ISBN: 978-989-8704-43-6 © 2022

THE PATTERN OF TEST-TAKING EFFORT ACROSS ITEMS IN COGNITIVE ABILITY TEST: A LATENT CLASS ANALYSIS

Hanif Akhtar
Faculty of Psychology, Universitas Muhammadiyah Malang, Indonesia
Doctoral School of Psychology, ELTE Eotvos Lorand University, Hungary

ABSTRACT

When examinees perceive a test as low stakes, it is logical to assume that some of them will not put out their maximum effort. This condition makes the validity of the test results more complicated. Although many studies have investigated motivational fluctuation across tests during a testing session, only a small number of studies have investigated motivational fluctuation across items within a single test. This study aims to examine the pattern of test-taking effort across items in cognitive ability tests when items are presented in a random order manner. Response Time Effort (RTE) was used as a measure of test-taking effort. This measure calculates the proportion of rapid responses in the test based on the response times for each item. Data from 213 university students completing the inductive reasoning test was examined using latent class analysis. The results suggested that examinees in low-stakes testing have different patterns of effort across items. Examinees who consistently provided a high level of effort across items had higher test performance, test-taking engagement, and RTE. Item position and item difficulty are also correlated negatively with test-taking effort. Implications of these results for researchers and practitioners are discussed.

KEYWORDS

Response Time Effort, Test-Taking Effort, Latent Class Analysis, Low-Stakes Assessment, Cognitive Ability

1. INTRODUCTION

Research in psychometrics often relies on the test taker's motivation to do well on low-stakes assessments (e.g., data collection for validation study). When examinees perceive a test as low-stakes, it is logical to presume that some individuals will not put out their maximum effort, especially if there are no personal consequences to their test performance. This circumstance makes the validity and interpretation of the test results more complicated. We define low-stakes testing as any testing that has no meaningful consequence for examinees (e.g., survey). Conversely, high-stakes testing has a meaningful consequence for examinees (e.g., personnel selection). It should be noted that even though there are no personal effects on test performance, many examinees appear to put up much effort when taking low-stakes tests (e.g., Barry & Finney, 2016; Pastor et al., 2019; Wise & Kingsbury, 2016; Wise & Kong, 2005). Thus, it is essential to identify those examinees who put in a low effort because that could lead to bias in the test data.

Several measures have been developed to measure test-taking effort. The most popular measures are those that use self-reports that should be completed right after taking the test (e.g., Freund et al., 2011; Knekta & Eklöf, 2015; Sundre & Moore, 2002). However, using self-report has three fundamental limitations. First, it is unclear how honestly examinees will inform their test-taking effort. Second, it is mildly intrusive and takes time. Third, self-reported measures only give a global index of effort during a test, making it difficult to investigate any variations in the effort that occur throughout a test event.

Another way of measuring test-taking motivation that involves the whole test-taking process is based on response times using Response Time Effort (RTE; Wise & Kong, 2005). RTE can be implemented in computer-based tests based on the assumption that unmotivated examinees will answer too quickly (i.e., before they have time to read and fully consider the item) when administered an item. This measure attempts to estimate the percentage of rapid guessing behavior in the test based on the reaction times for each item. Unlike

self-report, which provides a global measure of effort, RTE would allow researchers to investigate changes in the examinees' effort during a testing session because the data of response time is available for each item.

Several studies have found that test-taking efforts can rise or fall during testing sessions (Barry et al., 2010; Barry & Finney, 2016; Pastor et al., 2019; Penk & Richter, 2017). These studies illustrate how test-taking efforts change across tests within a single testing session but do not report how efforts change across items. In addition, several item characteristics might influence the test-taking effort. For example, item location (Pastor et al., 2019; Wise & Kingsbury, 2016), item difficulty (Asseburg & Frey, 2013), item length (Pastor et al., 2019), and item type (Sundre & Kitsantas, 2004) affect test-taking effort. By using RTE, researchers can explore the pattern of test-taking effort during a single test. In research-based testing, this is essential. Knowledge of how test-taking effort varies across items can enlighten the research design. For example, patterns characterized by engagement in the first half of a test might indicate reducing a test. Or, if the testing is used to estimate item parameters, it might be recommended to order the item randomly.

