Why should we worry about low examinee effort?

1

Section Learning Objectives

1 Why should we worry about low examinee effort?

Understand the prevalence of low examinee effort

Identify examinee/item characteristics associated with low effort

Articulate ways that low effort can affect parameter estimation

Articulate ways that low effort can affect common test-based inferences

Full Effort an Implicit Validity Assumption

- Valid uses of test scores (e.g., identifying struggling students, supporting learning needs in schools) implicitly assumes that examinees are providing full effort.

- If assumption is not met, validity can be undermined.

- According to the APA GUIDELINES for Psychological Assessment and Evaluation:

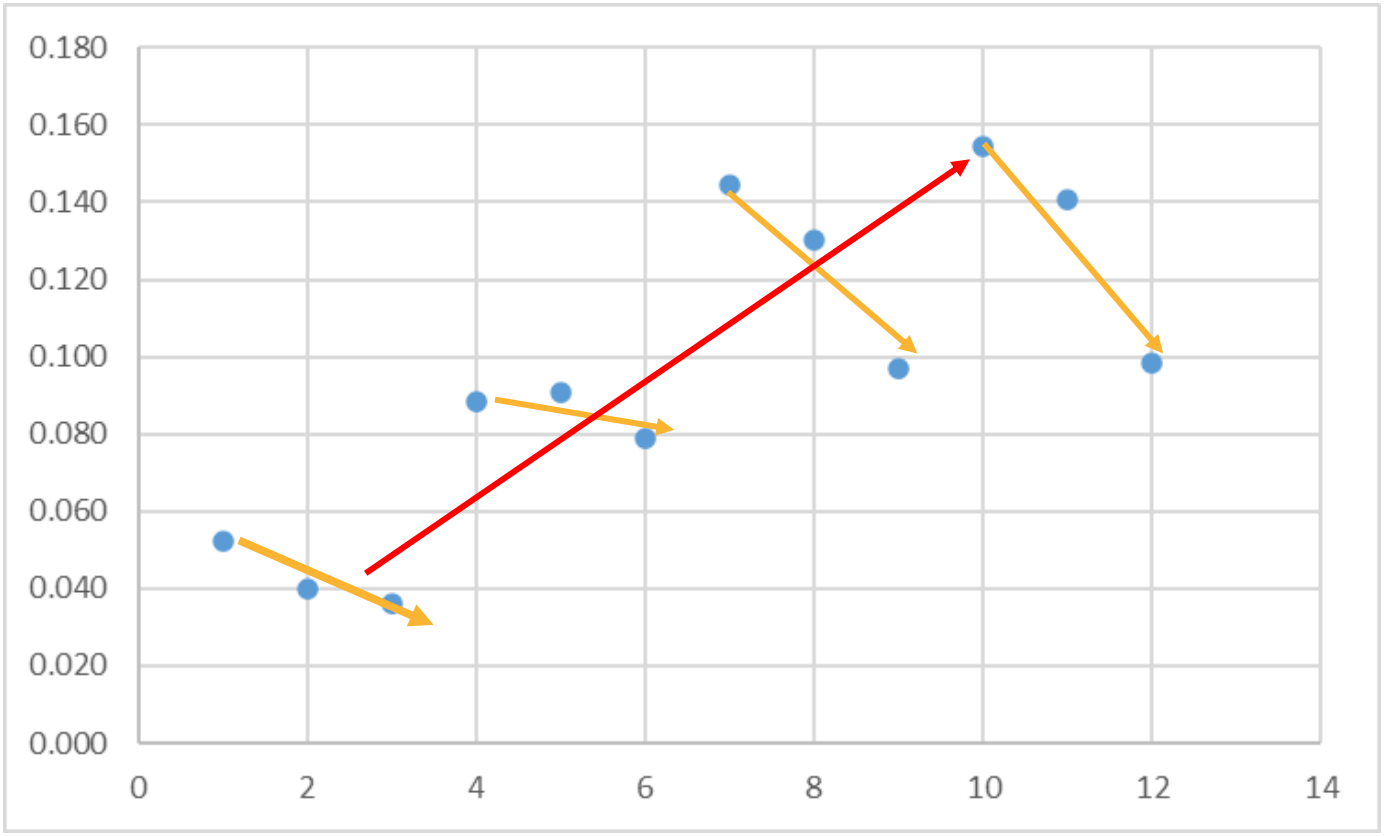
Examinees may underperform for many reasons, and not adequately assessing effort limits the interpretation of test results. Without systematically assessing effort, it becomes difficult to discern if variability and patterns of test results reflect actual performance or the influence of low effort, motivation, or some other factor besides ability. (Page 17).

Why might an examinee not try hard?

- The task at hand is too demanding
- Mental fatigue sets in as the test progresses
- The value of trying hard has not been made clear
- Students are disengaged not only on the test, but also in school
- Stereotype threat, anxiety, lack of self-belief

Low Effort Quite Common

Prop. Examinees with low enough effort to distort test scores



Grade 5 **Grade 8**

Low Effort Can Bias IRT Parameter Estimation

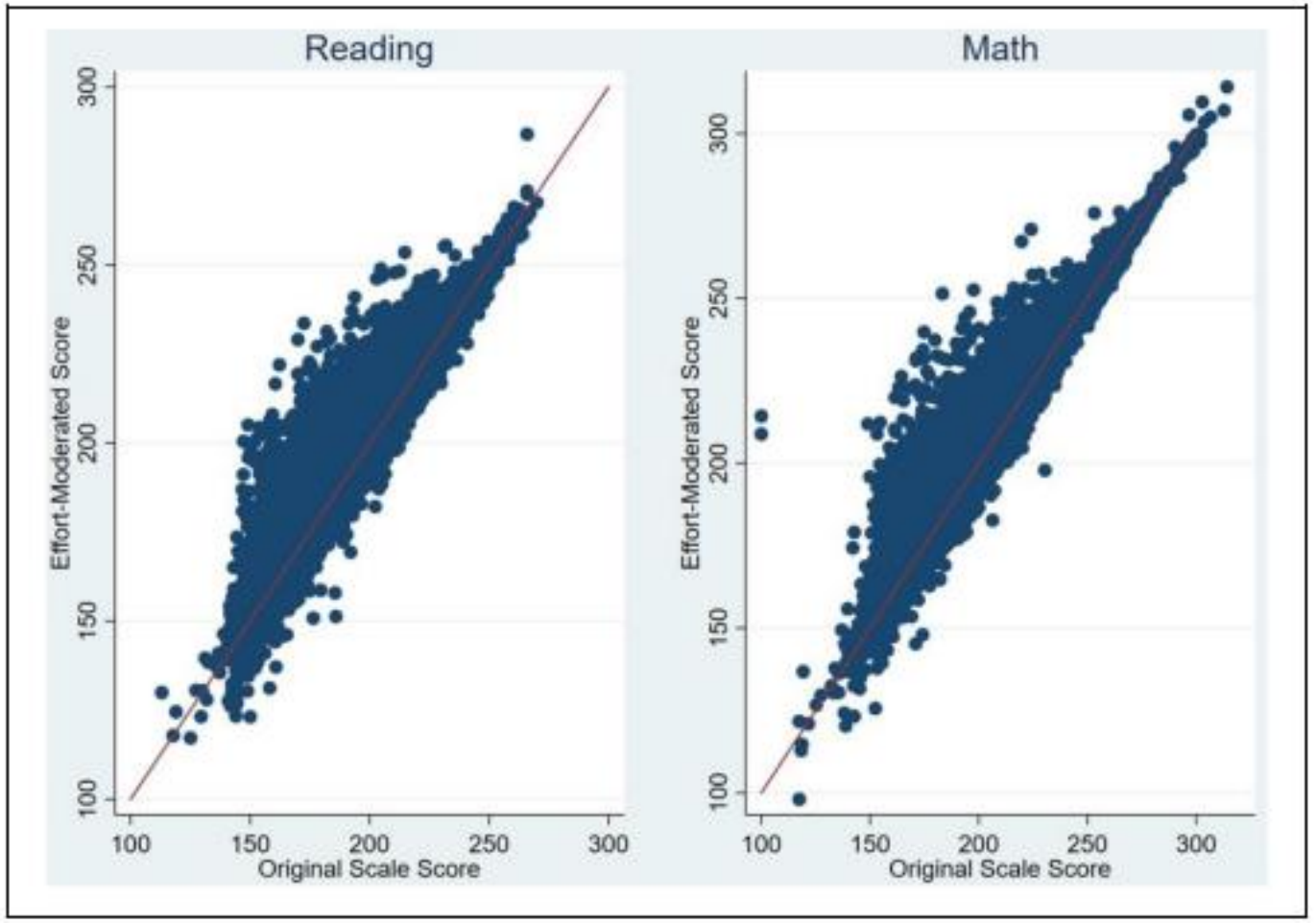
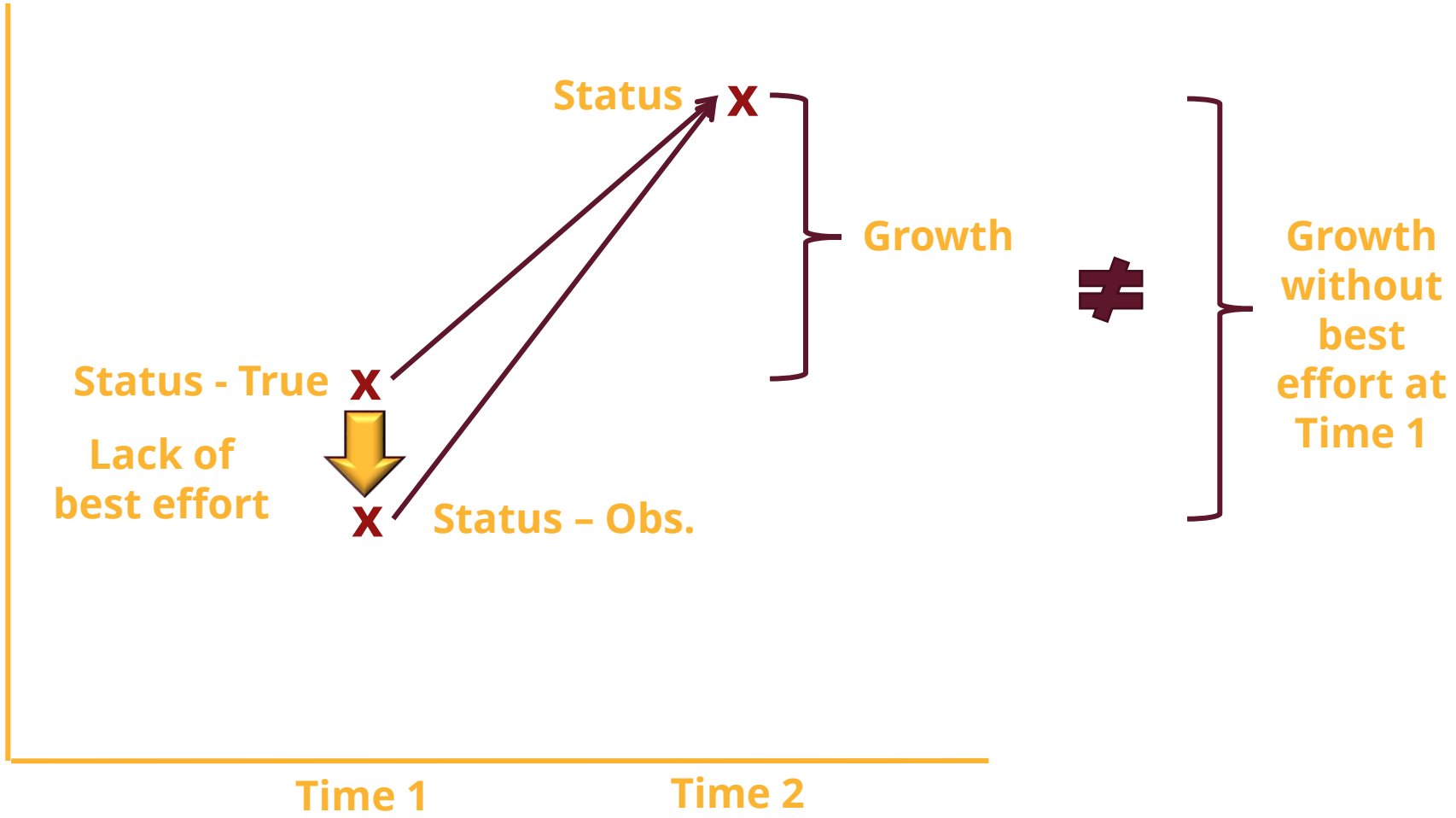


Figure 3. Rasch unit score comparison by model.

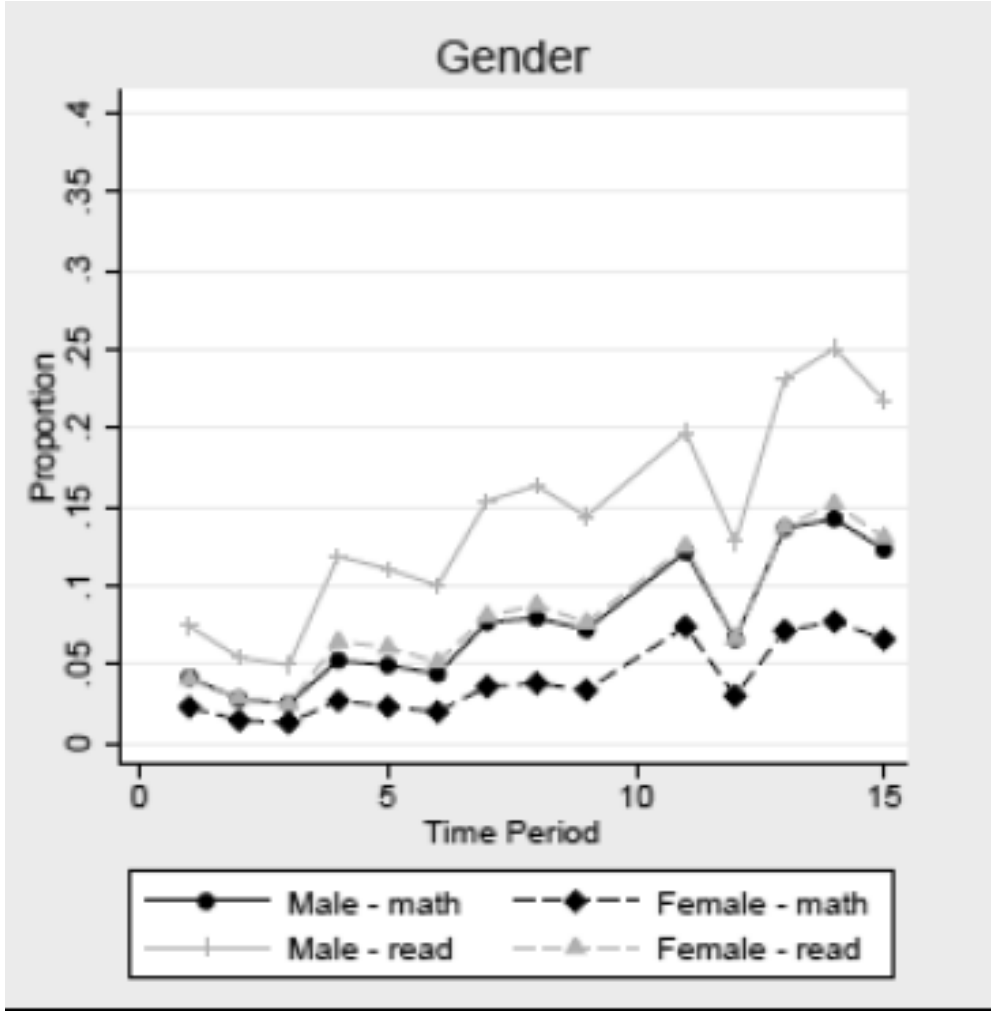
Note. An identity line is included to show where students with identical scores would fall.

Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: Can the Effort-Moderated IRT model serve as a solution?. *Applied Psychological Measurement*, *45*(6), 391-406.

