Response Time as a Predictor of Test Performance: Assessing the Value of Examinees' Response Time Profiles

by

Bin Tan

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation, and Data Science

Department of Educational Psychology
University of Alberta

Abstract

Response time data has gained extensive attention in recent years, thanks to the increasing use of computer-based assessments. Previous studies examined the relationship between response time and test performance, yielding inconsistent findings such as positive, negative, and no relationship. To comprehensively examine the complex relationship between examinees' response times and test performance, this study employs profile analysis to analyze assessment data from the Problem Solving and Inquiry tasks (PSI tasks) in the Trends in International Mathematics and Science Study (TIMSS) 2019 for grade-four mathematics. In the assessment, there were 29 items distributed across 17 screens, with response time recorded for each screen. The data used included responses from 27,682 fourth-grade examinees from 36 countries, including six benchmarking participants. The results of this study show that examinees' standardized response time varied throughout the test. Furthermore, the study shows the predictive power of screen response time for test performance, with varying strengths and directions of the regression coefficients for different screens. Notably, considering separate response times for individual screens and accounting for within-examinee variability in response time provide more accurate predictions of test scores than relying on total response time or the average response time per screen. Additionally, the analysis uncovers that the relationship between total or separate response times and test performance is also distinct across different achievement groups. Low achievers exhibit a stronger positive correlation between response time and performance, while advanced achievers show a negative correlation. Moreover, the study reveals the influence of item position and item difficulty on response time patterns and their relationship with test performance. These findings contribute to advancing our understanding of

the relationship between response time and test performance, with implications for the design and administration of future educational assessments.

Keywords: Response time, test performance, profile analysis, pattern analysis

Preface

This thesis is an original work by Bin Tan. No part of this thesis has been previously published.

Acknowledgement

I would like to express my sincere gratitude to everyone who has supported my educational and research endeavors.

First and foremost, I am deeply grateful to my supervisor, Dr. Okan Bulut, for his invaluable guidance throughout my academic journey and the completion of this thesis; we share closely aligned research interests and working styles, which have made my research journey exceptionally enjoyable. His emphasis on collaboration within our research group has allowed me to explore a diverse range of research topics. His insightful ideas and expertise in data analysis have been instrumental in my research.

I am also thankful to Dr. Alfred Sakyi for providing me with the opportunity to intern at Alberta Education since my first year as a Master's student. Together, we explored and conducted several research projects with a wide range of topics. Through this process, I honed my data analytical skills and made my research topics more relevant to practical educational challenges in Alberta and Canada. Dr. Sakyi's patience, openness to new ideas, and support throughout my work have been extremely valuable and influential to my future career.

Additionally, I extend my gratitude to my fellow research colleagues at the Centre for Research in Applied Measurement and Evaluation (CRAME), with whom I have had the opportunity to collaborate on various fascinating research projects. Our shared experiences have enriched my academic journey and also fostered enjoyable moments beyond our research pursuits.

Finally, I wish to acknowledge my parents, Yixin Tan and Lijun Wan, for their understanding and support in my educational journey; without them, it would have been impossible for me to pursue my studies abroad and reach this point in my academic career.

Table of Content

List of Tables

List of Figures

List of Abbreviations

ICT	Information and Communications Technology
IQR	Interquartile range
MANOVA	Multivariate Analysis of Variance
PSI	Problem Solving and Inquiry
RQ	Research question
TIMSS	Trends in International Mathematics and Science Study

1. Introduction

Response time usually refers to the time duration from the moment a task is presented to a person, to when they complete or exit it. In the field of educational measurement, the study of response time dates back to the time when Blommers and Lindquist (1944) counted high-school examinees' response time on each item on a reading comprehension test and calculated the correlation with item correctness. Collecting response time data in that era was apparently difficult; thus, traditional educational assessments focused primarily on the accuracy of responses to reveal what students know and can do. However, the recent increasing use of computer-based assessments allows for automatically collecting process data that occurred during the response processes, including the time for considering and answering each item. These data have attracted much attention because they offer invaluable insights into the complexities of test-taking behaviors and cognitive processes. It allows researchers and educators alike to consider not just what students know, but also how they engage with the testing process and its relationship with their test performance.

The relationship between time spent on a task and task performance, such as accuracy and correctness, has been extensively studied. Research has shown that the association between response time and test performance can be positive (Goldhammer et al., 2014), zero (Ratcliff et al., 2015), or negative (Sherbino et al., 2012; Castro et al., 2019). Many of the previous studies have solely focused on examining the correlation between response time and item correctness for individual items. For example, Castro et al. (2019) discovered that students who spent less time on time-constrained items (20 seconds) in an educational game had a higher probability of correctly answering the items better than others. However, since response time may vary across different items, the relationship between response time and item correctness for individual items

may vary as well. For example, in tests with time constraints, examinees' response time can be influenced by factors such as the need to speed up (Schnipke, 1995). Individual differences in ability levels and time management skills also can result in distinct patterns of test solution behavior and response time (van der Linden, 2009). Moreover, examinees' engagement and motivation may fluctuate throughout the test (Wise & Kingsbury, 2016). As such, considering the relationship between response time and individual item correctness may not effectively predict overall test performance.

Some researchers use total response time or the average response time across items to predict test performance (e.g., Sherbino et al., 2012). However, this practice, in fact, is based on the assumption that an examinee's response time, relative to other examinees, does not vary across items during a test event. In other words, if the response time for each item is standardized, the examinee's standardized response time is consistent across all items, thus there is no within-person variability. However, this assumption is untested and probably unrealistic. Therefore, it is important to test whether there is within-person variability and whether such variability provides valuable information in predicting test performance. Accordingly, a focus of this study is to quantify how much the variability in an examinee's response time across different items (a concept called "response time profile") contributes to predicting their test scores compared to using only the average response time.

The present study takes a profile analysis approach, examining the response time profiles in a holistic way and addressing the identified research gaps. This study utilizes data from an international computer-based assessment task, which was part of the Trends in International Mathematics and Science Study (TIMSS) 2019. The Problem Solving and Inquiry tasks (PSI

tasks) collected examinees' process data, including response time on each screen during test completion, enabling the examination of response time profiles.

This thesis begins with the introduction section, including a survey of the factors associated with changes in response time within a test event. The introduction then presents theories commonly used to explain the relationship between response time and test performance. This is followed by a review of the moderators of the relationship between response time and test performance, including item factors, person factors, and their interactions. In addition, a brief introduction to profile analysis is also included. The section concludes with an overview of the study and research questions (RQs).

1.1. Response Time

This section aims to provide an understanding of the changes in response time observed during a test event in literature, considering various influencing factors. These factors can be classified into item factors, which include characteristics like difficulty, and personal factors, such as engagement and ability. While item factors primarily account for within-person variations in response time, personal factors contribute to differences in response time patterns among individuals. Next, this section introduces popular theories applied to explain the changes in response time patterns during the test-taking process. The dual processing theory (Schneider & Shiffrin, 1977), Wise's test theory (2017), and the demands-capacity model of test-taking effort (Wise & Smith, 2011) were included because of their popularity in the research field of response time. These theories are also used to explain the other study findings presented in the introduction section.

1.1.1. Factors Influencing Response Time Changes

Within a test event, the response time in assessments varies across items, partially attributing to the influence of various item characteristics. For example, more difficult items generally require more time to answer (Kobrin, 2000). Supporting this idea, Yang et al. (2002) found a significant positive relationship between item difficulty and response time on the overall perceptual ability test items in a sample of 389 examinees. Moreover, certain item types require more time than others. For instance, multiple-choice questions may generally require less time compared to constructed-response questions. This is because multiple-choice questions are typically used to target lower cognitive abilities such as knowledge and memory, while constructed-response questions focus on higher-level cognitive abilities like designing and evaluating (Hancock, 1994). The position of items within a test can also impact response times in various ways. On one hand, items in the later part of the test may be answered faster, a phenomenon defined as response acceleration (Vida et al., 2021). On the other hand, early items may be answered more quickly due to a fresh mindset, while later items may take longer due to fatigue (Bolsinova et al., 2017).

A well-studied personal factor influencing response time patterns is motivation or test engagement. For the validity of achievement measurement, examinees' are generally expected to allocate sufficient time to each item in order to carefully consider their responses and demonstrate effort for generating correct responses. Thus, the investigation into the amount of time spent on each question can provide valuable insights into examinees' cognitive processes and test-taking behaviors. Schnipke (1995) conducted a study and discovered that as time expired during timed tests, some test takers would hastily fill in answers to the remaining items, hoping to guess correctly. She proposed two categories of responding behaviors: rapid guessing

and solution behavior. Rapid guessing refers to situations where test takers respond very quickly, indicating that they are unlikely to have fully considered the item before answering. In the current literature, rapid guessing behavior is either seen as an outcome of low motivation or engagement with the test (Wise & Kong, 2005), or it may be a result of time pressure when test takers are rushing to complete the test.

An examinee's ability also influences their response time patterns. In an Information and Communications Technology (ICT) literacy assessment, Deribo (2023) observed that examinees with higher or lower levels of ICT literacy tend to respond to items more quickly. This is because individuals with higher ICT literacy possess the necessary knowledge to solve the items and can retrieve that knowledge rapidly, while those with lower ICT literacy tend to quickly abandon a task when they realize they lack the necessary knowledge to solve it. Conversely, examinees with moderate ICT literacy levels demonstrated a slower pace and allocated more time to the tasks. This behavior can be attributed to their perception of a potential opportunity to solve the task through careful consideration and the utilization of effortful cognitive processes. Consistent with this finding, Rayner (1998) found that examinees with higher reading skills tend to exhibit more efficient eye movement patterns, including fewer and shorter fixations, longer saccades, and fewer regressions.

