Modeling Response Time: The F > C Phenomenon and the Distance-Difficulty Hypothesis

Nicole Bonge[a,*], Ronna C. Turner[a]

[a] *University of Arkansas,*

Abstract

Response time has long been thought to be closely related with intelligence via processing speed. If asked to describe a genius, one might describe a person who can do complex mental calculations quickly and with ease. This stereotype that students with high ability level tend to answer questions fastest has come under question, however, with research indicating that response time is a complex process, dependent on more than just ability level. In this study, we use multilevel linear models to analyze the responses and response times on the 2019 TIMSS mathematics achievement test for eighth-grade students in the United States. Our results demonstrate response time's dependency on student ability level, whether the student answered the item correctly (F > C phenomenon), and item difficulty in relation to the student's ability level (distance-difficulty hypothesis). We find evidence to support the F > C phenomenon, the distance-difficulty hypothesis, and an interaction between the two. Our results affirm the complexity of cognitive processes involved in item responses, and challenge the widespread use of response time as a stand-alone metric to assess item quality and respondent effort.

Keywords: Response Time, Item Response Theory, Multilevel Modeling, F > C Phenomenon, Distance-Difficulty Hypothesis

1. Introduction

Researchers have long contemplated the role of response time in achievement testing, and how time-to-completion can inform about an item functioning and participant performance. Among researchers, it is a widely held belief that response time is indicative of the cognitive effort required to answer an item (Höhne et al., 2007), so researchers use response time to gauge item quality and comprehensibility (Bassili et al., 1996; Lenzner, 2012; Lenzner et al., 2010), to

*Corresponding author

Email addresses: ngbonge@uark.edu (Nicole Bonge), rcturner@uark.edu (Ronna C. Turner)

detect respondent fatigue (Nguyen, 2017), effort (Bowling et al., 2023; Krosnick, 1991; Ulitzsch et al., 2022), and other characteristics such as persistence, motivation, and disengagement (e.g., Nagy & Ulitzsch, 2021; Wise & Demars, 2006). Researchers have also used response time to investigate other cognitive processes underlying item responses (De Boeck & Jeon, 2019), with recent research challenging the long-held "faster equals smarter" stereotype (Gernsbacher et al., 2020).

In this study, we extend previous models posed by Tancoš et al. (2023) to achievement data, arguing that response time depends on (1) item difficulty, (2) respondent ability level, (3) the distance between the item's difficulty and the respondent's ability level (referred to as the distance-difficulty hypothesis; Thissen, 1983), and (4) whether the respondent answered the item correctly (F > C phenomenon; Beckmann, 2000).

2. Theoretical Framework

2.1. The F > C Phenomenon

One approach to modelling response time is the F > C (False > Correct) phenomenon (Beckmann, 2000), also called the I > C phenomenon, which posits that respondents take longer to report incorrect answers than they take to provide correct ones. Formally, the F > C phenomenon is given by (Beckmann, 2000, as cited in Tancoš et al., 2023, p. 3):

$$t_{ij} = \mu + \gamma FC_{ij} + \epsilon_{ij}, \tag{1}$$

where t_{ij} is the response time of respondent j on item i; μ is an intercept representing the average response time across all items and respondents; FC_{ij} is a binary indicator representing whether respondent j answered item i correctly; γ is an unstandardized regression coefficient representing the mean difference in response time between false and correct answers; and ϵ_{ij} is a normally distributed residual.

2.2. The Distance-Difficulty Hypothesis

Another approach to modeling response time is the distance-difficulty hypothesis, based in item response theory, proposed originally by Thissen (1983). The distance-difficulty hypothesis states that response time decreases with increasing distance between an item's difficulty (b_i) and a respondent's ability level (θ_j), given by $\delta_{ij} = |\theta_j - b_i|$, called person-item distance. In other words, a respondent should take longer to answer a question that is close to their ability level. Conversely, a respondent should take less time to answer a question that is very easy or very difficult, relative to the respondent's ability level. Thissen's model is given by:

$$\ln(t_{ij}) = \mu + \tau_j + \beta_i - \gamma\delta_{ij} + \epsilon_{ij}, \tag{2}$$

where $\ln(t_{ij})$ is a logarithmic transformation of the response time for respondent j on item i (this transformation is meant to achieve normally distributed

errors ϵ_{ij}); μ is the intercept, representing the average response time across all respondents and items; τ_j represents respondent j's average response time across all items; β_i represents the response time of an average-ability respondent for item i; and γ is a coefficient representing the magnitude of the difference of the relationship between response time and the ability-difficulty distance, which is expected to be negative.

