
The Problem of Low Examinee Effort in Low-Stakes Tests

Steven L. Wise
Institute for Computer-Based Testing
Center for Assessment and Research Studies
James Madison University, USA

Paper presented at the 8th Annual Meeting of the Association for Educational Assessment – Europe, Stockholm, Sweden.

Three Important Requirements for Test Score Validity

- The test items should adequately cover the construct domain.
- The number of items should be sufficient to provide test scores with adequate reliability.
- Test scores should be reasonably free from construct-irrelevant variance (CIV).

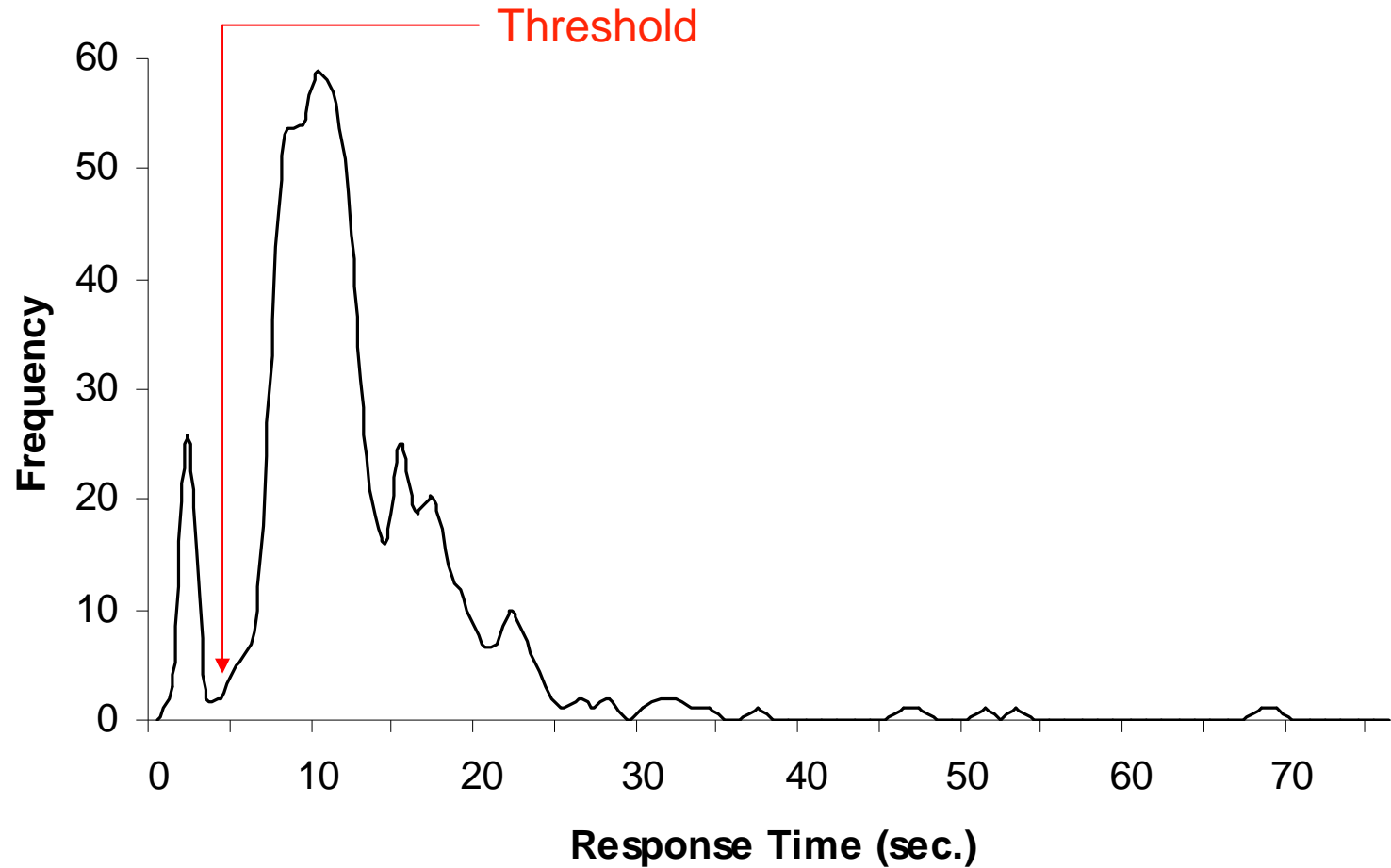
One Source of CIV: Examinee Effort

- The test giver assumes that examinees will give good effort.
- Low effort will lead to negatively biased proficiency estimates (which produces CIV).
- This leads to test scores with lower validity.
- This problem is most prevalent in testing situations without personal consequences for examinees (e.g., TIMMS, PIRLS).