In a study conducted by Thomas (2006), examinees were categorized into three groups based on their proficiency levels and pass/fail status: those who passed, those with indeterminate results, and those who failed. Interestingly, it was found that all groups displayed faster pacing or reduced response times, even when there was no time constraint. The group who passed had the quickest initial pacing, and their response time exhibited the most gradual decrease among the three groups. The indeterminate group displayed slower initial pacing than those who passed but

had a consistent reduction in response time throughout the test, even though there was no collective evidence indicating they were pressed for time. Conversely, the failed group started with the slowest pacing but gradually increased their pacing at a rate comparable to the indeterminate group.

1.1.2. Theories for Response Time and Cognitive Processes in Tests

Dual processing theory (Schneider & Shiffrin, 1977) suggests the existence of two cognitive processes involved in task completion: controlled processes and automatic processes. Controlled processes require cognitive resources and effortful mental operations, which increases cognitive load over time due to limited cognitive resources. Usually, when given an item, examinees employing controlled processes take some time to think before responding. On the other hand, automatic processes are executed effortlessly and fast and are unaffected by cognitive load, which usually reflects the familiarity or excellence of performing the task. In problem-solving tasks, individuals who cannot solve tasks automatically rely on controlled processes instead, which heavily rely on cognitive resources (Sweller et al., 1998). Due to the limitations of cognitive resources, individuals relying on controlled processes take more time to arrive at correct solutions and may struggle with challenging items due to increased cognitive load. Conversely, individuals with a higher proportion of automatized processes can solve items quickly and accurately, even difficult ones, as their working memory can handle higher cognitive loads with the aid of automatized processes.

Wise (2017) proposed a test theory for rapid guessing behaviors, hypothesizing three scenarios in which such rapid guessing may occur. First, motivated examinees may allocate excessive time to earlier items on the test, engaging in solution behavior. However, if they approach the end of the allotted time with a number of items remaining, they may quickly

complete the remaining items without careful consideration. This pattern of response time is commonly referred to as speeding. Second, when examinees perceive that the assessment results hold little consequence for them, they may lack motivation, leading them to either give up on responding to items or engage in rapid guessing behaviors. In the third scenario, rapid guessing may arise when examinees quickly realize that they lack the necessary knowledge, skills, or abilities to solve certain items. These scenarios suggest that response time and rapid guessing are influenced by many factors, including examinees' ability and task difficulty, demonstrating an idiosyncratic nature of response time.

To explain the relationship between item response time and test performance, the demands-capacity model of test-taking effort (Wise & Smith, 2011) provides valuable insights. The model defines two crucial constructs: resource demands and effort capacity. Resource demands refer to the level of effort required to correctly answer an item, while effort capacity pertains to the amount of effort an examinee is willing to exert in answering the items. According to Wise and Smith, examinees continuously assess their effort capacity in relation to the cognitive resources needed to answer an item, resulting in different cognitive strategies employed during test-taking. When examinees perceive that their effort capacity surpasses the resource demands of an item, they engage in solution behavior. On the other hand, if examinees have low effort capacity for an item, they may choose to rapidly guess or omit the item. The model specifies that the resource demands fluctuate throughout the test as they vary across different items. In addition, the examinees' effort capacity also changes due to changes in their motivation and cognitive resources (e.g., fatigue). Given both the changes in effort capacity and resource demand, examinees often exhibit idiosyncratic test-taking behaviors and response time patterns throughout the test. For instance, they may opt for rapid guessing on certain items while

engaging in solution behavior for others. Supporting the theory, Wise and Kingsbury (2016) found that after the occurrence of the fourth rapid guess, the probability of subsequent responses being rapid guesses dropped to only 40%, a significant deviation from 100% that would be expected under a state model.

1.2. Moderators of the Relationship between Response Time and Performance

There are inconsistent findings for the relationship between time spent on a task and task performance, such as accuracy and correctness. The most commonly reported association is a negative one, indicating that faster responses tend to result in better performance. For instance, Sherbino et al. (2012) conducted a study with 95 medical license examinees who were tasked with making diagnostic decisions for 25 cases. The researchers found that faster response time was generally associated with a higher probability of correctness for each case. In addition, for each respondent, a faster overall response time could predict a higher overall diagnostic accuracy. However, Ratcliff et al. (2015) reported that there was no correlation between university students' performance on a number ability test and their response time, whereas Goldhammer et al. (2014) identified a positive relationship between examinees' response time and test performance in problem-solving tasks. Moreover, Chen et al. (2018) analyzed data from multiple knowledge and reasoning tests varied by content and test types. Their findings revealed a curvilinear relationship between response time and response correctness across all tests: initially, as response time decreases, response correctness rapidly increases; however, as response time continues to decrease, the marginal improvement in response correctness slows down and eventually starts to decline.

Further investigation revealed that the relationship between response time and test performance is conditioned on various factors. Both item characteristics (e.g., item difficulty and

type) and personal factors (e.g., motivation, cognitive resources, ability) contribute to the diverse response time patterns and their relationship with test performance. In addition, many personal factors change throughout a test event. At the same time, these factors moderate the relationship between response time and test performance, which helps to explain the complex relationship between response time and test performance. This section focuses on reviewing the moderators from three perspectives: item factors, person factors, and their interactions. For each perspective, I present relevant research findings to illustrate how the factors influence the relationship between response time and test performance.

1.2.1. Item Factors as Moderators

The relationship between response time and accuracy is influenced by the cognitive demands of the items involved. It is reasonable to assume that more difficult items generally require more cognitive resources, the item difficulty can also affect the relationship between response time and performance. Wise and Ma (2012) found that the relationship between response time and accuracy is positively correlated with the difficulty of the items. That is, for easier items, a stronger negative correlation between response time and response accuracy is observed, while for more difficult items, the correlation can be positive. In addition, Goldhammer et al. (2014) conducted research on problem-solving tasks and identified a positive relationship between examinees' response time and test performance when controlled processing is required. This implies that longer response time was associated with better performance on the problem-solving tasks examined in their study. On the other hand, they found that in reading tasks that required more automated processing, the relationship between response time and test performance was negative. In order to explore the moderator role of cognitive resource demand, Krämer et al. (2023) conducted a study using three distinct reasoning tests and one test assessing

natural sciences knowledge. They examined 10 comparable subsamples, comprising a total of 2640 examinees. The findings consistently demonstrated a pattern of faster responses being associated with higher accuracy, indicating a relatively effortless processing style. However, as the difficulty of the items increased, and individual ability decreased, the effect reversed. In such cases, longer processing times were linked to higher accuracy, suggesting that more deliberate cognitive processes were required for better performance. This generalizability research provides further support for the notion that the relationship between response time and performance is influenced by the cognitive demands of the tasks and the examinees' ability level.

As introduced in the previous section, Chen et al. (2018) found a curvilinear relationship between response time and response correctness across six different tests; that is, correctness increases as response time decreases, but the marginal effects gradually diminish and eventually correctness decreases as response time decreases. They also found that the turning point in the curvilinear relationship occurs earlier in the knowledge tests than in reasoning tests. The researchers explained that answers heavily rely on information retrieval for knowledge-based tests in which individuals either possess the necessary knowledge and respond quickly, or they lack the knowledge and employ alternative strategies that consume more time but result in lower success rates. This explanation suggests a negative correlation between response time and correctness in knowledge-based tests. On the other hand, in tests that rely more on mental operations and intentional reasoning, it may take longer before a correct answer can be provided. Consistent with this research finding, Goldhammer et al. (2014) discovered a positive relationship between response time and performance in problem-solving tasks, while a negative relationship was observed in reading tasks.

A study by San Martin et al. (2006) illustrated that in knowledge-based tests, examinees are expected to recognize the correct answer immediately upon information retrieval. Otherwise, the examinees have to guess the correct answer using their partial knowledge, which often results in responses that are less likely to be correct but still take longer than both rapid guessing and solution behaviors. In another case, when items contain strong distractors, an examinee may rapidly select the distractor but still get an incorrect response. Meanwhile, carefully considering the item and a slower response time will result in a higher probability of answering that item correctly. Their study also showed the influence of item characteristics on the relationship between response time and correctness.

1.2.2. Person Factors as Moderators

When examining the relationship between response time and performance, it becomes evident that rapid guessing is a special case. It contaminates the hypothesized negative correlation between response time and performance because it leads to lower response accuracy. This is because faster response time is a result of guesswork rather than automatic cognitive processes (Deribo et al., 2023). For instance, while automatic cognitive processes suggest a negative linear relationship between response time and performance, a significant and overlooked portion of faster response time can be attributed to rapid guessing. These rapid-guessing behaviors can impact the true relationship between response time and performance since the test outcomes of rapid guessing are expected to be worse than those of automatic cognitive processes and solution behaviors.

In fact, numerous studies have demonstrated the difference in test performance between those who use rapid guessing and those who do not. For instance, Michaelides et al. (2020) found that PISA 2015 examinees who performed better overall on test items tended to engage in

less rapid guessing compared to their lower-performing peers. Another study by Wise (2006) revealed that 25.5% of responses produced under rapid-guessing behavior were correct, slightly higher than the expected accuracy rate of 25.1% for random responses. In contrast, responses generated under solution behavior exhibited a much higher accuracy rate of 72%. Therefore, to accurately assess the relationship between response time and accuracy, it is crucial to control or rule out the nuisance influence of rapid guessing. On the other hand, it is also important to differentiate fast automatic cognitive processes from rapid guessing. Rapid guessing is usually associated with a lower probability of correctness, whereas automatic cognitive processes often yield a higher probability of correctness as they occur as a result of examinees' excellence. Therefore, even if examinees who excel and employ automatic cognitive processes and those who exhibit rapid guessing behaviors spend the same amount of time on each item, It is important to identify who, when, and to what degree, rapid guessing presents during test events.