Although RTE is beneficial over self-report scales for discovering the pattern of test-taking effort across items within a single test, only a few researchers used it in research-based testing (e.g., Pastor et al., 2019; Wise & Kingsbury, 2016). Wise and Kingsbury (2016) conducted the study in the context of a single adaptive test administered to K-12 students. At the same time, Pastor et al. (2019) conducted the study in the context of three different low-stakes tests for higher education program assessment administered to undergraduate students. The present study was built on the previous work of Pastor et al. (2019) and Wise and Kingsbury (2016) by exploring whether individuals differ in their test-taking effort patterns across items within a test when items are presented in random order manner. Specifically, a latent class analysis (LCA; Lazarsfeld & Henry, 1968) was used with data collected from undergraduate students completing a cognitive ability test. There are three research questions in this study.

RQ1. Do participants show different patterns of effort?

RQ2. Do participants with different patterns of effort differ on other associated variables (e.g., test performance , self-reported effort) in expected ways?

RQ3. How do item characteristics (i.e., item position and item difficulty) relate to effort?

To pursue RQ1, LCA was performed to determine the presence of different patterns of test-taking effort. As RQ1 is fully exploratory, there is no hypothesis for this research question. To address RQ2, subsequent analyses were performed to provide evidence of the validity of the newly emerging LCA classes. The hypothesis for RQ2 was that examinees with high and relatively consistent effort across all items have higher test-taking effort (measured by RTE and self-report) and test performance. Spearman correlation was run in order to comprehend the variation in test-taking effort across items (RQ3). The results of this study together reveal whether various patterns of solution behavior exist when items are given in a random sequence and whether item characteristics (item difficulty or item location) explain variations in test-taking efforts. The hypothesis for RQ3 was that both item difficulty and item position were negatively correlated with test-taking effort.

2. METHODS

2.1 Data Source

Data were collected for research purposes aimed at developing a reasoning test. A total of 213 participants (154 females) participated in this study. Participants were undergraduate students in the Faculty of Psychology at the University of Muhammadiyah Malang, Indonesia, aged 18 to 23 years (M = 19.89, SD = 0.78). Participants were recruited through the instructor's information in the class during the course of psychological test construction. All participants completed the online inductive reasoning test.

2.2 Measures

The data from 'the odd one out' test were used to determine how motivational patterns fluctuate across items. The odd one out test is a cognitive-based multiple-choice measure intended to measure inductive reasoning. There are six pictures in an item, and one out of six pictures has the most different characteristics based on a

certain principle. Examinees were asked to find one picture that was most different from the others. Our research team developed the test. For online data collection, the platform PsyToolkit was used (Stoet, 2010, 2016). Each item was displayed on a separate page. Item reaction time was calculated as the number of milliseconds between an item appearing and examinees clicking "next" to move on to the next item.

Two measures of test-taking efforts were used: RTE and self-report. RTE was an average of Solution-Behavior (SB) of all items or scales answered. RTE values near 1 indicate high effort, and values near 0 indicate low effort. Examinees' response was classified as SB if they responded to the item above the threshold (scored 1). In contrast, if examinees responded to the item below the threshold, their response should be considered rapid-guessing behavior (scored 0). The 10% Normative Threshold (NT10) approach (Wise & Ma, 2012) was used to determine the threshold. This approach proposes that 10% of the average response time be used for the threshold. For example, if it takes participants an average of 50 seconds to respond to an item, a NT10 would be 5 seconds. If participants responded to the item above the threshold, their response should be SB.