Measuring Effort

- Self-report instruments
 - Short post-test scales using Likert-type items
 - Provides a global measure of effort
 - Respondent truthfulness may be a question
- Item response time-based measures
 - Requires computer-based testing
 - Useful with multiple-choice items
 - Unobtrusive
 - Provides item-by-item measure of effort.
 - Rapid-guessing behavior vs. solution behavior
 - Response time effort

An Item's Response Time Distribution



Response Time Effort (RTE)

$$SB_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } RT_{ij} \geq T_i, \\ 0 & \text{otherwise.} \end{array} \right\}$$

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

Available Strategies for Managing Low Examinee Effort

- Motivation filtering
- Effort-moderated IRT model
- Effort-monitoring computer-based test (CBT)

Motivation Filtering

- Discard data from examinees exhibiting low effort.
- Filtering can be done using either self-report or RTE.
- Requires assumption that examinee non-effort is unrelated to true proficiency.
- Impact:
 - Mean score increase of .2-.4 standard deviations
 - Increased convergent validity

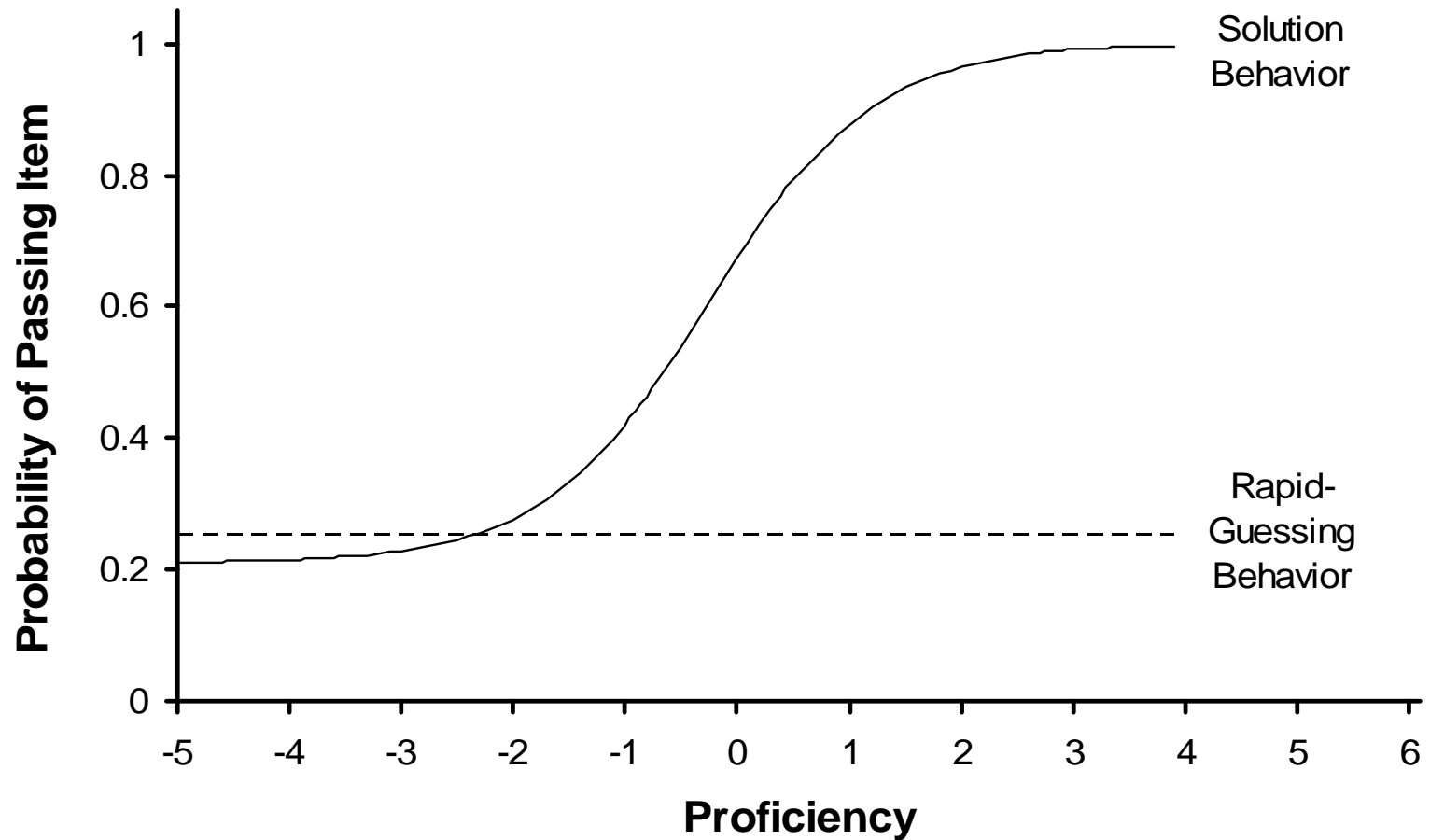
Effort-Moderated IRT Model

- Incorporates rapid guessing into the IRT scoring model. (SB = 0 if rapid guessing; 1 if solution behavior)

$$P_i(\theta) = (SB_{ij})(\text{solution behavior model}) \\ + (1 - SB_{ij})(\text{rapid - guessing behavior model})$$

