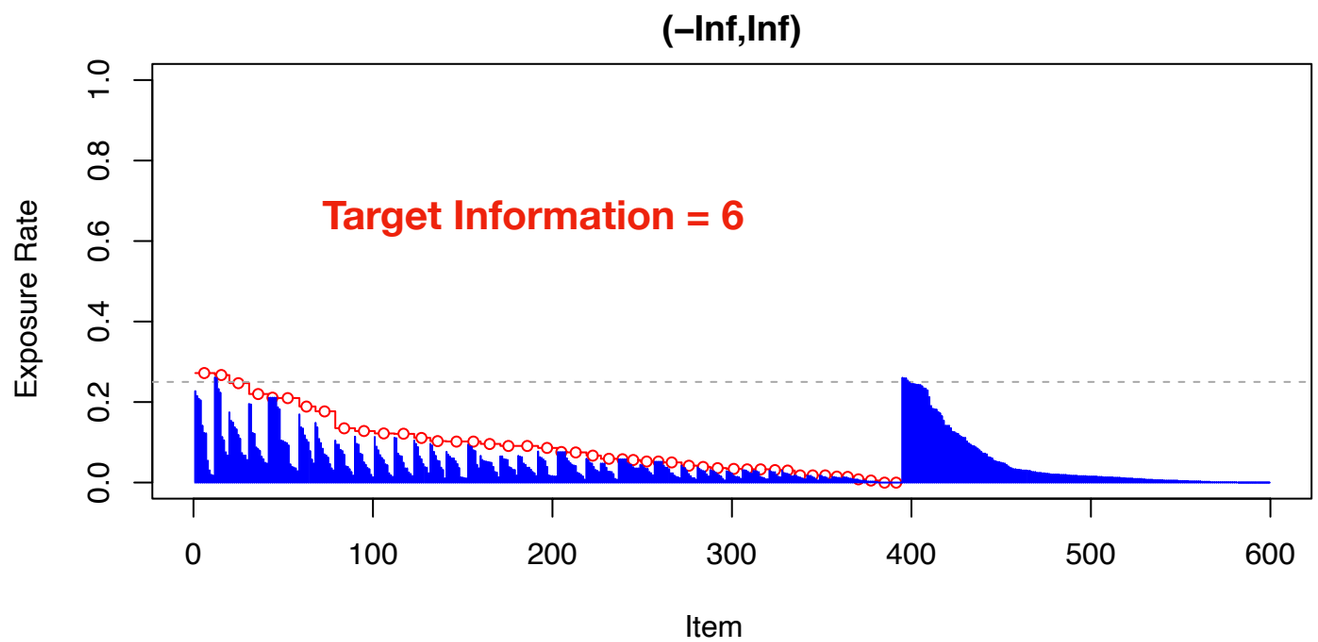


Results

Study 1

- Discrete -> Set
- Eligibility control





Results

Study 2

- No meaningful difference between item/set order conditions
- Slightly worse RMSE for exposure control
- Slightly worse RMSE for target information
- Slightly better exposure control for target information

Results

Study 2

Weave Order	Exposure Control	Information Type	Correlation	RMSE
D > S	None	Maximize	0.973	0.331
S > D			0.972	0.336
Interspersed			0.973	0.333
Set-based			0.968	0.362
D > S	Big-M	Maximize	0.966	0.371
S > D			0.968	0.366
Interspersed			0.967	0.369
Set-based			0.961	0.401
D > S	None	Target = 8	0.969	0.359
S > D			0.968	0.362
Interspersed			0.969	0.358
Set-based			0.966	0.373
D > S	Big-M	Target = 8	0.967	0.367
S > D			0.968	0.364
Interspersed			0.967	0.368
Set-based			0.962	0.397

