

# **A method for the comparison of item selection rules in Computerized Adaptive Testing**

J. R. Barrada<sup>1</sup>, J. Olea<sup>2</sup>, V. Ponsoda<sup>2</sup>, F. J. Abad<sup>2</sup>

<sup>1</sup>Universidad Autónoma de Barcelona

<sup>2</sup>Universidad Autónoma de Madrid

<http://www.uam.es/juanra.barrada>

Two main goals in CATs: accuracy and security.

The reasons for maximizing accuracy are clear.

We will consider an item bank as more secure the lower the probability of preknowledge of a part of the items that an examinee will face. An index for assessing this is the overlap rate.

Improving security increases scores validity.

There is an apparent trade-off between these goals.

Several item selection rules (ISRs) have been proposed. We will focus in 6 of them, 3 mainly related with maximazing accuracy and 3 with maximazing security.

**1. Point Fisher information (PFI):**

$$\max_{i \in B_p} I_i(\hat{\theta}) \rightarrow S_i$$

where  $B_p$  is composed by all the non-presented items.

**2. Fisher information weighted by likelihood (FI-L):**

$$\max_{i \in B_p} \int_{-\infty}^{\infty} I_i(\theta) L(\theta, x, g) d(\theta) \rightarrow S_i$$

**3. Kullback-Leibler function weighted by L (KL-L):**

$$\max_{i \in B_n} \int_{-\infty}^{\infty} KL_i(\theta \parallel \hat{\theta}) L(\theta, x, g) d(\theta) \rightarrow S_i$$

#### 4. Maximum information stratifying method with blocking (MIS-B):

$$\min_{i \in B_p} \left| \hat{\theta} - \theta_i^{\max} \right| \rightarrow S_i$$

where  $B_p$  is composed by all the items non-presented and whose  $I^{\max}$  does that item presentable in the  $q$ -th position of the test.

#### 5. Progressive method (PR):

$$\max_{i \in B_p} \left[ (1 - W_q) R_i + W_q I_i(\hat{\theta}) \right] \rightarrow S_i$$

#### 6. Proportional method (PP):

$$P(S_i) = \frac{I_i(\hat{\theta})^{P_q}}{\sum_{i=1}^n I_i(\hat{\theta})^{P_q}} \quad \sum_{j=1}^{i-1} P(S_j) < R \leq \sum_{j=1}^i P(S_j) \rightarrow S_i$$