Can Create Real Problem for Inferences at Examinee Level

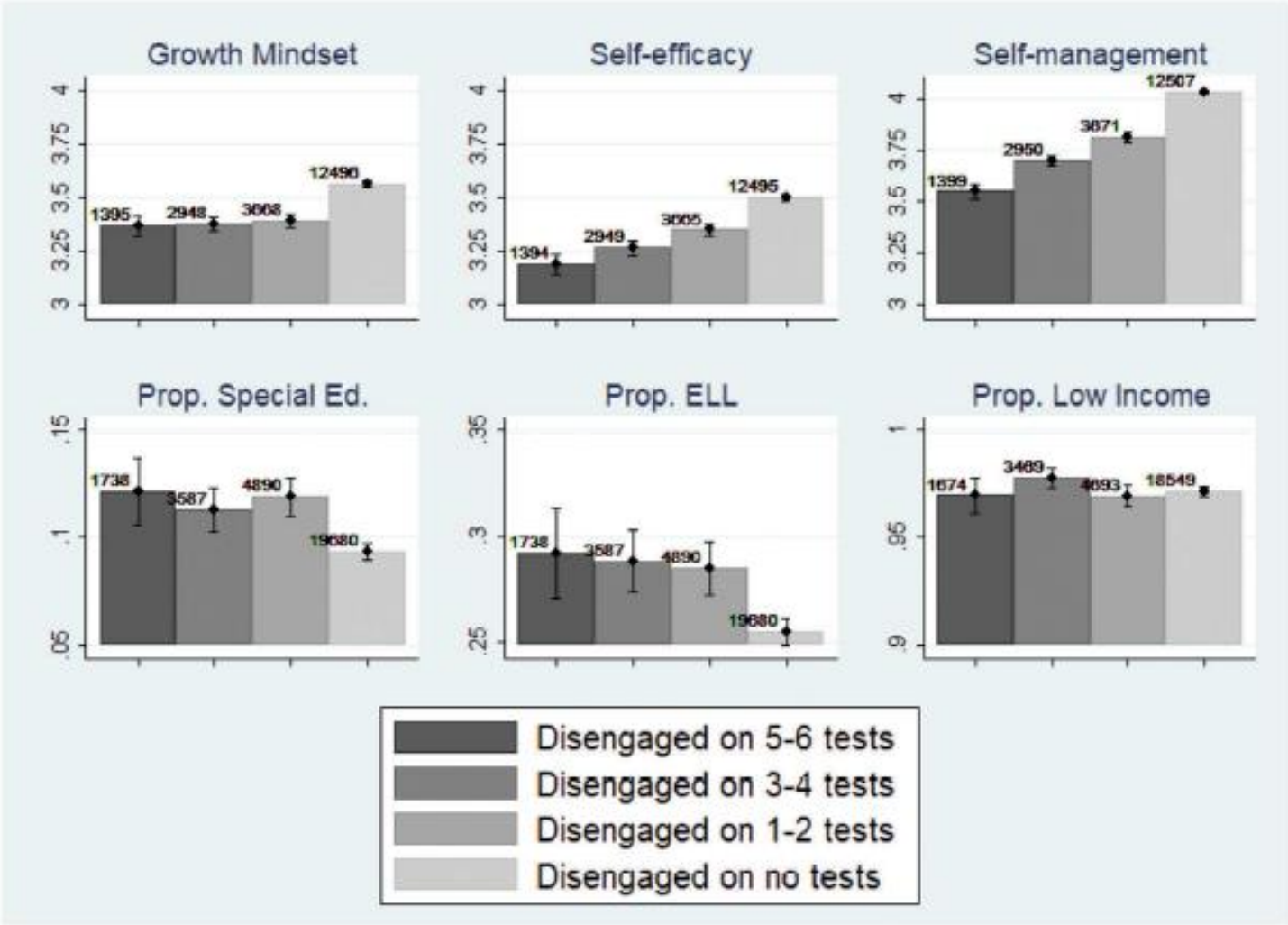


Effort Can Introduce Additional Issues for Aggregate Inferences



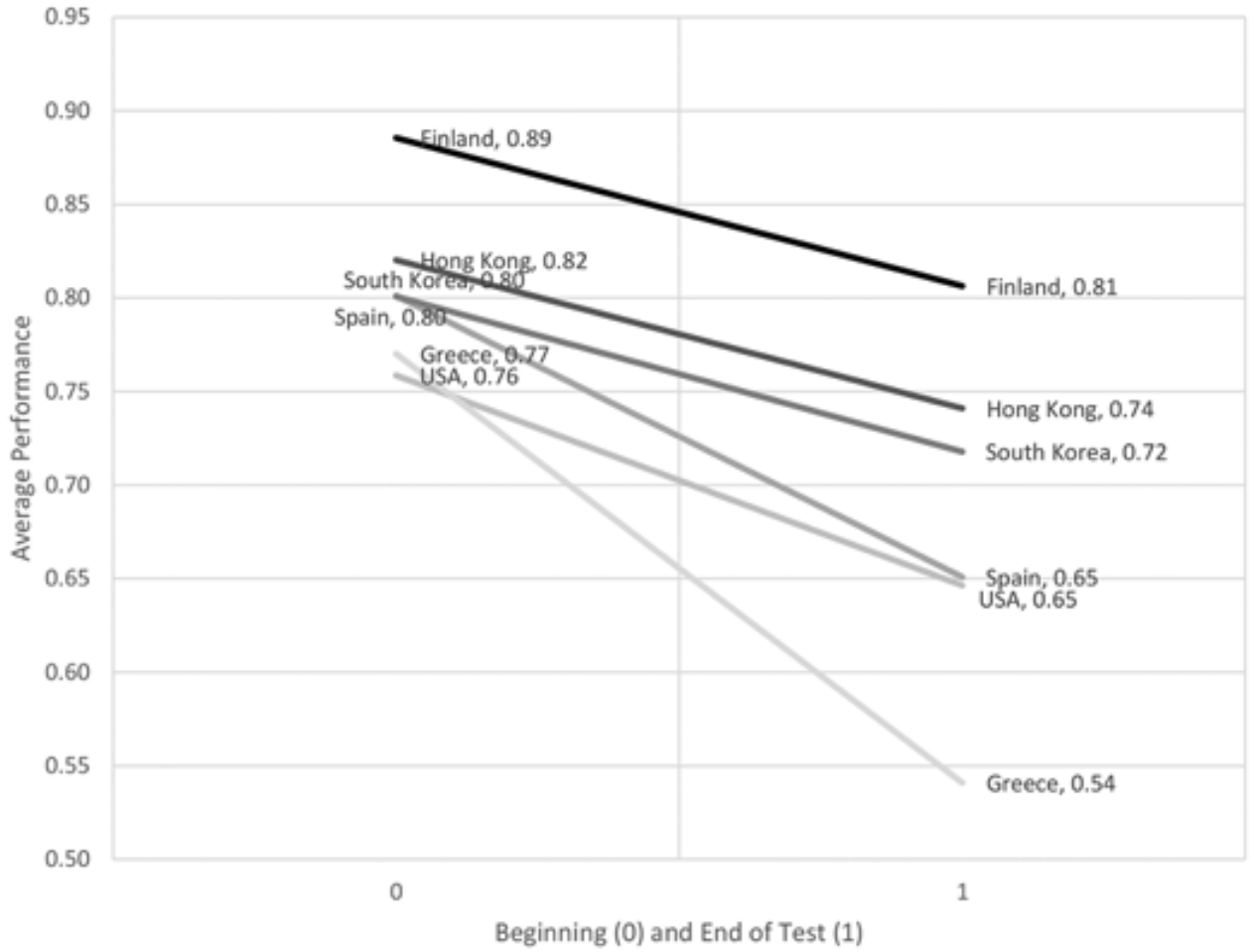
Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12), 1-26.

Effort Associated with Student Characteristics



Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment*, *24*(4), 327-342.

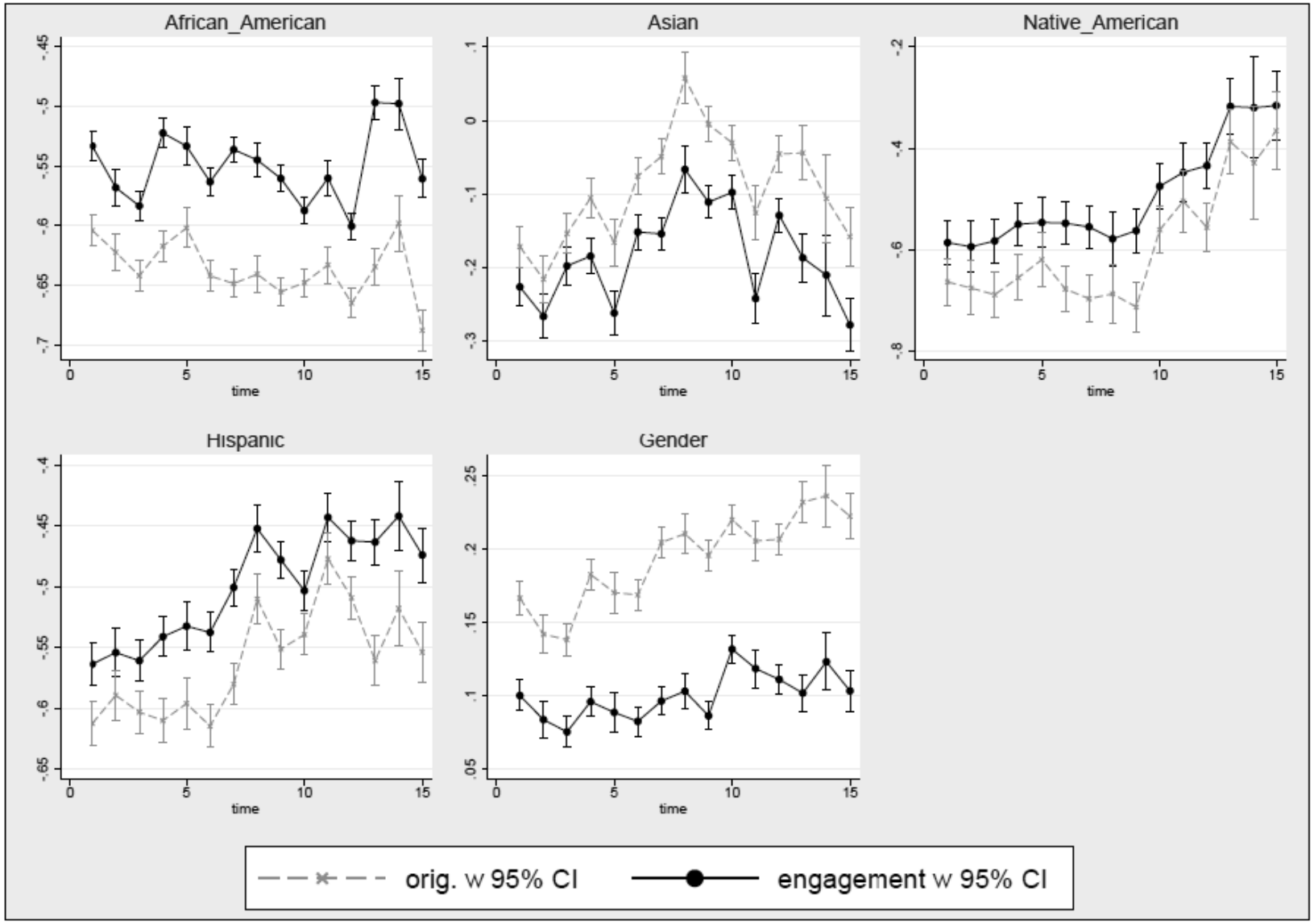
Effort Associated with Item and Test Features



Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, *13*(4), 519-552.

Undermine Student-level Inferences

Figure 3. Reading gaps in achievement conditional and unconditional on test effort



Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12), 1-26.

Undermine School-level Inferences

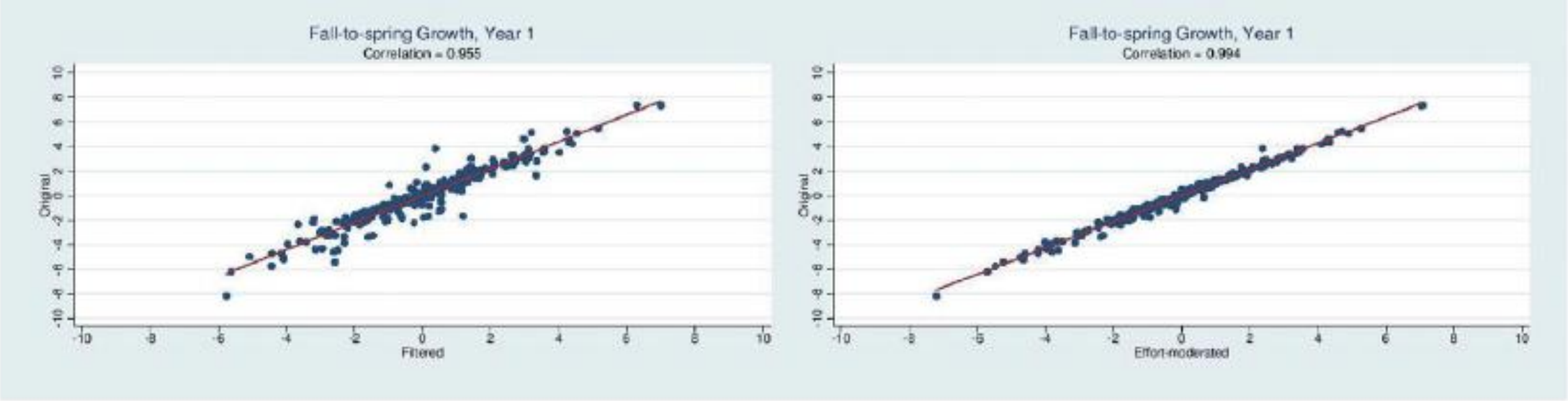


Figure 1. Comparisons of empirical Bayes estimates of school effectiveness.

Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, *13*(1), 147-175.

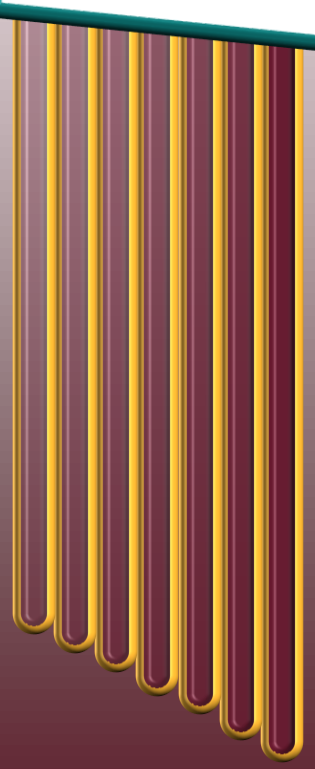
Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 40*(2), 267-284.

Low Effort Can Be Huge Threat to Validity

- Low examinee effort is
  - Common, especially in grades 6+
  - Associated with demographic factors like biological sex
  - Associated with student socio-emotional factors like self-efficacy, academic engagement
  - Associated with item characteristics (item position, item format)

- Low examinee effort can
  - Bias IRT-based person and item parameter estimates
  - Bias test-based inferences like achievement gap estimates

How is low effort defined and operationalized in testing contexts?



2

Section Learning Objectives

2 How is low effort defined and operationalized?

- Define common approaches to identifying low effort
- Understand the central role of item response times in many approaches to identifying low effort
- Identify sources of meta-data other than response times that can be useful in identifying low effort
- Articulate tradeoffs between using response times versus other sources of meta-data to identify low effort

How might low effort examinees behave?

- Should start by asking: what does low effort look like behaviorally? What is the mindset?
  - One option: examinees don't want to take the test and simply disengage?
  - What does that look like? They want the test to be over!
  - Start moving through the test very quickly, including simply clicking through items
  - Here, items are right at a rate no better than chance—the examinee is "rapidly guessing"
  - Truly disengaged examinee not doing this because the test is too hard and they simply can't do the items—there should be little relation between this behavior and true achievement (ideally)

Other mindsets behaviors

- An examinee moving very quickly might demonstrate other behaviors, like:
  - Providing nonsensical responses for open response items
  - Not clicking on necessary links
  - Not scrolling to the bottom of the page
- At the same time, there may be other behaviors related to low effort. For instance, someone might take a very long time to respond to an item if the test is not adaptive
- In short, there are many possibilities! Detecting low effort very much about understanding the psychology of test taking

Most Ways to Identify Low Effort Use Metadata

- Metadata refers to data other than an examinee's vector of item responses that happen to be captured when a test is given via computer

- Options include
  1. Percent correct
  2. Person fit indices
  3. Standard error of measurement (SEM)/test information
  4. Response times
  5. Additional sources of metadata
     a) Click data
     b) Scroll data
     c) Open response data