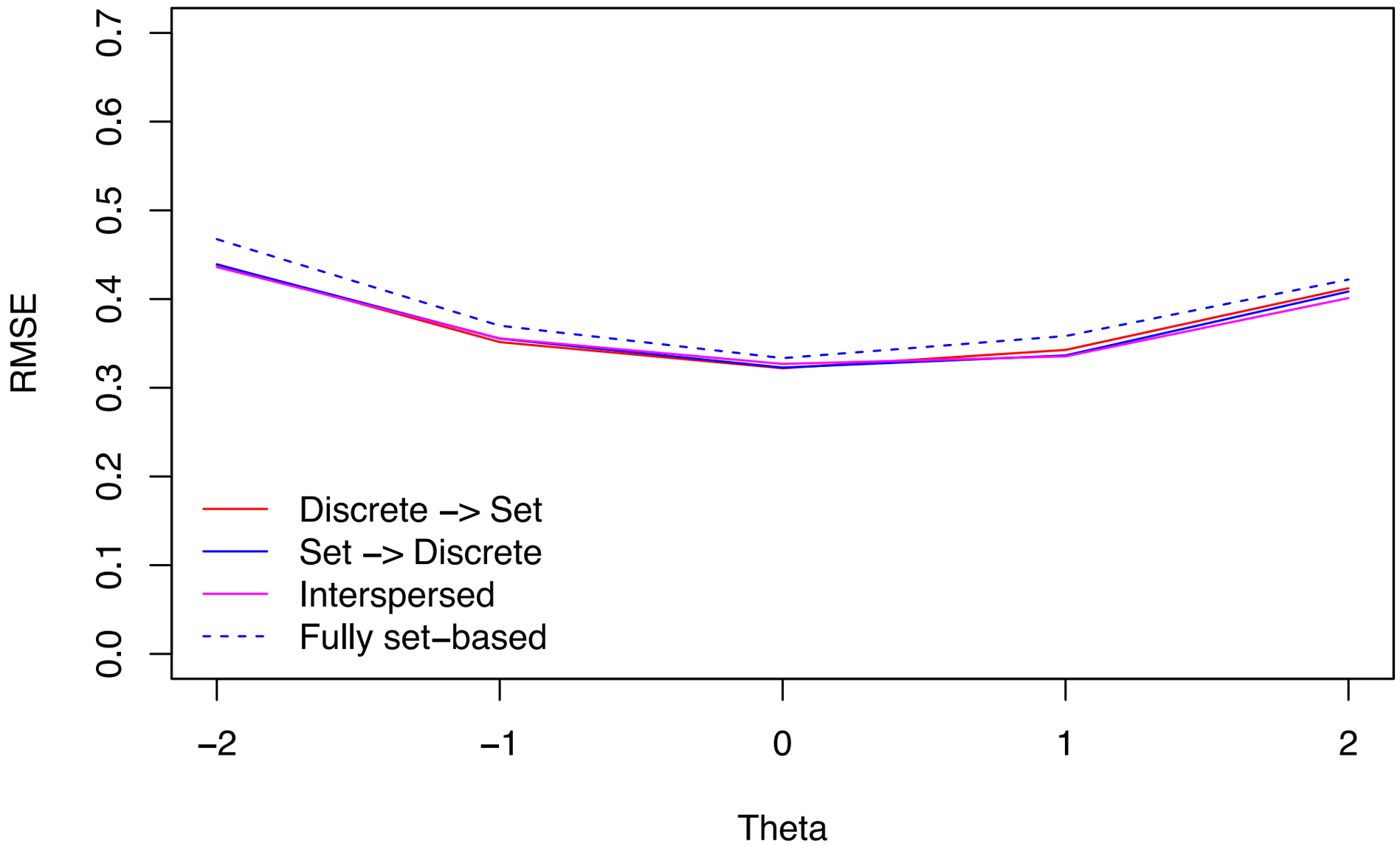
Results

Study 2

Weave Order	Exposure Control	Information Type	Correlation	RMSE
D > S	None	Maximize	0.973	0.331
S > D			0.972	0.336
Interspersed			0.973	0.333
Set-based			0.968	0.362
D > S	Big-M	Maximize	0.966	0.371
S > D			0.968	0.366
Interspersed			0.967	0.369
Set-based			0.961	0.401
D > S	None	Target = 6	0.959	0.413
S > D			0.959	0.411
Interspersed			0.960	0.409
Set-based			0.959	0.410
D > S	Big-M	Target = 6	0.958	0.416
S > D			0.960	0.411
Interspersed			0.961	0.404
Set-based			0.958	0.416