In addition to rapid guessing, the relationship between response time and response accuracy can be influenced by a broader range of factors, one of which includes high-capable examinees' time management skills (van der Linden, 2009). High-capable examinees can adjust their pace to the available time, leading to a negative correlation between response time and accuracy when sufficient time is provided. However, they can strategically allocate time when time is limited, which results in a positive correlation between response time and accuracy. Essentially, their time management skills allow them to maximize performance based on the task's timing conditions. In line with this notion, Bolsinova et al. (2017) argue that examinees adjust their balance of speed and accuracy throughout a test. In timed tests, examinees can allocate their cognitive resources to working relatively fast with more mistakes or working more accurately but slowly.

Bolsinova and colleagues (2017) explore other potential causes for the varying correlations observed between test performance and response time. For example, cognitive resources may decrease throughout a test due to fatigue, which may influence the relationship between response time and accuracy. Moreover, as the test is ending, the examinee may engage in speeding response behaviors, potentially adversely affecting accuracy. The authors also found that changes in concentration and motivation during the test could impact response time and accuracy. Specifically, they observed that an increase in both concentration and motivation could result in shorter response times and improved accuracy, while a decrease in these factors might lead to the opposite effect.

1.2.3. The Interactions between Item Factors and Personal Factors

The interaction of item factors and personal factors leading to diverse response time patterns and the idiosyncratic relationship between response time and test performance. Becker et al. (2016) pointed out that for very easy tasks, automatic processes exceed controlled processes, and thus the cognitive load is often low. As item difficulty increases, more capable examinees employ more automatic processes compared to less capable examinees, resulting in faster response times. On the other hand, less capable examinees may struggle to solve tasks correctly when the cognitive load becomes too high.

Similarly, Naumann and Goldhammer (2017) found that examinees with lower reading proficiency, relying on controlled processes, can still arrive at correct solutions for relatively easy items despite being at a slower pace. However, this reliance on controlled processes puts a burden on working memory. As cognitive load increases during the test, cognitive resources like working memory reach their limit, leading to fatigue and difficulties in solving tasks with high cognitive loads. According to Naumann and Goldhammer, examinees with lower reading

proficiency who rely on controlled processes may still be able to arrive at correct solutions for relatively easy items, although at a slower pace. However, this reliance on controlled processes places a burden on working memory. As cognitive load increases throughout the test, cognitive resources such as working memory eventually reach their limit, resulting in fatigue and the individual being unable to solve tasks with high cognitive loads correctly. The exhaustion of cognitive resources may result in unmotivated test-taking behaviors such as rapid guessing (Wise, 2005), leading to fast response time but low test performance.

Goldhammer and colleagues (2014) found opposite relationships between response time and performance in problem-solving and reading tasks. They observed a positive correlation between response time and performance for problem-solving tasks, whereas a negative correlation was noted for reading tasks. They attributed this disparity to the different cognitive demands of these tasks. Generally, problem-solving tasks necessitate more thoughtful and prolonged cognitive processing. On the other hand, reading tasks tend to engage more automatic cognitive processes, enabling proficient readers to finish the task more quickly and deliver better performance. Hence, for tasks such as reading that rely heavily on automatic cognitive processes, a quicker completion is a mark of proficiency and correlates with superior test performance.

1.3. Profile Analysis

This study employs profile analysis to examine the response time patterns of examinees and their relationship with test performance. Profile analysis encompasses a set of multivariate data analysis techniques (Stanton & Reynolds, 2000) focused on analyzing the shapes and patterns of profiles. Profiles are defined as vectors that contain an individual's or group's subscores from an assessment, such as scores on subdomains like algebra, geometry, and arithmetic in mathematics assessments. By analyzing students' scores on subdomains,

researchers can assess whether they performed equally well across all areas, providing insights into the relative strengths and weaknesses of their abilities based on variations in subscores.

There are various forms and applications of profile analysis. For instance, Giordano et al. (2020) examined the parallelism and flatness of profiles on four segments of adolescent coping among American adolescents who reported a history of nonsuicidal self-injury and those who did not. Parallelism uses one-way Multivariate Analysis of Variance (MANOVA) to assess whether each segment of the profile is identical across groups, while flatness measures the extent to which the profiles are flat across segments within any group. Additionally, criterion-related pattern analysis has been used to examine the relationship between subscores and a specific external criterion, quantifying the proportion of variability in the criterion that can be explained by the level or pattern effect (Davison & Davenport, 2002; Davison et al., 2015). The *level* effect is simply equivalent to the predictive value of the average subscores, while the *pattern* effect considers the variability of the subscores within each profile. An example of the applications of criterion-related profile analysis is provided by Biancarosa et al. (2019), who showed that students' subscores in one achievement test add more predictive value than their single total score in predicting students' scores in another achievement test.

In this study, response time serves as an analogous measure to subscores, with each examinee's response time on each screen or question item making up their profile. Utilizing profile analysis with response time data enables us to explore how much time examinees spent on each item, revealing their test solution behaviors and cognitive processes. Scrutinizing response times for each question allows an understanding of the duration spent on specific items relative to the examinees' average response time, thus quantifying within-person variations in response time. Comparisons of profiles can also be conducted between achievement groups,

testing whether examinees of different abilities spend the same amount of time on different sections of the test. Furthermore, by employing criterion-related profile analysis, this study aims to comprehensively comprehend the intricate relationship between examinees' response time profiles and overall test performance. This approach quantifies the value of within-person variation in response time in predicting examinees' test performance.

1.4. Study Overview and Research Questions

The literature review demonstrates that response time can be influenced by many item and person factors, thereby fluctuating throughout the test. For example, an examinee may answer the first item very rapidly compared to other examinees while answering the second item very slowly compared to other examinees. Therefore, in this study, I hypothesize that the standardized individual response times are not consistent throughout the test. This led to the first RQ:

RQ1: Do examinees' standardized response times fluctuate in the test event?

As the influence of item factors and personal factors on response time happens at the item level and the complex interplay results in the inconsistent association between response time and test performance, it highlights the necessity of examining what are the forms of response time profiles and to what extent the profiles can predict the test performance. This study aims to identify the profiles related to higher scores on test performance which can be referred to as criterion-related patterns. I hypothesize that individual response times are related to the test performance in an inconsistent manner, which further adds to the value of using response time for predicting test performance. For example, spending more time on one item on the test may be positively related to the examinee's overall test performance, while spending more time on another item may be negatively related to their test performance. This led to the second RQ:

RQ2: To what extent can examinees' standardized response time predict their overall test performance? What is the form and contribution of separate response times in predicting the test performance, beyond the overall or average response time of all items?

Another goal of this study focuses on the explainability of the relationship between response time and test performance. I examine the effects of examinee ability, item position, and item difficulty on the predictive value of response time for test performance. To achieve this goal, I differentiate examinees into groups of different levels of achievement. This differentiation allows me to analyze how their response time patterns change within and across test achievement groups and to identify the associated criterion-related patterns. In addition, I also summarize the relationship between criterion-related profiles and item difficulty, as well as item position. To this end, I aim to answer the following RQs:

RQ3: To what extent can examinees' standardized response time predict their overall test performance in each of the achievement groups? Are there any differences in response time profiles and criterion-related patterns across test achievement groups?

RQ4: What is the relationship between criterion-related patterns and item difficulty, as well as item position?

2. Methodology

2.1. Data

This study uses data from a computer-based assessment, the TIMSS 2019 PSI tasks. This section presents a description of the assessment design of the PSI tasks, introduces the data, and outlines the focal variables that were analyzed in this study.

2.1.1. TIMSS 2019 Problem-Solving and Inquiry Tasks

As an innovation in the 2019 assessment cycle of TIMSS, eight PSI tasks were designed as computer-based tests to assess students' higher-order mathematics and science knowledge and skills, such as reasoning. The eight eTIMSS PSI tasks were administered to grade four and grade eight students. Each grade had two PSI tasks, which were separated into two booklets and administered according to a rotated design. The PSI tasks were computer-based and featured visually attractive, interactive scenarios with narratives or themes that simulated real-world problems, aiming to engage students. Each PSI task consisted of a sequence of 4 to 16 multiple-choice or constructed-response items that addressed various topics outlined in the TIMSS 2019 Assessment Frameworks (LindenMullis & Martin, 2017). The development of the PSI tasks and items underwent a thorough and rigorous process (see the TIMSS Technical Report for further details).

This study focuses on the grade-four PSI mathematics assessment. In the grade-four mathematics test, a total of 29 items were distributed among three distinct scenarios (i.e., PSI tasks): School Party, Robots, and Little Penguins. The School Party scenario contained 11 questions. In this scenario, students were tasked with planning a school party, performing calculations for ticket prices, and determining the required quantities of food, drinks, and decorations. The Robots scenario included 6 questions that asked students to use a robot to solve mathematical problems based on input-output rules, subsequently identifying and determining the underlying rules governing the robot's behavior. Lastly, the Little Penguins scenario entailed 12 questions. Students were tasked with filling out a webpage on small penguins, which required them to answer mathematical problems linked to the creatures.

According to the *Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks* (Mullis et al., 2021), the grade-four students were provided with two booklets. In Booklet 15, the students were given three math PSI tasks in the same order (i.e., Penguins, Robots, and School Party), which they had to complete within 36 minutes. Then, they were presented with the science PSI tasks, which were completed in another 36 minutes. In Booklet 16, the students followed a different sequence. They first addressed the science PSI tasks before moving on to the three math PSI tasks. The math PSI tasks were in the same sequence as in Book 15 and needed to be completed within 36 minutes.

