2021

Application of a Theory-Driven Approach to Detect Cognitively Disengaged Test-Taker Behavior

Burcu Arslan
Educational Testing Service, barslan@ets.org

Blair Lehman
Educational Testing Service, blehman@ets.org

Jesse R. Sparks
Educational Testing Service, jsparks@ets.org

Jonathan Steinberg
Educational Testing Service, jsteinberg@ets.org

Application of a Theory-Driven Approach to Detect Cognitively Disengaged Test-Taker Behavior

Burcu Arslan, Blair Lehman, Jesse R. Sparks, & Jonathan Steinberg
Educational Testing Service, Princeton, NJ, USA

Abstract:

Bottom-up, data-driven response filtering methods that exclude unrealistically fast responses from calculating test scores have been successfully applied to improve test validity. We introduce a top-down, theory-driven method to detect cognitively disengaged behavior, compare it with a data-driven method using data from a nationally representative reading assessment, and discuss its potential and limitations.
Keywords: *test-taker disengagement, rapid guessing behavior, normative methods, theory-driven method*

1. Introduction

Cognitively disengaged test-taker behavior can be defined as a response given by a test-taker without performing the necessary mental operations to solve a problem or reason about it. For example, selecting a multiple-choice (MC) option without reading the item stem (i.e., rapid guessing behavior; RGB), or reading the item stem and starting to perform mental operations to reason about the question and then abandoning the process without completing all necessary mental operations (i.e., partial solution behavior). Detecting cognitively disengaged test-taker behavior is critical for test validity, especially for low-stakes assessments, such as the National Assessment of Educational Progress (NAEP), because disengaged behavior does not represent test-taker knowledge, skills, and abilities (Finn, 2015; Wise, 2017). Excluding responses that have unrealistically fast response times from calculating test scores has been successfully applied to improve test validity (Wise & Kong, 2005; Wise, 2017). Existing time-based methods for detecting RGB range from assigning a fixed threshold for all items to assigning item-specific thresholds based on a certain percentage (e.g., 10%; Normative Method (NM); Wise & Ma, 2012) of the average response time on each item (see Wise, 2017). Although applying the NM to filter RGB from scoring improves test validity, there are a couple of limitations of this method.

The first relates to the data-driven nature of defining NM-based thresholds. Being data-driven makes the threshold dependent on the sample's engagement within an item, which might be problematic when there is general disengagement in a specific item. The second limitation is that the NM detects *only* RGB, and responses that are not detected as RGB are classified as solution behavior although these responses are not necessarily indicative of *effortful* solution behavior (Finn, 2015; Lindner et al., 2019; Wise, 2017).

To mitigate these limitations, we introduce a theory-driven method to detect cognitively disengaged test-taker behavior, called Detecting Engagement Levels with Cognitive Modeling (DELCOM; Arslan et al., 2021) that makes it possible to define the threshold for each complex

or traditional item for the quickest possible solution behavior *a priori* to the data collection without sample bias so that cognitively disengaged behavior beyond RGB can be filtered from the test scores to further improve test validity.

2. Theoretical Framework

The DELCOM method is based on the cognitive architecture Adaptive Control of Thought–Rational (ACT-R; Anderson, 2007), which is a computational implementation of a unified theory of human cognition. This means that ACT-R includes the fixed mechanisms and structures that underlie human cognition, and it can simulate human cognition and behavior together with their timing. ACT-R has been used heavily in cognitive science research as well as Intelligent Tutoring Systems (Anderson et al., 1995) to predict and explain learner behavior.

ACT-R has different modules mapped onto the human brain, such as a visual module that perceives the environment and a manual module that acts on the environment. There are default parameters for timing of cognitive processes (i.e., micro-level timing values) based on decades of research (Anderson, 1990; Anderson & Lebiere, 1998; Anderson, 2007). For example, it takes 50 ms to encode a visual chunk of information or 250 ms to prepare for a motor movement after deciding to press a button. In addition to perception and action modules, ACT-R has cognition related modules, such as declarative and procedural memory.