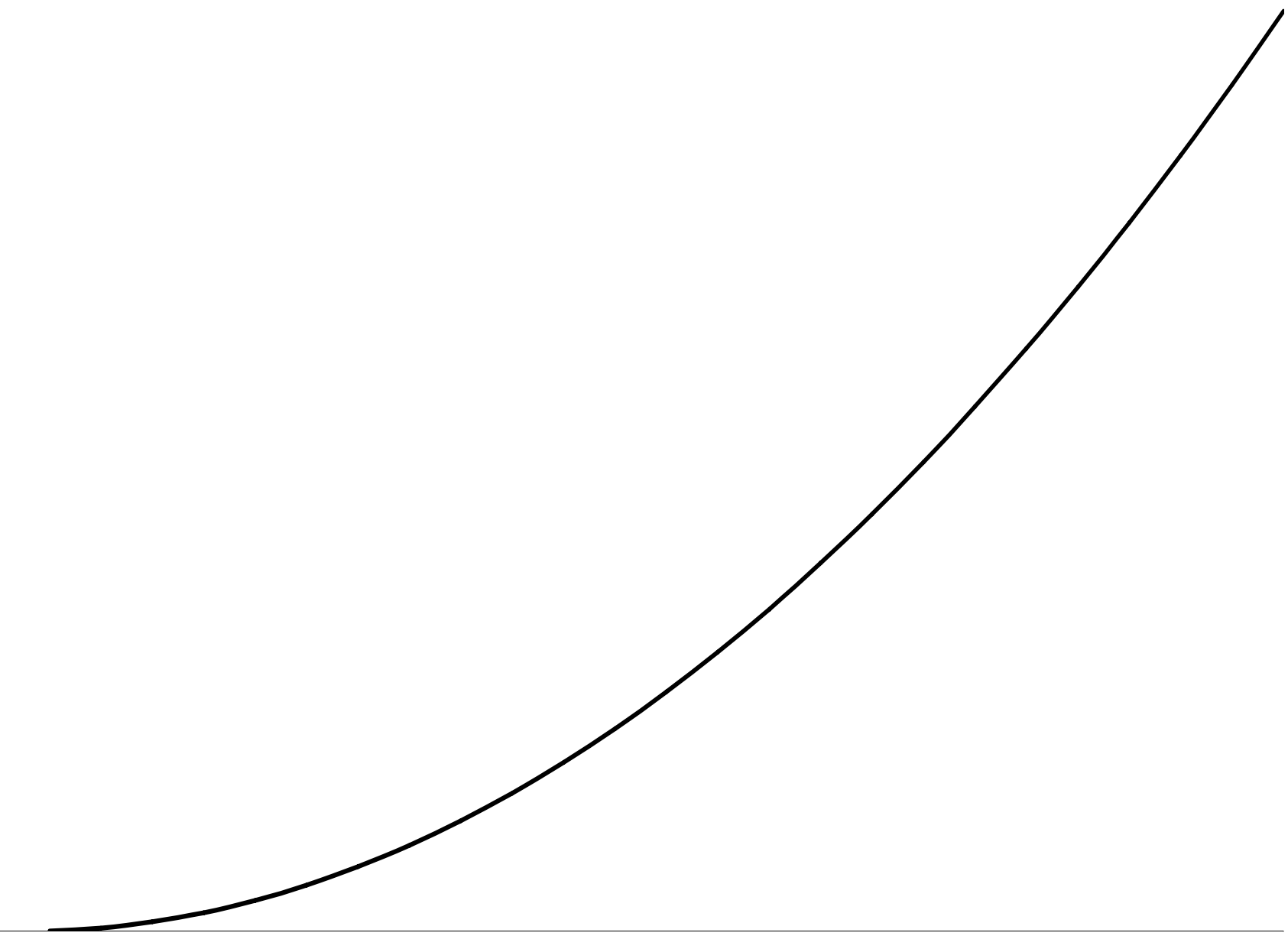
information

random

first item

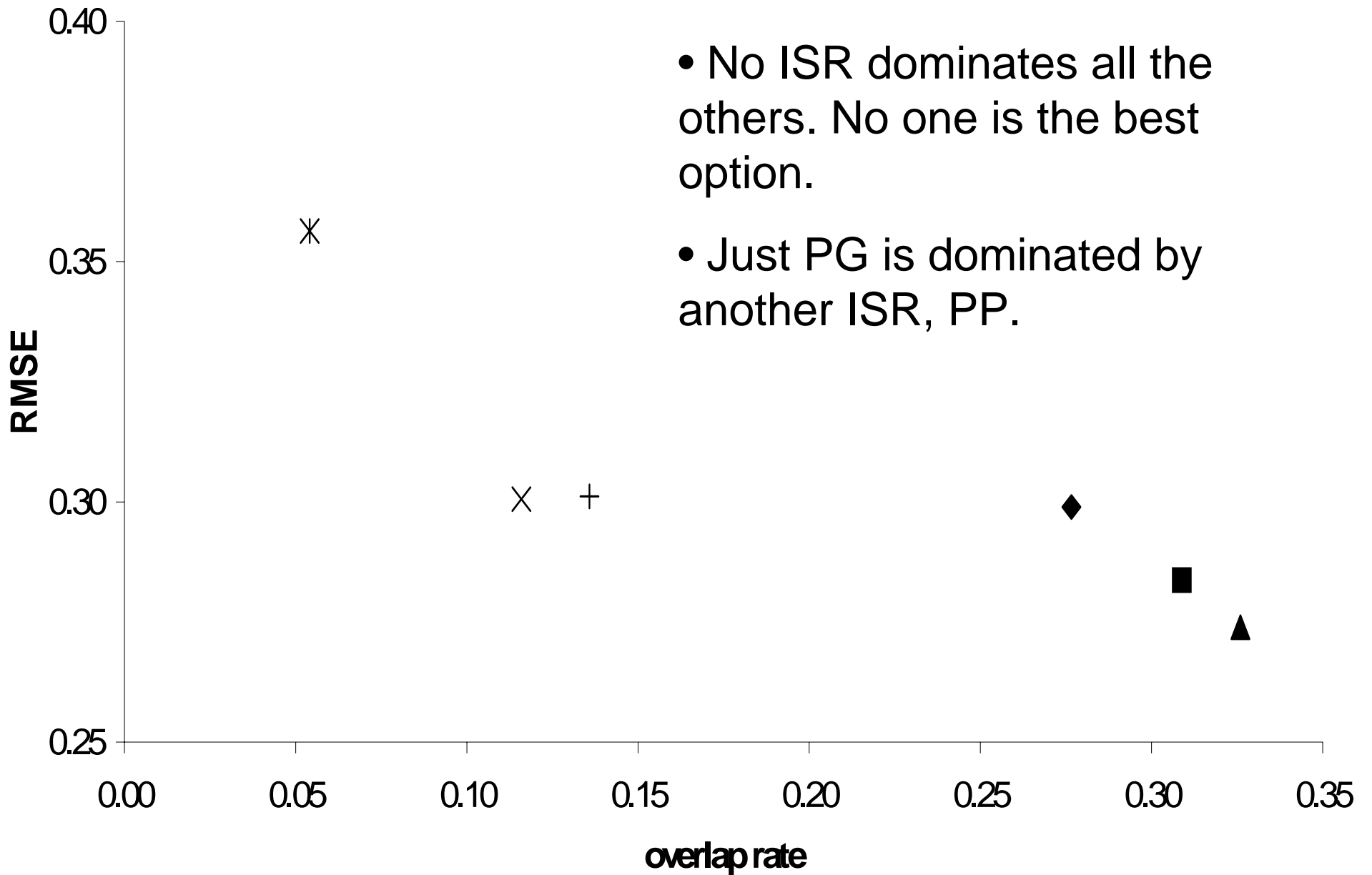
item position

last item



## Simulation study - Method

- Ten randomly generated banks with 500 items each one.
- Parameter distributions:  $a \sim N(1.2, 0.25)$ ;  $b \sim N(0, 1)$ ;  $c \sim N(0.25, 0.02)$ .
- 5000 examinees per item bank.