2.1.2. Participants

The fourth-grade TIMSS 2019 PSI tasks were administered in 30 countries or economies, involving a total of 27,682 fourth-grade students. To minimize the influence of factors like fatigue and maintain consistency, only students who completed booklet ID 15 were included in the analysis. As a result, the final sample size comprised 13,829 students. Among them, there were 6,724 girls and 6,755 boys. Based on the TIMSS 2019 international benchmarks for mathematics achievement (Mullis et al., 2020), the students' achievements were distributed as follows: 1,453 students (10.51%) did not reach low achievement, 2,522 students (18.24%) reached low achievement, 4,301 students (31.10%) reached intermediate achievement, 4088 students (29.56%) reached high achievement, and 1,465 students (10.59%) reached advanced achievement.

2.1.3. Focal Variables

Scoring and Achievement. The first plausible value of examinees' mathematics achievement, ASMMAT01, was used as the criterion variable. The imputation of examinees' plausible values was documented in the TIMSS 2019 Technical Report (Martin et al., 2020),

which comprises four stages. The first stage employs Item Response Theory (IRT) models to derive item parameters, such as difficulty, discrimination, and guessing effect, for each test item (von Davier, 2020). The three-parameter logistic model, the two-parameter logistic model, and the generalized partial credit model are applied to multiple-choice items, constructed response items worth 1 score point, and constructed response items worth up to 2 score points, respectively. In the second stage, the derived item parameters are combined with students' responses and selected background data in a latent regression model to estimate the examinees' latent ability. The third stage builds upon this model, utilizing the latent regression coefficients, examinees' responses, and background variables to generate five ability estimates for each examinee. In the fourth and final stage, these plausible values are linearly transformed, centered at a mean of 500 and a standard deviation of 100, to align with the previous TIMSS interpretations of students' achievement.

Response Time. In the fourth-grade math PSI tasks, a total of 29 items were distributed across three PSI tasks and 17 screens. There were six screens where only one item was presented to examinees, making the time spent on those screens equal to the item response time. However, in instances where a screen presented multiple items that shared a common stem but required separate responses, the response time was recorded as the same for all items since they appeared on the same screen. There were ten screens that presented two items to examinees, and one screen that displayed three items. Additionally, there were post-clue items that provided correct answers, enabling students to progress to the next question in a series under the same stem. These post-clue items were not included in the study's analysis. Overall, the study used response time data from 17 screens, referred to as 'screen response time' in subsequent sections. The screen response times were equal to, or closely approximated, the item response time.

Item Position and Difficulty. To address the last RQ of this study, I considered two specific attributes of the items: position and difficulty. However, since response time was collected based on the screen unit in this study, it was necessary to align the item position and difficulty with the screen position and screen difficulty, respectively. In this context, screen position refers to the fixed sequence in which screens were presented to the examinees. Additionally, the screen difficulty was directly obtained from the TIMSS Technical Report (Fishbein & Foy, 2020). For screens that contained multiple items, the screen difficulty was calculated as averaging the difficulty of all items within that screen.

2.2. Data Analyses

This section outlines the data analytical plan for addressing the proposed RQs. Firstly, data transformation and descriptive statistical analyses were conducted. Next, a profile analysis using Hotelling's T^2 was performed on the full sample to address the first RQ. The RQ2 was addressed through a criterion-related profile analysis. As for RQ3, the criterion-related profile analysis was conducted separately for the examinee groups of different levels of achievement. Finally, to answer the last RQ4, descriptive statistical analyses and correlational analyses were carried out to examine the influence of item position and difficulty on response time and the relationship between response time and test performance. This section also entails an introduction to the methodology used for the analyses.

2.2.1. Data Transformation

To enable meaningful comparisons between response times across screens and items, the z-score transformation was employed for each screen response time data. This process involved subtracting the raw response time from the mean response time of all examinees and then dividing the result by the standard deviation. This way, I could assess whether an examinee’s

response time differed in comparison to other examinees. In other words, it allowed me to determine if the examinee consistently maintained the same relative position, spending a similar amount of time on different items throughout the test when compared to other examinees.

To address the final RQ, the examinees were divided into five subsets based on the TIMSS international benchmark for mathematics achievement (Mullis et al., 2020). Examinees who scored below 400 were categorized as not reaching the low benchmark. Examinees who scored between 400 and 475 were classified as low achievers, whereas those who scored between 475 and 550 were intermediate achievers. Examinees who scored above 550 were classified as high achievers, while those who scored above 625 were considered advanced achievers.

2.2.2. Descriptive Statistics

I conducted a descriptive analysis of the focal variables, which included screen response time and examinees' test performance. The summaries of mean, standard deviation, and interquartile range (IQR) provide an overview of the central tendency and spread of the test scores and response times. I also calculated skewness to measure their asymmetry and kurtosis to evaluate the peakedness or flatness of the distributions. Additionally, I computed the percentage of missing values for each screen response time and for each examinee. To explore the relationships between the focal variables, I calculated and reported the correlation matrix. Tables and Graphs were used where appropriate.

2.2.3. One-Sample Profile Analysis with Hotelling's T^2

One-sample profile analysis can be understood as the analysis of repeated measures for testing within-subject equality of means. In this study, it was used to determine if examinees' standardized response time remains consistent across all screens. To assess the consistency, a

series of separate univariate t-tests could be performed to compare the mean response times on each pair of screens. However, this approach could become impractical when there are a large number of items due to the excessively large number of combinations, not to mention the increased risk of type I error. Therefore, I chose the one-sample profile analysis with Hotelling's T^2 as the alternative approach to assess the overall difference between screen response times. The one-sample profile analysis with Hotelling's T^2 tests the null hypothesis that the univariate means of standardized screen response times are equivalent. The null and the corresponding alternative hypotheses could be written as

$$H_0: \mu_1 = \mu_2 = ... = \mu_p, \text{ against } H_1: \text{at least one pair was not equal,}$$

where μ is the mean of the standardized response time for screens, the vector [1, 2, …, p] is the index of the screen. According to Bulut and Desjardins (2020), the null hypothesis could also be conceptualized as that the ratios of the means over their hypothesized means are all equal to one, against the alternative hypothesis that at least one of the ratios is not equal to 1. Mathematically, they could be expressed as

$$H_0: \frac{\mu_1}{\mu_1^0} = \frac{\mu_2}{\mu_2^0} = \cdots = \frac{\mu_p}{\mu_p^0} = 1$$

$$H_1: \frac{\mu_j}{\mu_j^0} \neq 1, for\ j \in \{1, 2, 3, \ldots, p\}$$

In addition to testing whether the ratios are all equal to 1, profile analysis can also be used to directly test whether all ratios are equal to each other. Thus, another pair of hypotheses could be proposed:

$$H_0: \frac{\mu_1}{\mu_1^0} = \frac{\mu_2}{\mu_2^0} = \cdots = \frac{\mu_p}{\mu_p^0}$$

H_1: at least one pair of ratios is not equal.

The two null hypotheses shown above could be tested using the paos function in the profileR package (Bulut & Desjardins, 2020) in R (v4.3.1; R Core Team 2023). The results addressed the first RQs proposed in this study (i.e., Do examinees' standardized response times fluctuate throughout the tests?)

2.2.4. Criterion-related Profile Analyses

Davison and Davenport (2002) developed a regression-based statistical method called criterion-related profile analysis to identify profiles (i.e., combinations of subscores) associated with a criterion variable. They asserted that if subscores are predictive, there is a profile pattern associated with high scores on the criterion variable with the least prediction error. This profile pattern can be described in terms of the linear regression coefficients and used to reveal the relationship between the individual subscores and the criterion variable. Additionally, Davison and Davenport (2015) demonstrated that criterion-related profile analysis has the advantage of quantifying the predictive value of subscores beyond the total score. Therefore, to address the RQs proposed in this study, I use the screen response times as the predictor subscores and test performance as the criterion variable to illustrate how separate response times explain the variations in examinees' test performance, beyond the predictive value of the total response time or mean response time. The following is a replication of Davison and Davenport's procedures for identifying the criterion pattern using multiple regressions.

In the first step, a multiple regression analysis is established to predict the criterion variable based on a composite of subscores. In general, the multiple regression can be written as

$$Y_p^{'} = \sum_v b_v X_{pv} + a$$

where $Y_p^{'}$ represents the predicted criterion score for person p, b_v denotes the regression coefficient for predictor subscore v, X_{pv} represents person p's subscore for predictor v, and a is

the intercept constant. To determine the criterion pattern, the regression coefficients identified in the multiple regression equation are subtracted from the mean regression coefficient. Let the criterion pattern be b*, the criterion pattern vector can be mathematically expressed as, $b^* = [b_v^* = b_v - \overline{b}]$, where $\overline{b} = \frac{1}{V}\sum_v b_v$.

The multiple regression analysis decomposes each person's profile of predictor subscores into two components: a profile level effect and a profile pattern effect. The level effect is defined as the mean of the subscores for person p, while the pattern effect is defined as a vector consisting of the deviations between a person's subscore and their mean score $X_{p.}$. Therefore, the level effect can be expressed as, $X_{p.} = \frac{1}{V}\sum_v X_{pv}$, where V is the total number of subscores, and the pattern effect can be denoted as $X_p = [X_{pv} - X_{p.}]$. According to Davison and Davenport (2002; see their Appendix A), the pattern effect can be re-expressed as $\mathrm{Cov}(b_v, X_{pv}) = \sum_v (b_v - b_{.})(X_{pv} - X_{p.})$, which is referred to as the profile fit score. The profile fit score represents the average covariance between a person's profile and the criterion-related pattern identified earlier. It measures the degree of similarity between a person's profile and the pattern that predicts a high score on the criterion variable. A higher covariance indicates a better match between the person's profile and the criterion-related pattern. If there are multiple groups, a separate criterion-related profile can be conducted for each group to determine the corresponding criterion-related patterns.

