The Expected Likelihood Ratio in CCT

Steven W. Nydick
April 9, 2016

Problems with Classification Testing

The problem in computerized classification testing (CCT):

How do we efficiently determine whether or not an examinee exceeds some cut-point, θ_0*, in the fewest number of items with a pre-specified accuracy rate?*

This is usually thought of as a **stopping-rule** issue.

- Stopping rule? Directly affects accuracy and efficiency.
- Item selection algorithm? Mostly consensus.
  - Select items at the cut-point separating categories.
  - Fisher information, KL divergence, mutual information.

Preliminaries

Assume the following for the remainder:

1. Items fit the unidimensional 3PL IRT model.

$$p_j(\theta_i) = \Pr(Y_{ij} = 1 | \theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-a_j(\theta_i - b_j)]},$$

2. Decisions are only mastery vs. non-mastery.

$$\mathrm{H}_0 : \theta_i = \theta_l = \theta_0 - \delta$$
$$\mathrm{H}_1 : \theta_i = \theta_u = \theta_0 + \delta,$$

3. Tests are variable-length with the SPRT decision rule.

The Sequential Probability Ratio Test

A commonly used stopping rule: The SPRT (e.g., Wald, 1947).

❶ Determine *simple* statistical hypotheses (Eggen, 1999):

$$\mathsf{H}_0 : \theta_i = \theta_l = \theta_0 - \delta$$
$$\mathsf{H}_1 : \theta_i = \theta_u = \theta_0 + \delta,$$

❷ Calculate log-likelihood ratio comparing the hypotheses.

$$\log\left[\mathsf{LR}(\theta_u, \theta_l | \mathbf{y}_i)\right] = \log\left[\frac{L(\theta_u|\mathbf{y}_i)}{L(\theta_l|\mathbf{y}_i)}\right]$$

❸ End test if log-likelihood ratio exceeds threshold.

How should we select subsequent exam items?

- Common knowledge: At the cut-point (θ_0).
- Are there alternative options?

Cut-Point Complications

Complication 1: Given a correct response, SPRT evidence depends on the model asymptote (Nydick, 2014).

The maximum of the log-LR given a correct response:

$$\hat{\theta}_0 = \frac{\log(c_j)}{2a_j} + b_j.$$

1. If $c_j = 0$, then more difficult items yield more evidence.
2. If $c_j > 0$, then more difficult items can yield less evidence.

How can this inform which items to select?

Cut-Point Complications

Complication 2: The expected increase in SPRT evidence depends on a person's true ability (Nydick, 2014).

The Expected log-Likelihood Ratio (ELR):

$$\mathbb{E}\left[\log\left[\text{LR}(\theta_u, \theta_l | Y_{ij})\right]\right] = p_j(\theta_i) \log\left[\frac{p_j(\theta_u)}{p_j(\theta_l)}\right] + [1 - p_j(\theta_i)] \log\left[\frac{1 - p_j(\theta_u)}{1 - p_j(\theta_l)}\right].$$

1. The ELR indicates the expected *increase* in the SPRT.
2. The ELR is dependent entirely on the IRT model and stopping rule.

Cut-Point Complications

Complication 2: The expected increase in SPRT evidence depends on a person's true ability (Nydick, 2014).

Which item maximizes the ELR?

- Assume fixed and constant *a*.
- Assume $c_j = 0$ for all items.

Then given a small δ, we find that

$$\lim_{\delta \to 0^+} \hat{b} = \frac{\theta_0 + \theta_i}{2}.$$

What would happen if we selected items at $\frac{\theta_0 + \hat{\theta}_i}{2}$?

Preliminary Simulation Method

Three item selection algorithms:

❶ FI-bound (the “recommended” algorithm).

