Journal of Educational and Behavioral Statistics

IF: 2.042

Optimizing the Use of Response Times for Item Selection in Computerized Adaptive Testing

Edison M. Choe
Graduate Management Admission Council

Justin L. Kern
University of Illinois, Urbana-Champaign

hua-hua Chang
University of Illinois, Urbana-Champaign

Reporter: Yingshi Huang

Introduction

- Computerized Adaptive Testing (CAT)

Efficiency?

− the number of items administered

information-based optimality criterion

maximum Fisher information criterion (MI)

both high discrimination and low time intensity

− the time it takes to complete the test

maximizes the ratio of Fisher information to expected response time (MIT)

a time-weighted version of a-stratification with b-blocking (ASBT)

the mean and variance of testing times & estimation error

Purpose: improve upon the innovative RT-based item selection methods

RT framework

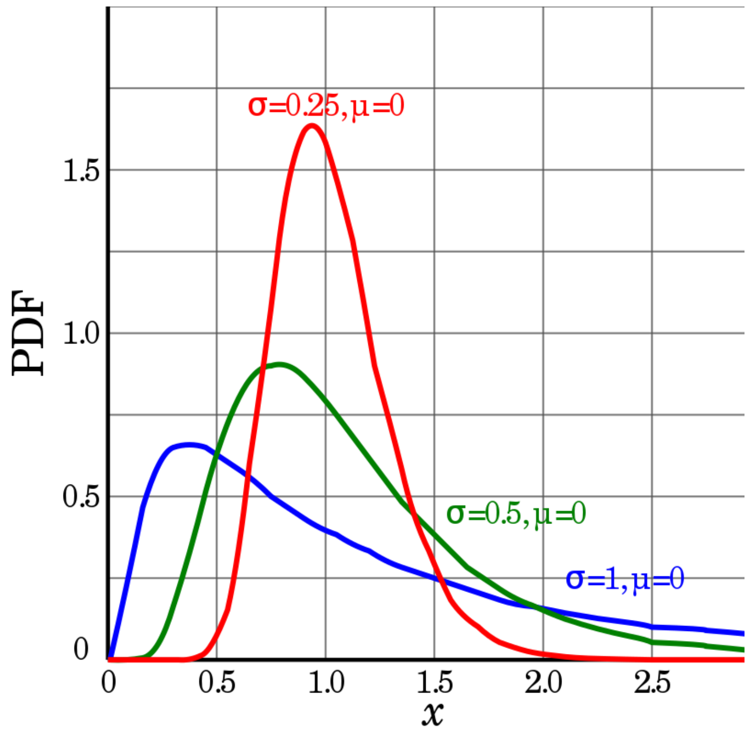
How to model response times?

- the lognormal model (van der Linden, 2006)
- the Box–Cox normal model (Klein Entink, van der Linden, & Fox, 2009)
- the Cox proportional hazards model (C. Wang, Fan, Chang, & Douglas, 2013)
- the linear transformation model (C. Wang, Fan, Chang, & Douglas, 2013)

RT framework

- the lognormal model: an idea of curve fitting

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$



[From *Wikipedia*]

Why lognormal?

➡ has the **positive support** and a **skew required** for response-time distributions

ability θ

$$f(t_{ij}|\tau_i) = \frac{1}{t_{ij}\sqrt{2\pi(1/\alpha_j)^2}} e^{-[\log t_{ij} - (\boxed{\beta_j} - \boxed{\tau_i})]^2/[2(1/\boxed{\alpha_j})^2]}$$

difficulty b

with $\mu = \beta_j - \tau_i$

$\sigma^2 = (1/\alpha_j)^2$

discrimination a

➡ analogy to the **two-parameter** logistic (2PL) response model

no need for guessing parameter
(time has a natural lower bound at t = 0)

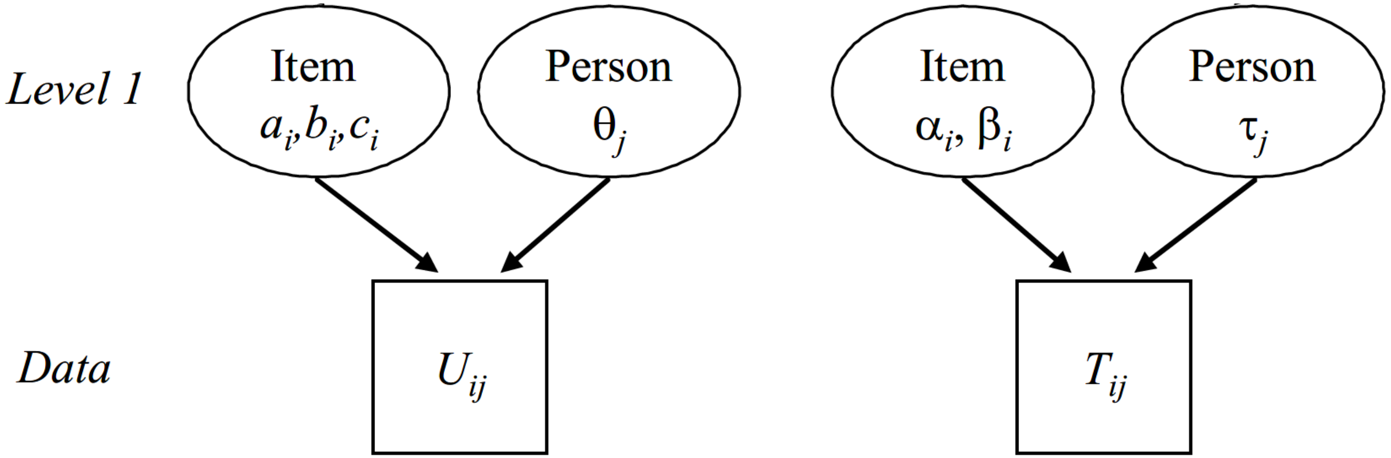
➡ expected RT: $E(T_{ij}|\tau_i) = e^{\beta_j - \tau_i + 1/(2\alpha_j^2)}$

van der Linden, 2006 *JEBS*

RT framework

How to model the relations between response and RT?