To detect cognitively disengaged behavior, DELCOM requires: a) calculating reading time based on prior literature (e.g., 371 words per minute for skimming; Carver, 1992); b) inferring proficient test-taker cognitive processes at a fine-grained level (e.g., retrieval of a story fact, comparing the retrieved fact to answer options, making a decision); c) determining time for the cognitive processes based on ACT-R; d) calculating time to perform the action to give a response; and e) summing all these values to calculate the threshold for the quickest possible effortful response to an item (i.e., a very conservative lower-bound estimate of proficient solution behavior).

3. Study Purpose

In another study (Lehman & Arslan, 2021), we showed that the DELCOM method was able to detect cognitively disengaged behavior beyond RGB in more complex items (i.e., drag-and-drop mathematics items). The purpose of the current study is to give an overview of the DELCOM method and validate it in application to traditional MC items in a nationally representative sample of grade 12 students who took a digitally delivered reading assessment. The reason for using MC items to validate DELCOM is that, in a less complex, traditional MC item, a theoretical time threshold of a very conservative proficient solution behavior assigned by DELCOM and a data-driven RGB threshold assigned by the NM are expected to be similar since proficient solution time is expected to be quick in simple MC items. Therefore, we expect that the DELCOM thresholds assigned for each item should not be very different than those from the NM.

4. Research Question

Are the theory-driven DELCOM thresholds assigned for each MC item comparable to the NM thresholds?

5. Methodology

Data from 7,355 grade 12 students were analyzed. Specifically, we sampled data for four 30-minute assessment blocks (students may have taken 1 or 2 of these target blocks). Each assessment block included 9-10 items, including MC and constructed response (CR) items; all MC items had four answer options. For each MC item, five different NM thresholds were established by calculating 10%, 15%, 20%, 25%, and 30% of the average time spent on each item. The DELCOM threshold was established as described above.

6. Results and Conclusions

As expected, the DELCOM threshold was similar to the NM thresholds in MC items (see Figure 1). Correlation analysis between each threshold and block score also confirmed this expected similarity (see Table 1).

Table 1

Correlations between the thresholds and block score

Threshold	*Block 1*	*Block 2*	*Block 3*	*Block 4*
NM 10%	.345**	.303**	.279**	.276**
NM 15%	.412**	.342**	.323**	.379**
NM 20%	.445**	.375**	.360**	.425**
NM 25%	.453**	.388**	.376**	.437**
NM 30%	.460**	.379**	.382**	.400**
DELCOM	.447**	.382**	.371**	.448**
Observations	1,793	1,824	1,836	1,793