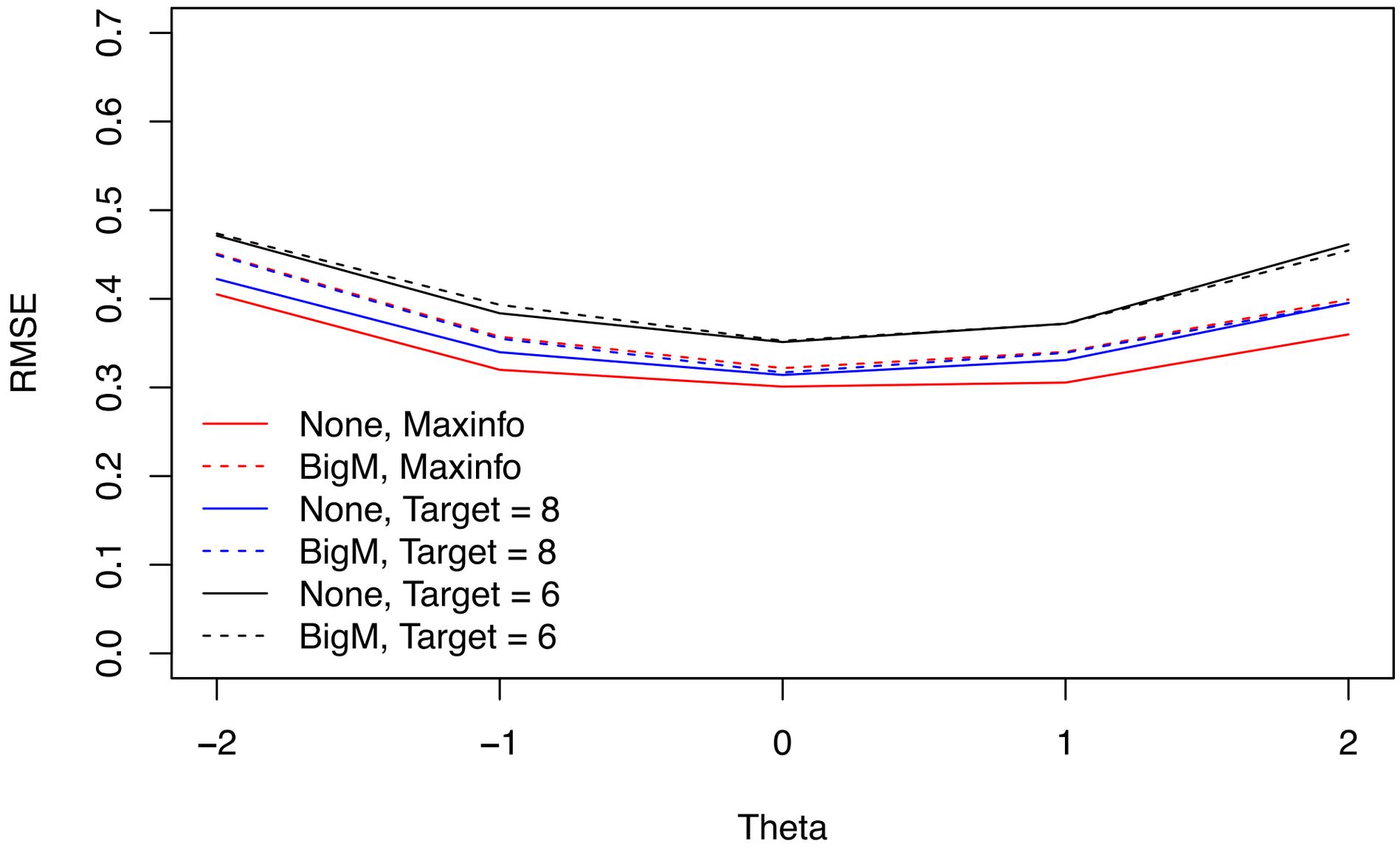
Results

Study 2 - Recovery of θ



Results

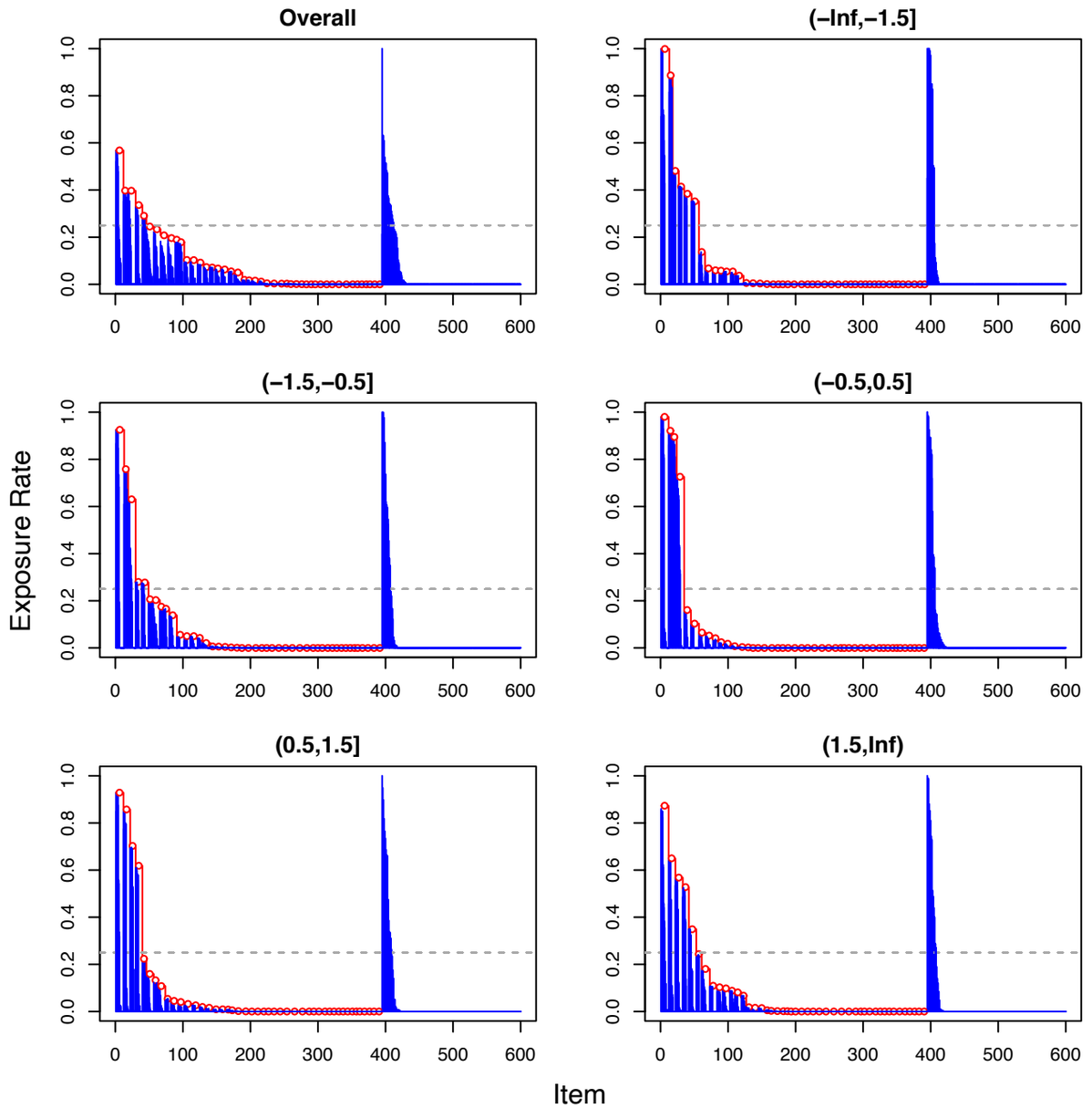
Study 2 - Recovery of θ