- a “plug-and-play approach”



1. response model & RT model

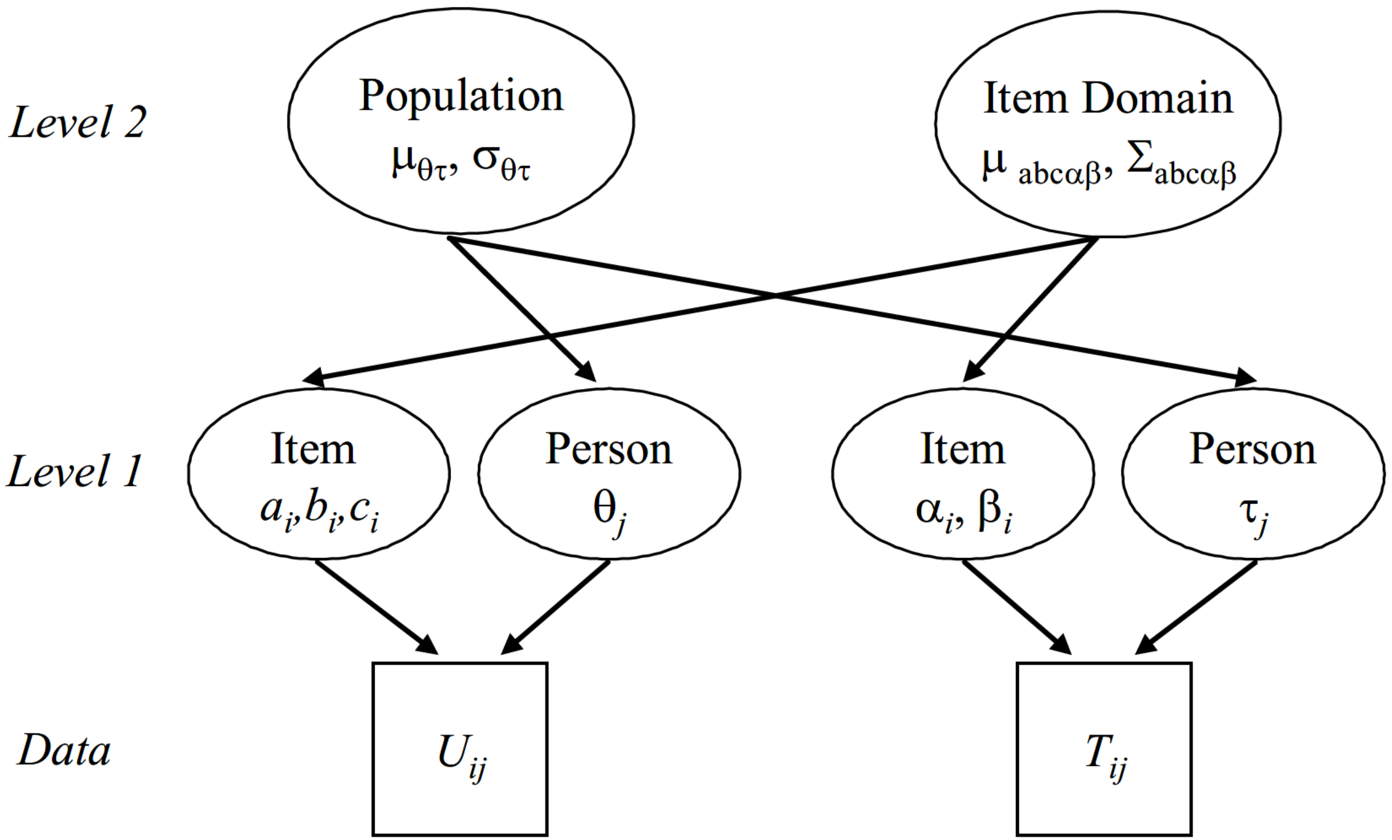
   e.g. 3PLM & lognormal model

van der Linden, 2007 *PSYCHOMETRIKA*

RT framework

How to model the relations between response and RT?

- a “plug-and-play approach”



2. population model & item-domain model

   e.g. multivariate normal distribution

   $$f(\boldsymbol{\xi}_i; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$$
   $$f(\boldsymbol{\psi}_j; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}})$$

1. response model & RT model

   e.g. 3PLM & lognormal model

van der Linden, 2007 *PSYCHOMETRIKA*

CAT framework

How to assemble a test?

- Maximum Fisher information method (MI)

prone to selecting items with high *a*

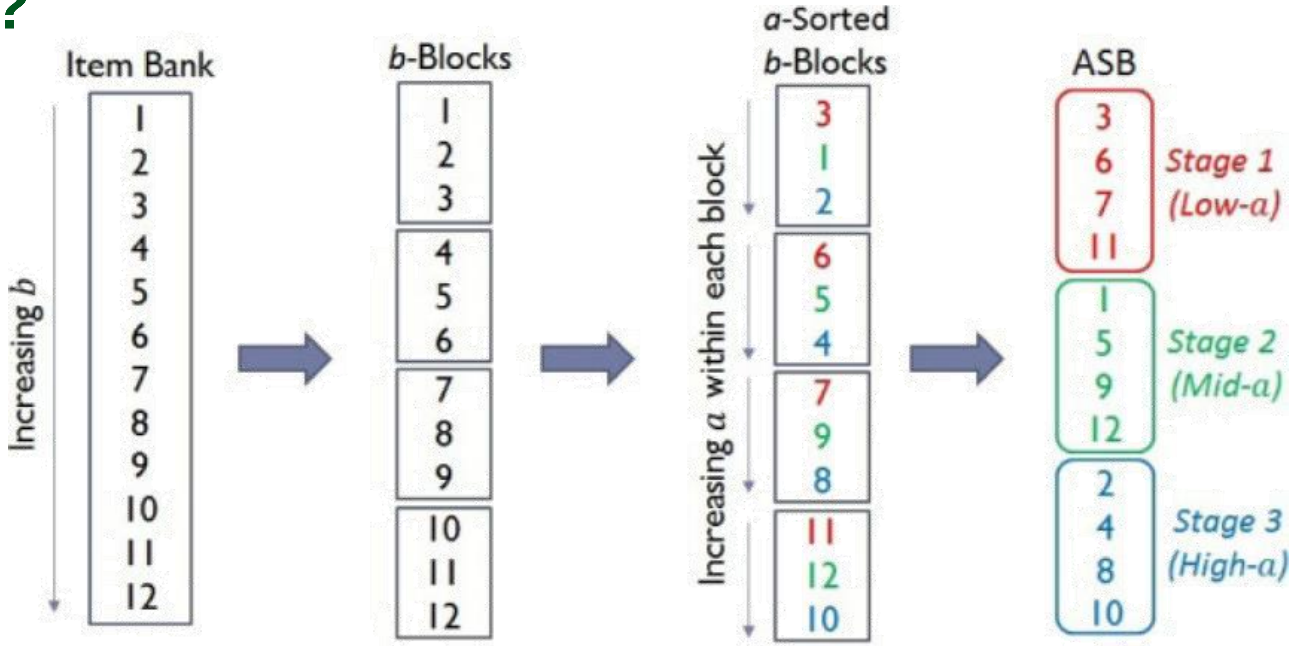
$$I_j(\theta_i) = \frac{(1-c_j)a_j^2 e^{a_j(\theta_i-b_j)}}{[1+e^{a_j(\theta_i-b_j)}]^2\{1-c_j+c_j[1+e^{a_j(\theta_i-b_j)}]\}} = \boxed{a_j^2}\left(\frac{1-P_j(\theta_i)}{P_j(\theta_i)}\right)\left(\frac{P_j(\theta_i)-c_j}{1-c_j}\right)^2$$

How to improve exposure balance?

- a-stratification with b-blocking (ASB)

at any given stage:

maximize $B_j(\hat{\theta}_i) = \dfrac{1}{|\hat{\theta}_i - b_j|}$



Motivation

How to use RT in item selection?