- Test length: 20 or 40 items.
- Initial estimation randomly generated in the interval  $(-.5, .5)$ .
- For MIS-B five strata with four items presented of each stratum.
- Maximum likelihood estimation in the interval  $[-4, 4]$ ; with constant response pattern, Dodd rule.



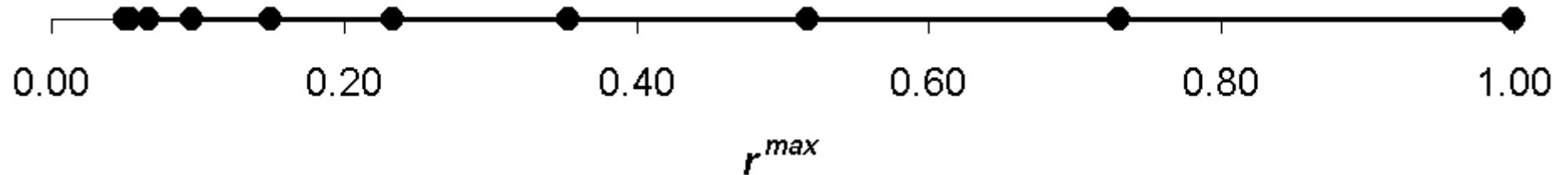
## Restrictions in maximum exposure rate for comparing ISRs