After identifying the criterion-related pattern in the first step, the second step involves estimating the variation of the criterion variable accounted for by the level and pattern effects. This is accomplished through another regression analysis, which can be written as

$$Y_p^{'} = b_1 X_{p.} + b_2 Cov(b_{p,} X_{pv}) + a$$

where $X_{p.}$ is the mean of the predictor subscores for person p (level effect), $\mathrm{Cov}(b_v, X_{pv})$ is the pattern effect, b1 is the regression coefficient of the level effect, and b_2 is the regression coefficient of the pattern effect. Then, the regression equation can be used to examine whether profile pattern effects incrementally explain the variation in the criterion variable over level effects. This is analyzed through a series of hierarchical regressions, where the criterion variable is predicted by (1) the level effect alone, (2) the pattern effect alone, (3) the increment of the pattern effect above and beyond the level effect, and (4) the increment of the level effect above and beyond the pattern effect.

The third and last step of criterion-related profile analysis is to conduct cross-validation, which involves replicating the profile pattern and level effects in another sample. To do so, the data can be randomly split into two subsets. The criterion pattern (b*) obtained by analyzing one subset is then used to predict the criterion for the other subset, and vice versa. This allows testing the generalizability of the results.

To perform the criterion-related profile analysis, I utilized the ProfileR package developed by Bulut and Desjardins (2018). I employed the *cpa* function with the default parameters, except for setting na.action to “na.omit”. This adjustment resulted in omitting missing values for the analysis, instead of the default setting, which terminates the analysis and reports errors.

2.2.5. Profile Analysis for Assessing Parallelism, Equality, and Flatness

To compare examinees of different achievement groups, I adopted profile analysis for assessing parallelism, equality, and flatness. By definition, parallelism refers to the similarity in the shape and pattern of the profiles across different groups. When profiles are parallel, it means that the relationship between predictors and the dependent variable, including the strength and

direction of the relationship, remains consistent across all groups. Parallelism can be tested using a one-way MANOVA, where the null hypothesis assumes no significant interaction between the subscore variables and the group. If the null hypothesis is rejected, it indicates that there are significant differences in the profiles among the groups.

Once parallelism is confirmed, further analysis can be conducted to examine equality and flatness in the profiles. The equality test aims to determine if there are significant differences in the mean scores across all subscore variables among the groups. To perform this test, the grand mean of the subscore variables is calculated for each group. For two groups, a univariate test can be employed, while a between-group one-way ANOVA is suitable for three or more groups. If the null hypothesis is rejected, it indicates that at least one group significantly differs from the others in terms of the mean subscore variables. On the other hand, flatness examines the extent to which profiles are similar within each group, which is similar to the profile analysis for one sample. The flatness test can be conducted using Hotelling's T^2 test with the null hypothesis that there are no differences in the mean subscore variables among the groups, given the profiles are parallel.

In this study, the profile analyses for assessing parallelism, equality, and flatness of screen response times were performed using the *pdg* function in the *profileR* package (Bulut & Desjardin, 2020) in R. The function tests parallelism first and then proceeds to test equality and flatness if parallelism is confirmed. In addition, to visualize the profiles and mean screen response times for each achievement group and compare them among groups, the R package ggplot2 was employed.

3. Results

The results section is divided into three parts. First of all, the descriptive statistics of focal variables, including the response time of examinees on the 17 screens of the PSI and their test scores, are reported in Section 3.1. Accompanying this, the graphical representation of missing values and the distributional characteristics of the focal variables are presented. Furthermore, a report of the bivariate correlations between these focal variables is also included. The second part presents the results of profile analyses conducted on the full sample (see Section 3.2). This includes the findings from a one-sample profile analysis with Hotelling's T^2, which was used to discern whether, on average, examinees spent equal amounts of time across the PSI screens. Additionally, the results from a criterion-related profile analysis of the full sample are detailed, showing the regression coefficients of individual screen response times for predicting test scores, the criterion-related pattern, and their predictive value in terms of full effect, level effect, and pattern effect. The cross-validation analysis results are also reported to demonstrate the generalizability of the earlier results. In the third part (Section 3.3), I report the results of criterion-related profile analyses conducted for each of the classified achievement groups. The results were similar to that in the second part – the corresponding regression coefficients, criterion-related patterns, and level/pattern/full effects are presented. In addition, the results of tests for parallelism, equity, and flatness are reported. Finally, in Section 3.4, the previous results are summarized to address the proposed RQs of this study.

3.1. Descriptive Statistics of Response Time and Test Score for the Full Sample

The percentage of missing values for each screen was visualized in Figure 1, suggesting that screens toward the end of the test tend to have more missing data. This reflects that some students were not able to finish the test within the time constraint. The histogram of the

percentage of missing values for each examinee is presented in Figure 2. It indicates that while most of the examinees were able to finish the test (i.e., the percentage of missing data was 0), the other examinees generally reached half of the total screens.

Figure 1

Missing Percentage of Response Time for Each Screen

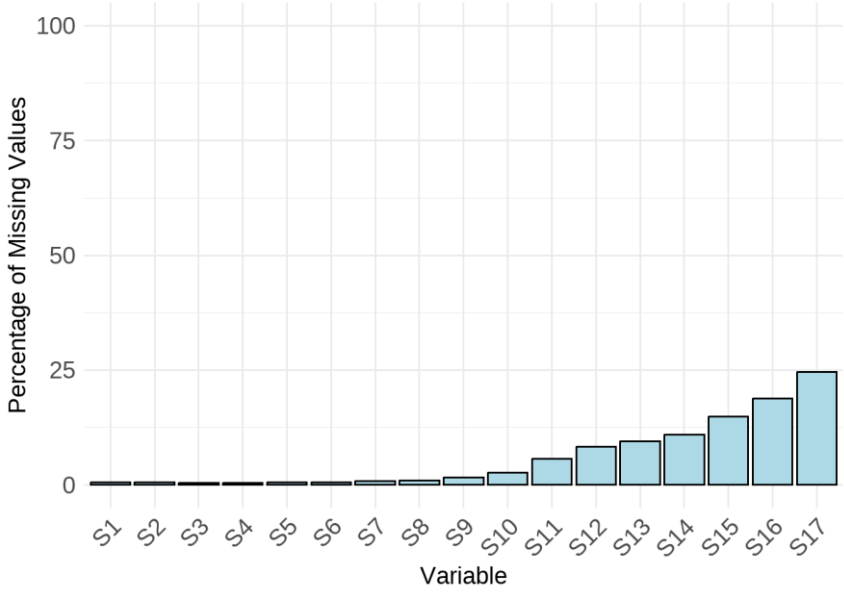
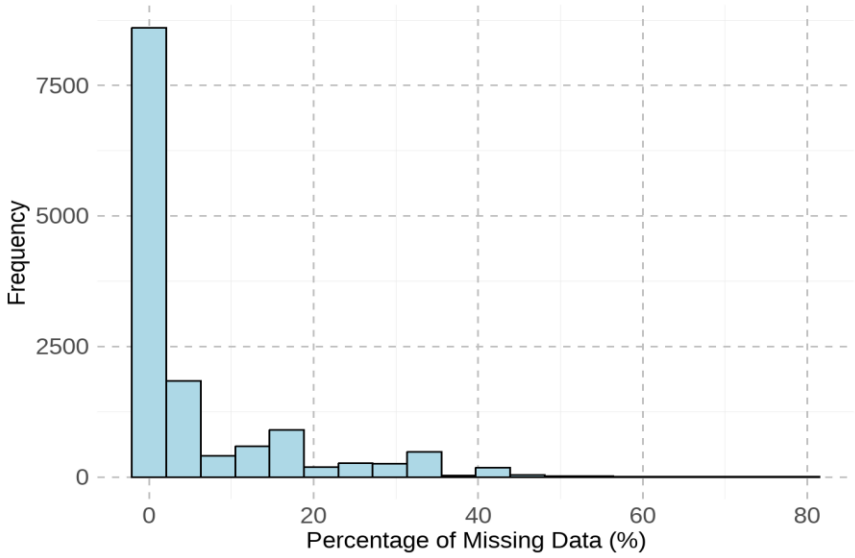


Figure 2

Histogram of Missing Data Percentage for Each Examinee



The descriptive statistics for the mean, standard deviation, IQR, skewness, and kurtosis of the screen response time and test scores for the full sample are presented in Table 1. This shows that the average screen response time ranges from 34.83 to 191.39, with standard deviations varying from 24.64 to 100.72. The IQR spanned from 22.39 to 121.52. The distributions of the focal variables are shown in Figure 3. All the distributions for screen response time demonstrated positive skewness and kurtosis values. There were some screen response times that showed very large kurtosis values (e.g., S7: 130.59; S11: 31.33), distributed as having heavier tails and a more pronounced peak. This means that a large proportion of examinees spent extremely longer or shorter time on those screens.