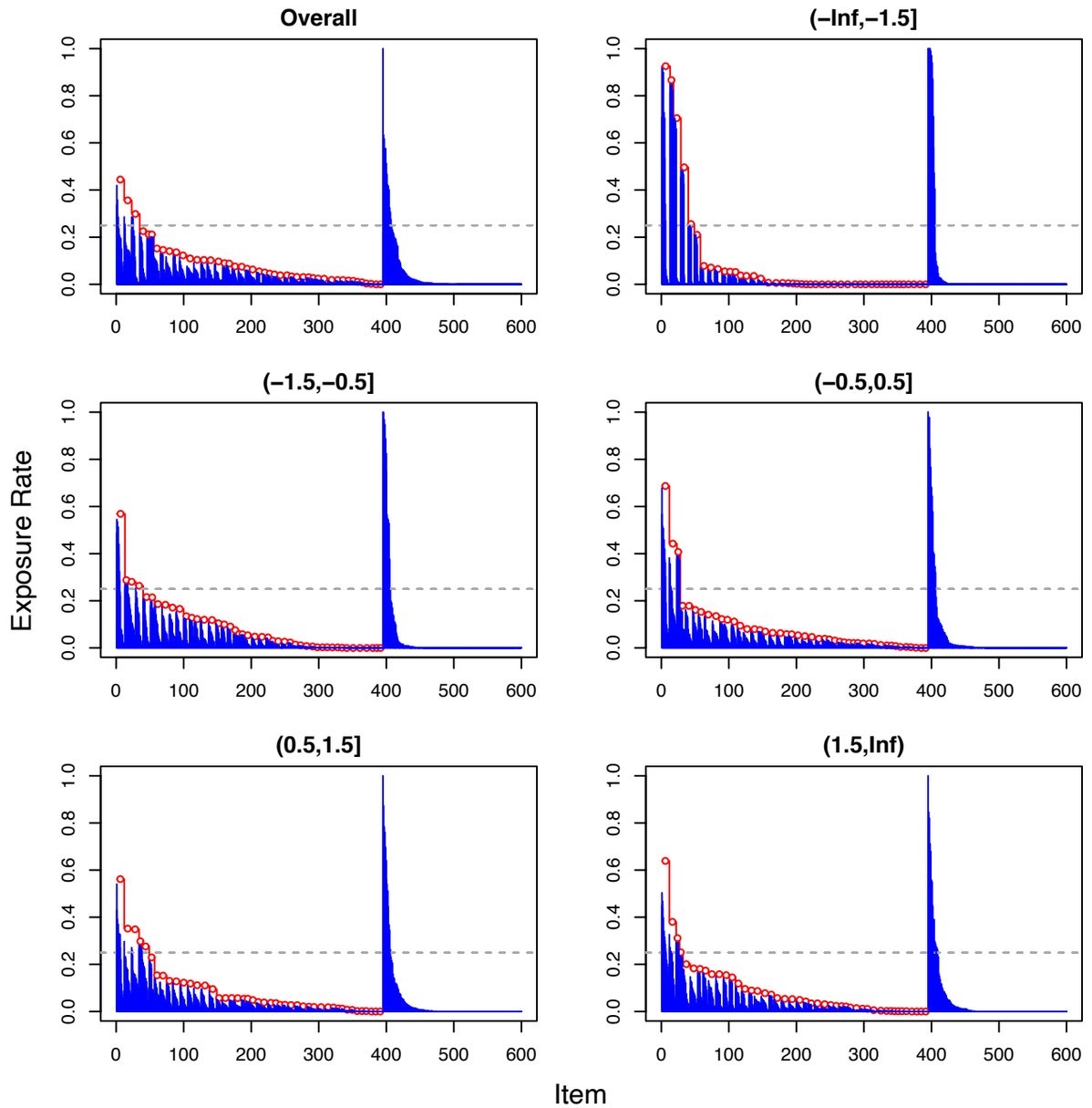
Results

Study 2 - Conditional Exposure Control (Discrete -> Set; No Exposure Control)