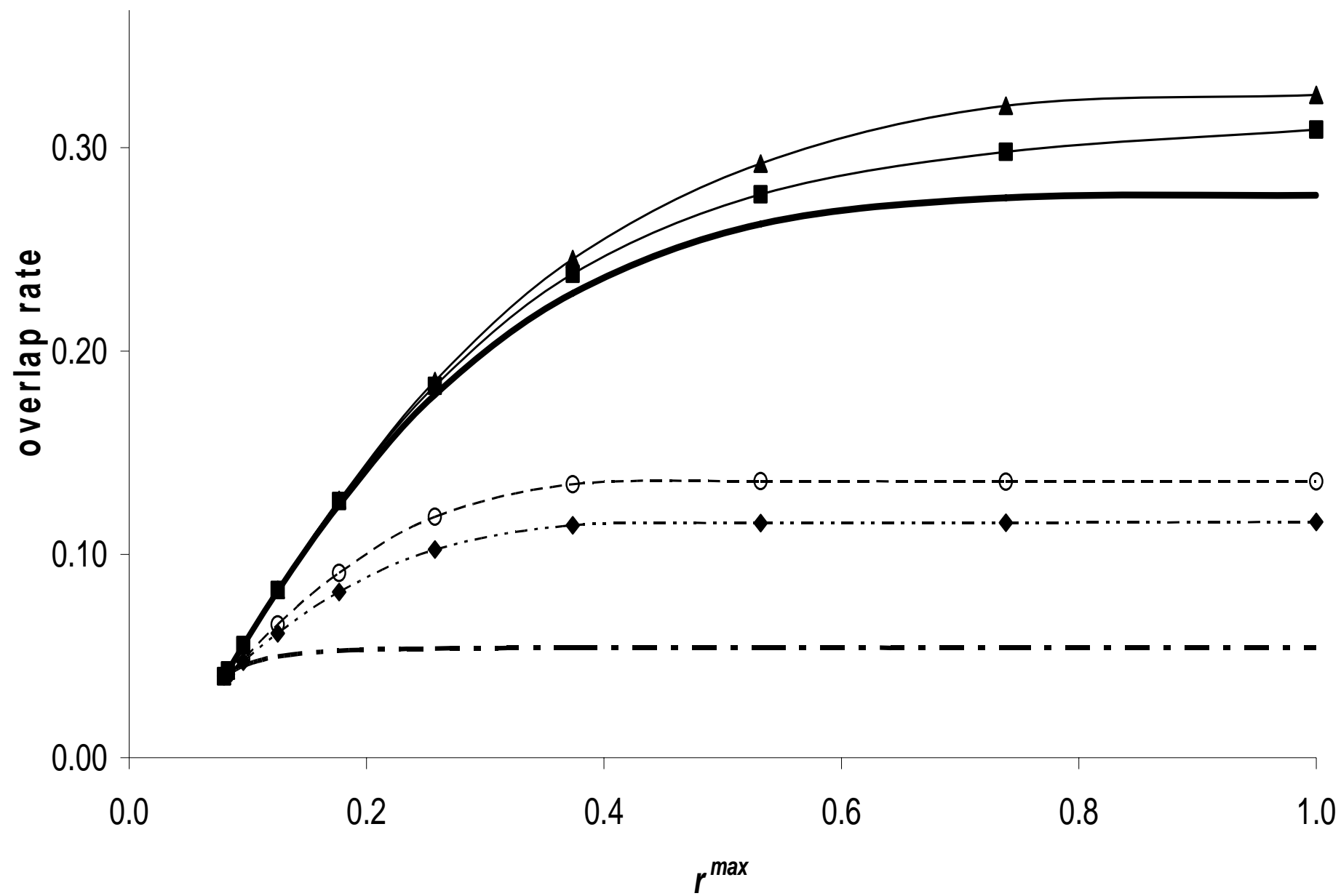
- Manipulating the maximum exposure rate acceptable ( $r^{\max}$ ) allows us to:
  - reduce overlap rate
  - as we can get the same overlap rate for two different ISRs we can compare them in accuracy, checking dominance.
- The best option for manipulating  $r^{\max}$  is the item-eligibility method:

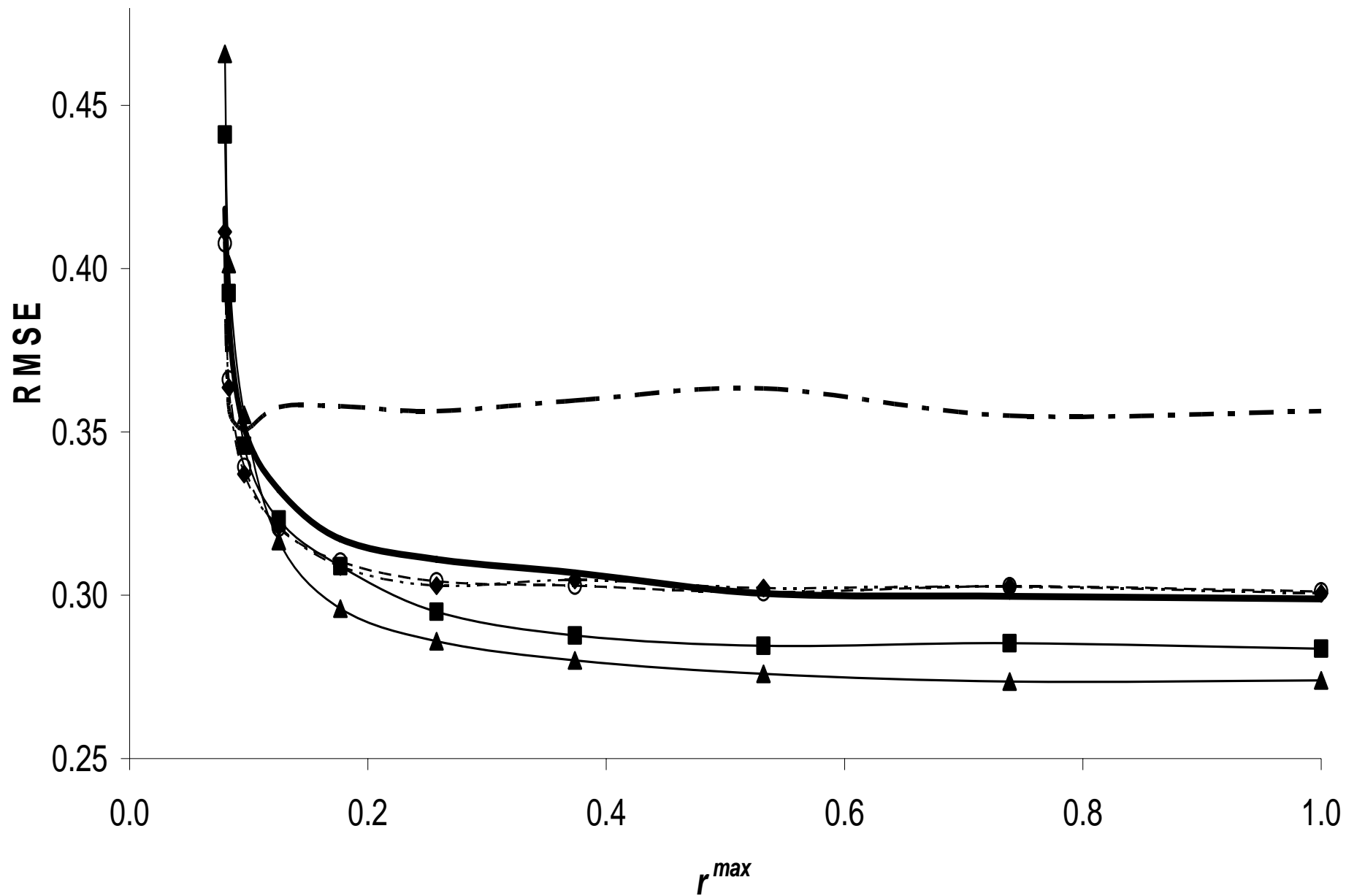
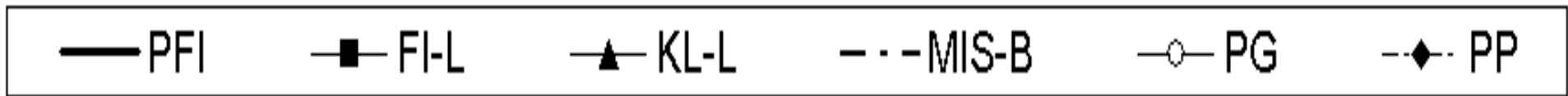
$$k_i^{(j+1)} = \begin{cases} 1 & \text{if } P^{(1..j)}(A_i) / k_i^{(j)} \leq r^{\max} \\ r^{\max} k_i^{(j)} / P^{(1..j)}(A_i) & \text{if } P^{(1..j)}(A_i) / k_i^{(j)} > r^{\max} \end{cases}$$