At the end of the testing session, two self-report items measuring examinees' attitudes toward the test were presented. One item measures test-taking efforts (i.e., "I took this test seriously"), and the other item measures test-taking engagement (i.e., "I enjoy doing tests"). These items were responded to on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

2.3 Data Analyses

To answer RQ1, latent class analysis (LCA; Lazarsfeld & Henry, 1968). LCA was performed using SB indices to investigate the number of different patterns of test-taking effort. Several LCA models were fit to the data, with each model specifying a different number of classes (K). Each LCA model has a log-likelihood (LL) value, with values closer to zero suggesting a higher likelihood of the data. Because the LL will always be closer to zero for models with more classes, The Bayesian information criterion (BIC; Schwarz, 1978) and The Akaike information criterion (AIC; Akaike, 1974) were used to evaluate model-data fit. The model with the lowest BIC and AIC values was deemed to be the optimal one. The preliminary analysis showed similar results to Pastor et al. (2019) study, indicating that for three-class models, the condition numbers were fewer than 10-6. As a result, only one- and two-class models were fit to the data in this investigation. LCA was performed using 'sirt' package (Robitzsch, 2021) in R software (R Core Team, 2013).

To answer RQ2, examinees were classified to the class with the highest posterior probability. The resultant class membership was employed in the t-test analysis to determine the association between class membership and each associated variable. Test performance, self-reported effort, self-reported engagement, and RTE were used as associated variables. Classes with higher and more consistent engagement in test-taking effort were expected to have higher test performance, self-reported effort, and RTE than classes with lower and less consistent engagement in solution behavior.

To answer RQ3, item difficulty (based on Rasch analysis), item position, and SB indices were correlated. SB index was a measure of test-taking efforts across items based on the average SB of all examinees' answers on a particular item. SB indices values near 1 indicate high effort, and values near 0 indicate low effort.

3. RESULTS

3.1 Preliminary Analysis

Preliminary analysis was conducted to examine descriptive statistics and intercorrelation among variables studied. A summary of the preliminary analysis is presented in table 1. RTE significantly positively correlated with test-taking engagement ($r = .22$, $p < .001$), test-taking effort ($r = .50$, $p < .001$), and test performance ($r = .38$, $p < .001$). It indicates that RTE was a valid measure of test-taking efforts. There are no gender differences in test-taking engagement, test-taking effort, RTE, and performance.

Table 1. Descriptive statistics and intercorrelation among variables

Variables	Female		Male		t	1	2	3
	N	Mean (SD)	N	Mean (SD)				
1. Engagement	154	3.3.2 (1.11)	59	3.44 (1.13)	0.68			
2. Effort	154	3.93 (0.99)	59	3.88 (1.04)	-0.31	.45***		
3. RTE	154	0.98 (0.09)	59	0.96 (0.11)	-1.43	.22**	.50***	
4. Performance	154	-0.01 (0.80)	59	0.03 (0.75)	0.31	.12	.25***	.38***

Note: RTE = Response Time Effort, ***P < .001

3.2 Pattern of Test-Taking Effort

Table 2 displays the outcomes of the one- and two-class LCA models. The AIC and BIC for the two-class models were lower than for the one-class models, indicating support for two-class solutions. The likelihood ratio tests provided similar support, indicating a considerably superior fit for the two-class models. Figure 1 depicts the estimated probabilities of engaging in test-taking effort for the two-class solutions based on item position and class. Class 1 is distinguished by lower probability and higher variability in solution behavior across items. Class 2 encompasses the majority of the examinee population and is distinguished by high and relatively constant engagement in solution behavior across all items.