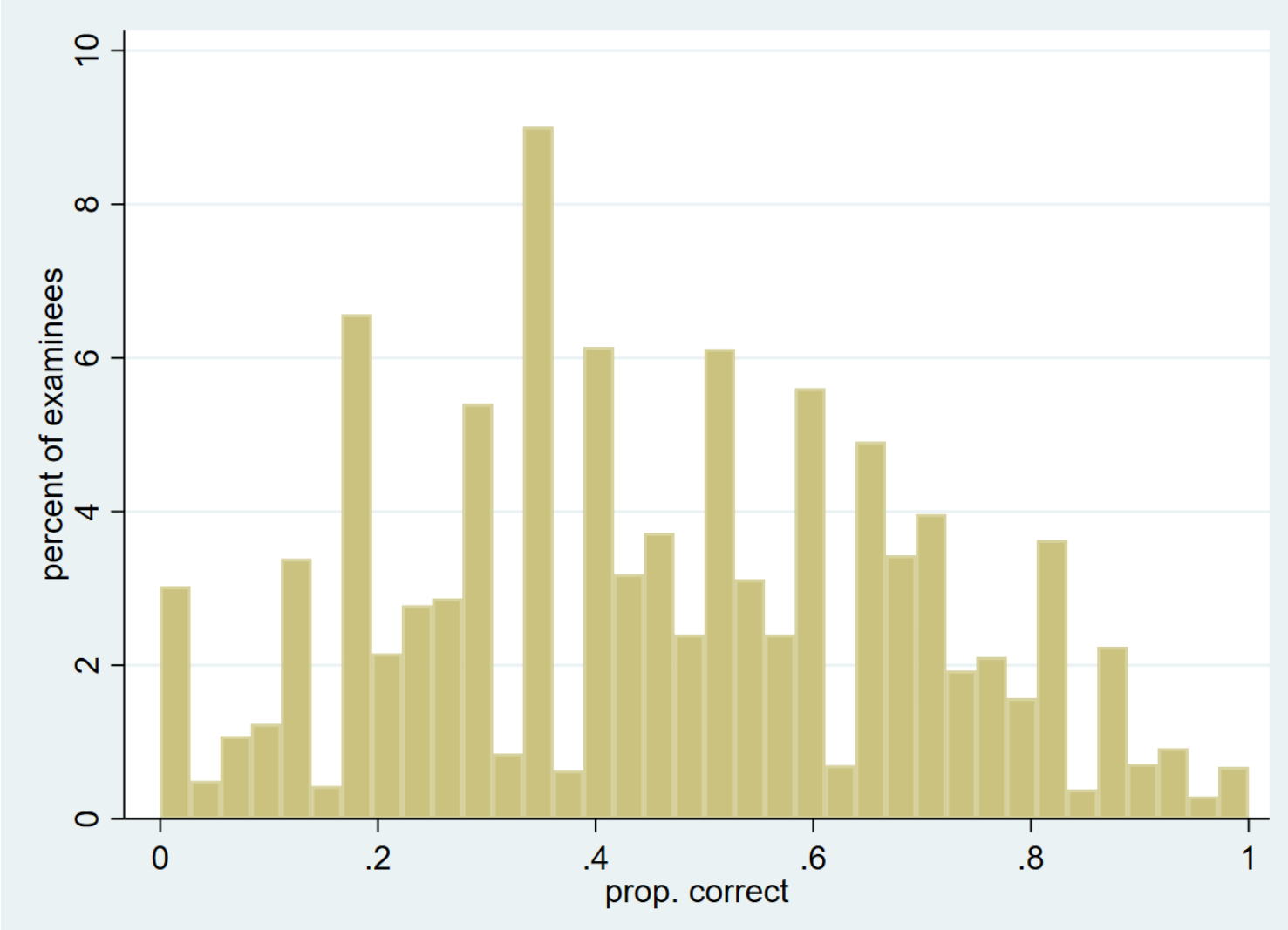
- In the remainder of this unit, I will define these sources of meta-data and weigh the pros and cons of using them

Sources of Data Other than Metadata Exist

- A primary example is examinee self-report
- For example, following a test, a student might be asked: “Did you give your full effort on this exam?”
- Two disadvantages.
  - First, tend to be global measures that ask students to report on their overall effort during the test. Difficult to detect instances in which non-effortful behavior occurs during only a portion of the test.
  - Second limitation is that it is difficult to assess how truthfully students respond (Wise, 2015)
- However, self-report can be useful as a piece of evidence to support the validity of other, non-self-report metrics and can therefore still be quite valuable

Option 1. Proportion Correct

- Can look at proportion of items examinee got correct
- Less useful on fixed form tests
- But can be very useful on CATs
- Still, might be several reasons other than low effort that the prop. correct strays from the target on a CAT



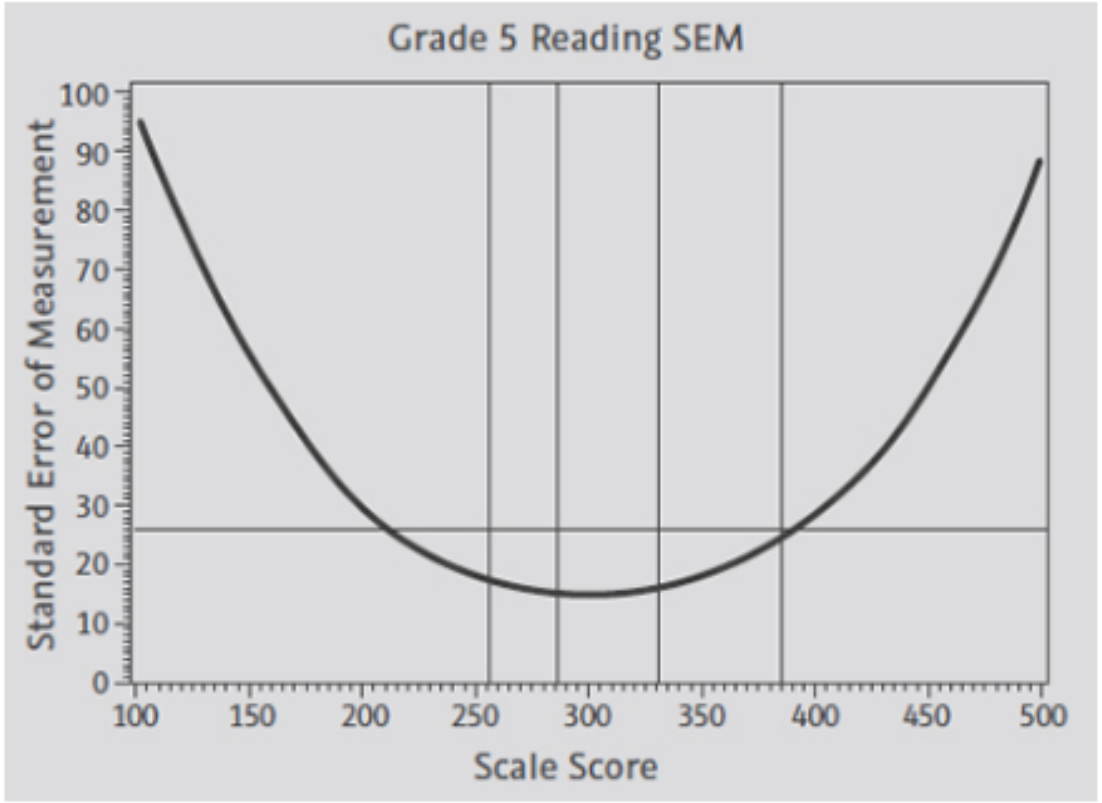
Option 2. Person Fit Indices

- Person fit indices compare a student's response pattern to what would be expected according to a theoretical measurement model
- Aberrant patterns indicate poor model fit
- The aberrant response pattern that is most congruent with unmotivated responding is random responding (Meijer, 2003).
- Limitation that they are sensitive to numerous sources of misfit, many of which have nothing to do with a lack of student effort

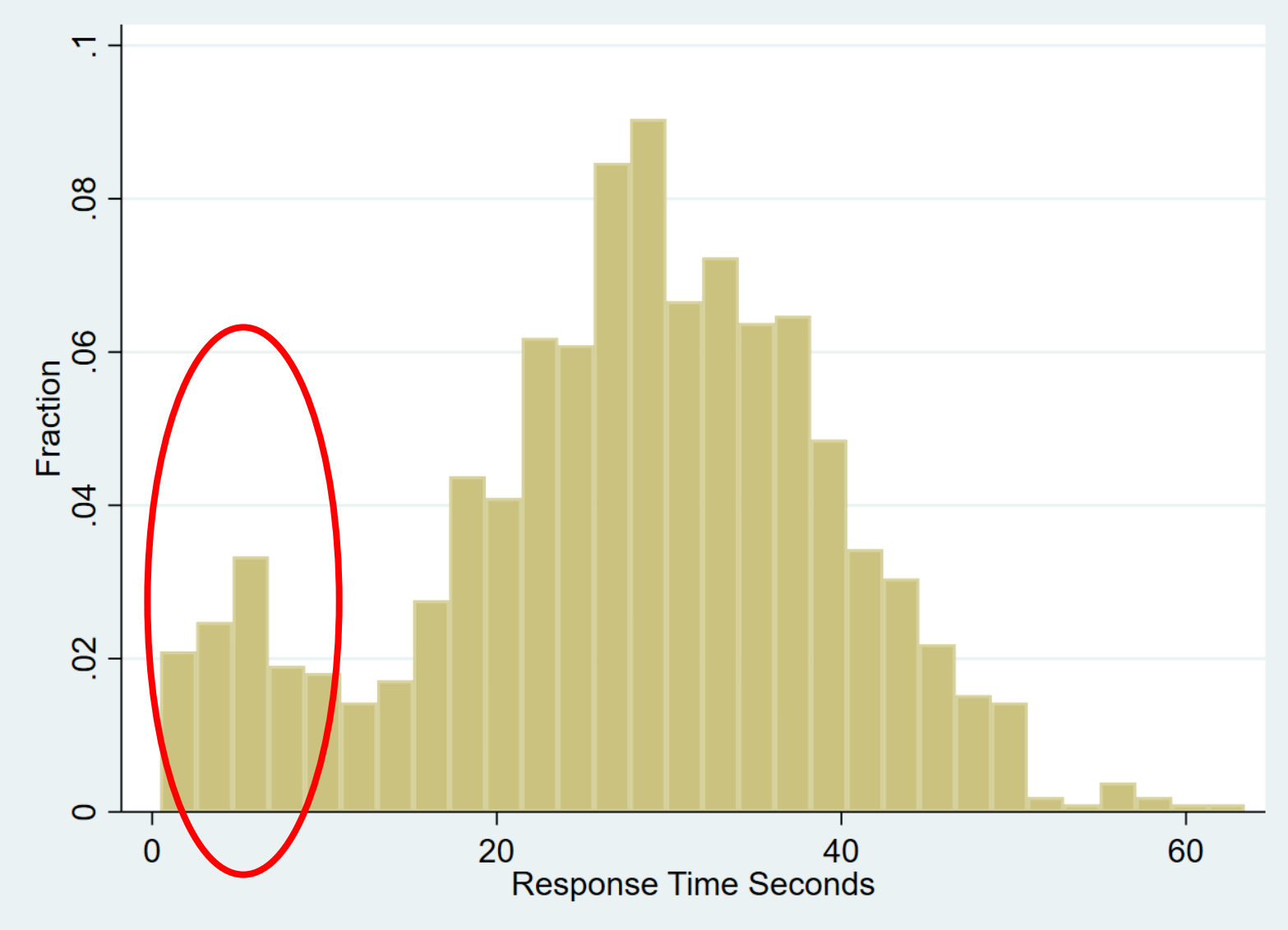
$$H^T(n) = \frac{\sum_{m=1, m \neq n}^{N} \left(\left[\sum_{i=1}^{L} X_{ni} X_{mi} \right] / L - P_n P_m \right)}{\sum_{m=1, m \neq n}^{N} \left(\min \left[P_n (1 - P_m), P_m (1 - P_n) \right] \right)}$$

Option 3. Measurement Error/Information

- Typically, the SEM on a test is higher at the extremes
- A very high SEM nearer the middle of the scale could be due to low effort
- However, it could be due to other factors as well
- Also does not allow one to detect low effort at the item response level
- Nonetheless, SEM can be used as corroborating evidence



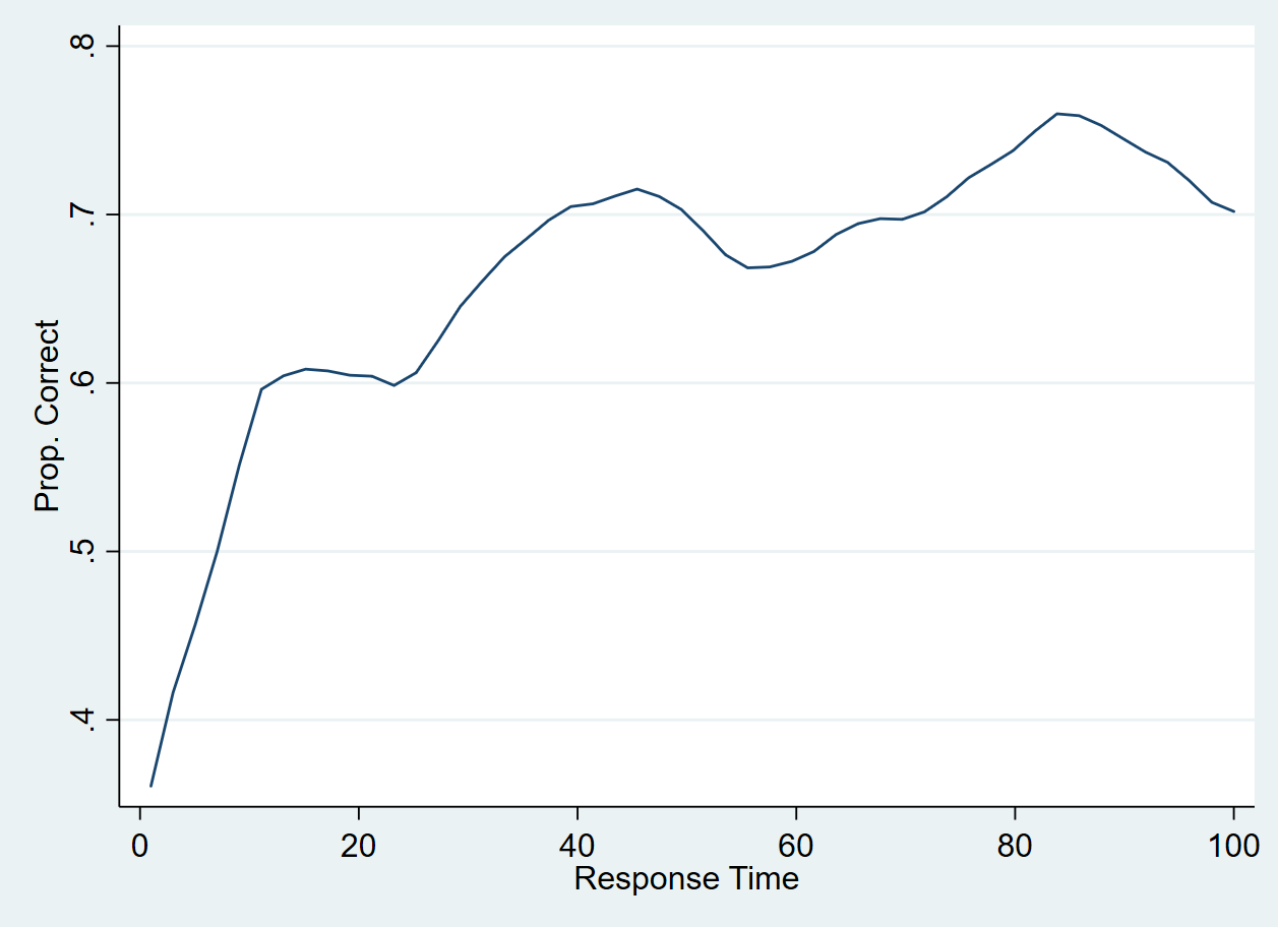
Option 4. Response Times



Solution Behavior Versus Rapid Guessing

- Examinees begin a test in what we will call “solution behavior”: They actively try to determine the solution (correct answer) to every item (Schnipke & Scrams, 1997)

- Yamamoto (1995) assumed that examinees might switch to a random-response strategy (what Schnipke and Scrams call ***rapid-guessing behavior***) as time elapses, which can be identified in part by the lowered accuracy of the responses



Response Time Effort (RTE)

- Wise and Kong (2005) developed their response time effort (RTE) index
- RTE equals the proportion of items that a test taker responds to with solution behavior.
- For example, an RTE value of .90 indicates that the student exhibited solution behavior on 90% of his items, and rapid-guessing behavior on 10%

RTE – validity argument (Wise, 2015)

1. RTE should demonstrate adequate levels of reliability

2. RTE should be correlated with other measures of test-taking effort

3. Those engaged in rapid guessing should get those items right at a rate no better than chance

4. RTE should not be correlated with measures of academic ability

Option 5. Additional Sources of Metadata

- As mentioned previously, there are other possible behaviors related to low effort

- For example, did the examinee:
  - Scroll to bottom of the page explaining the item?
  - Respond appropriately in text box?
  - Click on the necessary links?
  - Actually use tools like highlighters and calculators?
  - The list goes on…

- Still note, that these behaviors remain indicative of a common mindset: the examinee has decided that the rewards of working hard on the test are not worth it given the demands of the items

Advantages and Disadvantages

- Self-report data and metadata other than response times
  - Can be useful to help corroborate additional evidence on effort, but…
  - Can be due to factors other than low effort
  - Oftentimes do not allow detection at the item level

- Response times, meanwhile
  - Allow for inferences at the item level
  - Do not suffer from self-report bias since examinees often unaware
  - Lend themselves to metrics like RTE, supported by validity evidence for intended use
  - However, could argue that they often lend themselves to crude approximations of effort