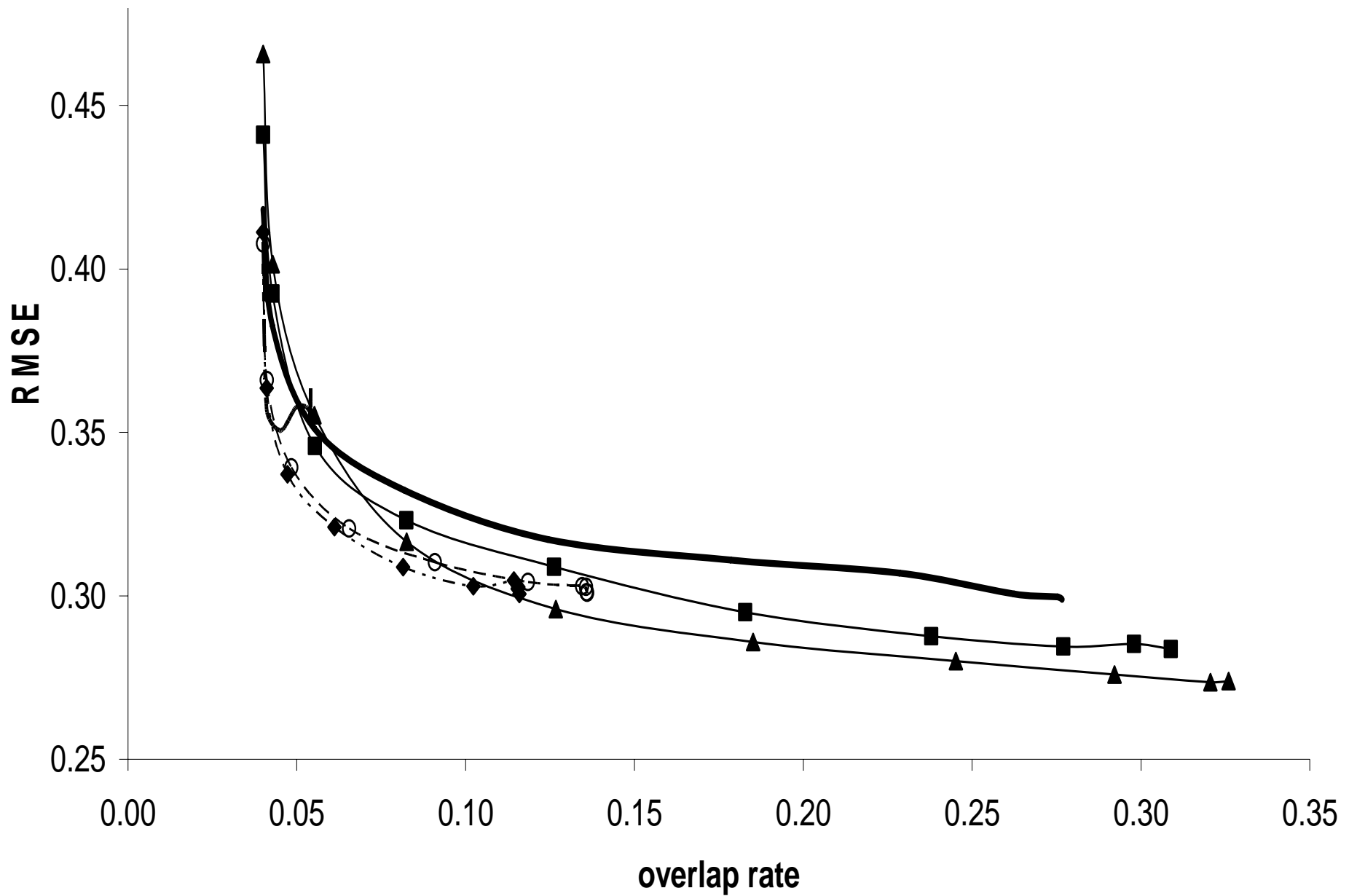
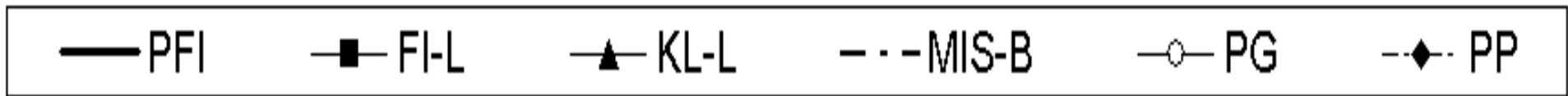
## Simulation study - Method