Ferrando and Lorenzo-Seva (2007) proposed an alternative person-item distance measure according to the two-parameter logistic (2PL) model. This model is given by (Ferrando & Lorenzo-Seva, 2007, p. 528):

$$\delta_{ij} = \sqrt{a_i^2(\theta_j - b_i)^2}, \tag{3}$$

where a_i is the discrimination for item i. Well-designed items have positive discrimination, in which case, one can simplify Equation 3 to $\delta_{ij} = a_i|\theta_j - b_i|$. We call this discrimination-adjusted person-item distance the "ability-difficulty distance."

2.3. The Tancoš Models

Combining these approaches, Tancoš et al. (2023) examined the time children took to complete fluid-reasoning tasks in a game-based application, using item difficulty, respondent ability level, and answer correctness as predictors in multilevel regression models with response time as the outcome variable.

Tancoš et al. (2023) proposed two series of multilevel models relating fluid intelligence in children to response time. The first series of models focused on the F > C phenomenon, controlling for item difficulty (b_i) and respondent ability (θ_j) with an interaction between answer correctness (FC_{ij}) and respondent ability. This series of model culminates with the following model (Tancoš et al., 2023, p. 7):

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_1 FC_{ij} + \gamma_2 b_i + \gamma_3 \theta_j + \gamma_{13} FC_{ij}\theta_j + \epsilon_{ij}, \tag{4}$$

where $\ln(t_{ij})$ is a logarithmic transformation of the response time for person j to item i; μ is the fixed intercept, representing the transformed response time of an average-ability respondent to an item of average difficulty; ν_j is the random intercept for person j, representing the general speediness of respondent j (the average transformed response time of respondent j across all items); β_i is the random intercept for item i, representing the expected transformed response time required by a respondent of average ability to item i; γ_1, γ_2, γ_3, and γ_{13} are fixed effects of the corresponding predictors; and ϵ_{ij} is the normally distributed residual.

The second series of models Tancoš et al. (2023) proposed assessed the distance-difficulty hypothesis, and investigates the incremental validity of the distance-difficulty hypothesis over the F > C phenomenon. This model is given by (Tancoš et al., 2023, p. 8):

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_4 \delta_{ij} + \gamma_1 FC_{ij} + \gamma_{14} FC_{ij}\delta_{ij} + \epsilon_{ij}, \tag{5}$$

where $\ln(t_{ij})$, μ, ν_j, β_i, and FC_{ij} are as defined above; $\delta_{ij} = |\theta_j - b_i|$ is the absolute distance between respondent j's ability level and item i's difficulty; γ_1, γ_2, γ_3, and γ_{13} are the fixed effects of the corresponding predictors; and ϵ_{ij} is the normally distributed residual.

In their study, Tancoš et al. (2023) found that the F > C phenomenon remained significant after controlling for item difficulty and person ability, with a significant interaction between ability level and item correctness. Moreover, they found ability level to be a significant predictor of transformed response time (though, only in items with moderate or high difficulty), while item difficulty was not. In sum, their results indicate that on items with moderate-to-high difficulty, children with higher ability levels took longer to report incorrect answers than children with lower ability levels. However, Tancoš et al. (2023) failed to find a relationship between response time and ability level in correctly answered items, challenging the "faster equals smarter" stereotype.