Setting Response Time Thresholds

3

3 Setting Response Time Thresholds

Section Learning Objectives

Define and implement threshold-setting methods

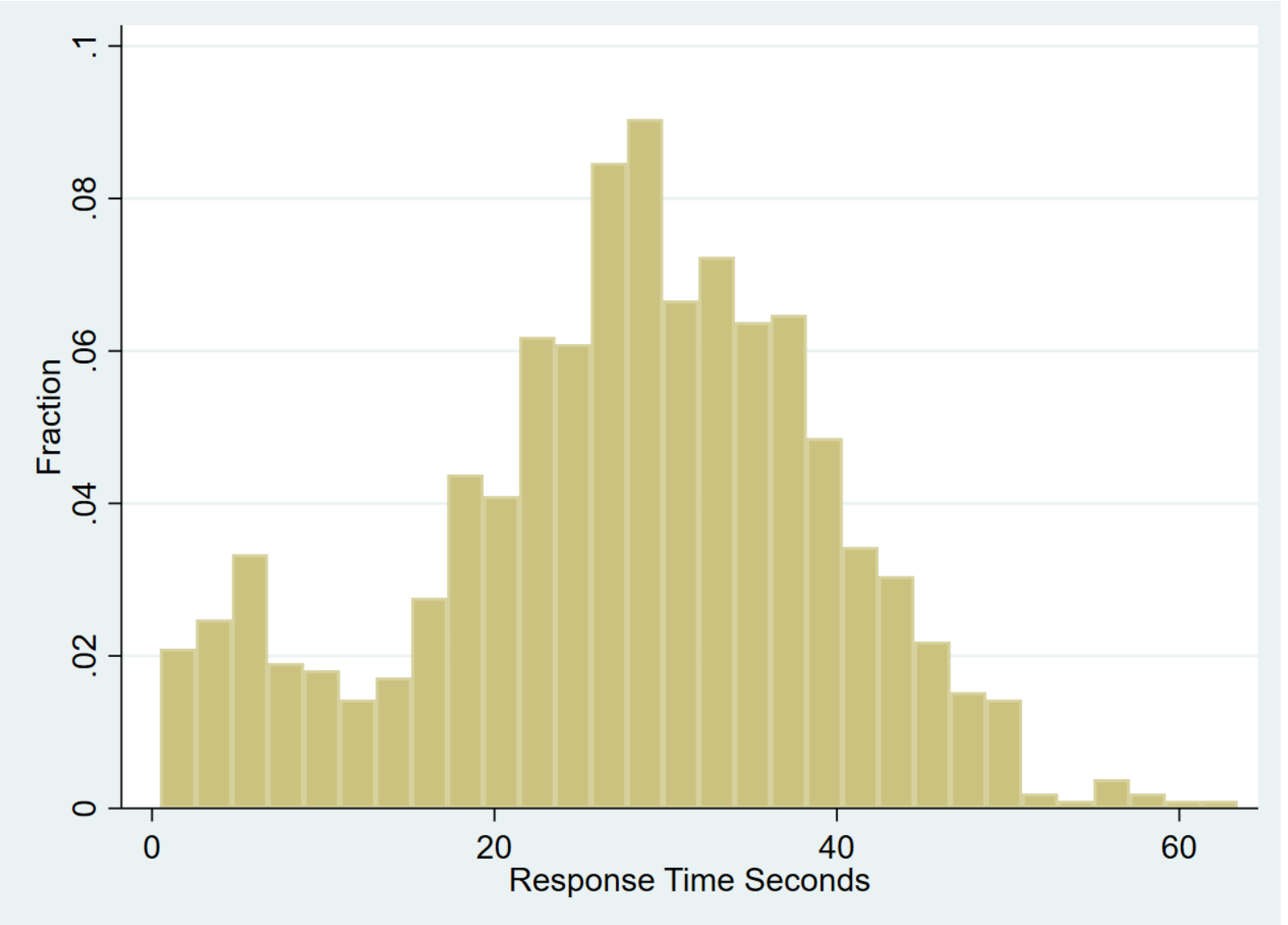
Articulate why the method chosen likely depends on the intended use of scores

Understand limitations and strengths of each approach

Four Main Options to set RT thresholds

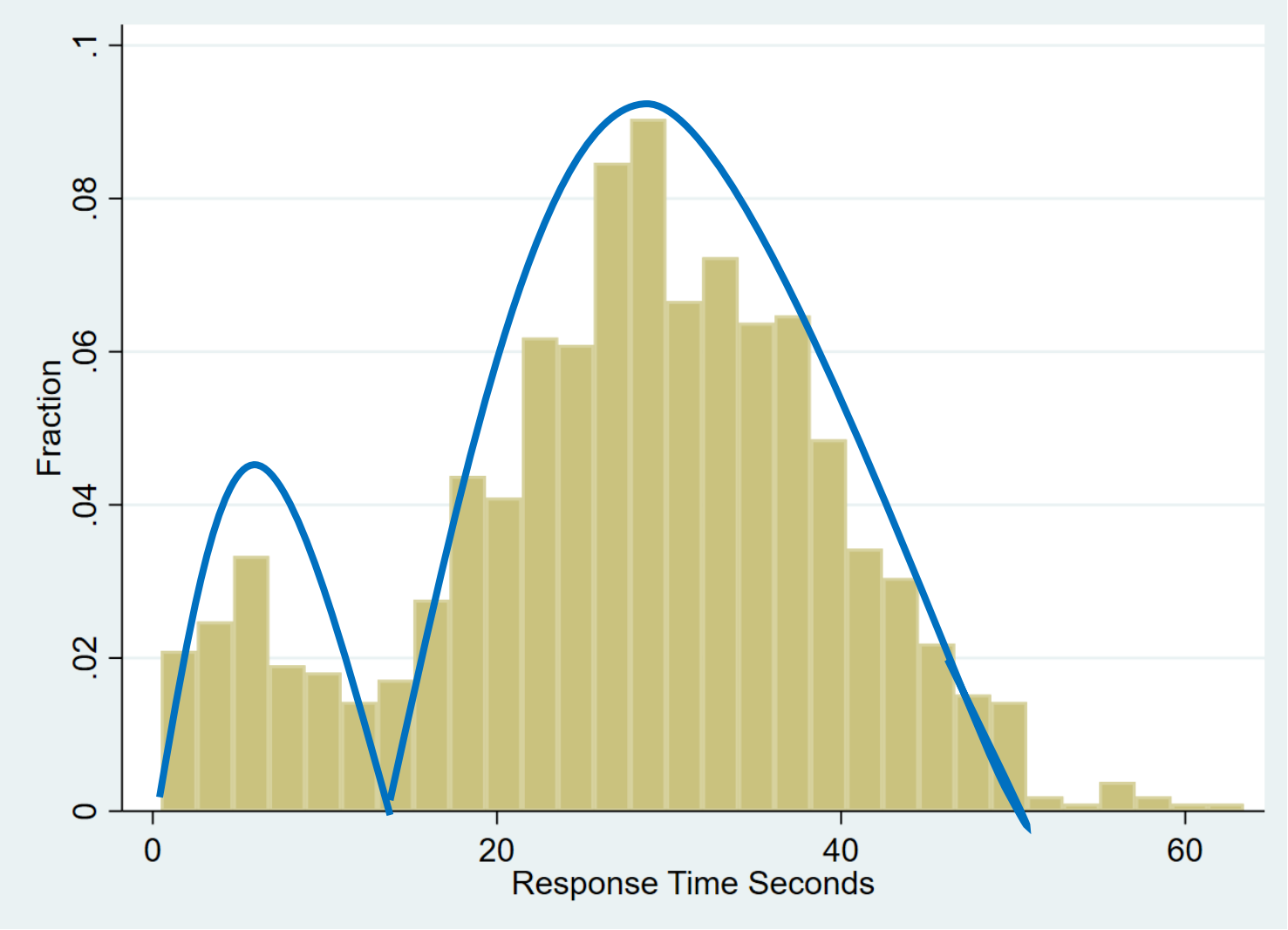
- Visual inspection
- Mixture model (mixture log normal, MLN)
- Cumulative proportion correct (CUMP)
- Normative threshold (NT)
- Test information-based

Visual inspection



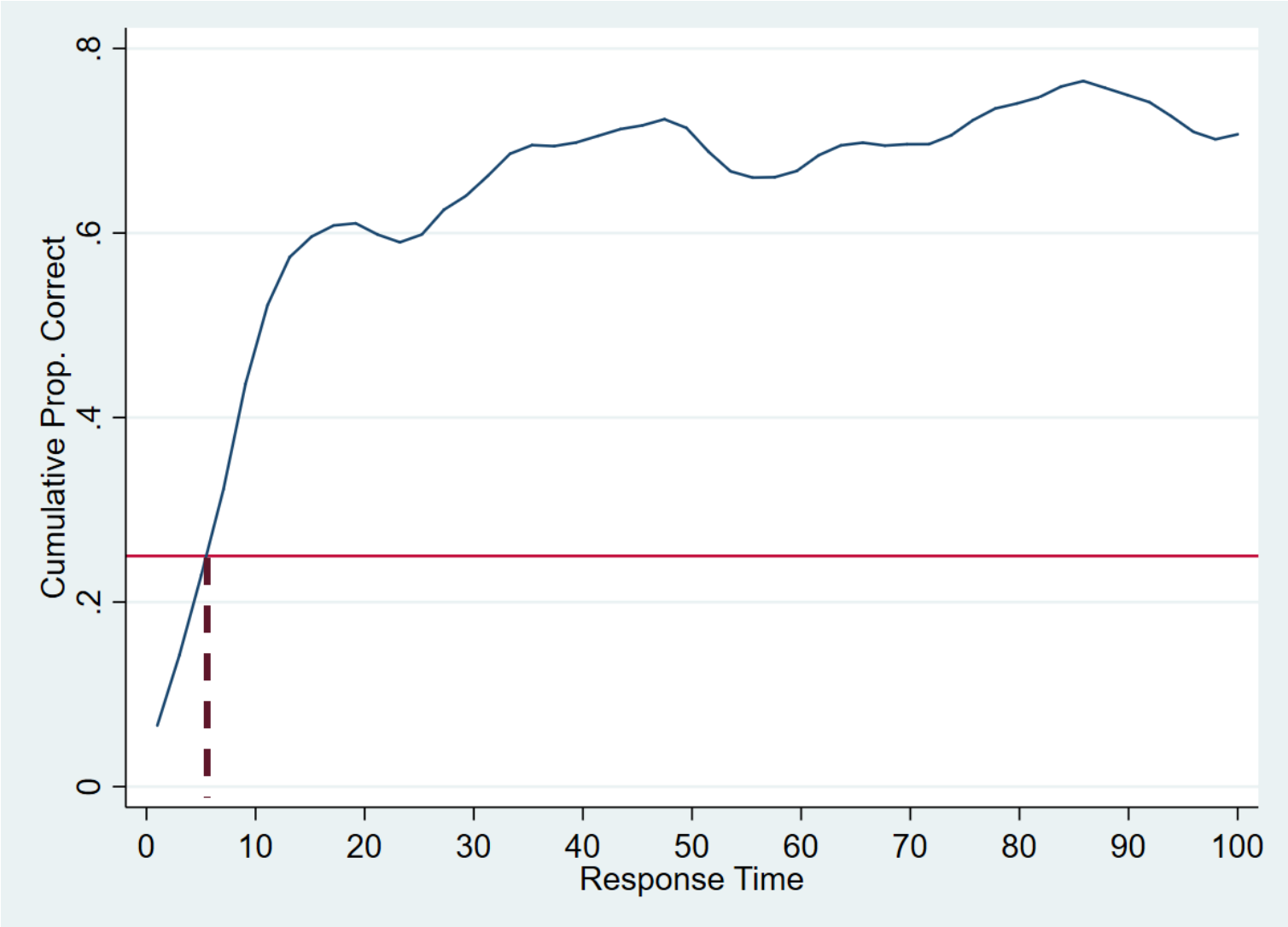
Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593-619.

Mixture model (mixture log normal, MLN)



Cumulative proportion correct (CUMP)

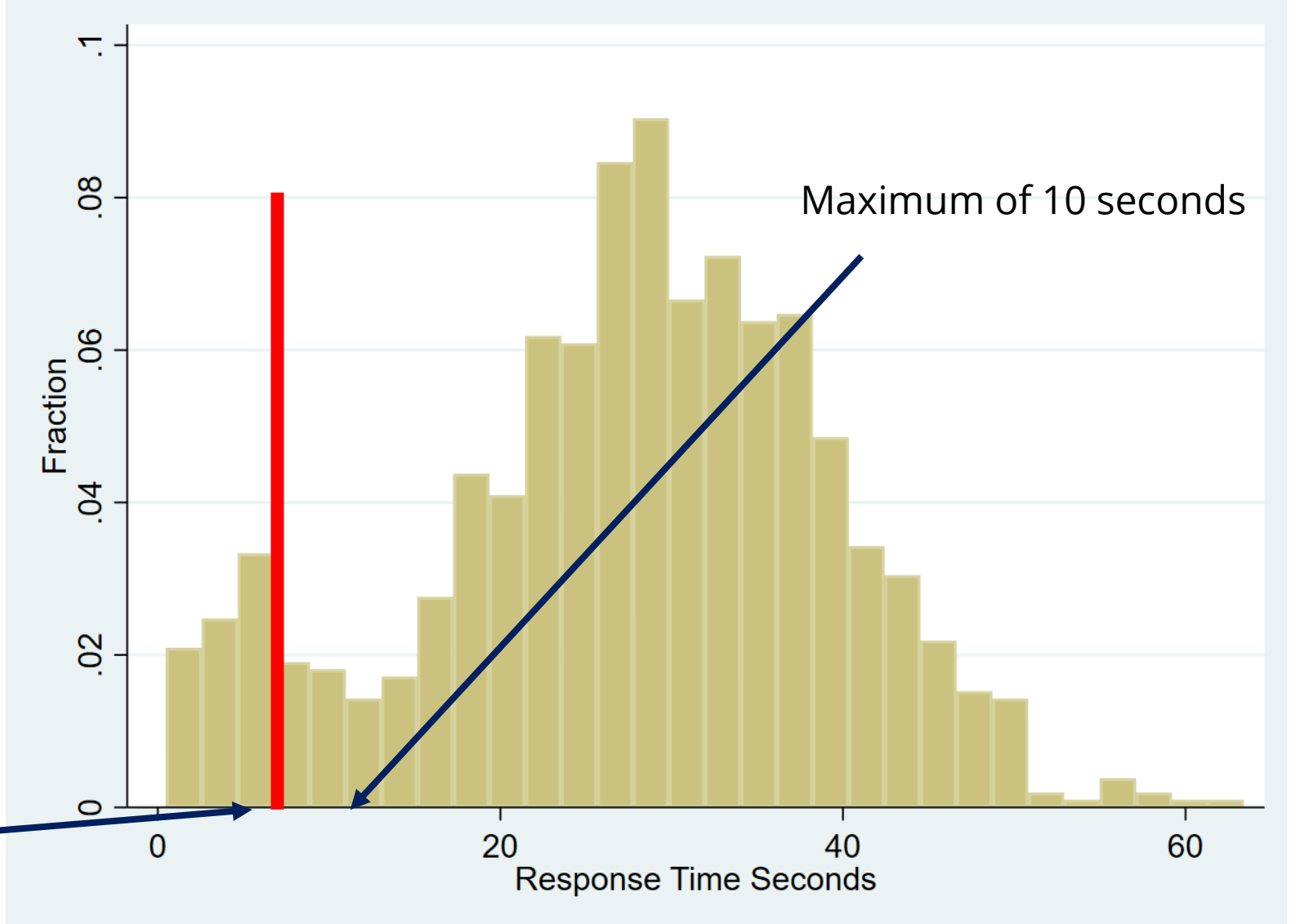
Normative threshold (NT)



Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173-183.