Maximum Information



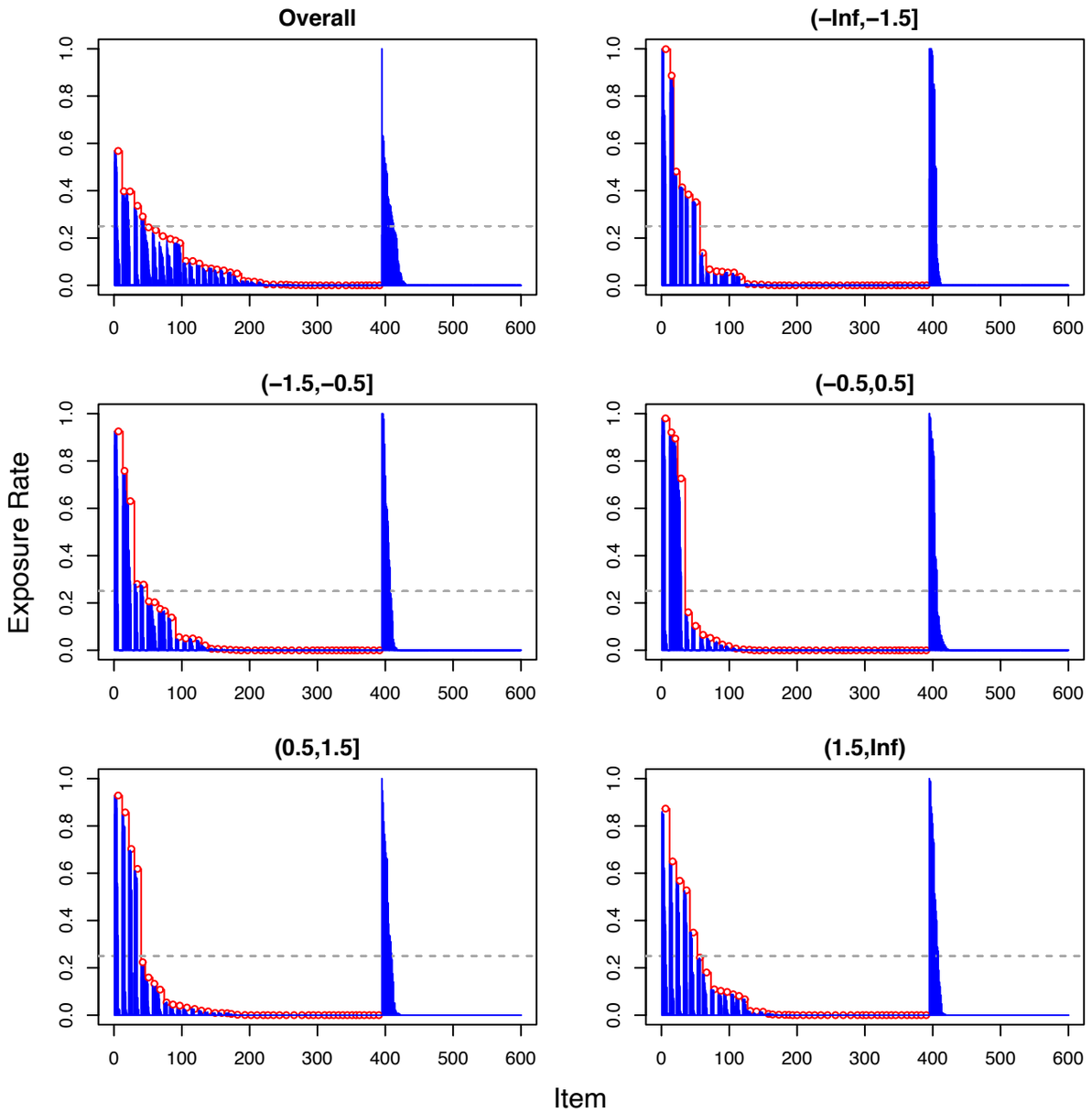
Target Information = 8



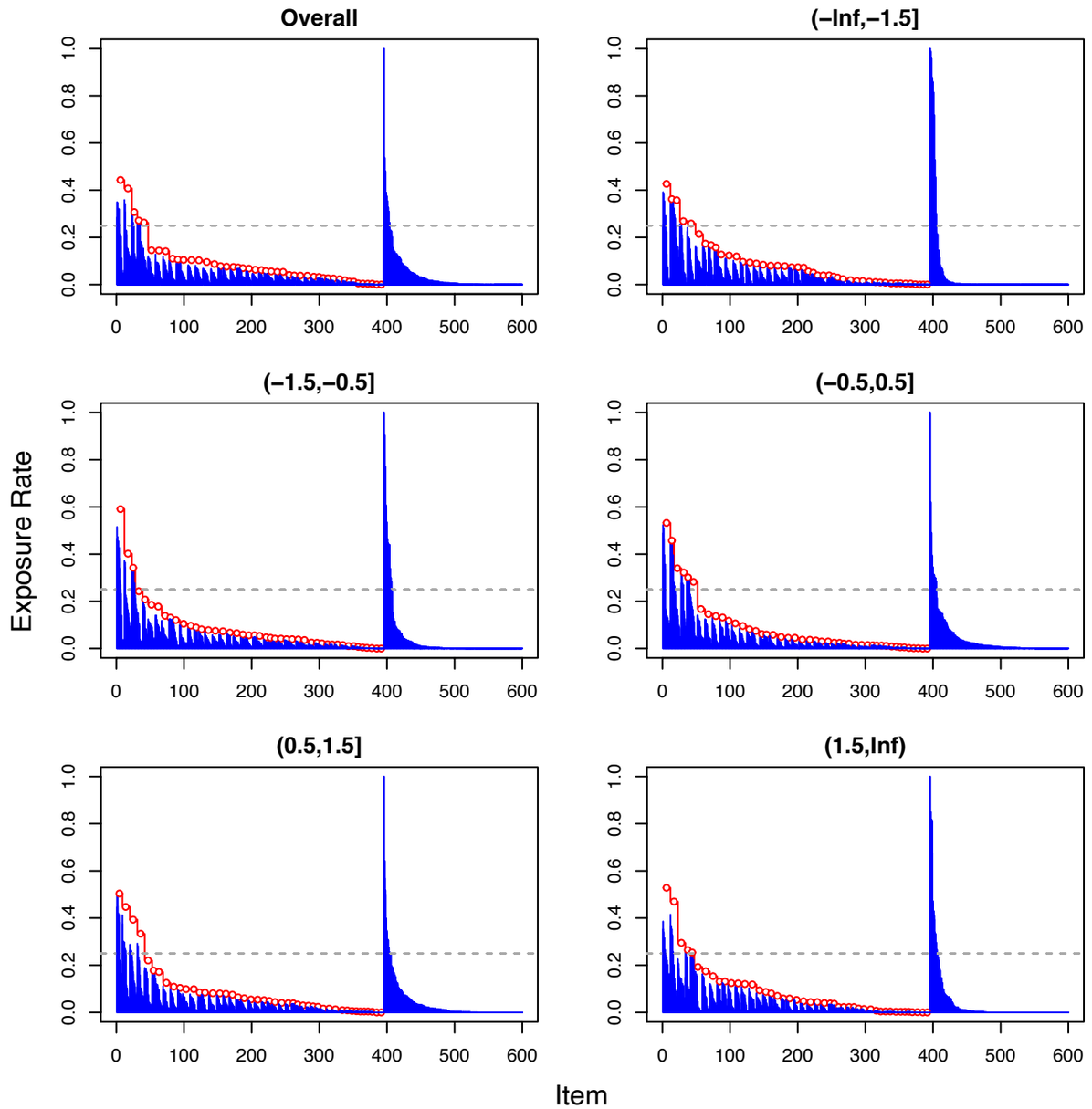
Results

Study 2 - Conditional Exposure Control (Discrete -> Set; No Exposure Control)