*Note: NM = normative method; ** indicates p < .01*

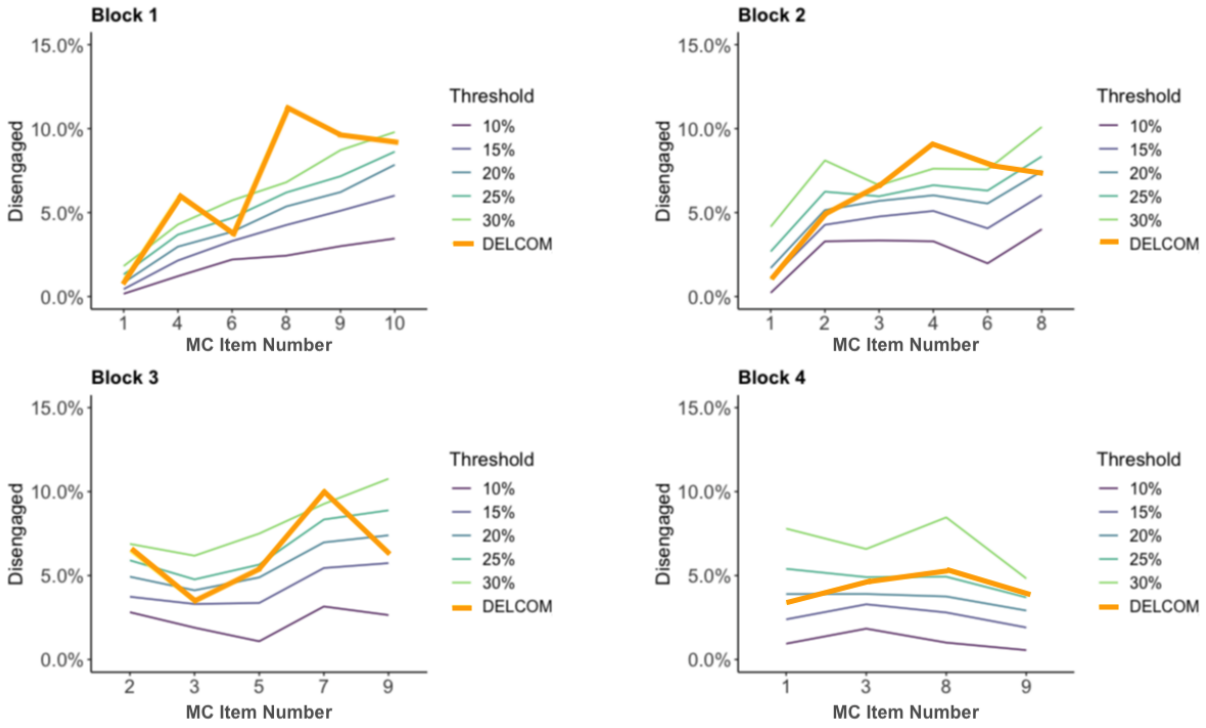
The DELCOM threshold represents a very conservative estimate of the quickest possible solution behavior based on a unified theory of cognition that has been empirically validated with decades of research. Therefore, when the percentage of disengaged behavior classified by the NM thresholds is higher than the disengaged behavior that is classified by the DELCOM threshold, it indicates that the assigned NM threshold for that item is stringent (e.g., see Figure 1, Block 1, NM thresholds 25% and 30% for item number 6; or Block 4, NM threshold 30%, item numbers 1, 3, 8, and 9). Moreover, when the percentage of disengaged behavior classified by the NM thresholds is lower than the disengaged behavior that is classified by the DELCOM threshold, it indicates that the NM threshold that is assigned for that item is too lenient (e.g., see Figure 1, Block 1, NM thresholds 10%, 15%, 25% and 30% for item numbers 4, 8, and 9; or Block 4, NM thresholds 10% and 15%, item numbers 1, 3, 8, and 9). This is because DELCOM thresholds are based on a well-established cognitive theory whereas

NM thresholds are not, and it is *not* possible to respond to an item faster in a cognitively engaged way than the DELCOM threshold.

Despite its advantages compared to existing methods, DELCOM cannot currently be directly applied to CR items. In addition to item response times, incorporating students' actions is important to further improve test validity (e.g., Sahin & Colvin, 2020). We are planning to tackle these issues in future work. Overall, DELCOM is a valid, theory-driven method to detect cognitively disengaged test-taker behavior beyond RGB and can be applied to different item formats (e.g., MC, drag-and-drop) in different domains (e.g., reading, mathematics, science).

Figure 1

Comparison of DELCOM threshold with the different normative method (NM) thresholds in multiple choice (MC) items in four different blocks



References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, *4*(2), 167-207.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195324259.001.0001

Anderson, J. R., & Lebiere, C. J. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Arslan, B., Lehman, B., & Lindner, M. (July, 2021). *Introducing a theory-driven method to detect different levels of engagement in technology-enhanced items*. [Paper presentation]. The 12th Conference of the International Test Commission, Virtual.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*(2), 84-95.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report (RR–15-19)*, 1–17. doi: 10.1002/ets2.12067

Lehman, B., & Arslan, B. (July, 2021). *Test-taker engagement levels when responding to drag-and-drop mathematics items.* [Paper presentation]. The 12th Conference of the International Test Commission, Virtual.

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, 10, 1533.

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessments in Education*, 8, 1-24.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L., & Ma, L. (2012, April). Setting response time thresholds for a CAT item pool: The normative threshold method. *In annual meeting of the National Council on Measurement in Education*, Vancouver, Canada (pp. 163-183).

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52-61.