Table 1

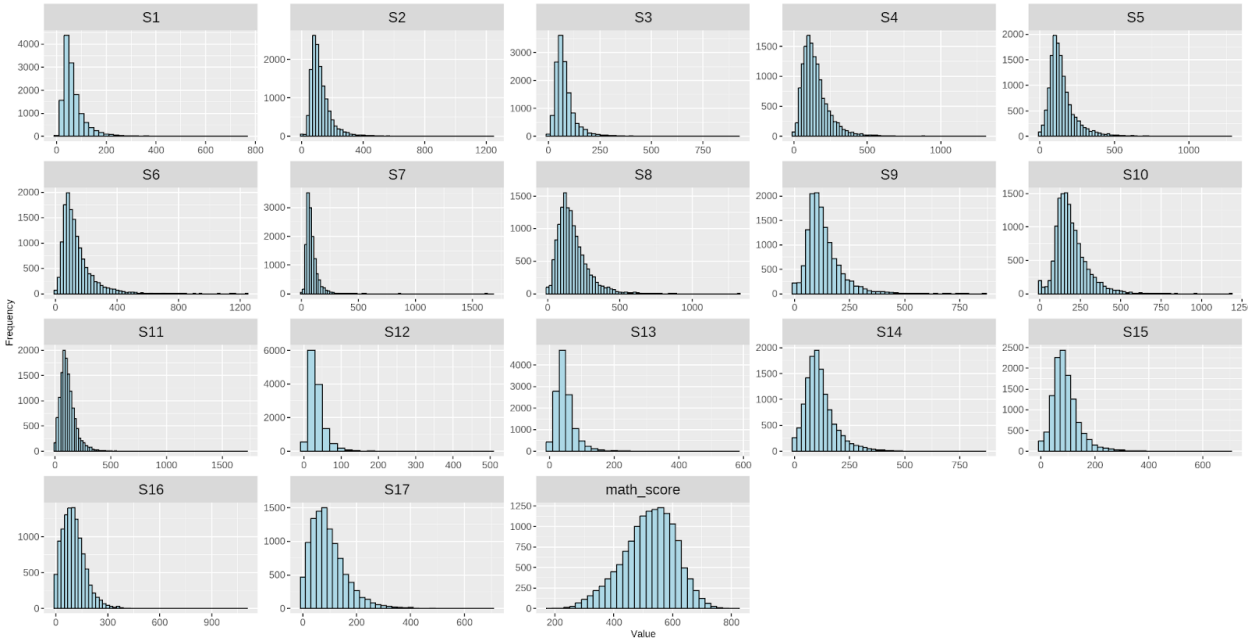
Descriptive Statistics of Focal Variables (Full Sample).

Screen	*Mean*	*SD*	*IQR*	Skewness	Kurtosis
S1	68.79	51.32	43.07	3.50	21.65
S2	123.12	74.53	67.80	3.23	21.23
S3	81.30	55.27	46.30	3.58	24.92
S4	148.90	98.20	100.43	2.34	11.32
S5	149.38	97.96	86.65	2.58	11.91
S6	137.13	100.70	96.82	2.48	10.84
S7	91.66	57.77	49.63	6.85	130.59
S8	170.21	100.72	114.15	1.56	4.76
S9	130.57	78.06	80.81	1.88	6.88
S10	191.39	97.52	107.37	1.43	4.40
S11	113.98	72.94	77.62	2.89	31.33
S12	34.83	24.64	22.39	3.95	37.48
S13	48.47	31.18	29.80	2.81	19.20

S14	115.44	70.93	74.22	1.83	7.44
S15	91.35	52.65	54.68	1.93	8.72
S16	103.27	68.27	86.46	1.54	9.83
S17	95.40	69.03	80.63	1.61	5.01
Score	520.17	89.91	121.52	-0.36	-0.06

Figure 3

Histograms of the Focal Variables



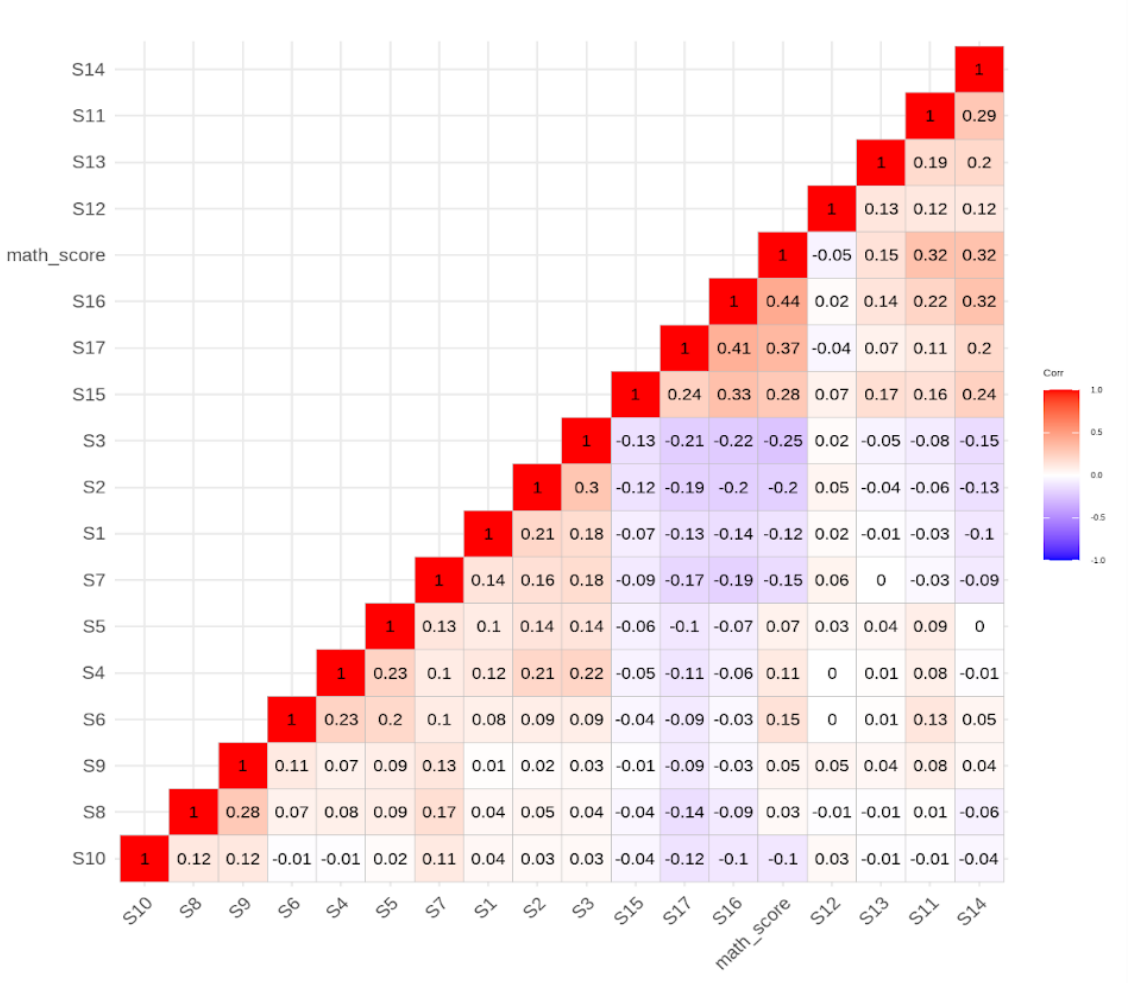
Note: S1-S17 represents the response times for the 17 screens.

The bivariate Pearson correlations between the focal variables are presented in Figure 4. With regard to the correlations between test scores and screen response times, there were four moderate positive correlations (S11, S14, S16, and S17), four small positive correlations (S4, S6, S13, and S15), five small negative correlations (S1, S2, S3, S7, and S10), and four trivial correlations (S5, S8, S9, and S12). Notably, with more time spent on later screens in the test

(e.g., Screens 14, 16, and 17), examinees' math scores were higher. These screens also exhibited negative correlations with the response times on the other screens, demonstrating a pattern that students who spent less time on earlier screens tended to spend more time on these latter screens. In addition, the response times on the first few screens had positive correlations among each other. However, these response times had trivial to small positive or negative correlations with the test score.

Figure 4

Pearson Correlations Coefficients among Focal Variables



3.2. Results of Profile Analysis for the Full Sample

A one-sample profile analysis using Hotelling's T^2 was conducted to examine if examinees spent the same amount of time on each of the 17 screens on average. According to the results, both null hypotheses specified in the analysis were rejected. The first test rejected the null hypothesis that the ratio of the mean response time is 1 for each screen, $T^2 = 152906.62$, $F(17, 10309) = 8980.57$, $p < .001$. Moreover, the second test rejected the null hypothesis that the ratio of the mean response time is the same for each screen, $T^2 = 1452.63$, $F(16, 10310) = 1452.63$, $p < .001$. Therefore, it can be concluded that the examinees' response time did not remain constant across screens. As this study had no interest in determining which screen response time differs, post hoc comparisons were not performed.

A multiple regression analysis was conducted to predict all examinees' test scores based on their response times on 17 screens. Table 2 shows the regression coefficients and the associated criterion-related pattern. The criterion-related pattern was calculated by subtracting each regression coefficient from the mean regression coefficients. The results showed that response times on separate screens are moderately strong predictors of test scores, with each of the separate response times contributing independently to the prediction. Figure 5 visualizes the criterion-related pattern across the 17 screens, revealing nine positive values and eight negative values. The positive values indicate that spending more time on those screens is associated with higher test performance. Conversely, negative values indicate that spending more time on those screens is associated with lower test performance. It can be found that examinees with the criterion-related pattern tended to spend less time on the first few questions and more time on the last few questions.