Furthermore, Tancoš et al. (2023) found incremental validity of the distance-difficulty hypothesis above and beyond the F > C phenomenon, providing evidence to support that answer correctness moderates the effect of ability-difficulty distance, (δ_{ij}), on response time. Taken together, these results suggest that items near a respondent's ability level take the longest amount of time to answer, with even more time required to answer items incorrectly. Moreover, as δ_{ij} increased the difference in response time narrowed, eventually changing direction. That is, when an item was very easy or very difficult relative to the respondent's ability level, respondents tended to take longer to report correct answers than false ones.

2.4. The Proposed Models

In this study, we will extend the Tancoš et al. (2023) methodology to mathematics achievement data from the 2019 TIMSS (Trends in International Mathematics and Science Study) for fourth- and eighth-grade students. Using these data, we will compare results from two series of multilevel linear models to assess the F > C phenomenon, the distance-difficulty hypothesis, and any interaction between the two.

The first series of models will assess the F > C phenomenon, controlling for ability level and item difficulty. With these models, we will determine whether the relationship between ability level and response time is moderated by answer correctness. Then, we will assess the distance-difficulty hypothesis using the second series of models, controlling for the discrimination-adjusted ability-difficulty distance. If we find the F > C phenomenon and distance-difficulty hypothesis both hold, we will investigate whether there is a significant interaction between the (discrimination-adjusted) ability-difficulty distance, which we call "the Tancoš model." A significant interaction would suggest a difference in the relationship between the time needed to answer an item and the ability-difficulty distance, according to whether the respondent answered the item correctly. See Appendix 1 for the list of models used.

3. Method

3.1. Participants and Measures

We examined data from the 2019 TIMSS mathematics achievement assessment. The TIMSS is a set of international assessments of fourth- and eighth-grade students' mathematics and science achievement and attitudes, along with survey responses from teachers and principals to gather information related to the background contexts for learning. The TIMSS is conducted every four years, with 64 countries participating in the 2019 assessment cycle (Mullis et al., 2020). In the United States, data were collected from 9,944 eighth-grade students in 273 schools. Beginning in 2019, the United States opted into the eTIMSS, a new computer-based version of the assessment, allowing response time data to be collected alongside participating student responses. Students answered items from one of fourteen booklets, composed of item block combinations. Response times were computed as time in seconds spent on a page with a single item, or a set of connected items (such as multi-part questions); to maximize the validity of our response time records, we dropped items that were presented as a set on a page and kept only those items that were presented in isolation on a page. We call these "isolated items." After dropping items presented together on a page, we analyzed responses from each booklet to determine which booklet(s) had the most isolated items and the most responses per booklet. A list of each booklet's number of isolated items and number of respondents for eighth grade students are shown in Appendix 2. Of the 14 booklets, we chose Booklet 14. Our final sample contained 552 complete responses to 13 isolated items.

3.2. IRT Analysis

All analyses were conducted using R Statistical Software (v4.4.1; R Core Team, 2024). The TIMSS mathematics assessments were validated by the creators using a two-parameter logistic (2PL) model, so we performed a 2PL IRT analysis using the *mirt* R package (v1.41; Chalmers, 2012) to obtain each item's difficulty (b_i), discrimination (a_i), and each respondent's estimated ability parameter (θ_j). See Table 1 for descriptive statistics and 2PL parameters. Empirical reliability was sufficiently high for analysis ($r_{xx} = .85$).

3.3. Main Analysis

Our main analysis involved two series of nested linear multilevel regression models. The first series of models tested the F > C phenomenon, controlling for item difficulty and respondent ability; the second series of models assessed the distance-difficulty hypothesis and its interaction with the F > C phenomenon. We used the *lme4* package (v1.1-35.5; Bates et al., 2015) to estimate each model, and after estimation, we computed the variance inflation factor (VIF) to check each model for multicollinearity using the *car* package (v3.1-3; Fox & Weisberg, 2019). Finally, we used the *lmerTest* package (v3.1-3; Kuznetsova et al., 2017) to compare models.

We began our analysis by estimating a null model (Model 0) to serve as a baseline for both series of models. Consistent with Tancoš et al. (2023), Model 0 included only fixed and random intercept terms for respondents and items; all models used a logarithmically transformed response time for the dependent variable to linearize the relationship between the predictor variables and response time.