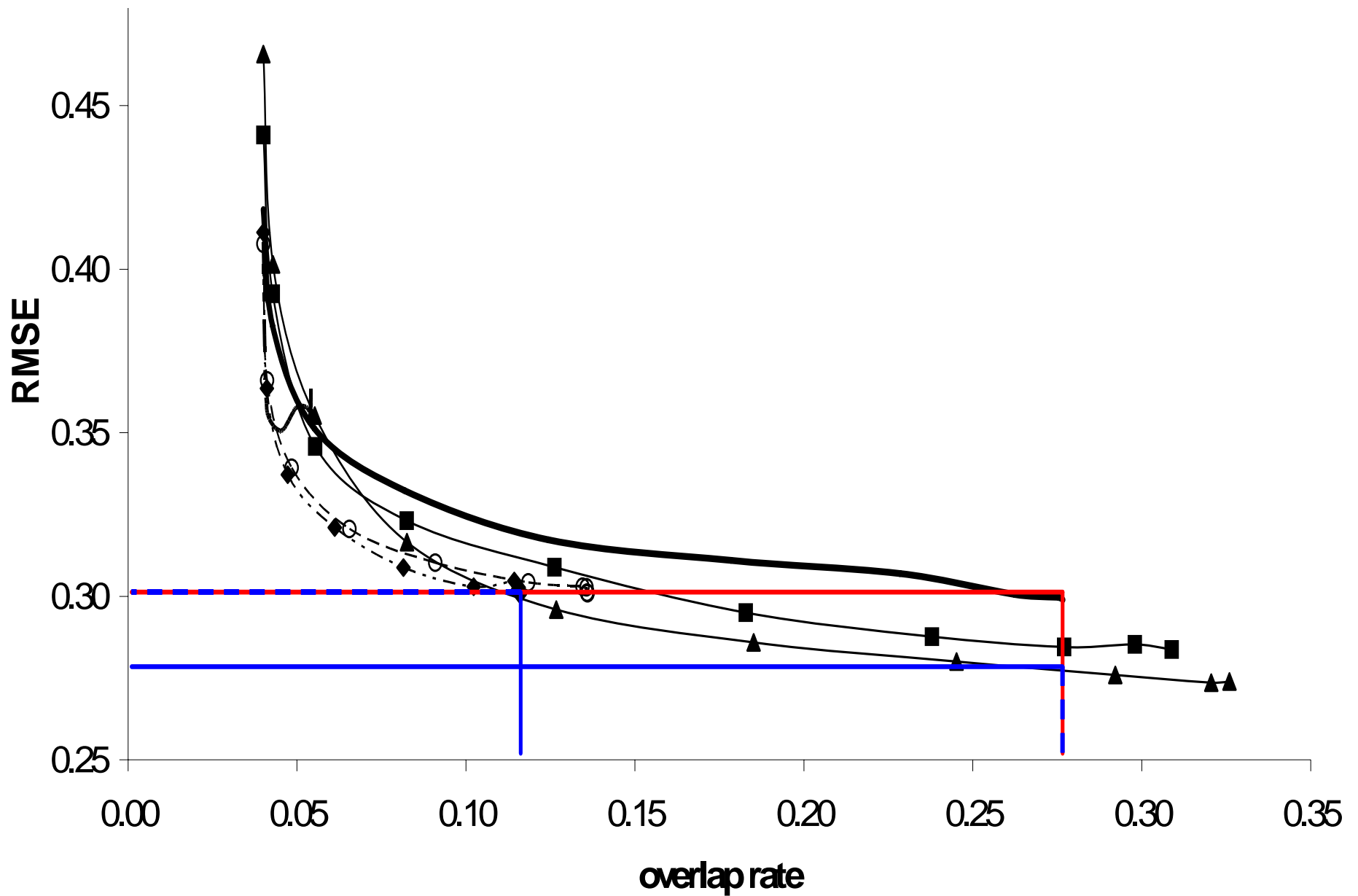
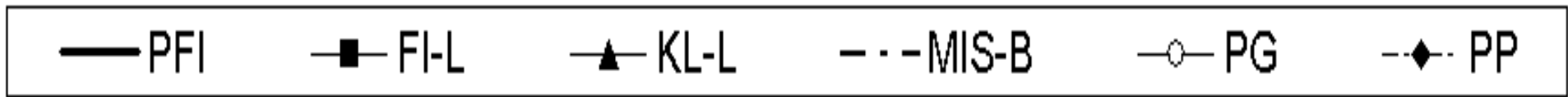
- The same than in the previous study.
- Ten different values for  $r^{max}$  were used, from the minimum possible to no restriction.