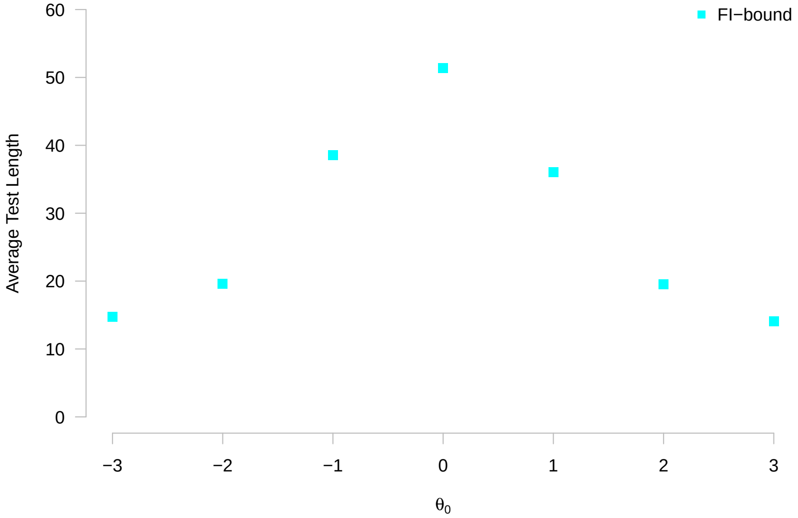
❷ FI-ability (the “not-recommended” algorithm).

❸ FI-middle (a new option).

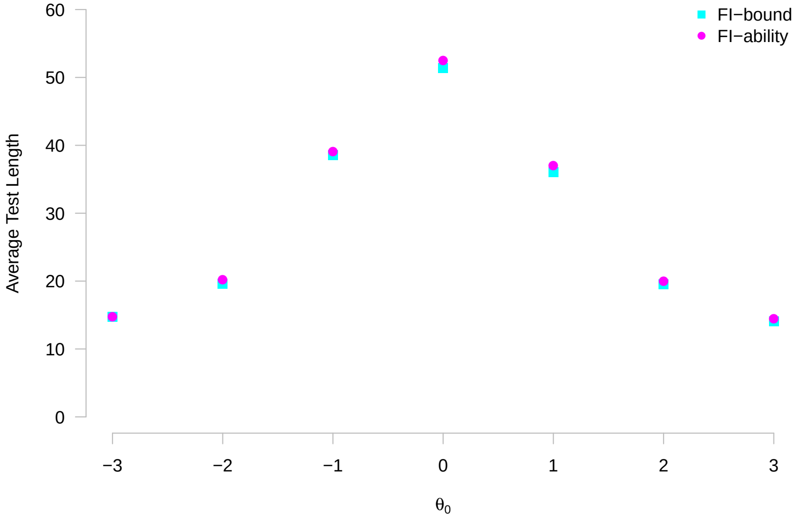
Other specifications of this simulation:

- 10,000 simulees with $\theta \sim N(0, 1)$.
- 750 size item bank according to the 2PL IRT Model
- Classification bounds at $\theta_0 \in \{-3, -2, -1, 0, 1, 2, 3\}$.
  - Results aggregated across all simulees at a bound.