Maximum Information



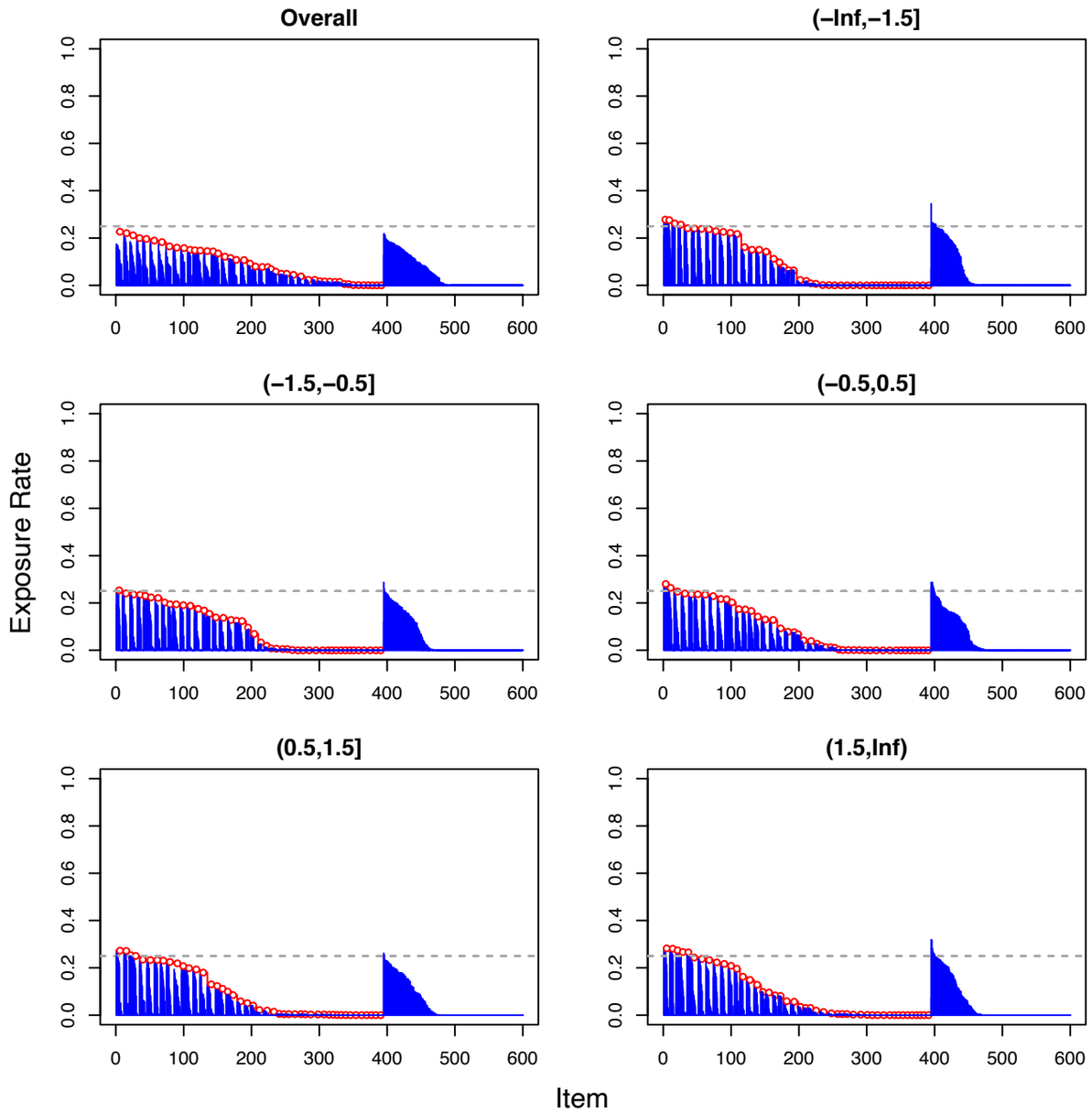
Target Information = 6



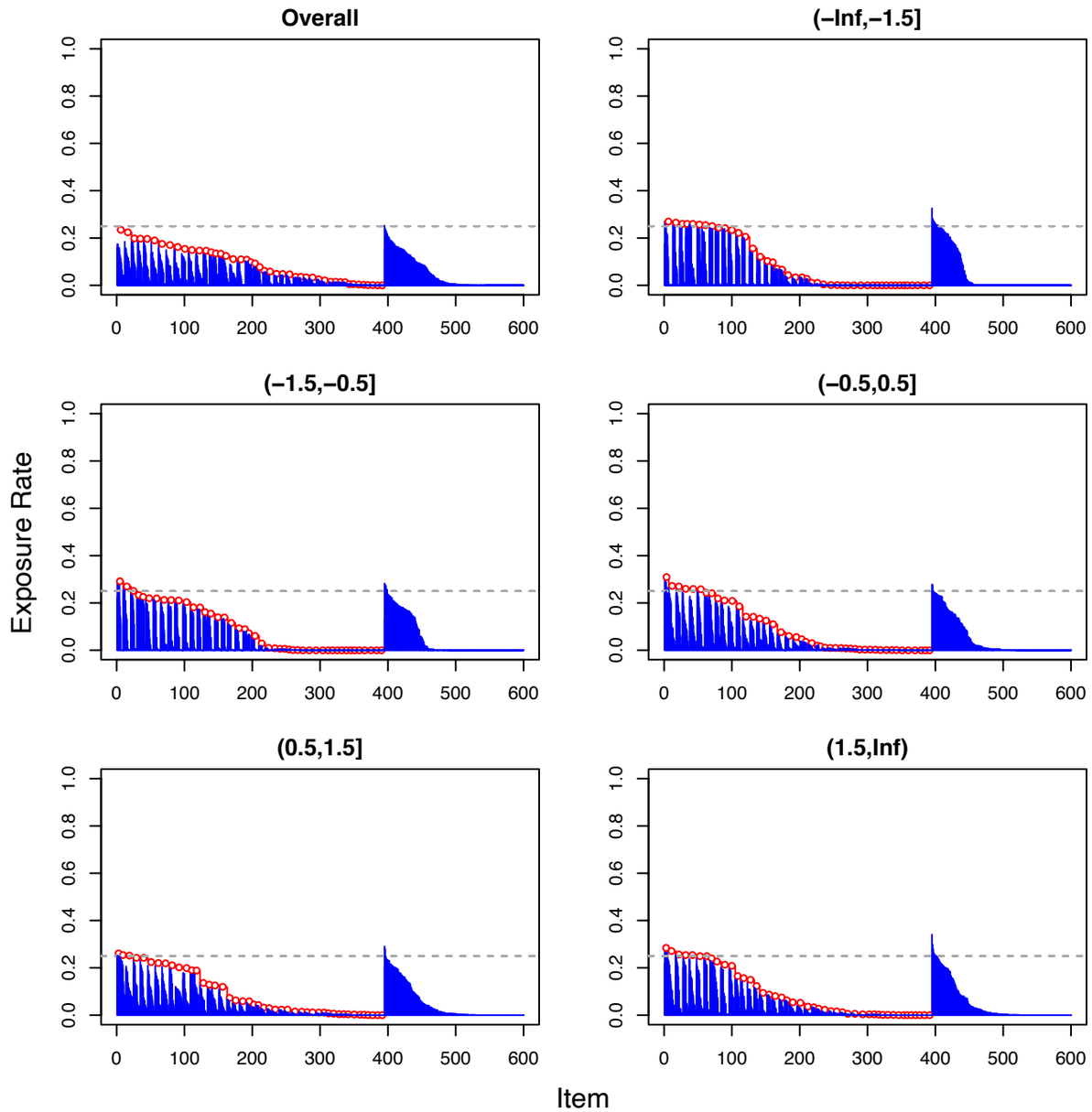
Results

Study 2 - Conditional Exposure Control (Discrete -> Set; BigM Exposure Control)