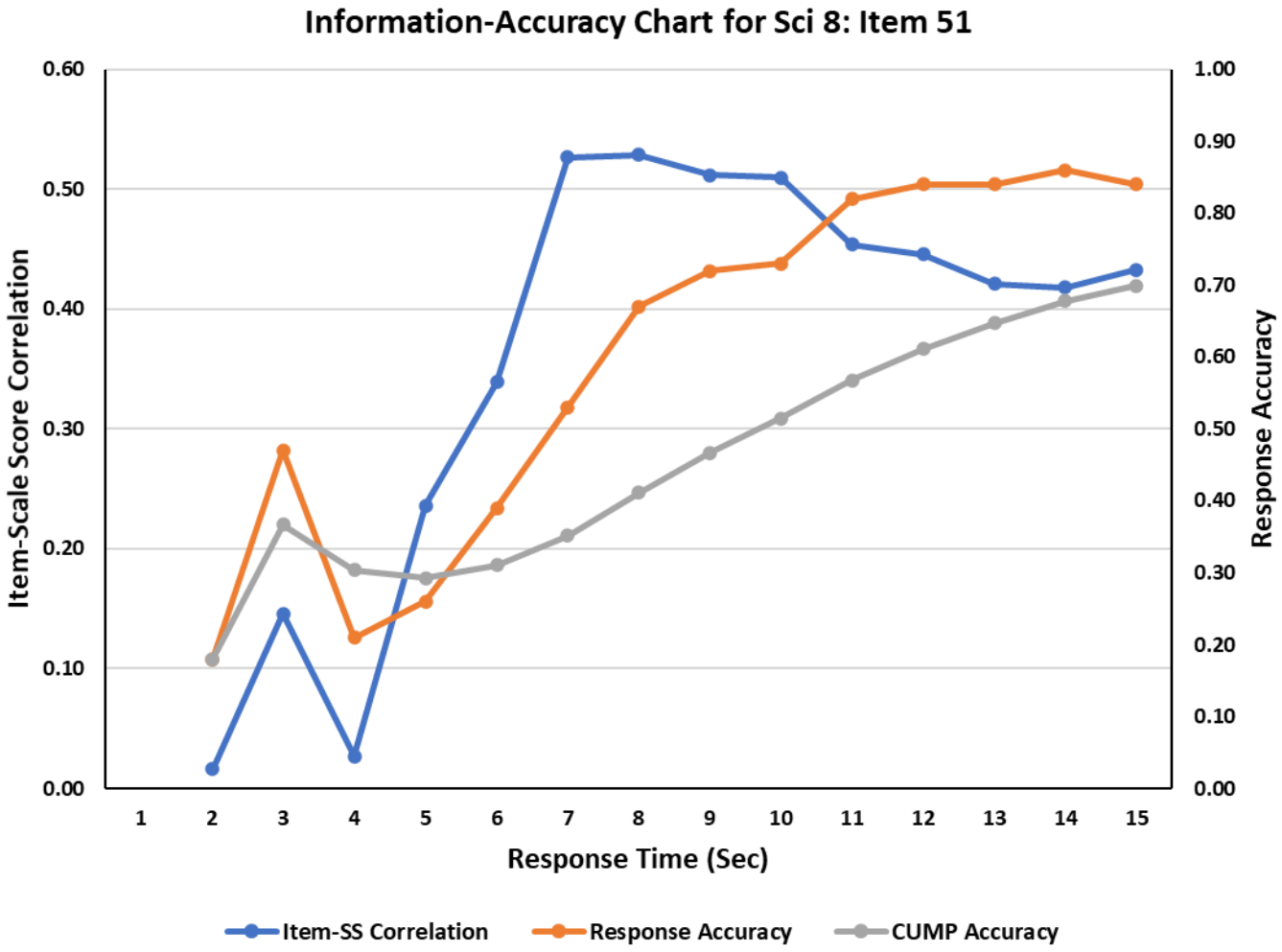
Cumulative proportion correct (CUMP)

Normative threshold (NT)



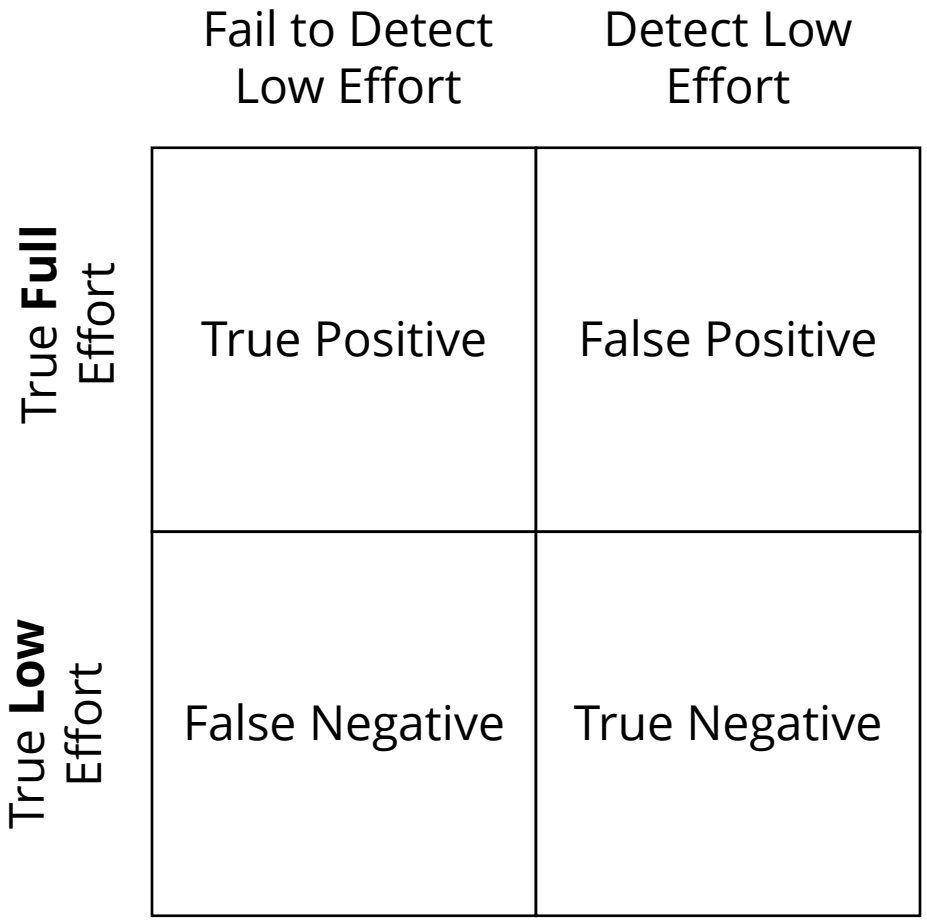
Bottom 10% of the RT distribution

Test information-based

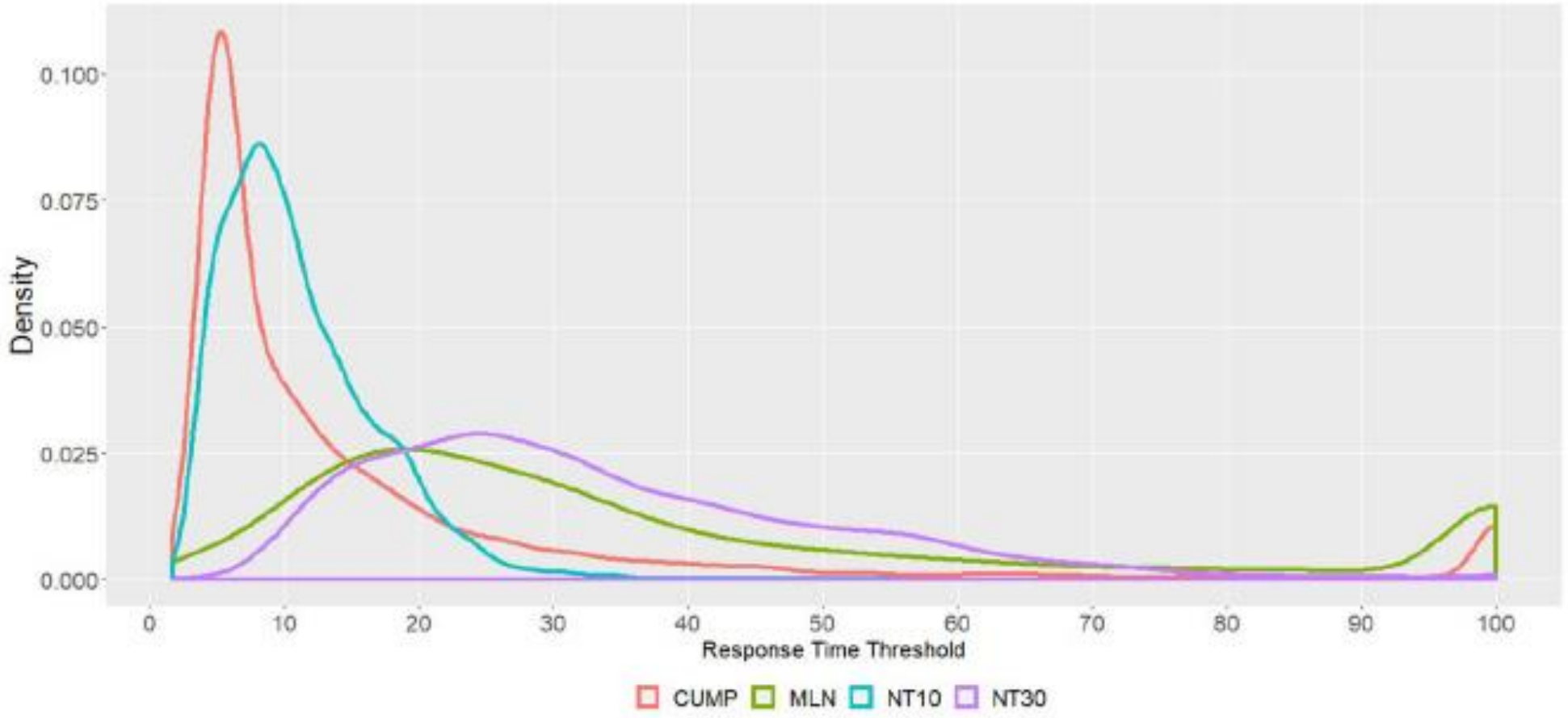


Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, *32*, 325-336.

A Brief Note on False Positive/Negatives

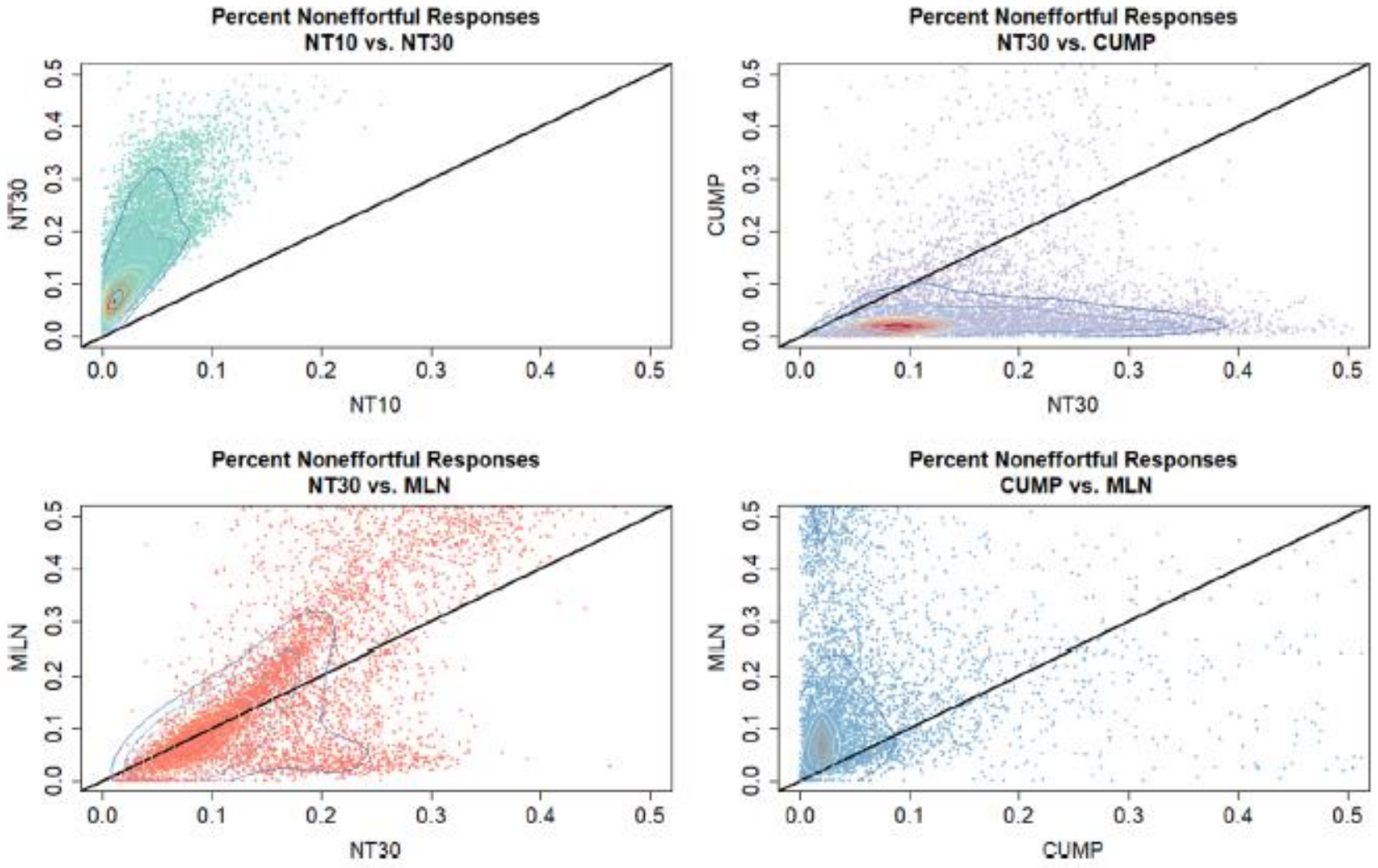


Comparing Approaches – Thresholds



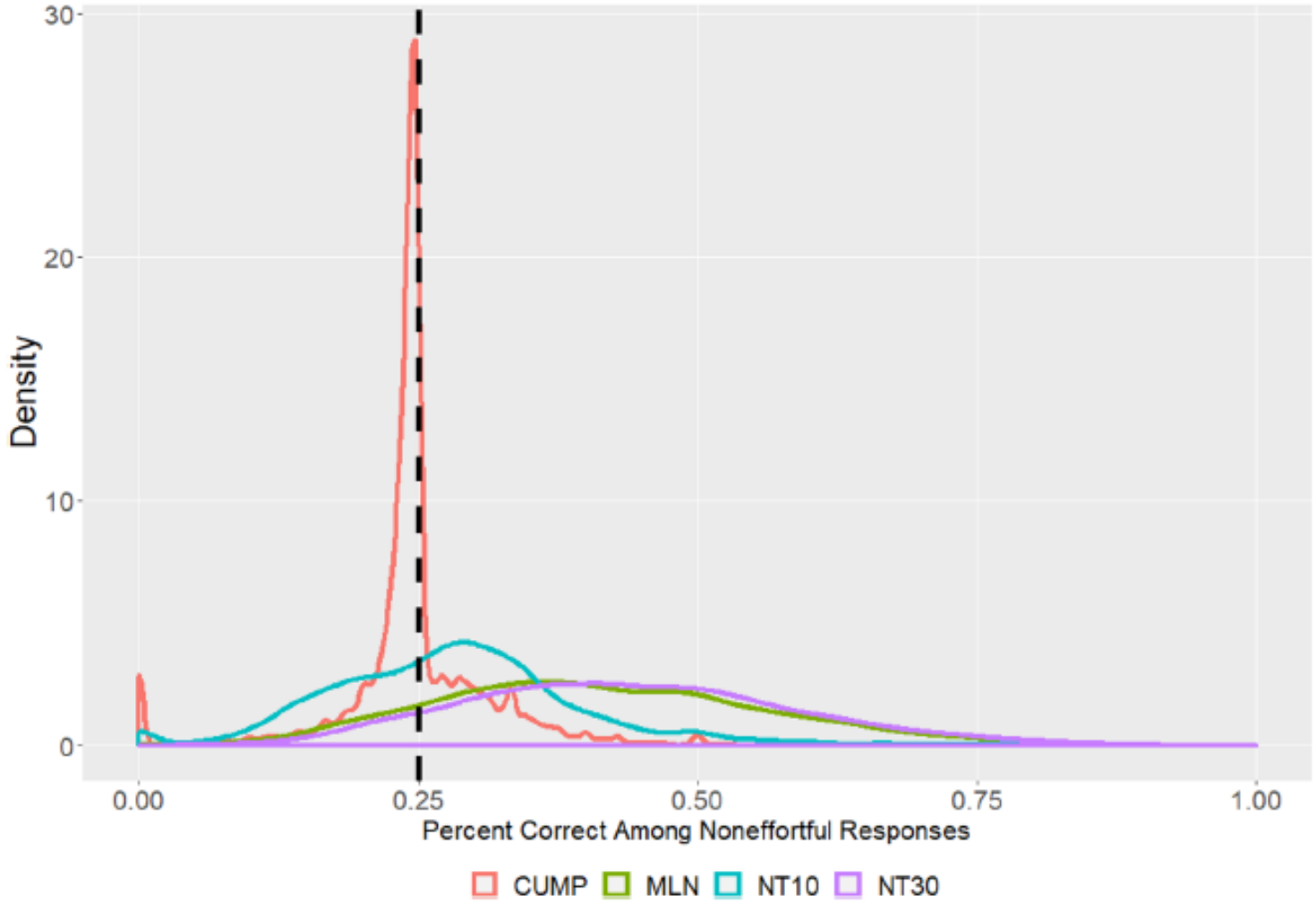
Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education*, *9*(1), 1-21.

Comparing Approaches – Prop. Low Effort



Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education*, *9*(1), 1-21.

Comparing Approaches – Correct Responses



Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education*, *9*(1), 1-21.

Conclusions

- First, various threshold-setting methods produce viable thresholds for at very different rates.
- Second, having the highest proportion of items with viable thresholds is not synonymous with being supported by the strongest validity evidence for those thresholds.
- Third, there are often inconsistencies in the thresholds and, therefore the item responses identified as noneffortful by method.
- Choice probably comes down to intended use, especially if using individual scores or in the aggregate

Addressing Low Effort

4

4 Addressing Low Effort

Section Learning Objectives

- Outline ways to prevent low effort before or as it happens
- Define person- versus item-level filtering
- Gain conceptual understanding of how several baseline IRT models to address low effort are specified
- Identify which IRT model might be best in your context

Two Broad Options

1. Front-end options: addressing low effort before or when it's happening
    a) Motivating students
    b) Monitoring effort in real time

2. Back-end options: addressing low effort after data are collected
    a) Removing low-effort examinees
    b) Removing low-effort item responses

Option 1: Front-end approaches

- One way to try to address low effort is to stop it before it happens. There are many ways educators could approach this

- However, another approach is to monitor effort in real time
  - E.g., we studied an operational testing program that notified proctors when students disengaged
  - After a sufficient number of rapid guesses, proctors were notified
  - We found that engagement increased and performance improved post-notificaiton

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183-192.