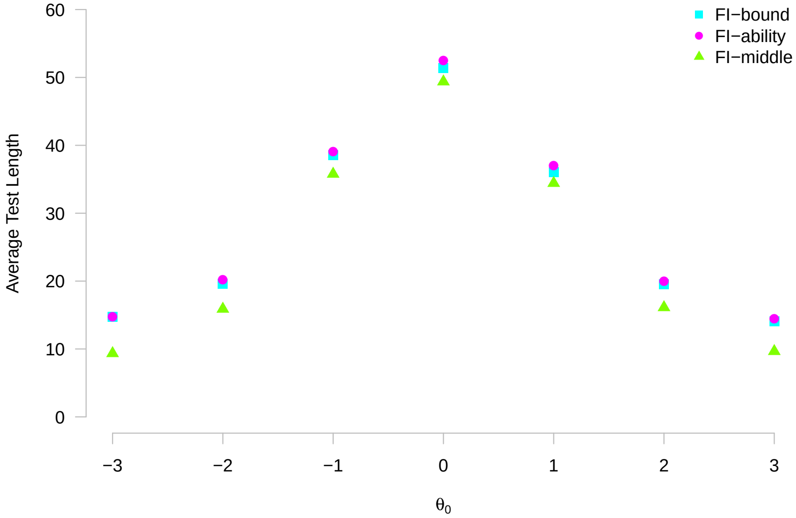
Preliminary Simulation Results: Length



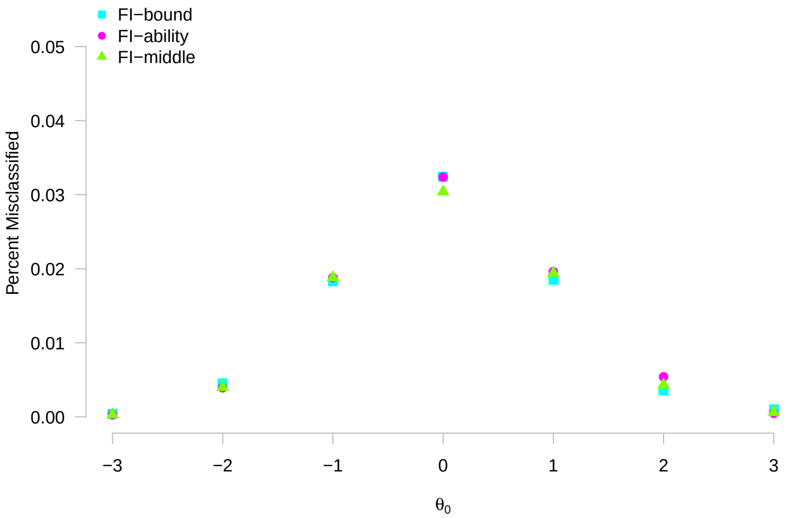
Preliminary Simulation Results: Length



Preliminary Simulation Results: Length



Preliminary Simulation Results: Accuracy



Fixed Specifications

❶ Latent Trait

- $N = 10,000$
- $\theta \sim N(\mu = 0, \sigma = 1)$

❷ Classification Bounds

- $\theta_0 \in \{-3, -2, -1\}$
- $\theta_0 = 0$
- $\theta_0 \in \{+1, +2, +3\}$

❸ Stopping Rules

- SPRT
  - $j_{\text{min}} = 5$
  - $j_{\text{max}} = 200$
  - $\delta = 0.1$
  - $\alpha = \beta = .05$

Item Banks and IRT Models

❶ Size of Item Bank

- $J = 750$
- $J = 1,500$

❷ 3PL IRT Model

- b-parameters
  - $b \sim \text{Unif}(\text{min} = -4, \text{max} = 4)$ (Flat)
  - $b \sim N(\mu = 0, \sigma = 1.500)$ (Moderate)
  - $b \sim N(\mu = 0, \sigma = 0.707)$ (Peaked)
- c-parameters
  - $c = .25$ (Fixed)
  - $c = .00$ (None)
  - $c \sim \text{Beta}(\alpha = 19.8, \beta = 79.2)$ (Random)
- a-parameters
  - $a \sim \log N(\mu_{\text{log}} = 0.38, \sigma_{\text{log}} = 0.25)$

Item Selection Algorithms

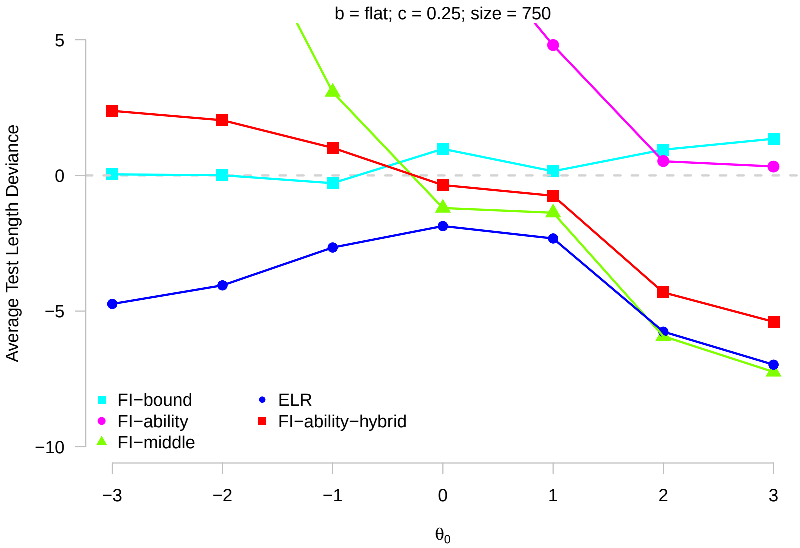
1. FI-bound
2. FI-ability
3. FI-middle
4. KL-bound
    - In paper *only*.
5. KL-estimated
    - In paper *only*.
6. ELR
7. FI-ability-hybrid
    - FI-ability until $s_{\hat{\theta}_i} < .5$.
    - ELR for remainder of exam.
8. KL-estimated-hybrid
    - In paper *only*.