Table 2. Fit indices for latent class analysis models

Model	loglike	Deviance	Npars	Nobs	AIC	BIC	LRT *p*
One-class	-1164.21	2328.43	50	213	2428.43	2596.50	-
Two-class	-622.93	1245.86	101	213	**1447.87**	**1787.36**	< **0.001**

Note: Npars = number of parameter estimated, Nobs = number of sample, AIC = Akaike information criterion, BIC = Bayesian information criterion, LRT *p* = *p*-value of Likelihood Ratio Test

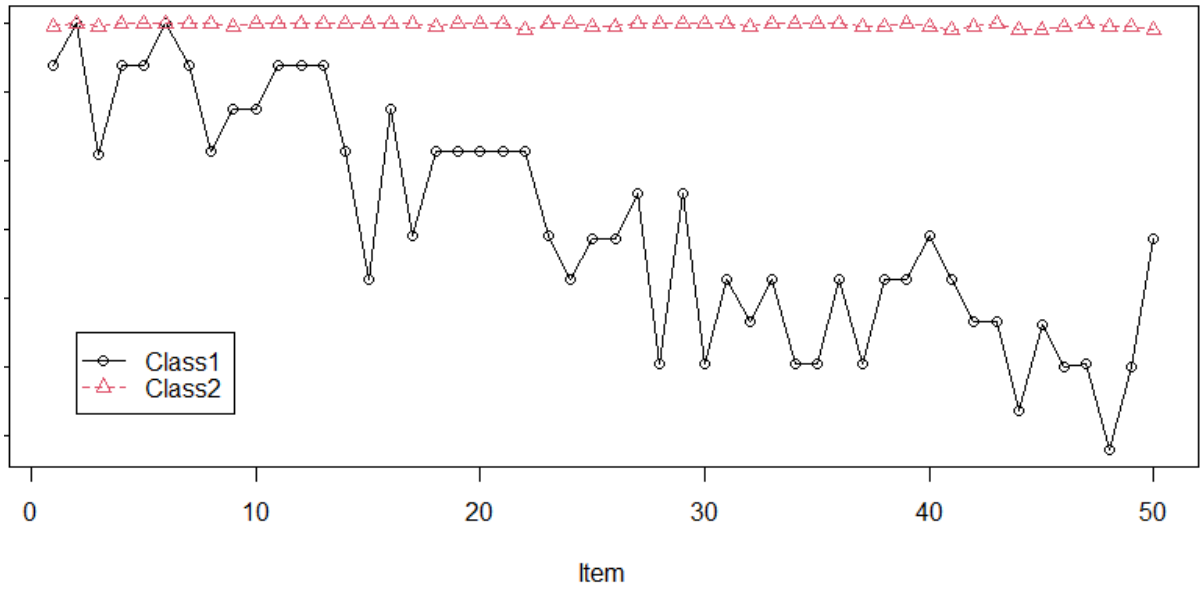

Figure 1. Estimates of the probability of engagement in solution behavior by item position using the two latent-classes model

3.3 Association between Class Membership and Auxiliary Variables

In the t-test analysis, the resultant class membership was employed to capture the relationship between class membership and each auxiliary variable. The results of the analyses using auxiliary variables (test-taking engagement, test-taking efforts, RTE, and test performance) are shown in Table 3. There were statistically significant differences across groups for all auxiliary variables. Every comparison has a substantial effect size regarding practical significance, with Cohen's d greater than 0.80 (Sawilowsky, 2009). Not only were these disparities notable, but they also pointed in a direction that supported the notion that Class 1 is less motivated than Class 2. Specifically, Class 1 had a substantially lower test-taking engagement, test-taking effort, RTE, and test performance relative to Class 2.

Table 3. T-test results across class membership

	Class	N	Mean (SD)	*t*	*d*
Engagement	1	17	2.53 (0.87)	3.25***	1.82
	2	196	3.43 (1.11)		
Effort	1	17	2.47 (1.07)	6.80***	1.72
	2	196	4.04 (0.90)		
RTE	1	17	0.72 (0.21)	18.96***	4.79
	2	196	1.00 (0.01)		
Performance	1	17	-0.85 (0.95)	4.86***	1.23
	2	196	0.07 (0.73)		

Note: RTE = Response Time Effort, ***$p < .001$

3.4 Item Characteristic Analyses

Spearman correlation analysis found that SB indices is significantly negatively correlated with item difficulty ($r = -.44$, $p < .01$) and item position ($r = -.88$, $p < 0.01$). The test-taking effort appears to decline significantly as the test progresses and items become difficult. Figure 2 shows the scatterplot for the correlation between SB indices and item characteristics.