Two Broad Options

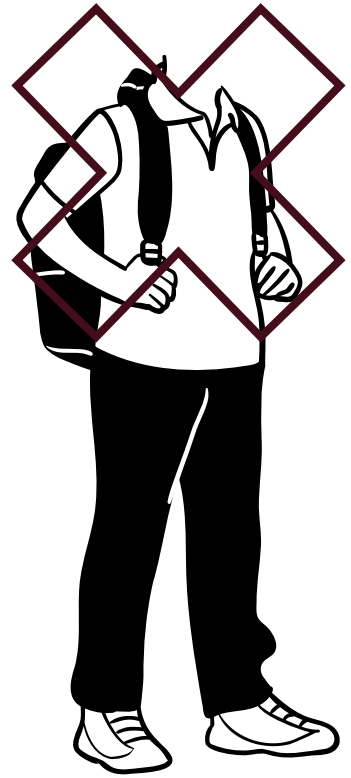
1. Front-end options: addressing low effort before or when it's happening
   a) Motivating students
   b) Monitoring effort in real time

2. ***Back-end options: addressing low effort after data are collected***
   a) Removing low-effort examinees
   b) Removing low-effort item responses

Option 2(a). Person-level Filtering

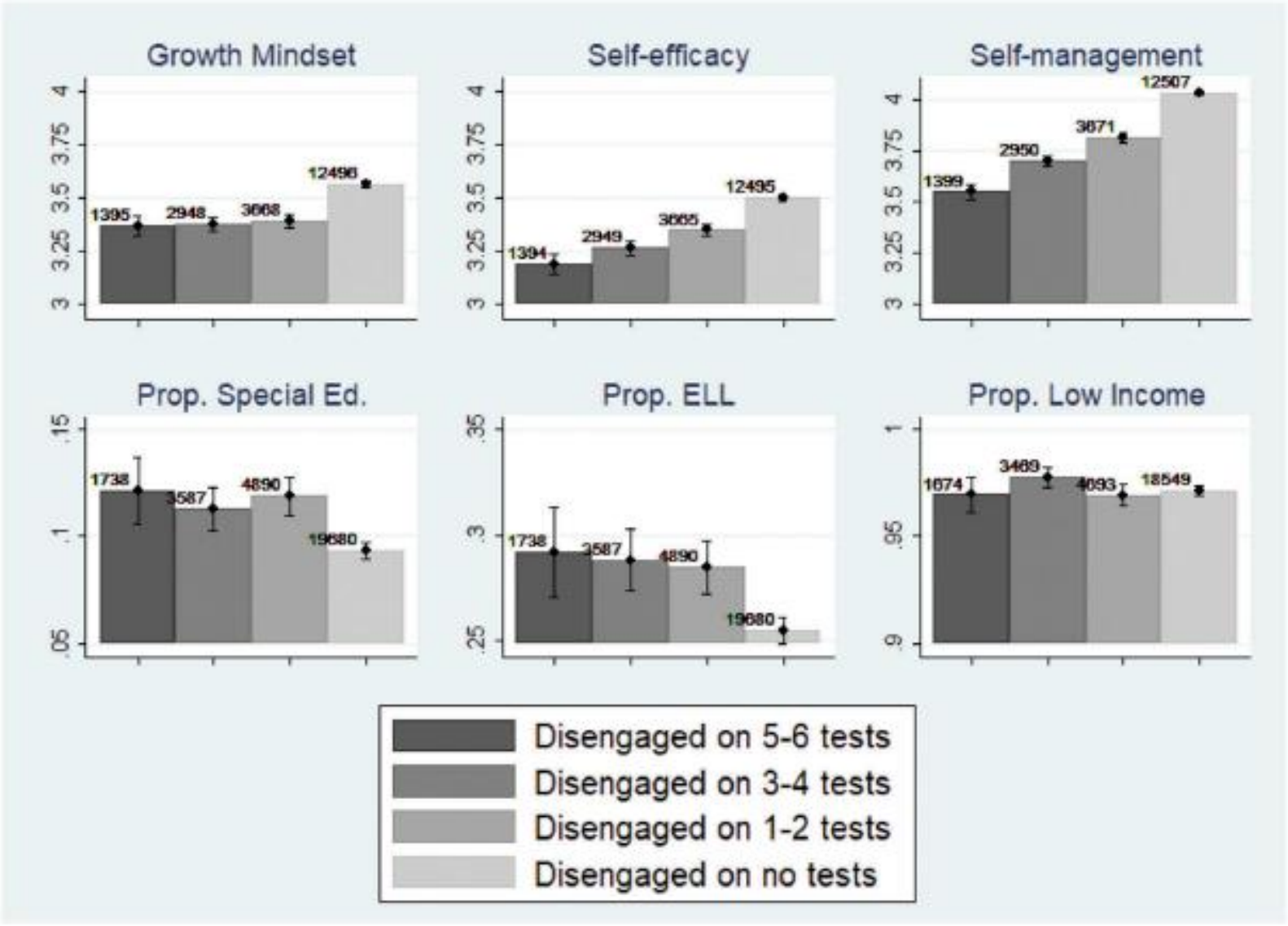
- Works much as it sounds
- Identify examinees with less than perfect effort
- (Could use RTE<.90)
- Remove those examinees from the sample
- Conduct analyses

Effortful



Non-Effortful

Bias due to demographics, other non-observables



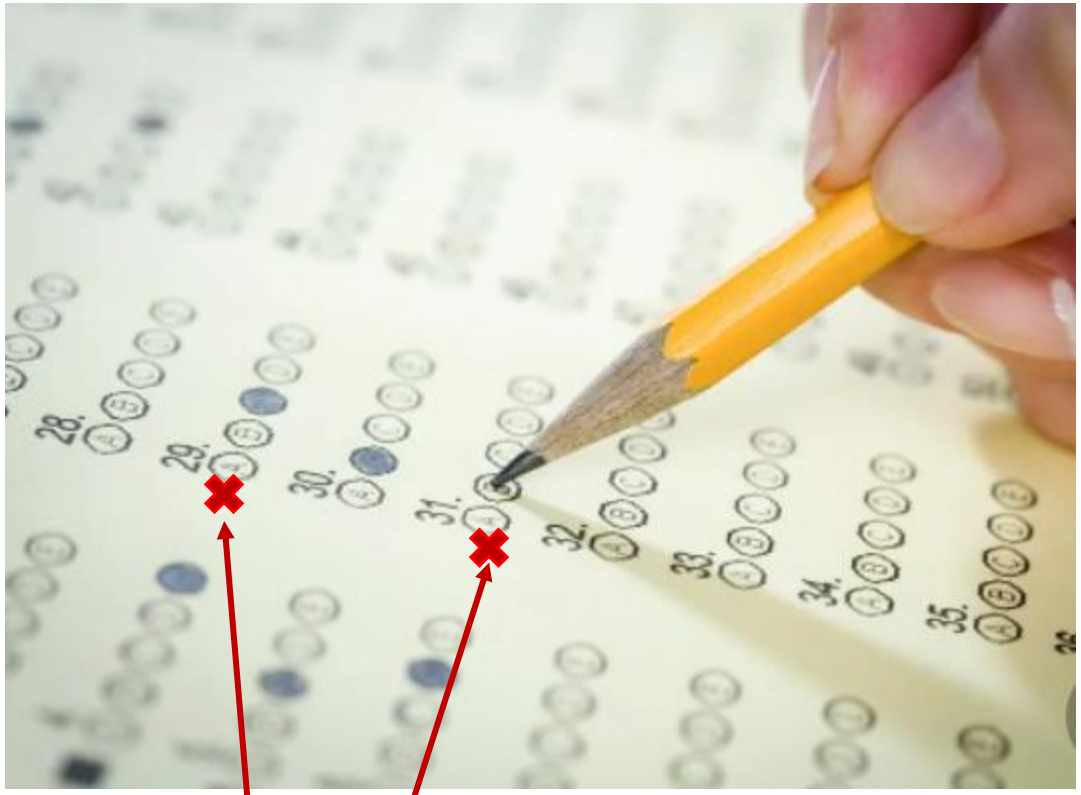
Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment*, *24*(4), 327-342.

Bias due to correlation between effort, true ability

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014*(161), 69-82.

Option 2(b). Item-level Filtering

- Like examinee filtering, item filtering is similar
- Select a threshold, and identify item responses that are RGs
- Remove RGs from the matrix of item responses
- Score using an IRT model



Below the RT Threshold

Effort Moderated IRT Model – Set Up

Let's say we have item i with a response time thresholds T_i and examinee j with a response time RT_{ij}. A dichotomous index of solution behavior, SB_{ij} could be computed as

$$SB_{ij} = \begin{Bmatrix} 1 \; if \; RT_{ij} \geq T_i \; , \\ 0 \; otherwise \end{Bmatrix}$$

The index of overall response time effort for examinee j on a particular test of k items is given by

$$\frac{\sum_{i=1}^{k} SB_{ij}}{k}$$

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19-38.

Effort Moderated IRT Model

Conceptually, the effort moderated IRT model can be thought of as

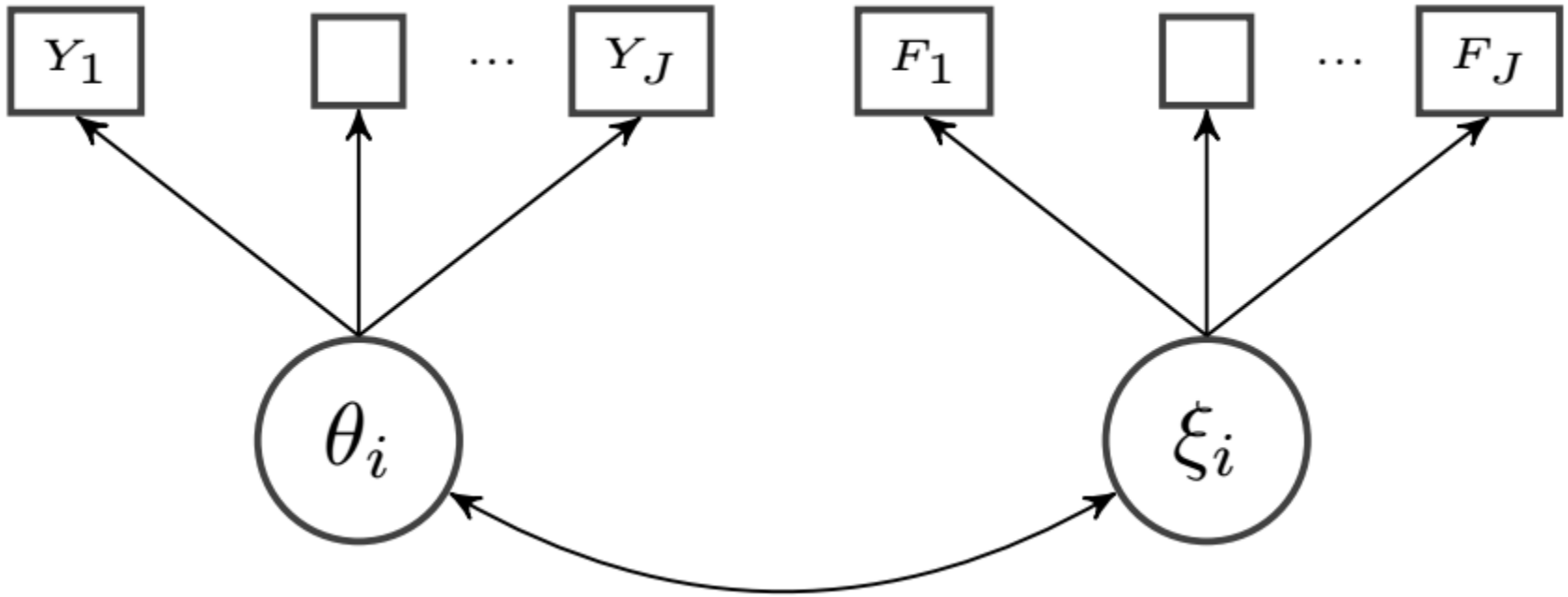
$$P_i(\theta) = \left(SB_{ij}\right)(solution\ behavior) + \left(1 - SB_{ij}\right)(rapid\ guessing\ behavior).$$

For example, the 3PL model would be:

$$P_i(\theta) = \left(SB_{ij}\right)\left(c_i + (1 - c_i)\left(\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}\right)\right) + \left(1 - SB_{ij}\right)(g_i)$$

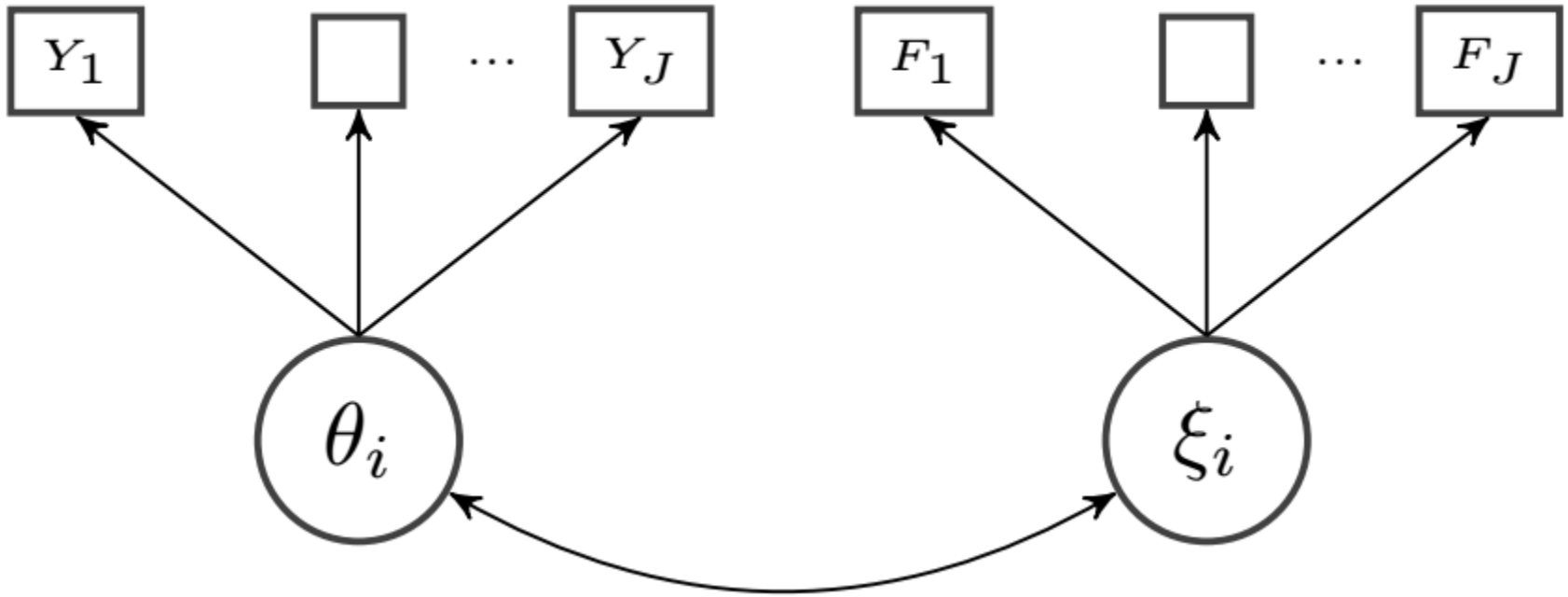
Where $P_i(\theta) = g_i$ is a constant probability model with g_i equal to the reciprocal of the number of response options for item i.