Misc and Conditions Table

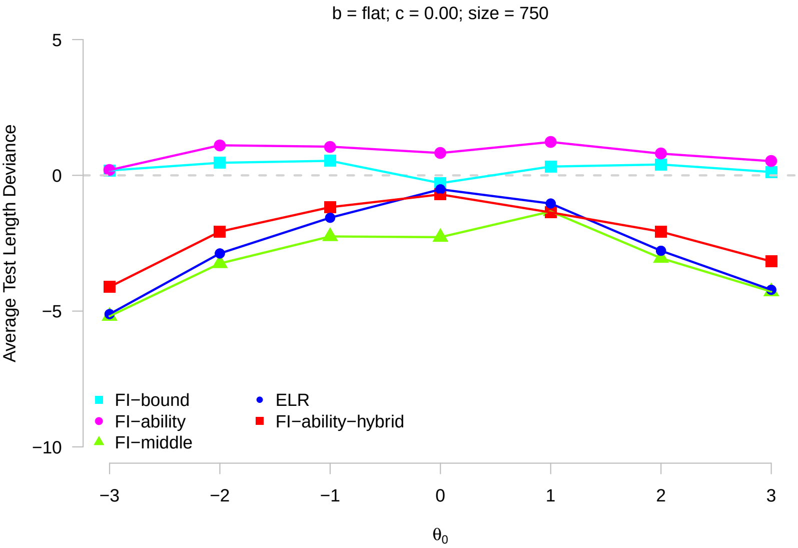
Conditions Table

Variable	Number of Conditions
b	3 (Flat, Moderate, Peaked)
c	3 (None, Fixed, Random)
J	2 (750, $1,500$)
θ_0	7 ($-3, -2, \ldots, 3$)
Item Selection	8
Overall	1008