For the first series of models, we began by defining Model A1, which includes only a binary predictor for item correctness (FC_{ij}); Model A2 includes two additional predictors, item difficulty b_i and respondent ability θ_j for control variables; finally, Model A3 includes an interaction term between item correctness and respondent ability to investigate whether item correctness moderates the relationship between ability level and the F > C phenomenon. Model A3 is given by Equation 4.

The second series of models begins with Model B1, which includes only the discrimination-adjusted ability-difficulty distance, $\delta_{ij} = a_i|\theta_j - b_i|$ ("ability-difficulty distance") to test the distance-difficulty hypothesis. Model B2 includes FC_{ij}, representing item correctness, to assess the incremental validity of the F > C phenomenon beyond the distance-difficulty hypothesis. Finally, Model B3, given by Equation 5, includes an interaction term between the ability-difficulty distance δ_{ij} and item correctness FC_{ij} to investigate whether the distance-difficulty hypothesis is moderated by item correctness.

3.3.1. Null Model

The null model, Model 0, served as our baseline model for both series of models. The fixed intercept was significant, (μ = 3.92, 95% CI [3.67, 4.18]). Transforming this parameter, we see the average response time for a student of average ability on an item of average difficulty was 54.16 seconds. Table 2 displays complete results for Model 0.

3.3.2. Models Assessing F > C Phenomenon

We began with Model A1. We found it took students significantly more time to answer an item correctly than it did to answer incorrectly (γ_1 = 0.21, 95% CI [0.17, 0.25]). Transforming this parameter, we see students took on average, 1.23 seconds longer to provide correct responses than incorrect ones. Table 3 displays complete results for Model A1.

In Model A2, the F > C phenomenon was suppressed but remained significant, (γ_1 = 0.18, 95% CI [0.14, 0.22]). Students with higher ability levels took significantly longer to answer items (γ_3 = 0.13, 95% CI [0.09, 0.17]), and students took significantly longer to answer more difficult questions (γ_2 = 0.36, 95% CI [0.08, 0.62]). Table 4 displays complete results for Model A2.

In Model A3, the F > C phenomenon remained significant (γ_1 = 0.19, 95% CI [0.15, 0.23]), as did item difficulty (γ_2 = 0.35, 95% CI [0.07, 0.62]) and student ability level (γ_3 = 0.22, 95% CI [0.18, 0.27]). Moreover, we found a significant

interaction between item correctness and ability level (γ_{13} =-0.21, 95% CI [-0.26, -0.16]). Taken together, we can interpret this to mean that students with lower ability levels took longer to answer items correctly than incorrectly, and students with higher ability levels took longer to provide incorrect answers than correct ones. Moreover, we found a positive relationship between ability level and response time for incorrect answers, but no substantial relationship between ability level and response time for correct answers. This pattern is shown in Figure 1, and Table 5 displays complete results for Model A3.

Overall, every model fit better than the null model, and each model fit better than the preceding model in the series Table 6. Given the information criteria and the goodness-of-fit measures, we chose to retain Model A3 as our final model.

3.3.3. Models Assessing the Distance-Difficulty Hypothesis

Model B1 assessed the distance-difficulty hypothesis as a predictor for the logarithmic transform of response time, which was significant (γ_4 = -0.09, 95% CI [-0.10, -0.08]). From this, we see that for each one-unit increase in δ_{ij}, response time is decreased by 8.6%. For example, when $\delta_{ij} = 0$, the expected response time is 60.46 seconds; when $\delta_{ij} = 1$, the expected response time is 55.26 seconds; when $\delta_{ij} = 2$, the expected response time is 50.50 seconds, and so on. Table 7 displays the complete results for Model B1.

For Model B2, we added item correctness as a fixed effect, which was significant (γ_1 = 0.22, 95% CI [0.18, 0.26]). The ability-difficulty distance remained significant but was suppressed (γ_4 = -0.09, 95% CI [-0.11, -0.08]). From this, we see that students took longer to provide correct answers than incorrect ones, and regardless of item correctness, as the ability-difficulty distance increased, response time decreased. See Table 8 for complete results for Model B2.