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

Two Types of Item Response Functions



Impact of the Effort-Moderated IRT Model When Rapid Guessing is Present

- Compared to standard IRT model:
 - Better model fit
 - More accurate estimation of item parameters
 - Higher convergent validity

A Proactive Alternative: The Effort-Monitoring CBT

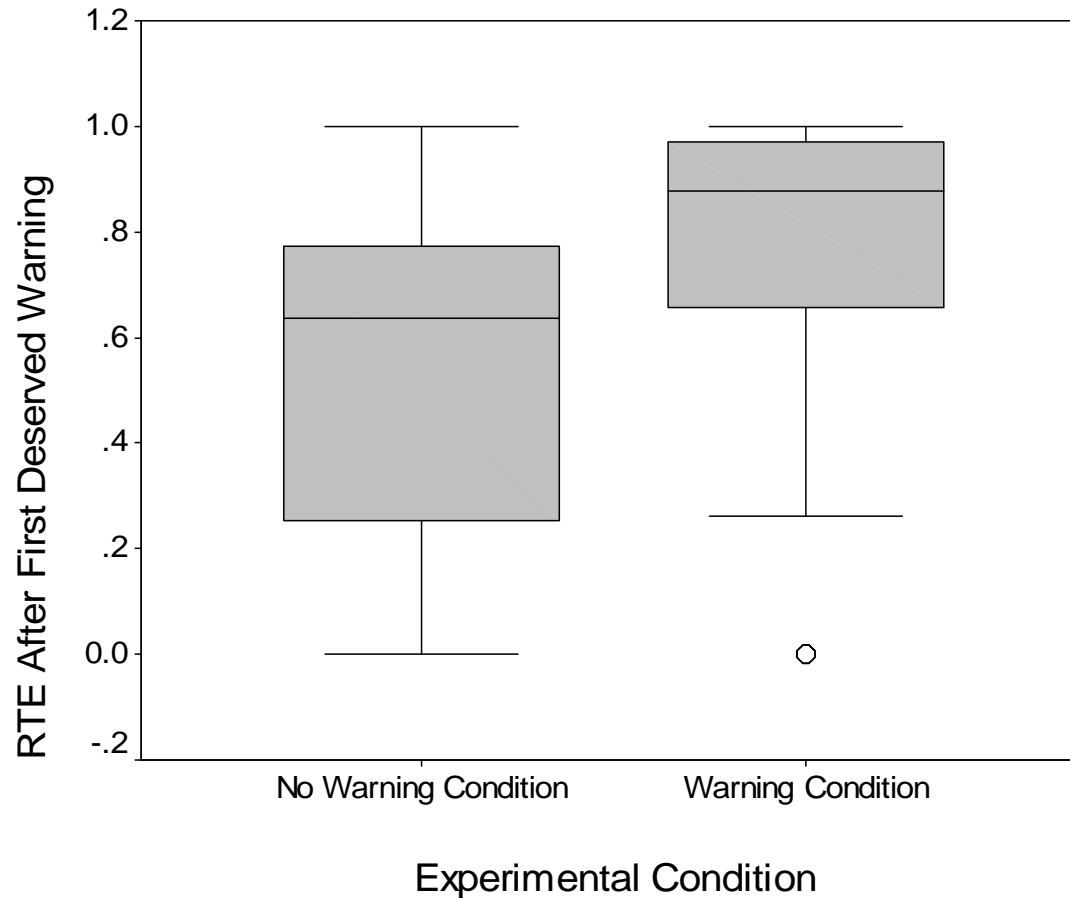
- Rapid guesses could be monitored by the computer as the CBT is administered.
- What if the computer intervened during the test sessions of examinees it detects were not giving effort, and warned them that they should try harder?
- This is called an effort-monitoring CBT.
- Impact:
 - Higher RTE scores
 - Improved test performance
 - Higher convergent validity

Example of an Effort-Monitoring CBT Warning Message

- Your responses to this test indicate that you are not giving your best effort.

It is very important that you try to do your best on the tests you take on Assessment Day. These assessment data are used by the university to better understand what students learn at JMU, and what improvements need to be made. In addition, JMU's assessment data are reported to the state as evidence of what JMU's students know and can do.

Impact of Effort-Monitoring CBT on Examinees Deserving a Warning Message



Effect of Effort-Monitoring CBT on Convergent Validity

	Correlation with Test Performance (No Warning)	Correlation with Test Performance (Warning)	Change in Shared Variance
SAT- Verbal	.22	.33	+.06
SAT- Math	.11	.43	+.17
GPA	.22	.35	+.07

Summary

- Examinee non-effort can degrade the validity of test score-based inferences.
- There are several new procedures that can be used when non-effortful responses are present in the test data.
- These procedures have been shown to improve test score validity.