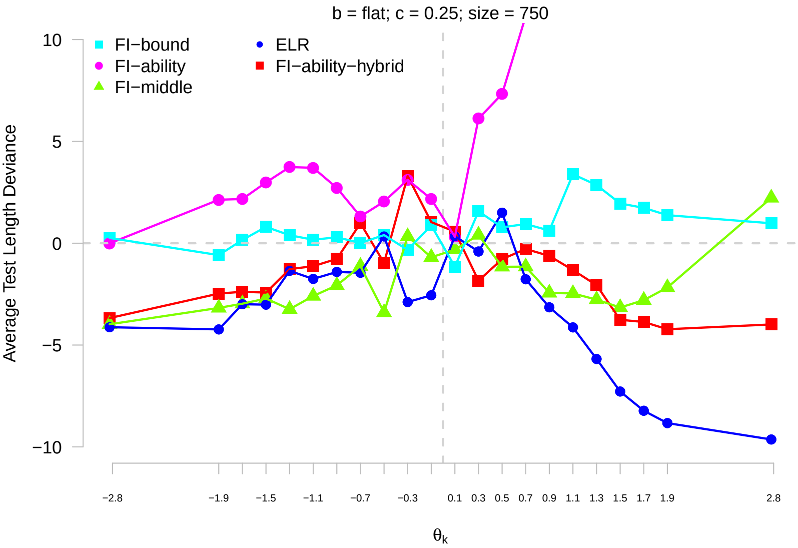
Overall ($J = 750; c = .25$): Length



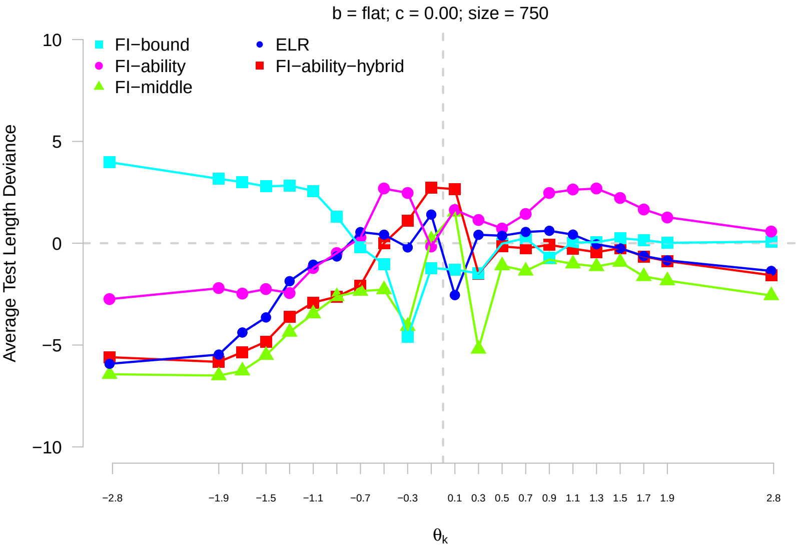
Overall ($J = 750; c = .00$): Length



Conditional ($J = 750; c = .25$): Length



Conditional ($J = 750; c = .00$): Length



Summary of Results

What are the answers to the following questions?

1. **Do different item selection algorithms perform differently for various cut-points relative to the ability distribution?**
2. Do different item selection algorithms yield different average test lengths for different groups of simulees?
3. Can we decrease test length by considering ability as well as the classification bound in CCT item selection?
4. Should we build tests by selecting items with difficulty close to the classification bound?

Yes. Maximizing Fisher information based on the ability estimate works worse if $c > 0$ and low θ_0.

Summary of Results

What are the answers to the following questions?

1. Do different item selection algorithms perform differently for various cut-points relative to the ability distribution?
2. **Do different item selection algorithms yield different average test lengths for different groups of simulees?**
3. Can we decrease test length by considering ability as well as the classification bound in CCT item selection?
4. Should we build tests by selecting items with difficulty close to the classification bound?

Yes. Bound-based algorithms performed better near the bound. Modified algorithms performed better elsewhere.

Summary of Results

What are the answers to the following questions?

❶ Do different item selection algorithms perform differently for various cut-points relative to the ability distribution?

❷ Do different item selection algorithms yield different average test lengths for different groups of simulees?

❸ **Can we decrease test length by considering ability as well as the classification bound in CCT item selection?**

❹ Should we build tests by selecting items with difficulty close to the classification bound?

Yes. ELR and FI-middle yielded the shortest tests for most classification bounds, item banks, and simulees.

Summary of Results

What are the answers to the following questions?

❶ Do different item selection algorithms perform differently for various cut-points relative to the ability distribution?

❷ Do different item selection algorithms yield different average test lengths for different groups of simulees?

❸ Can we decrease test length by considering ability as well as the classification bound in CCT item selection?

❹ **Should we build tests by selecting items with difficulty close to the classification bound?**

Probably not. The most efficient tests would have items with a (relatively) wide distribution of difficulties.

Extensions

How can we better consider uncertainty in θ?

$$\text{ELR}_j(\theta|w_{ij}) = \int_\Theta w_{ij}\text{ELR}_j(\theta)d\theta$$

- Posterior ELR (or FI-middle)
  - $w_{ij} = \pi(\theta|\mathbf{y}_{i,j-1})$
- Likelihood-weighted ELR (or FI-middle)
  - $w_{ij} = L(\theta|\mathbf{y}_{i,j-1})$

How do these results generalize?

- Polytomous models
- Multidimensional models
- Alternative stopping rules
- Curtailment

Thank You!

References

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–260.

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of Educational and Behavioral Statistics*, *39*, 203–230.

Wald, A. (1947). *Sequential analysis*. New York, NY: John Wiley.