MIRT Model for Response Accuracy & Effort



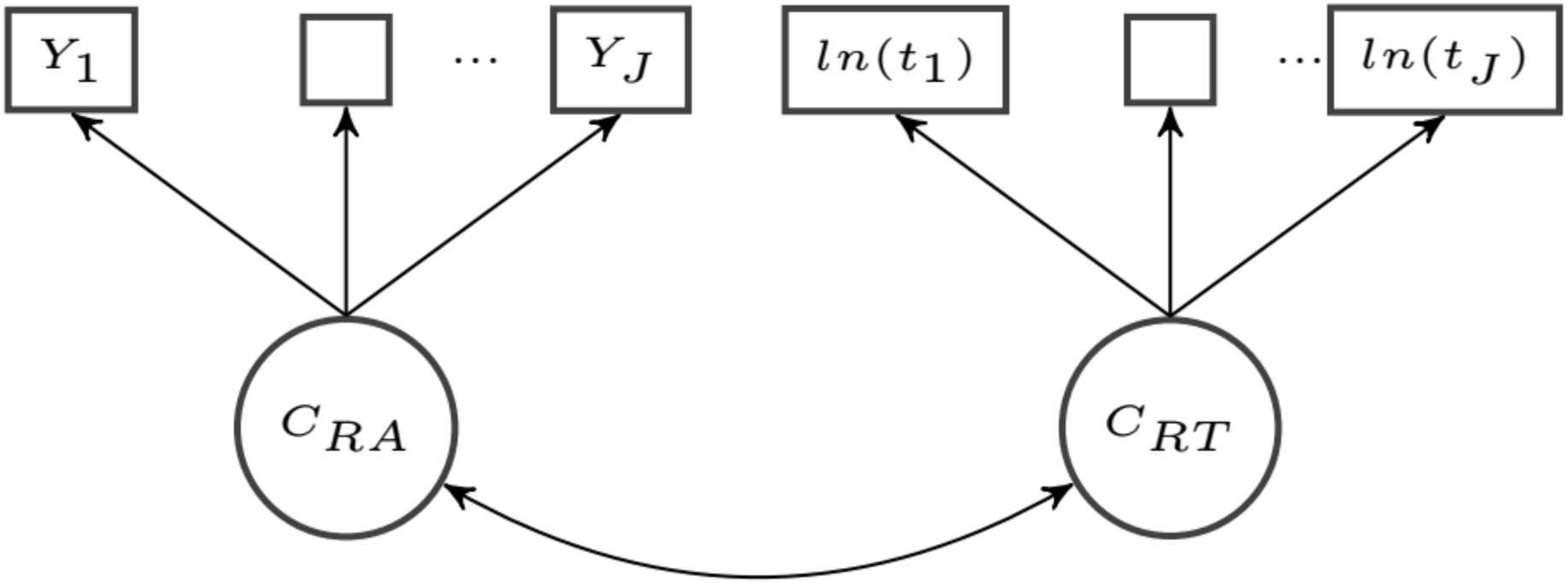
Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in psychology*, *10*, 145.

MIRT Model for Response Accuracy & Effort



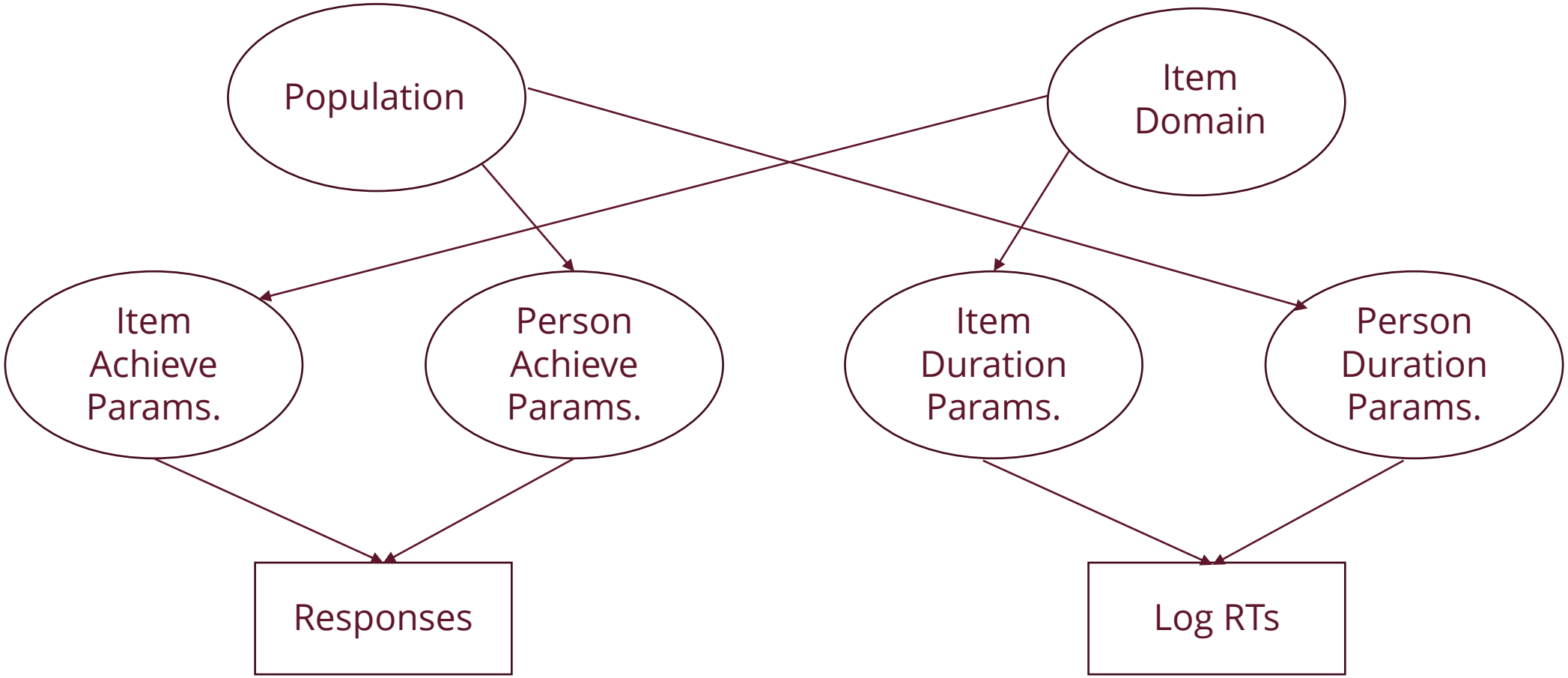
Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in psychology*, *10*, 145.

Mixture Model for Response Accuracy & Time



Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, *80*(4), 775-807.

Other RT-based Models



van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327-347.

Conclusions

- While addressing low effort before or while it happens is ideal, it is not always feasible
- Addressing low effort by removing low-effort examinees from the sample is not ideal because it can introduce substantial bias into aggregate estimates
- There are several modeling approaches available to produce achievement estimates that, in some way, account for low effort
- Perhaps the most straightforward is the effort moderated IRT model, which essentially treats item responses deemed noneffortful as missing
- However, there are also more sophisticated models that provide additional flexibility and require fewer assumptions
- Again, the choice of model will likely come down to intended use. For example, some of the mixture approaches may be more difficult to implement in largescale CAT contexts

How is low effort defined and operationalized in self-report contexts?

Section Learning Objectives

5 How is low effort defined in self-report contexts?

Name common approaches to identifying low effort on surveys

Identify tradeoffs involved in using each approach

Compare with response time approaches used for achievement

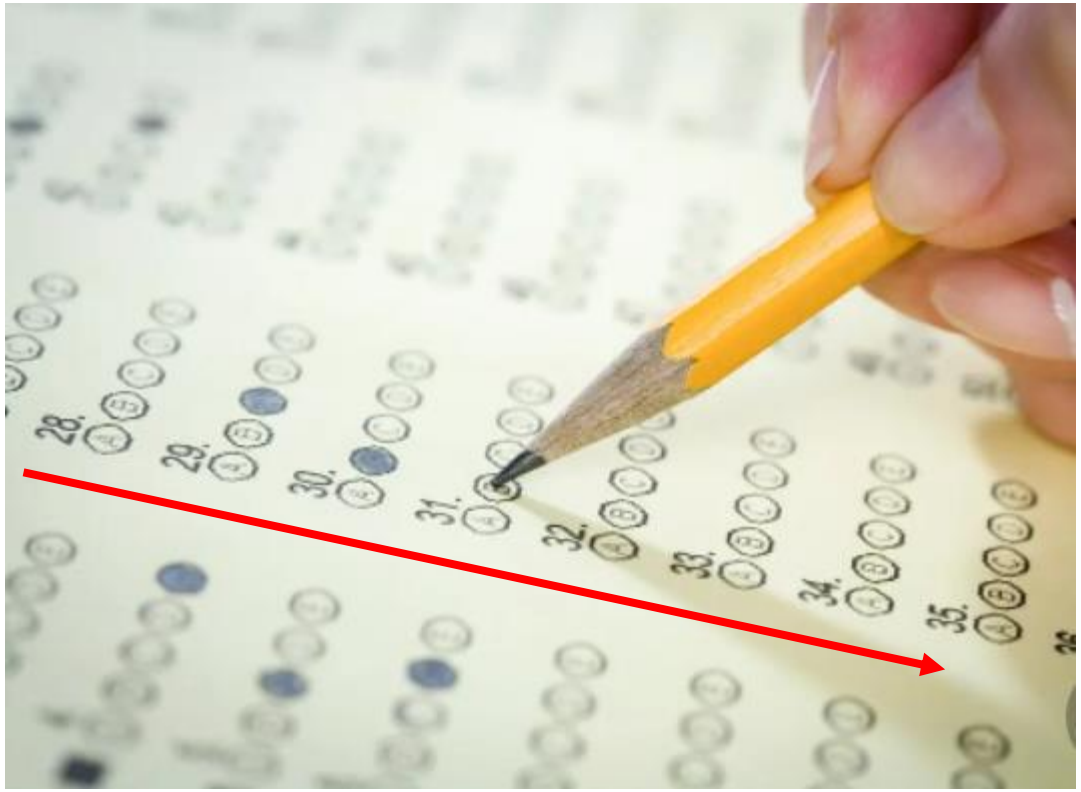
Know how to use multiple approaches in tandem to improve validity of inferences

Options for Identifying Low Effort on Surveys

1. Straight-lining
2. Synonym/antonym
3. Reverse coding
4. Fit/distance indices
5. Response time
6. Using response time in conjunction with other approaches

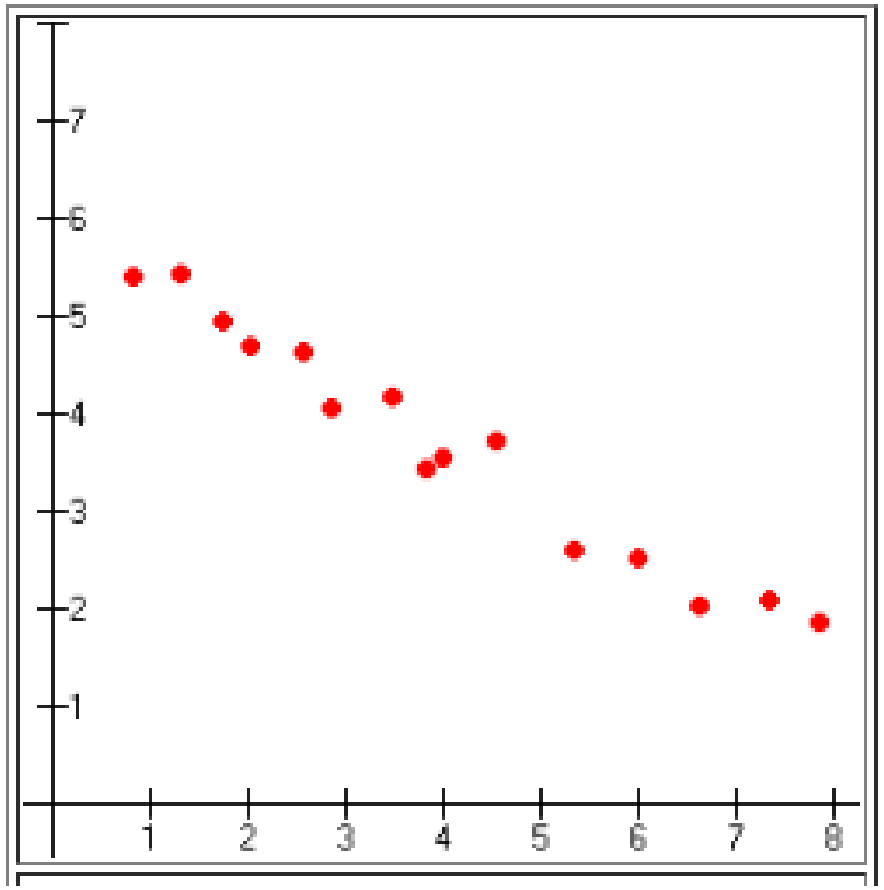
Option 1. Straightlining

	Strongly Disagree	Disagree	Agree	Strongly Agree
Item 1			X	
Item 2			X	
Item 3			X	
Item 4			X	
.				
.				
.				
Item N			X	



Option 2. Synonym/Antonym

- Would expect some item responses to be strongly correlated, others to be negatively correlated
- The degree to which an individual's responses follow this pattern can provide a clue about effort