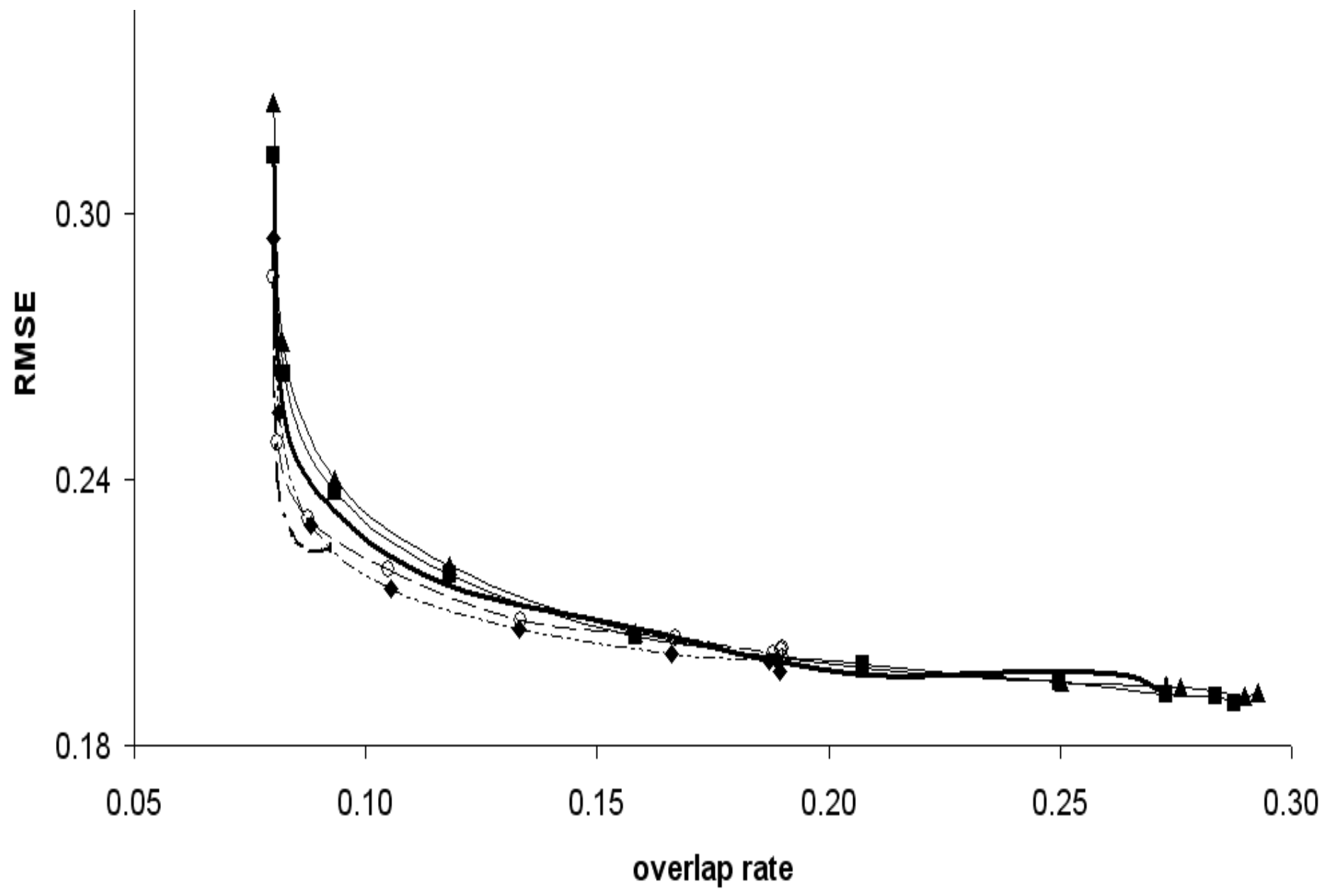












# Conclusions

- With this method, we can choose the best ISR for a desired accuracy or security.
- The decision will depend on our goals.
- There is no ISR that can be described as the best one: the ranking depends on the level of the variables that you look at.
- This means that it is necessary an exhaustive manipulation of  $r^{max}$  when comparing ISRs.
- We can say that some ISRs are dominated all through the possible values of RMSE or overlap, so they are never interesting options. PFI, for instance.
- When no strict exposure control is applied, it seems that the best option is KL\*L. When the control is thick, PP. Thicker and test length equal to 40, MIS.

## A tricky study? Future questions

- We have been working with item parameters considered as without measurement error. It is quite probable that not all the ISR react in the same way to the capitalization of chance when calibrating item parameters.
- We have been considering just two goals of CATs. Importantly, we have not considered bank maintenance. The more similar is the distribution of the parameters of the items used to the original distribution of the bank parameters, the easier to maintain the bank.
- We have used as measure of bank risk only test overlap. Although this is the commonly employed variable, it is not clear if this variable is convenient for this purpose. Is the same sharing an item of low a parameter than sharing an item of high a parameter? Is the same sharing items at the beginning of the test than sharing them at the end?

# References

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (In press). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational & Behavioral Statistics, 29*, 439-460.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational & Behavioral Statistics, 29*, 273-291.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational & Behavioral Statistics, 22*, 203-226.