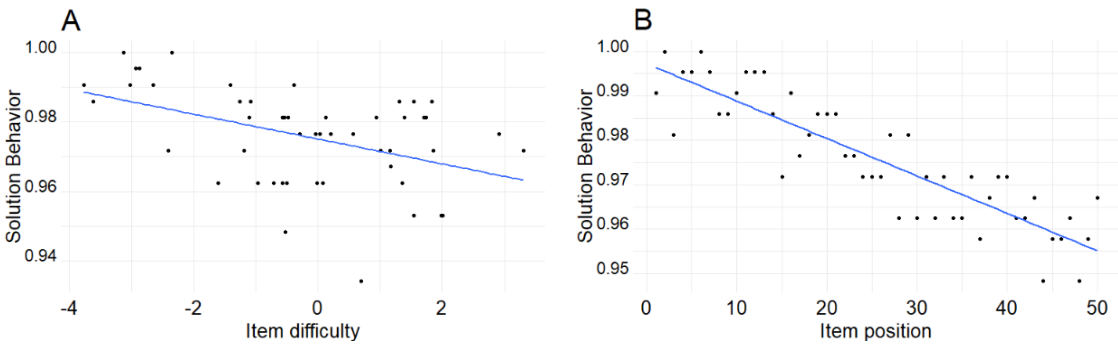

Figure 2. Correlation between SB indices and item difficulty (A) and item position (B)

4. DISCUSSION

Although some studies have examined the motivational change between tests during a testing session, few have focused on motivational change across test items. LCA was used in combination with fixed-item testing to investigate if there are patterns of test-taking effort across items and which item properties explain variations in test-taking effort. Two types of examinees with distinct patterns of test-taking effort across items were identified. The first class, which comprised 8% of the population, exhibited considerably less and more varied effort throughout the tests. The second class, which included 92% of the population, demonstrated high and consistent effort throughout the tests. This result is consistent with previous studies (Barry & Finney, 2016;

Pastor et al., 2019; Wise & Kingsbury, 2016; Wise & Kong, 2005) that even though there are no personal consequences for test performance, many examinees appear to put up a lot of effort when taking the test.

The two-class solution was supported by additional analysis. Examinees who consistently engaged in solution behavior across items showed greater levels of test-taking engagement, test-taking effort, RTE, and test performance than examinees who were prone to rapid-guessing behavior. Item characteristic analyses indicated that the probability of solution behavior decreases as item position increases. This finding is consistent with previous studies (Pastor et al., 2019; Wise & Kingsbury, 2016) despite the difference in the order of item presentation. Item difficulty was also negatively correlated with test-taking effort. However, the correlation coefficient was smaller than the correlation between test-taking effort and item position. This result indicates that item position has a greater influence on test-taking effort.

The findings of this study have several implications for testing practices. First, when data collection aims to estimate item parameters on a cognitive ability test (e.g., item difficulty level), randomly ordering the item may be a good practice to reduce the item-order effect. This result suggested that items presented at the end of the test tend to be answered carelessly by the examinees, which causes the estimated parameter to be misleading. It implies that the incorrect answers given by the examinees to the items presented at the end of the test do not necessarily indicate that the items are difficult. It could be the result of low test-taking effort. Thus, randomly ordering the item might reduce this item-order effect. Based on these findings, practitioners could consider using shorter tests in low-stakes testing contexts.