- maximizes the ratio of Fisher information to expected response time (MIT)

$$\mathrm{IT}_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)}$$

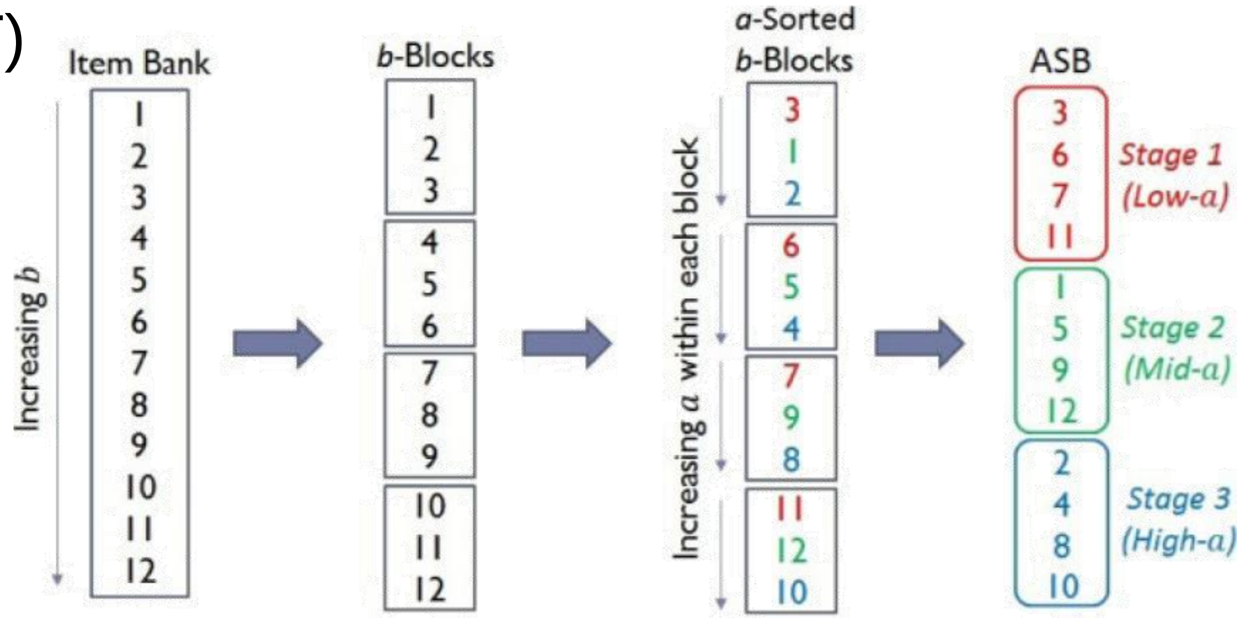
favors items with high information and low expected RTs

- a time-weighted version of ASB (ASBT)

at any given stage:

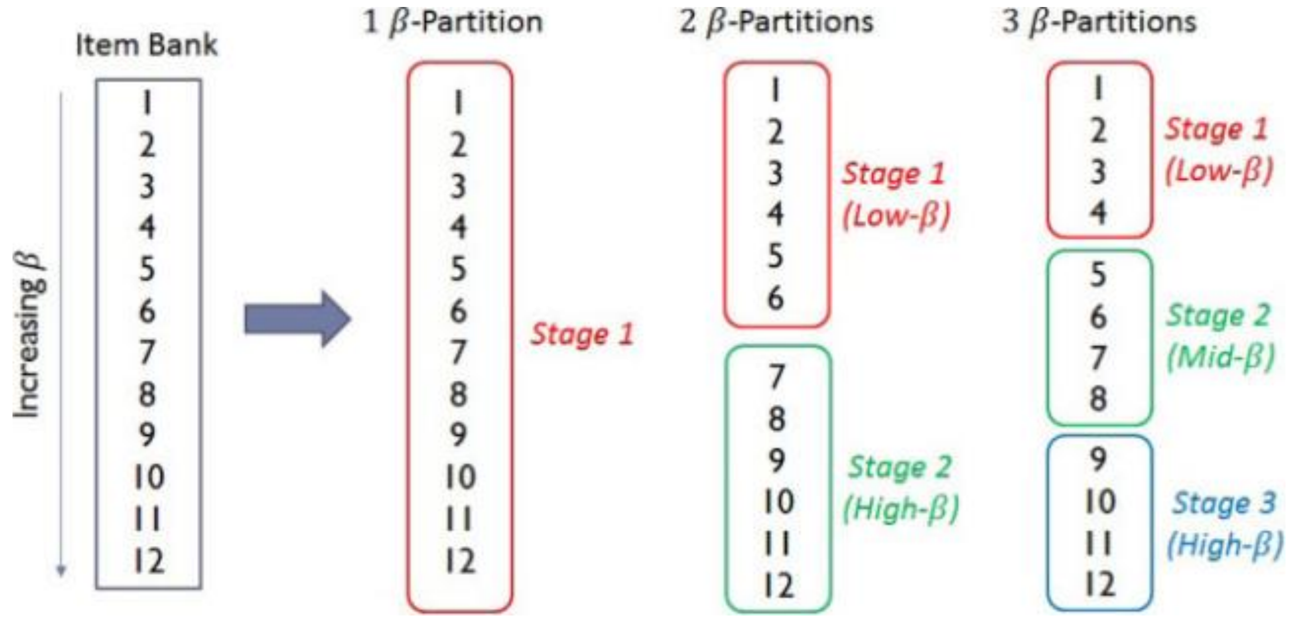
maximize $\mathrm{BT}_j(\hat{\theta}_i, \hat{\tau}_i) = \dfrac{B_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)}$

sacrifice the benefits of time weighting



Proposed Item Selection Procedures

1. β-partitioned MIT (BMIT)



→ select items from each successive stage using MIT

2. MI with β-matching (MIB)

$$\mathrm{IB}_j(\hat{\theta}_i,\, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{|\beta_j - \hat{\tau}_i|}$$

→

- **less restrictive** than perpetually selecting items with the lowest βj and highest αj
- **lower RT variability** across examinees $E(T_{ij}|\tau_i) = e^{\beta_j \cancel{- \tau_i} + 1/(2\alpha_j^2)}$

Proposed Item Selection Procedures

3. Generalized MIT (GMIT)

$0.5^{0.5} \approx$ **0.71 > 0.59** $\approx 0.5^{0.75}$?

vary the influence of the centered expected RT

In MIT: $\mathrm{IT}_j(\hat{\theta}_i, \hat{\tau}_i) = \dfrac{I_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)}$ VS $\mathrm{IT}_j^G(\theta_i,\ \tau_i) = \dfrac{I_j(\theta_i)}{|E(T_{ij}|\tau_i)\ -\ v|^{\boxed{w}}}, \quad \{v, w\} \in \mathbb{R}^2_{\geq 0}$

$$E(T_{ij}|\tau_i) = 0$$

$$e^{\beta_j - \tau_i + 1/(2\alpha_j^2)} = 0$$

the least time intensive items

substantial variability of testing times

$$E(T_{ij}|\tau_i) = v$$

$$e^{\beta_j - \tau_i + 1/(2\alpha_j^2)} = v$$

$$\beta_j + 1/(2\alpha_j^2) = \tau_i + \ln v$$

stabilize testing times

vary from person to person

Simulation studies

- investigate the performance of three new RT-informed criteria for item selection