Maximum Information



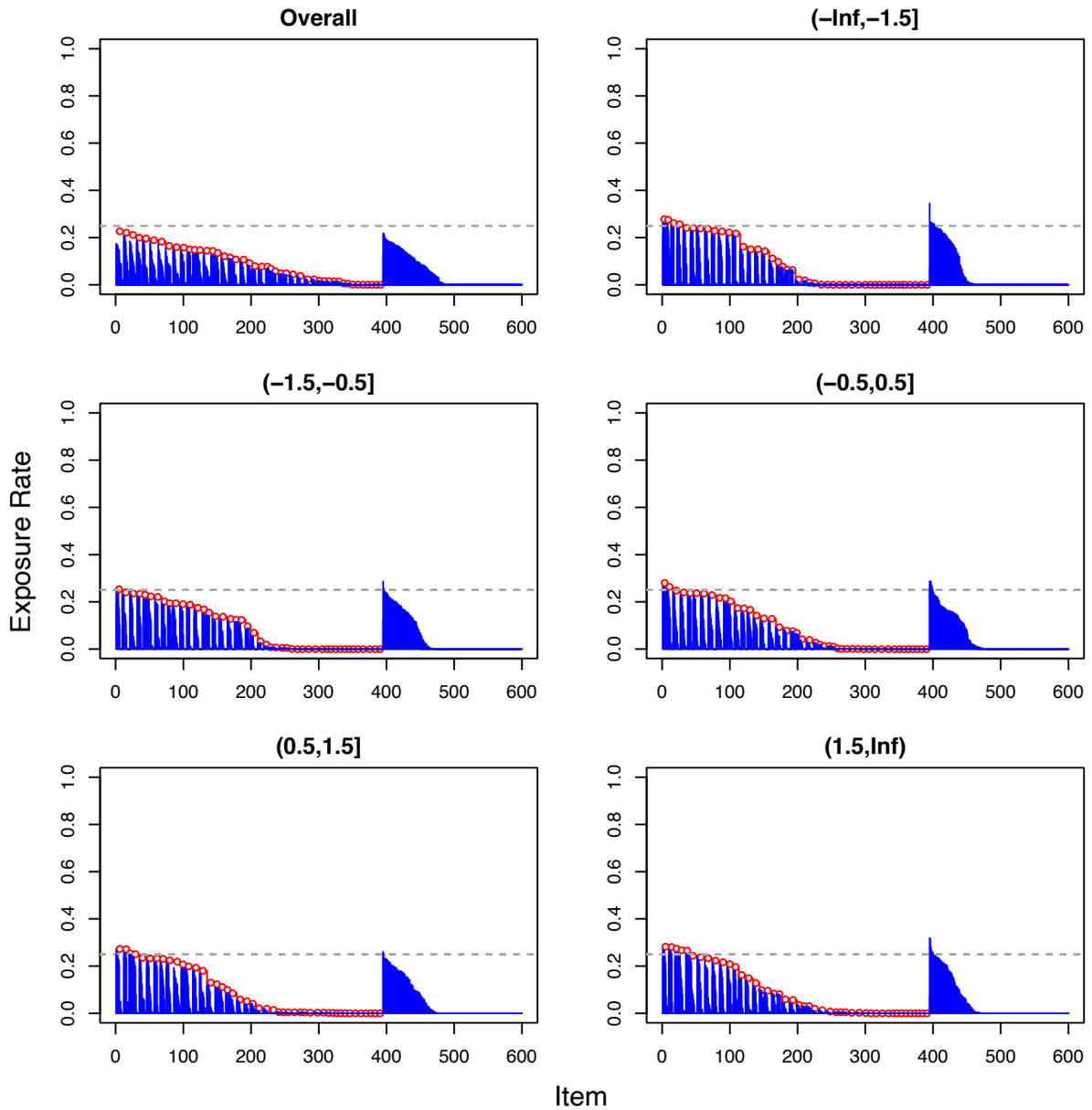
Target Information = 8



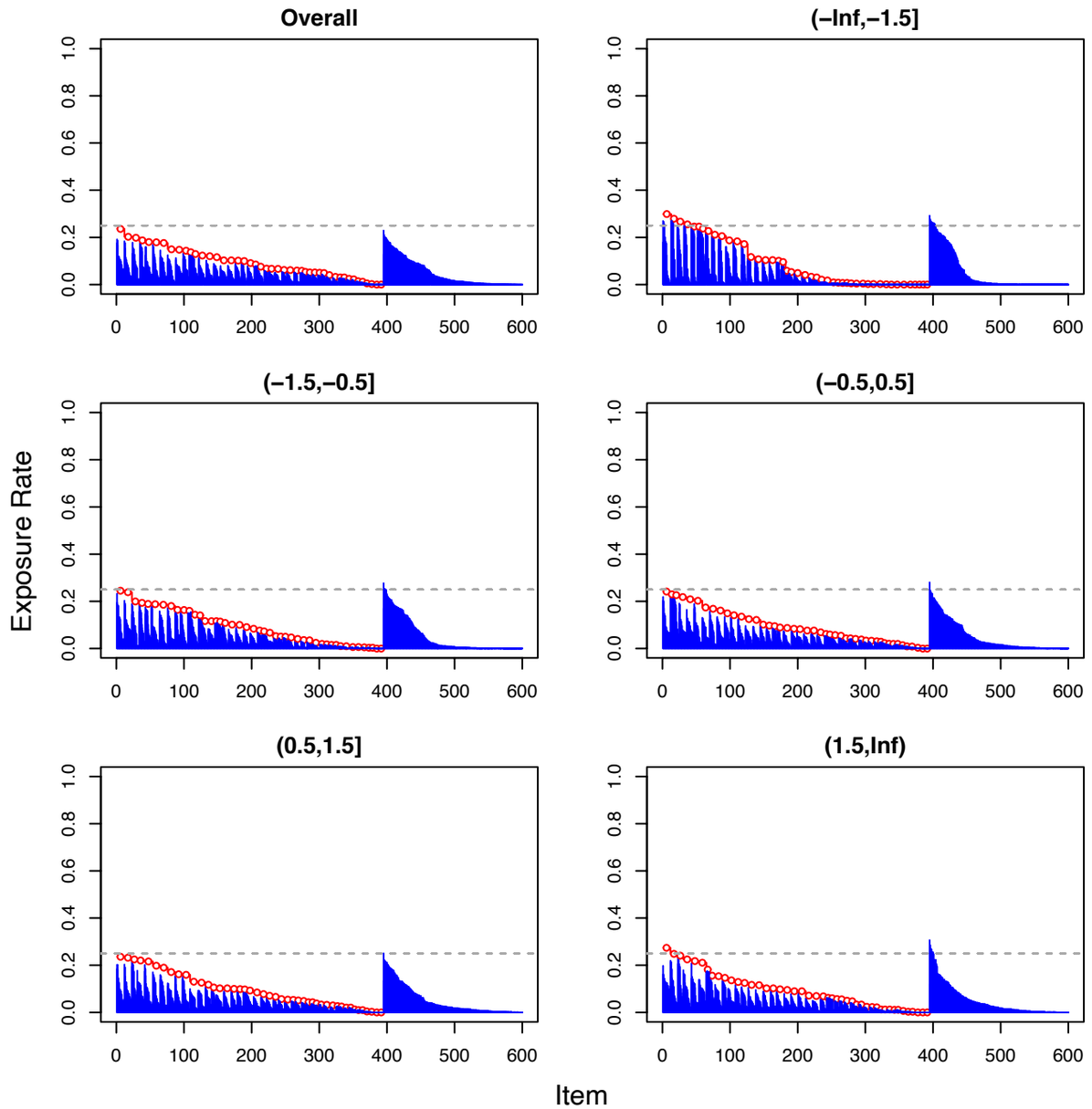
Results

Study 2 - Conditional Exposure Control (Discrete -> Set; BigM Exposure Control)

Maximum Information



Target Information = 6



Conclusion

Take-aways

- The universal shadow test assembler implemented in **TestDesign** can effectively handle complex test specifications and constraints.
- We illustrated how a new item selection method, the **target information selection criterion**, can be used in the shadow-test approach.
- It is possible to change the target information as the test progresses, e.g., for an effect kin to α-stratification without having to partition item pools.
- We can also switch the selection criterion, e.g., from target information to maximum information, at some point within a test.

*https://cran.r-project.org/web/packages/TestDesign/index.html

References

Choi, S. W., Lim, S., & van der Linden (in preparation). TestDesign: An optimal test design approach to constructing fixed and adaptive tests in R.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270.

van der Linden,W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

van der Linden, W. J., & Diao, Q. (2014). Using a Universal Shadow-Test Assembler with Multistage Testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications*. Chapman and Hall/CRC Press.

van der Linden, W. J., & Choi, S. W. (2019). Improving item-exposure control in adaptive testing. *Journal of Educational Measurement*, 57, 405-422.

Thank you!