Finally, in Model B3, we added an interaction term between item correctness and the ability-difficulty distance, which was significant (γ_{14} = -0.08, 95% CI [-0.11, -0.05]). Item correctness and ability-difficulty distance remained significant (γ_1 = -0.07, 95% CI [-0.08,-0.05] and γ_4 = 0.29, 95% CI [0.24,0.34], respectively). Taken together, we interpret these results as follows: students take longer to provide correct answers to items near their ability level. As the ability-difficulty distance increases, the relationship switches around $\delta_{ij} = 3.8$, such that students beyond this point take longer to provide incorrect answers. So, students take longer to provide incorrect answers to items that are very difficult or very easy relative to their ability level. Moreover, response time decreases as the ability-difficulty distance increases, regardless of item correctness; that is, students respond quicker to items that are very difficult or very easy relative to their ability level. This pattern is shown in Figure 2, and Table 9 displays complete results for Model B3.

Overall, each model fit the data better than the null model, and every model in the series fit the data better than the preceding model (Table 10). Per this

evidence, along with information criteria for each model, we chose to retain Model B3 as our final model.

4. Conclusion

The results of our study show that eighth-grade students' responses to a mathematics achievement assessment support the F > C phenomenon after controlling for ability level, item difficulty, and item correctness. Moreover, students' responses provide evidence to support the distance-difficulty hypothesis after controlling for ability-difficulty distance and item correctness, with a significant interaction between item correctness and the ability-difficulty distance.

The interaction between the F > C phenomenon and item correctness may have several explanations. Students with higher ability levels may have higher levels of grit and may be more likely to try multiple approaches to a problem if their initial strategy/approach does not yield a solution (Credé et al., 2017; Duckworth et al., 2007; Steinmayr et al., 2018). In contrast, students with lower ability levels may not see a clear path to a solution when beginning a problem. As we saw, lower-ability students took longer to answer questions correctly, which could be indicative of students pausing to develop a plan of attack before answering the question. Researchers (e.g., Chan et al., 2022) have found that longer pauses before beginning a problem can lead to more efficient problem-solving, and, ultimately, faster response times.

The observed patterns in the relationship between the ability-difficulty distance and response time align with the findings from the Tancoš et al. (2023) study; students took longest to answer questions that matched the student's ability level, and response time decreased as the ability-difficulty distance increased. One possible explanation for the extended time taken to answer questions that match a student's ability level is that the question may be difficult enough to be challenging, but not so challenging that they have little chance of answering the question correctly. When items are very easy relative to a student's ability level, this pattern seems intuitive: easier items take less time. However, the idea that students take less time to answer very difficult questions is less intuitive but may be a result of an initial misjudgment of the item's difficulty, which would explain our finding that students took longer to answer very difficult questions incorrectly.

Our study is limited by the low-stakes nature of the TIMSS mathematics assessments. Students may not have been highly motivated to perform well in this achievement assessment, particularly toward the end of the test, so our estimates of student ability and item difficulty may be biased by a lack of sufficient effort. Moreover, the magnitude of errors may be compounded by the initial IRT estimate and subsequent multilevel regression model. Finally, there may be unanalyzed differences in the observed relationships across students of different demographic groups (such as gender, race/ethnicity, or socio-economic status). Future research could examine the magnitude of the fixed effects we analyzed

in this study by demographic group. Future research may also explore whether the results from this study hold with attitudinal scales or achievement scales in other academic disciplines. This study extended the response time models found by Tancoš et al. (2023) in a fluid-reasoning task to a mathematics achievement assessment involving crystallized intelligence. Researchers and educators alike utilize response time as a catch-all measurement of item suitability, respondent effort, and ability-level, but this study's results imply response time is not appropriate as a stand-alone diagnostic tool, as response time is influenced by a variety of factors.