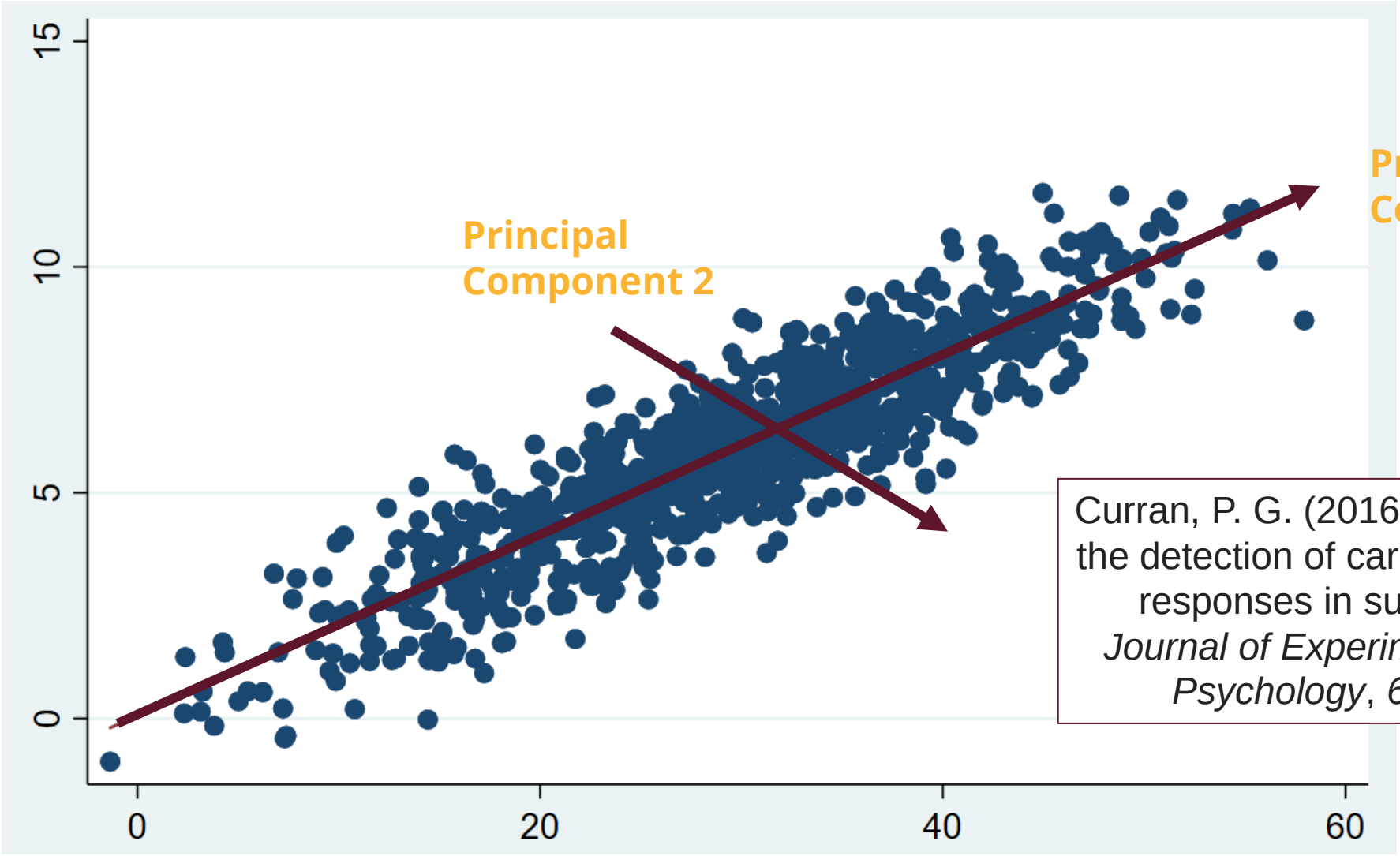


Option 3. Reverse Coding

1. Strongly Agree
2. Agree
3. Disagree
4. Strongly Disagree

1. Strongly Disagree
2. Disagree
3. Agree
4. Strongly Agree

Option 4. Fit and Distance Indices



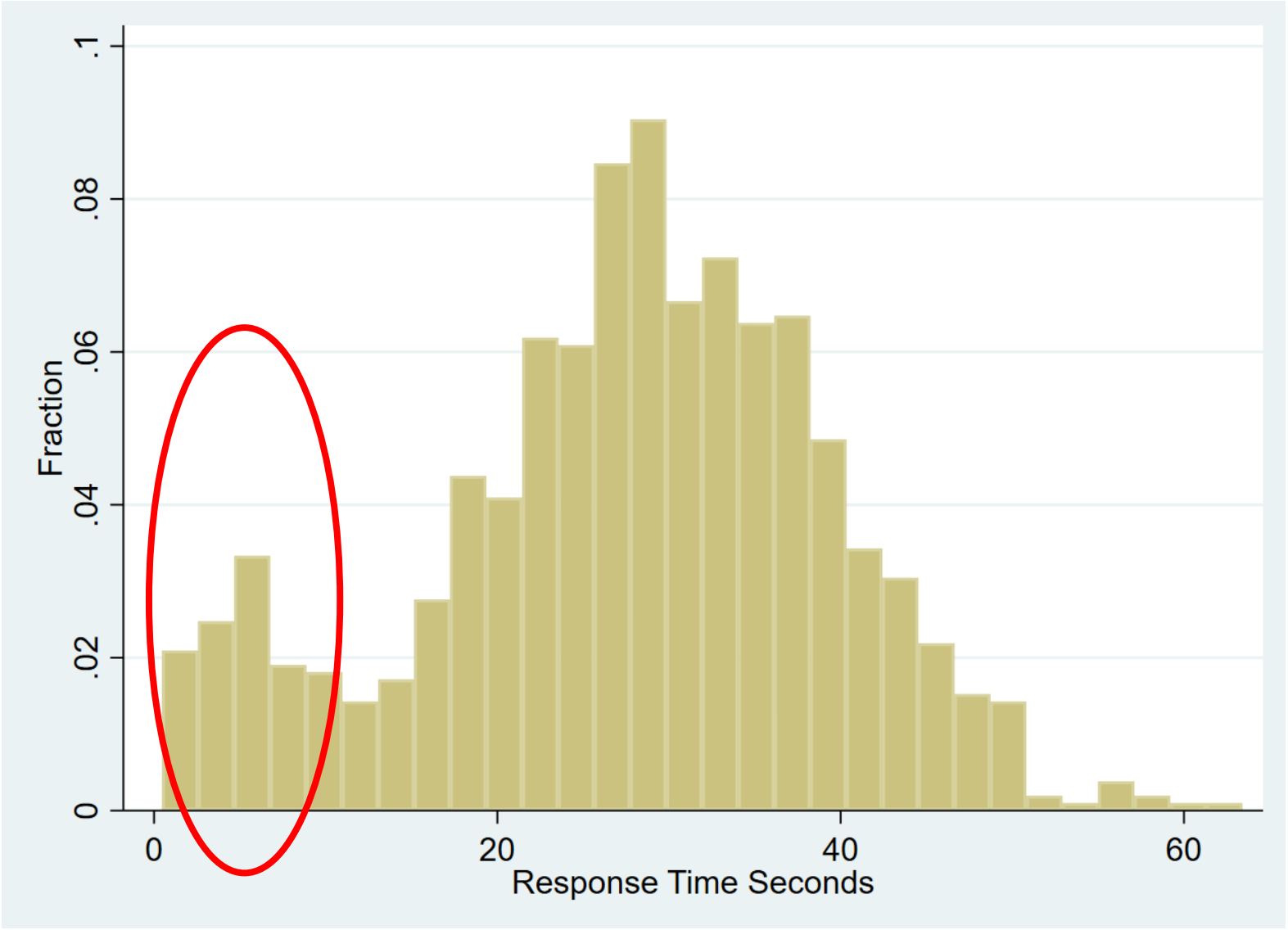
Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19.

Limitations of the Options Thus Far

- The options discussed so far have many benefits, including being directly linked to behaviors a disengaged respondent might demonstrate
- However, they also have important limitations, including that they:

1. Are at the person level, not the item level (filtering issues)
2. Could be due to issues other than low effort
3. Provide fewer options to check the validity of inferences like using proportion correct on an achievement test

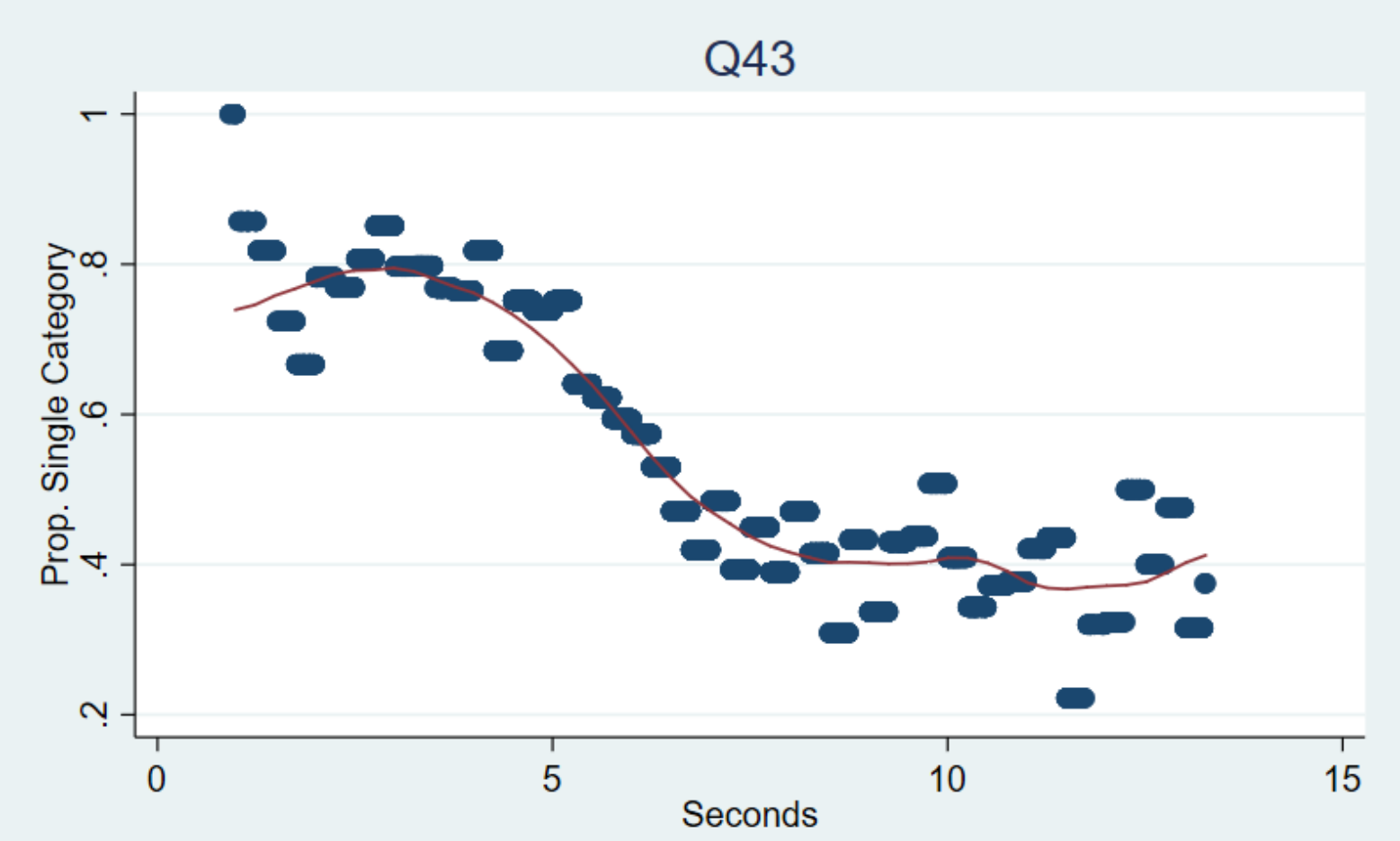
Option 5. Response Times



Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99-114.

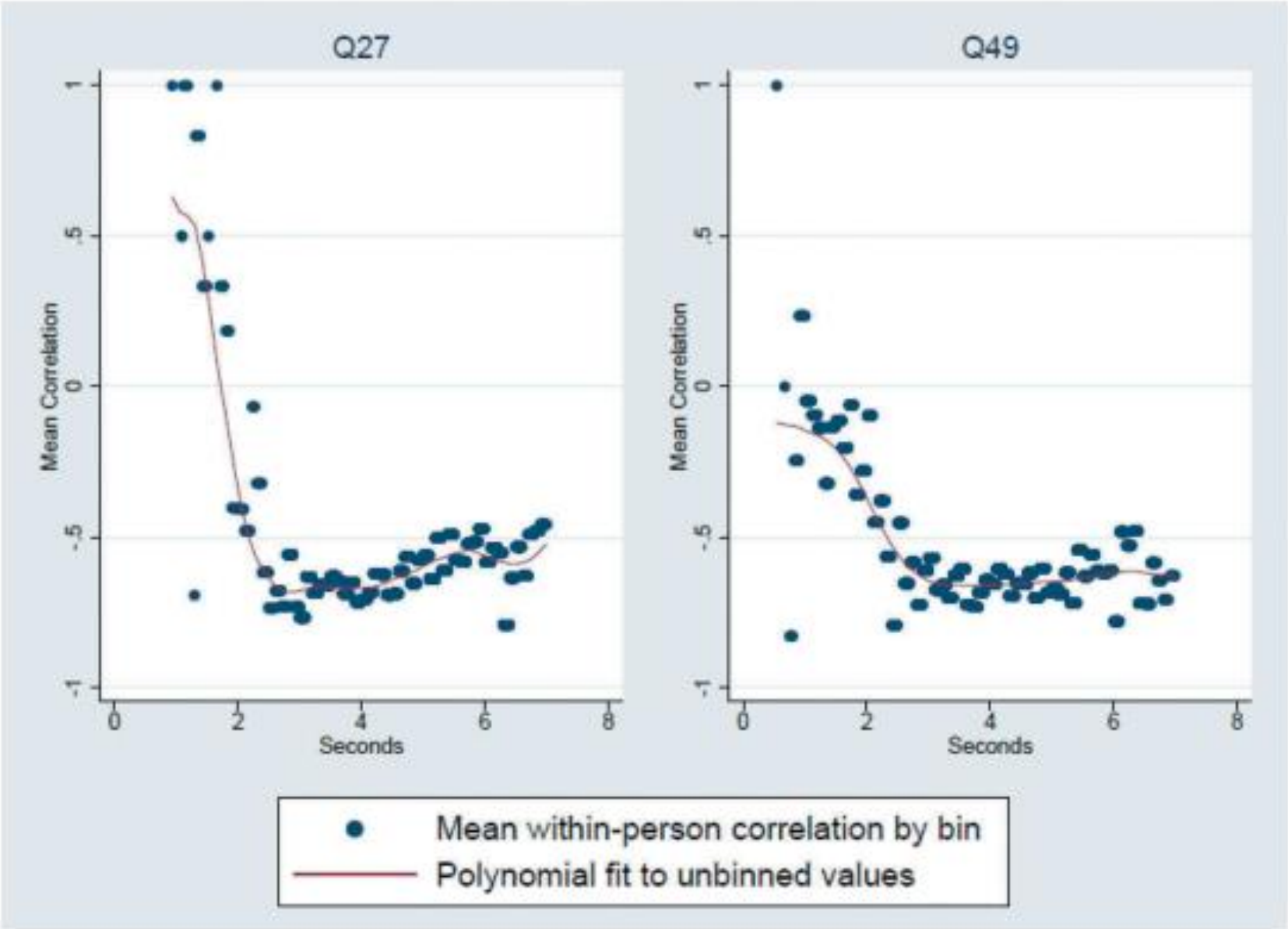
Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, *32*(2), 151-165.

Option 6. Using RT with Other Methods

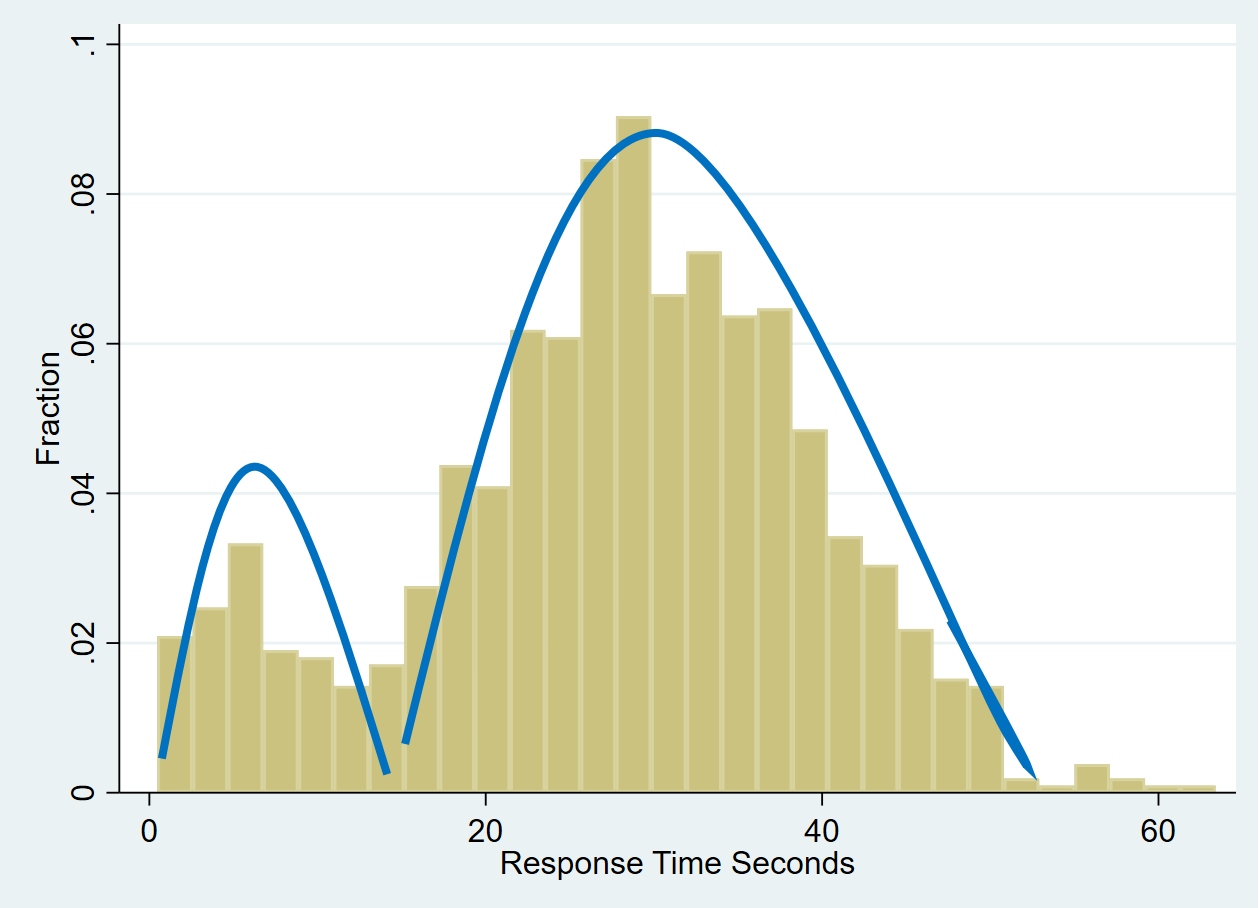


Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, *32*(2), 151-165.

Option 6. Using RT with Other Methods

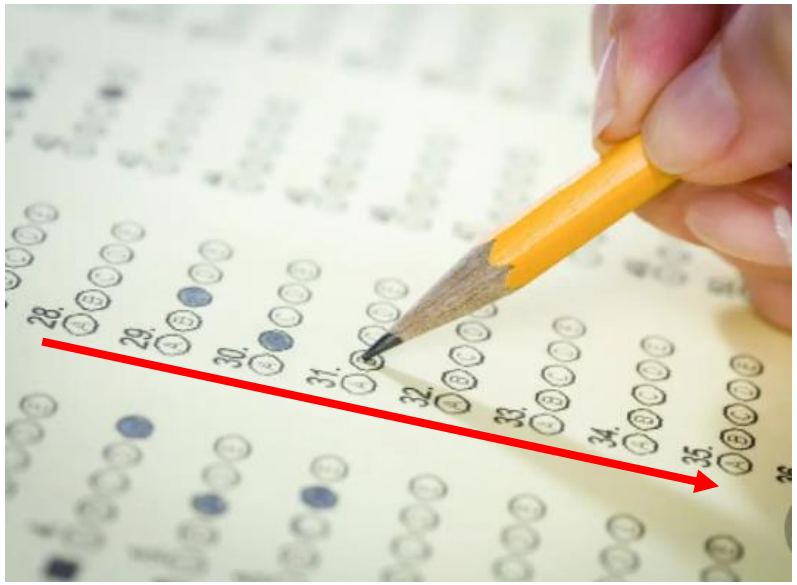


Mixture models also growing in use



Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*, 83-112.

Disengagement as a source of data in its own right?



Soland, J., Zamarro, G., Cheng, A., & Hitt, C. (2019). Identifying naturally occurring direct assessments of social-emotional competencies: The promise and limitations of survey and assessment disengagement metadata. *Educational Researcher*, *48*(7), 466-478.

Conclusions

- There are many ways that researchers have tried to identify low survey effort – major challenge not having correct/incorrect items
- However, unlike response times, many do not allow one to detect low effort on an item-by-item basis
- While one could use response times for survey items, those distributions are often not bimodal
- Therefore, a promising approach is to use response times in conjunction with other detection methods

Module Citation

Soland, J. (2023). Understanding and mitigating the impact of low effort on common uses of test and survey scores [Digital ITEMS Module 32]. *Educational Measurement: Issues and Practice, 42*(2), 75-76.