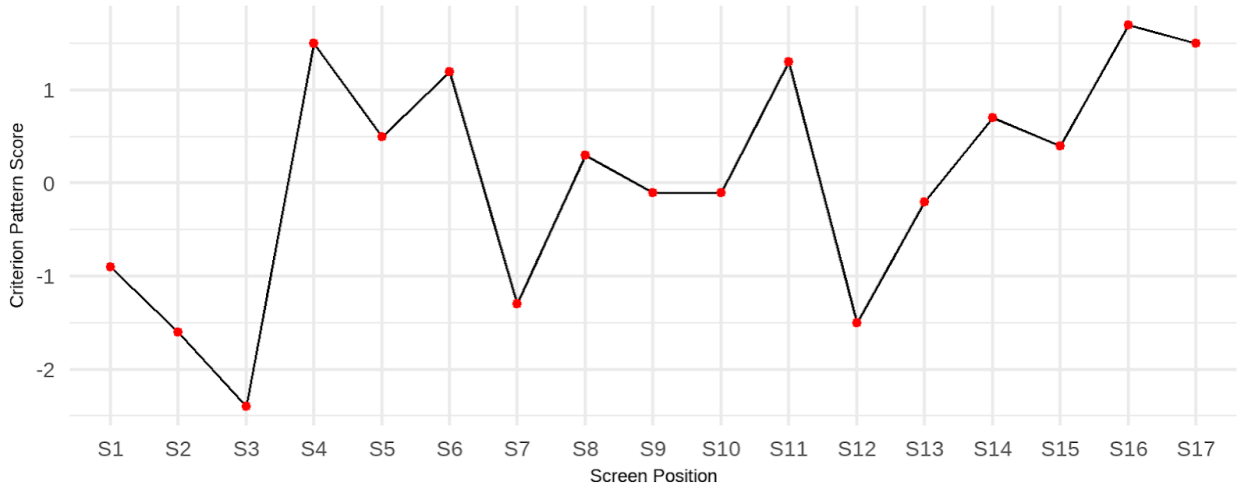
Table 2

Regression Coefficients and the Associated Criterion-Related Patterns (Full Sample)

Screen	b_v	*S.E.*	*t*	*p*	*b**
S1	-0.04	0.01	-3.50	< .001	-0.09
S2	-0.10	0.01	-9.06	< .001	-0.16
S3	-0.19	0.01	-16.79	< .001	-0.24
S4	0.21	0.01	17.50	< .001	0.15
S5	0.10	0.01	8.82	< .001	0.05
S6	0.18	0.01	14.78	< .001	0.12
S7	-0.07	0.01	-6.46	< .001	-0.13
S8	0.09	0.01	8.71	< .001	0.03
S9	0.04	0.01	4.09	< .001	-0.01
S10	-0.05	0.01	-5.34	< .001	-0.01
S11	0.19	0.01	19.08	< .001	0.13
S12	-0.09	0.01	-10.78	< .001	-0.15
S13	0.03	0.01	3.96	< .001	-0.02
S14	0.12	0.01	12.95	< .001	0.07
S15	0.09	0.01	10.61	< .001	0.04
S16	0.23	0.01	24.05	< .001	0.17
S17	0.21	0.01	22.75	< .001	0.15

Figure 5

Criterion-Related Patterns for the Full Sample



Once the criterion-related pattern is identified, another multiple regression analysis, $Y_p^{'} = b_1 X_{p.} + b_2 Cov(b_{p,} X_{pv}) + a$, decomposed the proportion of variance explained in the test score attributable to the level and profile pattern effects. A set of hypothesis tests examined the significance of each of the pattern, level, and full effects (i.e., combined effect). The null hypothesis was that the proportion of variability in test scores explained by the pattern/level/full effect was zero. Two additional tests were conducted to examine if the pattern or level effects have significant incremental effects on each other with the null hypothesis that the proportion of variability in test scores explained by the level/pattern effect was equal to the full effect. The results are shown in Table 3, suggesting that all null hypotheses were rejected. Specifically, in the full sample, the profile pattern effect alone accounts for 34.21% of the variance in test scores, while the level effect alone only explains 10.03% of the variance. Additionally, there was an incremental effect of the profile pattern effect and the level effect on each other, as the combined effects significantly explain more variability in test scores ($R^2 = 0.40$; combined effect vs. level effect: $\Delta R^2 = 0.30$, $p < .001$; combined effect vs. pattern effect, $\Delta R^2 = 0.06$, $p < .001$). The results

suggested that the criterion-related pattern identified for the full sample performed significantly better in predicting examinees' test scores than the profile and level effects.

Table 3

Hypothesis Testing for the Changes of R^2 (Full Sample)

Hypotheses	ΔR^2	*df1*	*df2*	*F*	*p*
R^2.full = 0	0.40	17	10308	41.58	<.001
R^2.pattern = 0	0.34	16	10308	334.94	<.001
R^2.level = 0	0.10	1	10308	1150.07	<.001
R^2.full = R^2.level	0.30	16	10308	319.19	<.001
R^2.full = R^2.pattern	0.06	1	10308	965.77	<.001

Note: R^2.full: the proportion of variance in the criterion variable explained by the full model. R^2.pattern: the proportion of variance explained by the pattern effect. R^2.level: the proportion of variance explained by the level effect.

Finally, cross-validation was used to examine if the results produced in the criterion-related profile analysis were generalizable. The original full sample was randomly split into two equally sized subsets, and then the test scores were predicted by each set of screen response times in the two randomly drawn subsets. After that, the regression coefficients of each of the screen response times in one subset were used to generate the criterion-related patterns, profit fit scores, and the estimation of pattern/level/full effects for the other subset, and vice versa. The differences in the obtained criterion-related patterns for the two subsets were visualized in Figure 6. The results of the regression analysis examining the pattern, level, and full effects in the cross-validation were reported in Table 4. According to the results, the criterion-related patterns and the proportion of variance in test scores explained were comparable between the two subsets. Moreover, those results were also similar to the results of the criterion-related analysis conducted

for the full sample. Thus, it can be concluded that the results of the criterion-related profile analysis were generalizable.

Table 4

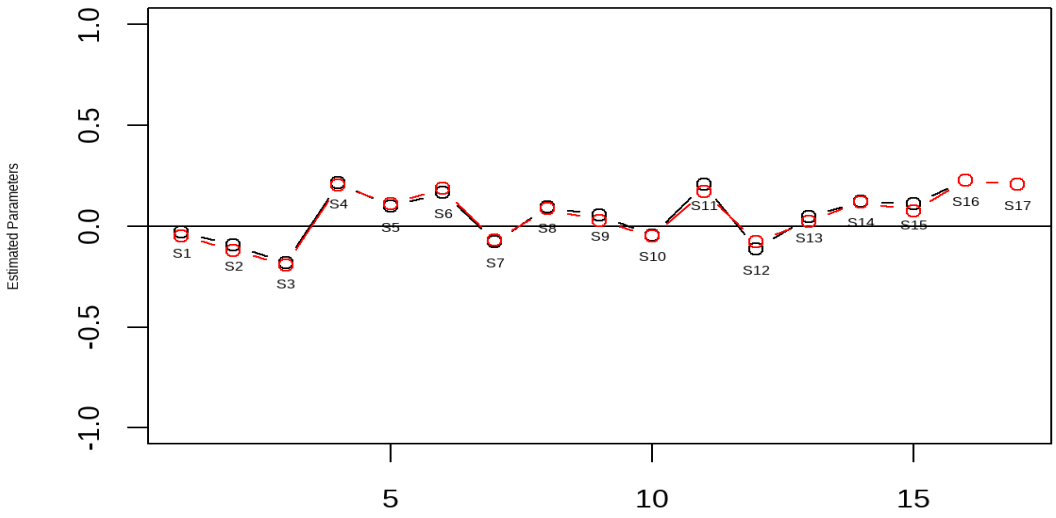
Hypothesis Testing for the Changes of R^2 in Cross Validation

Hypotheses	ΔR^2	*df1*	*df2*	*F*	*p*
$R^2.full = 0$					
Random Sample 1	0.41	1	5161	35.42.16	<.001
Random Sample 2	0.39	1	5159	3230.03	<.001
$R^2.pattern = 0$					
Random Sample 1	0.34	1	5162	2677.83	<.001
Random Sample 2	0.34	1	5160	2624.81	<.001
$R^2.level = 0$					
Random Sample 1	0.11	1	5162	643.44	<.001
Random Sample 2	0.09	1	5160	510.83	<.001
$R^2.full = R^2.level$					
Random Sample 1	0.30	1	5162	2578.05	<.001
Random Sample 2	0.30	1	5160	2474.82	<.001
$R^2.full = R^2.pattern$					
Random Sample 1	0.09	1	5162	569.55	<.001
Random Sample 2	0.09	1	5160	401.57	<.001

Note: R^2.full: the proportion of variance in the criterion variable explained by the full model. R^2.pattern: the proportion of variance explained by the pattern effect. R^2.level: the proportion of variance explained by the level effect.