References

Tables

Table 1. Descriptive Statistics

Item	Sample	Item Difficulty	Item Discrimination	Response (Proportion Correct)		Response Time (S
				M	SD	M	SD
1	552	-0.13	1.01	0.53	0.50	40.90	47.73
2	552	0.00	1.75	0.50	0.50	72.68	58.69
3	552	0.34	1.67	0.41	0.49	56.62	59.44
4	552	-0.84	2.09	0.74	0.44	34.77	29.84
5	552	-0.38	2.41	0.62	0.49	70.38	81.01
6	552	-0.16	2.57	0.55	0.50	102.51	61.66
7	552	0.08	1.93	0.48	0.50	55.06	47.03
8	552	1.51	2.21	0.12	0.32	78.65	62.14
9	552	1.57	2.85	0.09	0.29	115.56	80.58
10	552	1.05	2.54	0.20	0.40	58.11	39.61
11	552	-0.61	1.31	0.65	0.48	67.59	43.22
12	552	-0.77	1.23	0.68	0.47	40.25	31.85
13	552	0.65	1.74	0.32	0.47	143.72	101.17

Table 2. Parameters for Model 0

				95% CI	
	Coef.	Est.		LL	UL
Fixed Effects					
Intercept		3.92	***	3.67	4.18
Random Effects					
Person intercept variance	var($_j$)	0.18	***		
Item intercept variance	var($_i$)	0.20	***		
Residual Variance	var($_{ij}$)	0.41			

Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 3. Parameters for Model A1

	Coef.	Est.		95% CI LL	95% CI UL
Fixed Effects					
Intercept		3.83	***	3.56	4.10
Correct Answer (FC)	$_1$	0.21	***	0.17	0.25
Random Effects					
Person intercept variance	var($_j$)	0.17	***		
Item intercept variance	var($_i$)	0.22	***		
Residual Variance	var($_{ij}$)	0.41			

Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 4. Parameters for Model A2

	Coef.	Est.		95% CI LL	95% CI UL
Fixed Effects					
Intercept		3.78	***	3.56	4.00
Correct Answer (FC)	$_1$	0.18	***	0.14	0.22
Item Difficulty	$_2$	0.35	***	0.08	0.62
Person Ability	$_3$	0.13	***	0.09	0.17
Random Effects					
Person intercept variance	var($_j$)	0.16	***		
Item intercept variance	var($_i$)	0.15	***		
Residual Variance	var($_{ij}$)	0.41			

Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 5. Parameters for Model A3

	Coef.	Est.		95% CI LL	95% CI UL
Fixed Effects					
Intercept		3.83	***	3.60	4.05
Correct Answer (FC)	$_1$	0.19	***	0.15	0.23
Item Difficulty	$_2$	0.35	***	0.07	0.62

Table 5. Parameters for Model A3

	Coef.	Est.		95% CI LL	UL
Person Ability	$_3$	0.22	***	0.18	0.27
FC x Ability	$_{13}$	-0.21	***	-0.26	-0.16
Random Effects					
Person intercept variance	var($_j$)	0.15	***		
Item intercept variance	var($_i$)	0.16	***		
Residual Variance	var($_{ij}$)	0.40			

Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 6. Model Fit For Series 1 Models

	Model 0	Model A1	Model A2	Model A3
Conditional R^2	.48	.50	.50	.50
Marginal R^2	.00	.01	.11	.12
Log-likelihood	-7,558.04	-7,503.89	-7,483.22	-7,446.96
AIC	15,124.07	15,017.77	14,980.44	14,909.92
BIC	15,151.58	15,052.16	15,028.59	14,964.95
$\Delta\ ^2$(df)		108.30(1)***	31.33(2)***	72.53(1)***

Table 7. Parameters for Model B1

	Coef.	Est.		95% CI LL	UL
Fixed Effects					
Intercept		4.10	***	3.82	4.39
Ability-Diffiulty Distance	$_4$	-0.09	***	-0.10	-0.08
Random Effects					
Person intercept variance	var($_j$)	0.17	***		
Item intercept variance	var($_i$)	0.24	***		
Residual Variance	var($_{ij}$)	0.40			

Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 8. Parameters for Model B2

	Coef.	Est.		95% CI LL	95% CI UL
Fixed Effects					
Intercept		4.01	***	3.71	4.31
Correct Answer (FC)	$_1$	0.22	***	0.18	0.26
Ability-Diffiulty Distance	$_4$	-0.09	***	-0.11	-0.08
Random Effects					
Person intercept variance	var($_j$)	0.16	***		
Item intercept variance	var($_i$)	0.28	***		
Residual Variance	var($_{ij}$)	0.40			

Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 9. Parameters for Model B3

	Coef.	Est.		95% CI LL	95% CI UL
Fixed Effects					
Intercept		3.98	***	3.69	4.27
Correct Answer (FC)	$_1$	0.29	***	0.24	0.34
Ability-Diffiulty Distance	$_4$	-0.07	***	-0.08	-0.05
FC x Distance	$_{14}$	-0.08/td>	***	-0.11	-0.05
Random Effects					
Person intercept variance	var($_j$)	0.17	***		
Item intercept variance	var($_i$)	0.26	***		
Residual Variance	var($_{ij}$)	0.39			

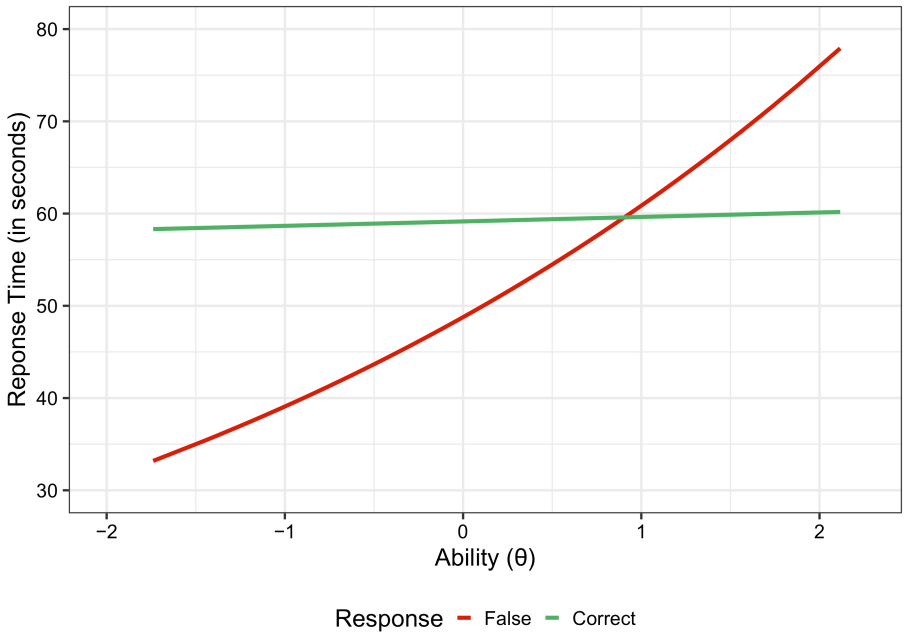
Note. Coef – coefficient; Est – estimate; CI – confidence interval; LL – lower limit; UL – upper limit; Var – variance. * p < .05, ** p < .01, *** p < .001.

Table 10. Model Fit For Series 2 Models

	Model 0	Model B1	Model B2	Model B3
Conditional R^2	.48	.53	.55	.54
Marginal R^2	.00	.03	.05	.04
Log-likelihood	-7,558.04	-7,450.04	-7,385.87	-7,374.85
AIC	15,124.07	14,910.08	14,783.73	14,763.69
BIC	15,151.58	14,944.47	14,825.00	14,811.84
$\Delta\ ^2$(df)		215.992(1)***	128.346(1)***	22.041(1)***

Figures

Figure 1. From Model A3, students' predicted response time according to ability level and whether the student answered the question correctly.