  (under the hierarchical framework: 3PLM + lognormal models)

Item Selection Methods	
MI	Maximum information
MIT	MI with time
ASB	*a*-stratification with *b*-blocking
ASBT	ASB with time
MIB	MI with β-matching
BMIT	β-partitioned MIT
GMIT	Generalized MIT

Performance baseline: MI
Ideal item pool usage but worst accuracy: Random

– Study 1.

  hundreds of **simulations** were conducted with a broad range of parameter values

  ➡ two representative sets

– Study 2.

  further validate the effectiveness of GMIT

  ➡ real data (high-stakes)

Simulation studies

- Evaluation Criteria

1. RMSE

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}$$

$$\mathrm{RMSE}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2}$$

2. M and SD of testing times

$$\overline{\mathrm{tt}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{tt}_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j \in R_i} t_{ij}$$

$$s_{\mathrm{tt}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{tt}_i - \overline{\mathrm{tt}})^2}$$

3. M and SD of test overlap rates

$$\overline{\mathrm{tor}} = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n}\mathrm{tor}_{ii'} = \frac{n}{L(n-1)}\sum_{j=1}^{m}\mathrm{er}_j^2 - \frac{1}{n-1}$$

$$s_{\mathrm{tor}} = \sqrt{\left[\binom{n}{2} - 1\right]^{-1}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n}(\mathrm{tor}_{ii'} - \overline{\mathrm{tor}})^2}$$

Study 1

- Set 1
  - item parameters

$$(a_j^*,\, b_j,\, \boxed{\beta_j}) \sim \mathcal{N}_2[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1] \Rightarrow a_j^* = \log a_j$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.3 \\ 0.0 \\ \boxed{0.0} \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.25 \\ 0.00 & 0.25 & \boxed{0.25} \end{bmatrix}$$

$$c_j \sim \beta[2,\ 10]$$

$$\boxed{\alpha_j} \sim U[2,\ 4]$$

  - person parameters

$$(\theta_i,\, \boxed{\tau_i}) \sim \mathcal{N}_2[\boldsymbol{\mu}_2,\, \boldsymbol{\Sigma}_2]$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.25 \\ 0.25 & \boxed{0.25} \end{bmatrix}$$

- Set 2
  - item parameters

$$(a_j^*,\, b_j,\, \boxed{\beta_j}) \sim \mathcal{N}_2[\boldsymbol{\mu}_1,\, \boldsymbol{\Sigma}_1]$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.30 \\ 0.00 \\ \boxed{-0.25} \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.20 \\ 0.00 & 0.20 & \boxed{0.16} \end{bmatrix}$$

$$c_j \sim \beta[2, 10]$$

$$\boxed{\alpha_j} \sim U[0.5,\ 2.5]$$

  - person parameters

$$(\theta_i,\, \boxed{\tau_i}) \sim \mathcal{N}_2[\boldsymbol{\mu}_2,\, \boldsymbol{\Sigma}_2]$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 0.00 \\ 0.25 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.20 \\ 0.20 & \boxed{0.16} \end{bmatrix}$$

Study 1

- For each set
  - 500 items
  - 1000 examinees
  - 50 test length (first item chosen randomly)
  - Estimation: MLE + EAP (as an interim substitute)

- For ASBT
  - five strata of 100 items each (10 items each stage)

- For BMIT
  - One β-partition: equivalent to no β-partitioning
  - Two β-partitions: **low** 250 items (first 25); **high** 250 items (next 25)
  - Three β-partitions:

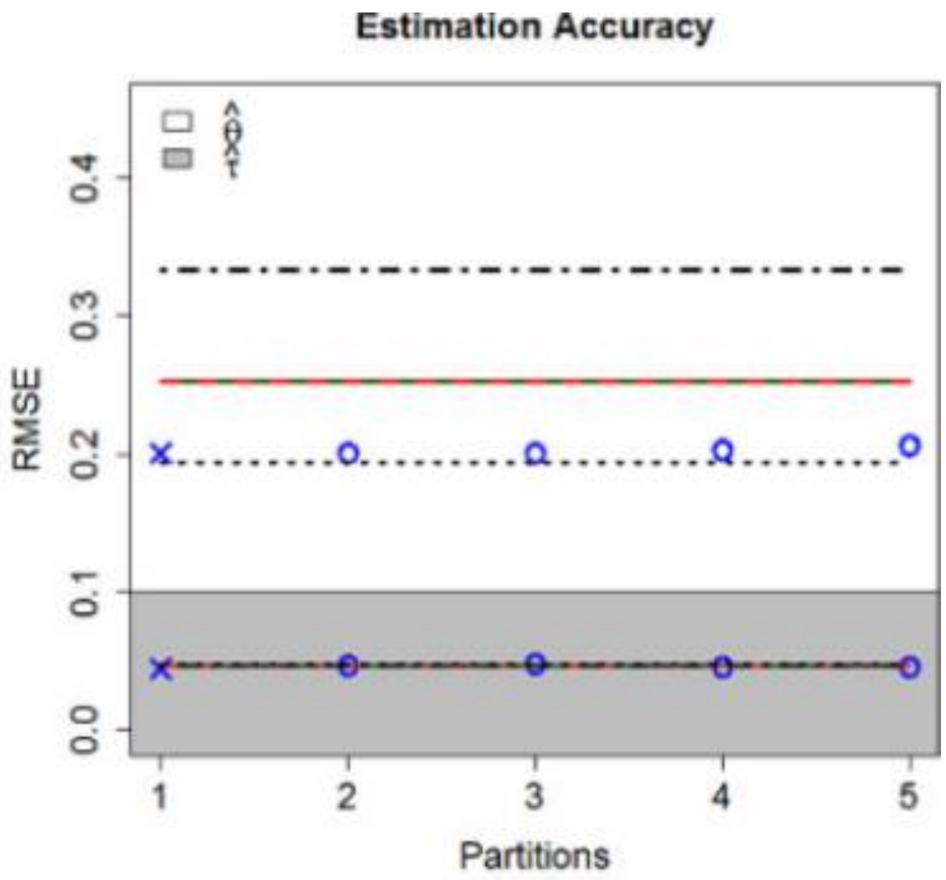
    low 167 items (first 17); **mid** 167 items (next 17); **high** 166 items (final 16)