Second, providing a motivator item at the beginning of the test might be beneficial to maximizing examinees' test-taking effort. A motivator item is a relatively easy item. From the expectancy-value theory (Wigfield & Eccles, 2000) perspective, an easy item could lead to high expectancy, resulting in a high effort. In addition, these findings indicate that easier items resulted in a higher test-taking effort. Thus, using an easier item in the context of low-stakes testing might reduce the source of test score invalidity due to low test-taking efforts. This practice is important especially if test results are used for group comparison studies (e.g., comparing cognitive abilities across nations).

Third, it is found that the two classes differ significantly in their test performance. When a test is regarded as low-stakes for examinees yet high-stakes at higher levels, the validity of the interpretation of the findings is jeopardized. When less motivated examinees score below their real ability, the test-taking effort may be a cause of bias. This issue is not only on an individual level but also on a group level, especially when different groups have different motivations and their outcomes are compared. Based on these findings, practitioners may choose to include some sort of motivational intervention to conclude their assessments. The article from Rios (2021) provides a good review based on empirical studies about how to increase test-taking effort in low-stakes assessments.

This study has several limitations. First, this study only examined the pattern of test-taking effort in one kind of cognitive ability test. A different test might have different results as item type affects test-taking behavior (Pastor et al., 2019; Sundre & Kitsantas, 2004). Second, gender was not equally distributed in this study, with twice the number of females. As males were found rapidly guess nearly twice as often as females (Soland, 2018), the results of this study might be influenced. Third, examinees are limited to university students. Fourth, this study used NT10 only to determine the threshold for RTE.

5. CONCLUSION

In summary, this study found that examinees in low-stakes testing have different patterns of engagement in test-taking efforts across items. Examinees who engaged in solution behavior consistently across items had higher test-performance, test-taking engagement, test-taking effort, and RTE than examinees who had a lower engagement in solution behavior. The results of item characteristic analyses indicated that item position and item difficulty are negatively correlated with test-taking effort. Due to several limitations of the study mentioned previously, this study should be considered a preliminary examination of the pattern of test-taking effort across items in the cognitive ability test. Further investigation with a different test, sample characteristics, and analysis procedures should be performed to examine the generalizability of these findings.

REFERENCES

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEICE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1093/ietfec/e90-a.12.2762

Asseburg, R., & Frey, A. (2013). Too hard , too easy , or just right ? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92–104.

Barry, C. L., & Finney, S. J. (2016). Modeling Change in Effort Across a Low-Stakes Testing Session: A Latent Growth Curve Modeling Approach. *Applied Measurement in Education*, *29*(1), 46–64. https://doi.org/10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? a high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629–634. https://doi.org/10.1016/j.paid.2011.05.033

Knekta, E., & Eklöf, H. (2015). Modeling the Test-Taking Motivation Construct Through Investigation of Psychometric Properties of an Expectancy-Value-Based Questionnaire. *Journal of Psychoeducational Assessment*, *33*(7), 662–673. https://doi.org/10.1177/0734282914551956

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton-Mifflin.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of Solution Behavior across Items in Low-Stakes Assessments. *Educational Assessment*, *24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.r-project.org

Rios, J. (2021). Improving Test-Taking Effort in Low-Stakes Group-Based Educational Testing: A Meta-Analysis of Interventions. *Applied Measurement in Education*, *34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741

Robitzsch, A. (2021). *sirt: Supplementary Item Response Theory Models*.

Sawilowsky, S. S. (2009). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *121*(12), 1–26. https://doi.org/10.1177/016146811812001202

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods 2010 42:4*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2016). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Https://Doi.Org/10.1177/0098628316677643*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, *14*(1), 8–9.

Sundre, Donna L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-talking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L., & Kingsbury, G. G. (2016). Modeling Student Test-Taking Motivation in the Context of an Adaptive Achievement Test. *Journal of Educational Measurement*, *53*(1), 86–105. https://doi.org/10.1111/jedm.12102

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. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). Setting Response Time Thresholds for a CAT Item Pool: The Normative Threshold Method. In *Annual meeting of the National Council on Measurement in Education*.