Appendix

Appendix 1. List of Models Used

Null Model (Model 0):

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \epsilon_{ij}$$

Model A1:

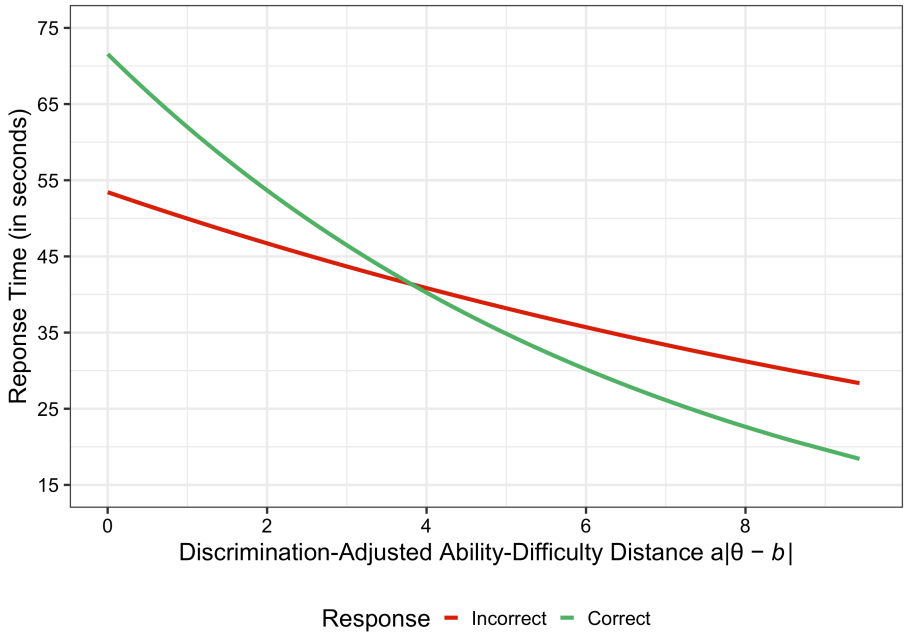
$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_1 FC_{ij} + \epsilon_{ij}$$

Model A2:

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_1 FC_{ij} + \gamma_2 b_i + \gamma_3 \theta_j + \epsilon_{ij}$$

Model A3:

Figure 2. From Model B3, students' predicted response time according to the discrimination-adjusted ability-difficulty distance and whether the student answered the question correctly.



$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_1 FC_{ij} + \gamma_2 b_i + \gamma_3 \theta_j + \gamma_{13} FC_{ij}\theta_j + \epsilon_{ij}$$

Model B1:

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_4 a_i|\theta_j - b_i| + \epsilon_{ij} \tag{6}$$
$$= \mu + \nu_j + \beta_i + \gamma_4 \delta_{ij} + \epsilon_{ij} \tag{7}$$

Model B2:

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_4 a_i|\theta_j - b_i| + \gamma_1 FC_{ij} + \epsilon_{ij} \tag{8}$$
$$= \mu + \nu_j + \beta_i + \gamma_4 \delta_{ij} + \gamma_1 FC_{ij} + \epsilon_{ij} \tag{9}$$

Model B3:

$$\ln(t_{ij}) = \mu + \nu_j + \beta_i + \gamma_4 a_i|\theta_j - b_i| + \gamma_1 FC_{ij} + \gamma_{14} FC_{ij} a_i|\theta_j - b_i| + \epsilon_{ij} \tag{10}$$
$$= \mu + \nu_j + \beta_i + \gamma_4 \delta_{ij} + \gamma_1 FC_{ij} + \gamma_{14} FC_{ij}\delta_{ij} + \epsilon_{ij} \tag{11}$$

Appendix 2. Booklet Overview for 2019 TIMSS Mathematics Achievement Assessment

Booklet Number	Isolated Items	Sample Size
1	10	624
2	12	621
3	12	626
4	5	606
5	12	619
6	11	633
7	12	623
8	8	617
9	11	606
10	13	619
11	13	620
12	11	625
13	10	637
14	13	622

Note. Isolated items refer to items that are shown on a webpage as a standalone item. Adapted from Trends in International Mathematics and Science