- For GMIT
  - $V = \{0.0, 0.1, ..., 3.0\}$
  - $W = \{0.50, 0.75, 1.00\}$

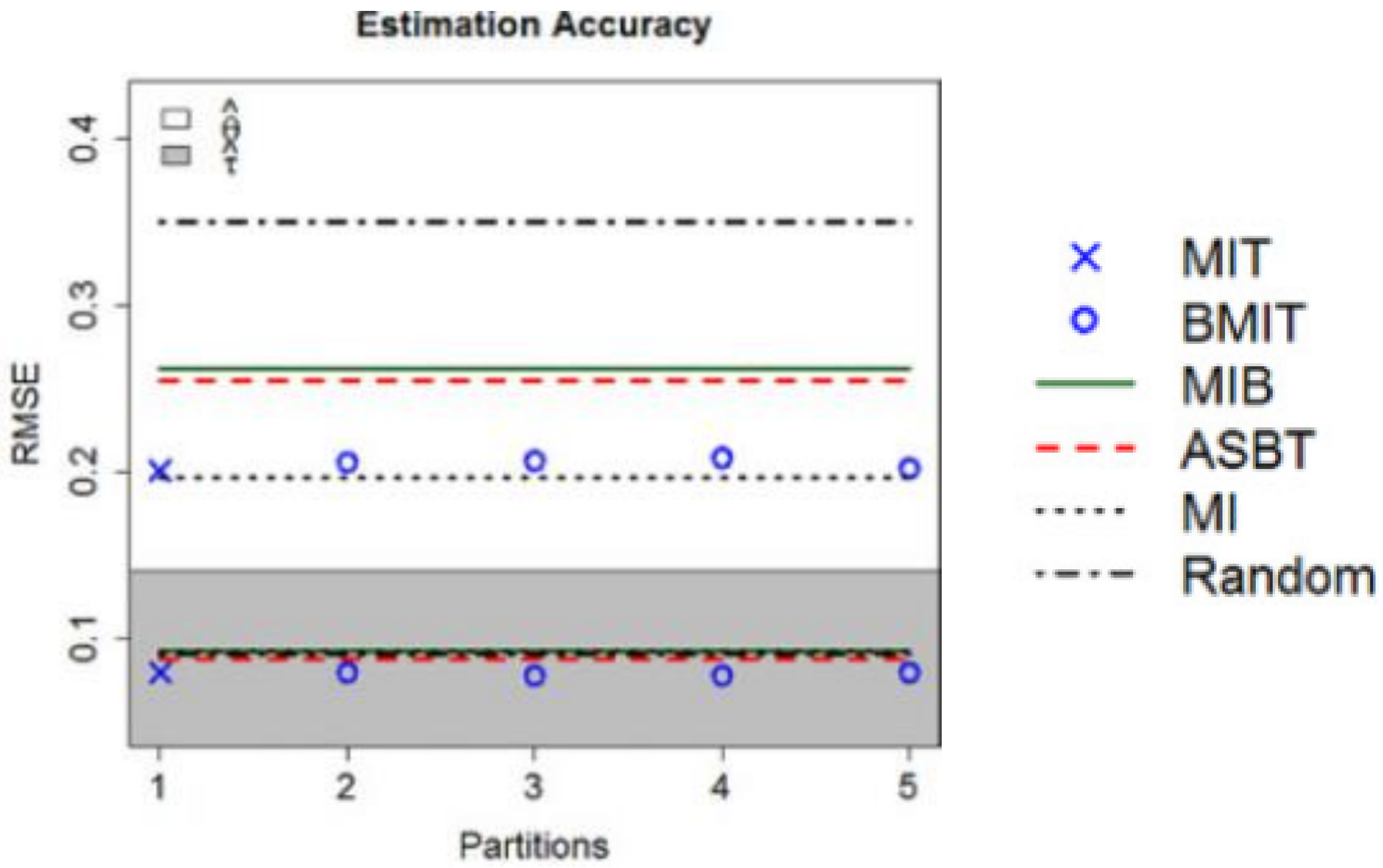
    $|V \times W| = 93$

Results – Study 1

Set 1



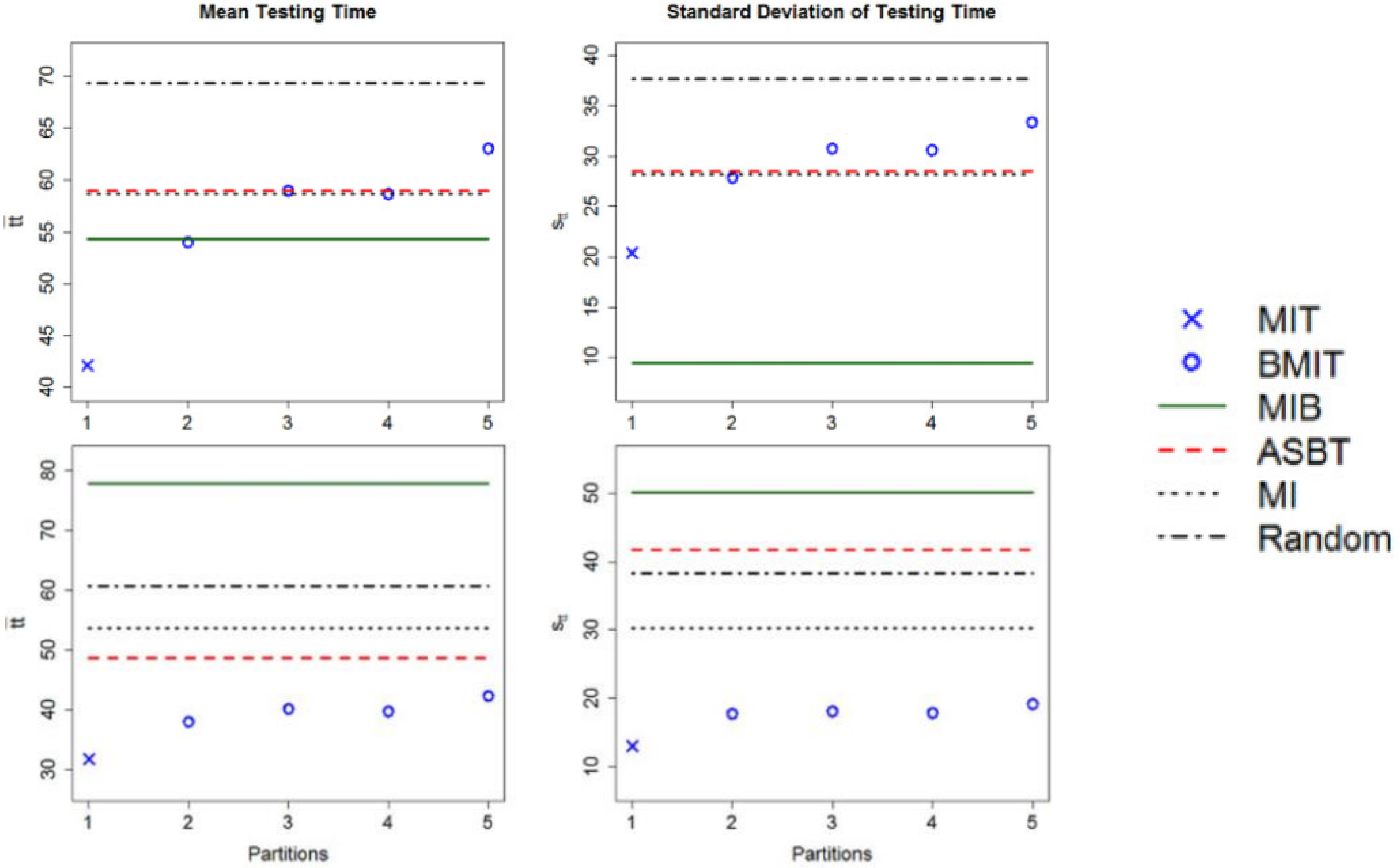
Set 2



Results – Study 1

Set 1

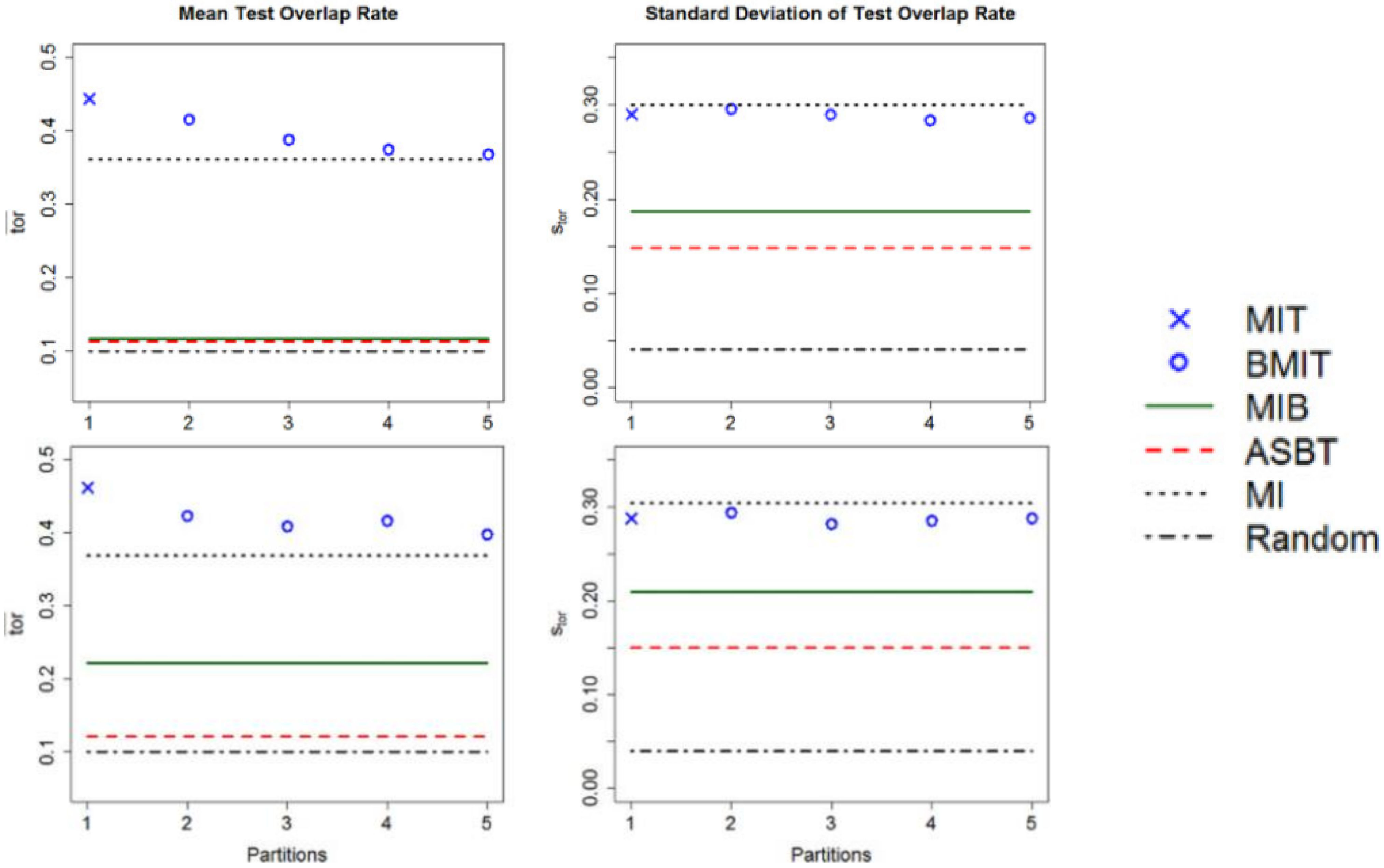
Set 2



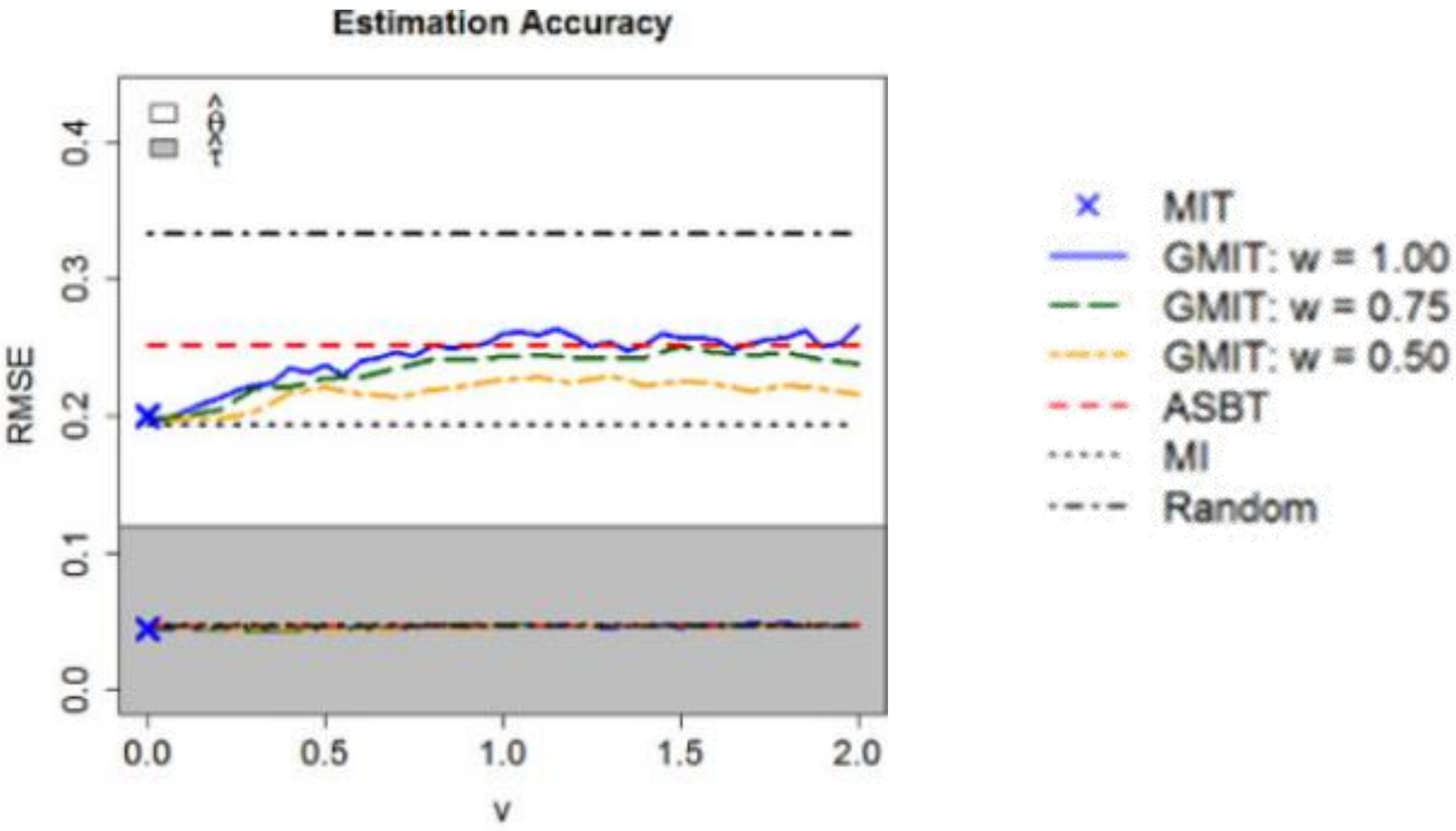
Results – Study 1

Set 1

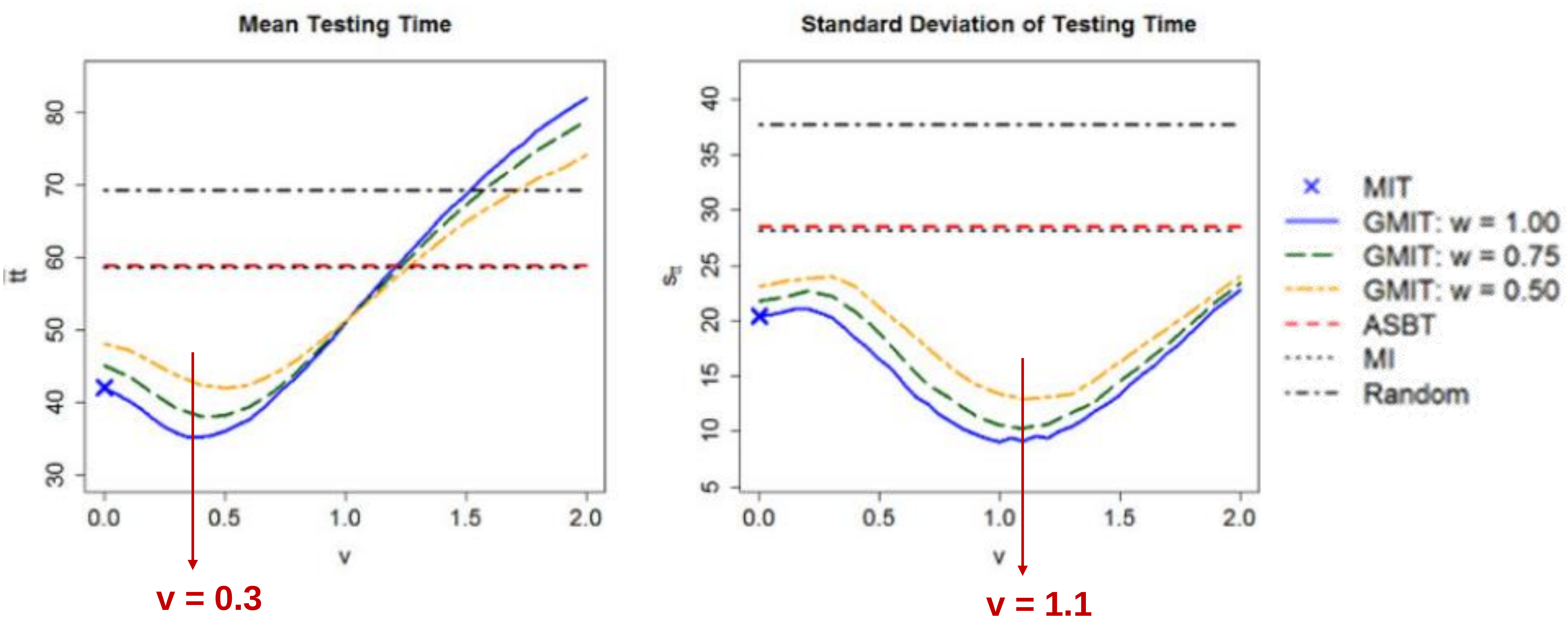
Set 2



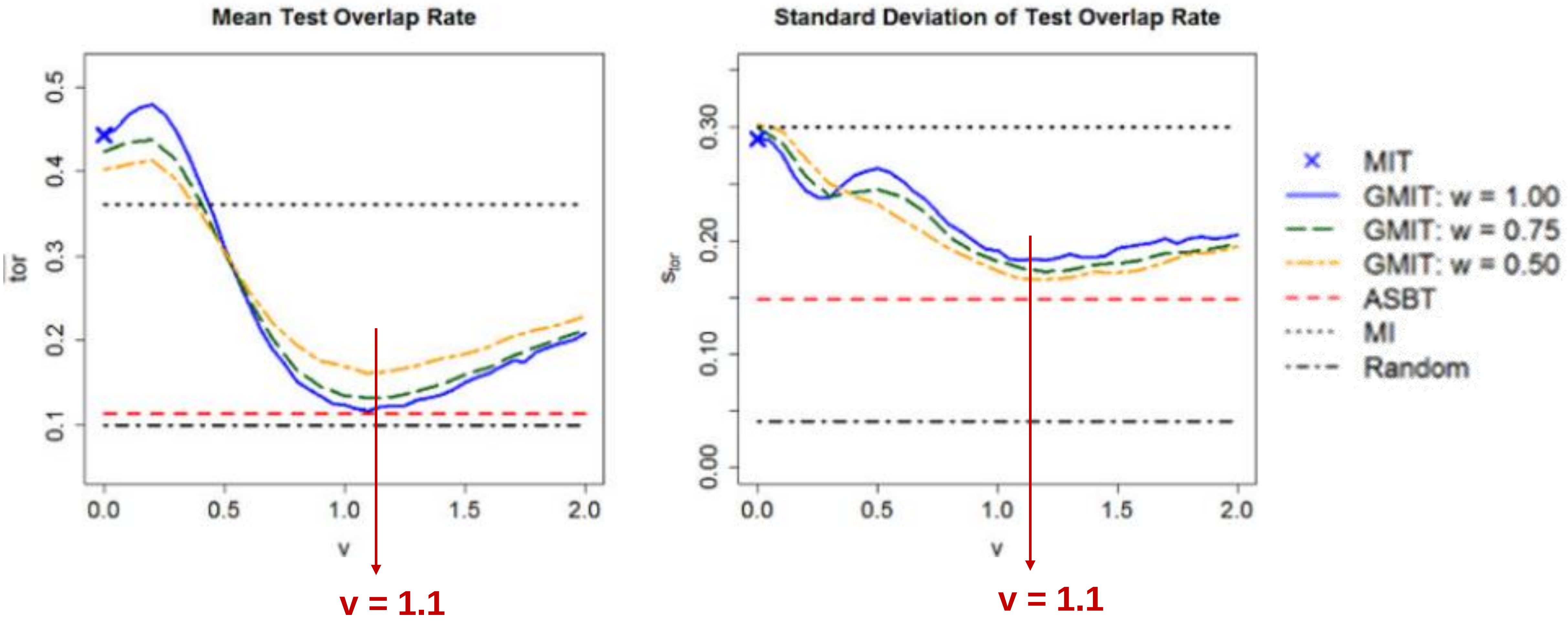
Results – Study 1



Results – Study 1



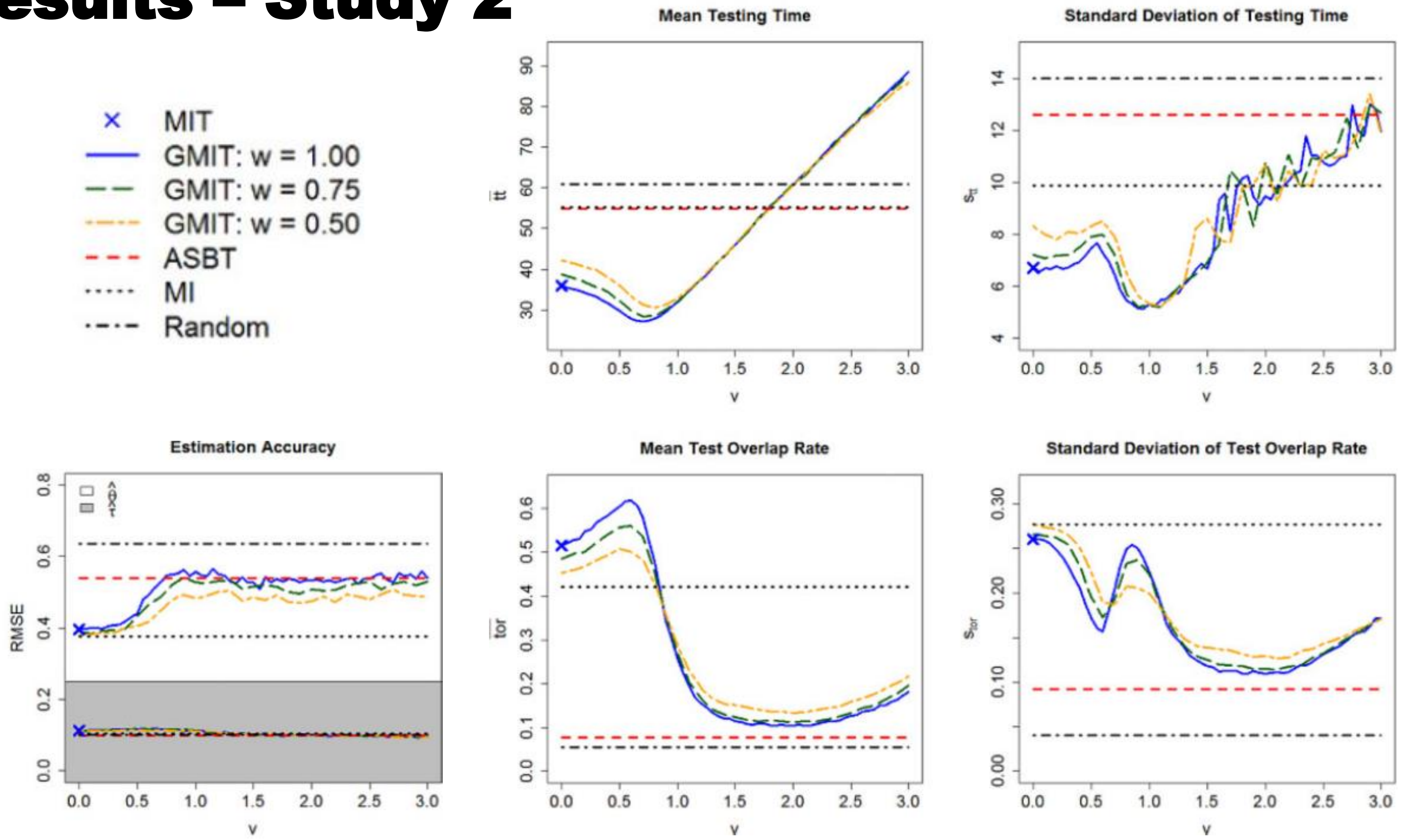
Results – Study 1



Study 2

- real data from a high-stakes, large-scale standardized CAT
  - 2000 examinees
  - item pool:
    500 multiple-choice items (3PLM)
  - α & β:
    a modified version of van der Linden’s (2007) MCMC routine
    ➡ fixed a, b, c to the precalibrated values, and mean(τ) = 0
  - 30 test length (first item chosen randomly)
  - Estimation: MLE + EAP (as an interim substitute)
- For ASBT
  - five strata of 100 items each (6 items each stage)

Results – Study 2



Discussion

- provide strong evidence for the overall superiority of GMIT
  - **increase the validity of test scores**
    - ✓ markedly reducing the mean and variance of testing times
      - ➡ curtail the likelihood of time pressure–induced rapid guessing
    - ✓ dramatically reducing the mean and variance of test overlap rates
      - ➡ decrease the chances of item preknowledge
  - **the truly remarkable feature:**
    - ✓ **without imposing** explicit item exposure **controls** or RT **constraints**

Discussion

- the initialization of GMIT for **use in practice**:

  1. calibrating the item pool
  2. generating examinees
  3. establishing a set of evaluation criteria
  4. conducting a series of CAT simulations with a range of v and w values
  5. selecting the optimal {v, w}

  – **two or more criteria:**

    depend on the minimally acceptable levels

    the user’s rational judgment

Discussion

- the initialization of GMIT for use in practice:
  - **objective measure:**

$$\Omega_{\{v,w\}} = \boldsymbol{\gamma}^T \mathbf{Z}_{\{v,w\}}, \quad \{v, w\} \in V \times W$$

a weighted average of the standardized criteria

(if the values of γ are nonnegative and sum to 1)

Rank	$\{v, w\}$	$\Omega_{\{v,w\}}$
1	$\{1.4, 0.75\}$	$-.4746$
2	$\{1.5, 1.00\}$	$-.4537$
3	$\{1.5, 0.75\}$	$-.4436$
4	$\{1.4, 0.50\}$	$-.4182$
5	$\{1.3, 1.00\}$	$-.4070$
6	$\{1.6, 0.50\}$	$-.4027$
7	$\{1.6, 0.75\}$	$-.3935$
8	$\{1.3, 0.75\}$	$-.3865$
9	$\{1.9, 0.75\}$	$-.3758$
10	$\{1.3, 0.50\}$	$-.3708$
⋮	⋮	⋮
93	$\{3.0, 0.75\}$	$.5055$

placed more emphasis on ability estimation accuracy

Future directions

- implementation and evaluation under **a wide variety of schemes**
- confirm the usefulness of the technique **in operational CAT**
- compare GMIT to **other RT-based methods** not considered in this article
- β-partitioning may have potential in **substantive applications**

Thanks for Listening!

Reporter: Yingshi Huang