Figure 6

Criterion-Related Patterns for the Random Samples in Cross-Validation



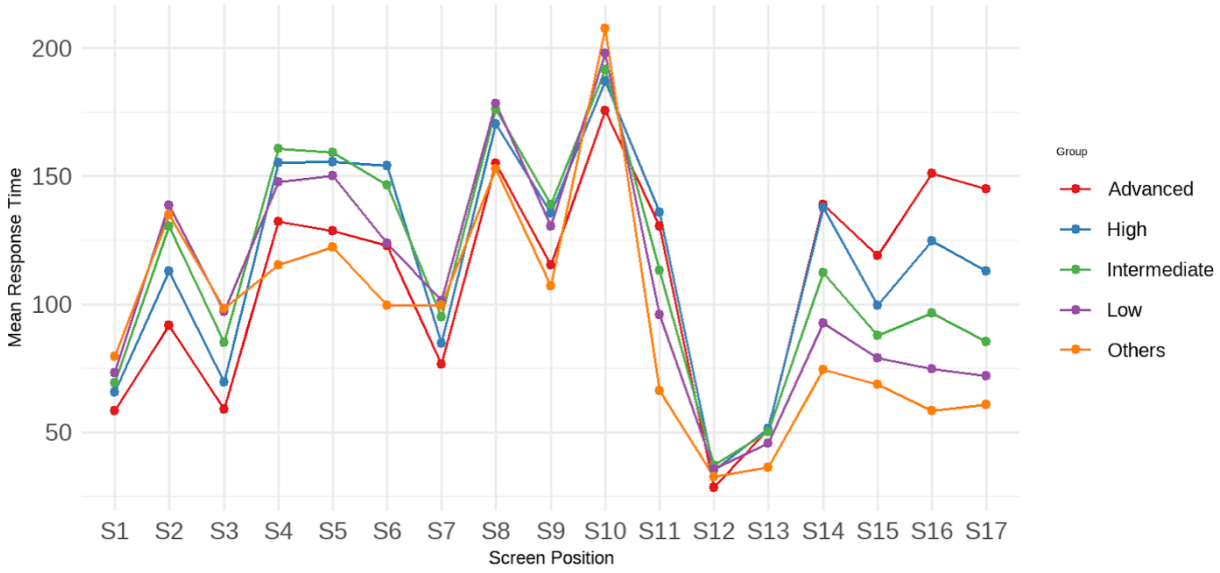
3.3. Results for Comparing Achievement Groups

The significance of the differences in the profiles of response time across achievement groups was examined. Not surprisingly, the null hypothesis of MANOVA for the test of parallelism was rejected, which revealed that the profiles of the response time across achievement groups were not parallel, $F(32, 20616) = 154.42$, $p < .001$, *Wilks'* $\lambda = 0.65$. As parallelism was not present, I did not proceed to tests for equality and flatness. As shown in Figure 7, the response time profile for each achievement group was visualized in the line chart. The chart displays similar but unique response time patterns for examinees of different performance groups. It was observed that examinees with lower test performance tended to spend more time on early items and less time on later items. Additionally, the response time spent on Screens 4-6 by examinees with the lowest test performance deviated dramatically from the response time of other achievement groups. While all achievement groups spent the longest time on Screen 10, the group of examinees with the lowest performance spent the longest time

among all examinees. They dedicated much more time to later items compared to students with lower performances. On the other hand, examinees with higher test performances spent less time on early items but more time on later items, compared to other groups with lower test performance. Regarding students in the middle ranges, their average response times were higher on Screens 4-9 and Screens 11-14 compared to students with either the highest or lowest performance.

Figure 7

Profile Means of Response Time across Screens for Achievement Groups



The regression coefficients of screen response time in predicting test scores of examinees in each achievement group are presented in Table 5. For the examinees who did not reach low achievement or those who reached high achievement, there were 11 positive regression coefficients and six negative regression coefficients. For the low and intermediate achievers, there are nine positive regression coefficients and eight negative regression coefficients. Lastly, the advanced achievement group saw only six positive regression coefficients and 14 negative regression coefficients. The associated criterion-related patterns for the achievement groups are

also presented in Table 5 and are further visualized in Figure 8. It can be observed that in the advanced achievement group, investing more time in the first three screens is associated with higher scores, while investing more time on screens in the middle is linked to lower test scores. Specifically, spending more time on Screens 6, 9, and 12 shows the strongest negative association with test scores. In contrast, for examinees who did not achieve low scores, spending more time on Screens 4 and 11 can result in relatively higher scores within that achievement group. The patterns related to the criteria exhibit some fluctuations for the other groups, but to a lesser extent compared to the groups with the highest and lowest achievers.

Figure 8

Criterion-Related Pattern by Achievement Group

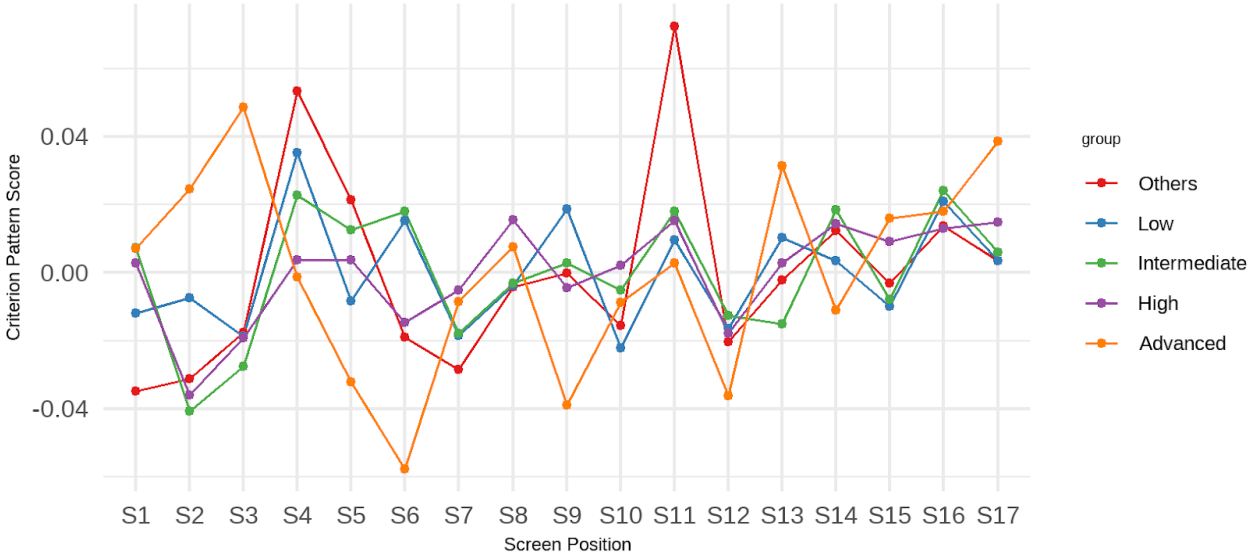


Table 5

Regression Coefficients and the Criterion-Related Patterns by Achievement Group

	Below Low		Low		Intermediate		High		Advanced	
	b_v	b^*	b_v	b^*	b_v	b^*	b_v	b^*	b_v	b^*
S1	-0.02	-0.03	-0.01	-0.01	0.01	0.01	0.01	0.00	0.00	0.01
S2	-0.01	-0.03	0.00	-0.01	-0.04	-0.04	-0.03	-0.04	0.01	0.02
S3	0.00	-0.02	-0.01	-0.02	-0.02	-0.03	-0.01	-0.02	0.04	0.05
S4	0.07	0.05	0.04	0.04	0.03	0.02	0.01	0.00	-0.01	0.00
S5	0.04	0.02	0.00	-0.01	0.02	0.01	0.01	0.00	-0.04	-0.03
S6	0.00	-0.02	0.02	0.02	0.02	0.02	-0.01	-0.01	-0.07	-0.06
S7	-0.01	-0.03	-0.01	-0.02	-0.01	-0.02	0.00	-0.01	-0.02	-0.01
S8	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.01
S9	0.02	0.00	0.02	0.02	0.01	0.00	0.00	0.00	-0.05	-0.04
S10	0.00	-0.02	-0.02	-0.02	0.00	-0.01	0.01	0.00	-0.02	-0.01
S11	0.09	0.07	0.01	0.01	0.02	0.02	0.02	0.02	-0.01	0.00
S12	0.00	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.02	-0.05	-0.04
S13	0.02	0.00	0.01	0.01	-0.01	-0.02	0.01	0.00	0.02	0.03
S14	0.03	0.01	0.01	0.00	0.02	0.02	0.02	0.01	-0.02	-0.01
S15	0.01	0.00	-0.01	-0.01	0.00	-0.01	0.01	0.01	0.00	0.02
S16	0.03	0.01	0.03	0.02	0.03	0.02	0.02	0.01	0.01	0.02
S17	0.02	0.00	0.01	0.00	0.01	0.01	0.02	0.01	0.03	0.04

The results of examining the level and pattern effects are presented in Table 6. Overall, the findings indicated that the proportion of variation in test scores explained by the full effect for all achievement groups was significantly smaller than the full effect in the full sample. The R^2 values for the full effect ranged from 5% to 9%, suggesting that the relatively high R^2 value for predicting test scores based on response time primarily resulted from the differences among achievement groups. However, for all achievement groups, the pattern effects surpassed the level effects, demonstrating the value of using examinees' separate screen response time to predict their test scores. This implies that the predictive value of response time for test scores was mostly attributed to the individual's within-person variability in response time patterns. With the exception of the achievement group identified as below low achievement, the level effect for other achievement groups was nearly zero and negligible, although still statistically significant.

Table 6

Hypothesis Testing for the Changes of R^2 by Achievement Group

Hypotheses	*ΔR²*	*df1*	*df2*	*F*	*p*
Below Low Achievement					
R^2.full = 0	0.09	17	1152	6.66	<.001
R^2.pattern = 0	0.04	16	1152	2.98	<.001
R^2.level = 0	0.04	1	1152	43.61	<.001
R^2.full = R^2.level	0.05	16	1152	4.20	<.001
R^2.full = R^2.pattern	0.05	1	1152	63.02	<.001
Low Achievers					
R^2.full = 0	0.06	17	1869	7.11	<.001

R^2.pattern = 0	0.05	16	1869	6.71	<.001
R^2.level = 0	0.00	1	1869	10.51	<.001
R^2.full = R^2.level	0.06	16	1869	6.86	<.001
R^2.full = R^2.pattern	0.01	1	1869	12.83	<.001
Intermediate Achievers					
R^2.full = 0	0.06	17	2999	12.83	<.001
R^2.pattern = 0	0.06	16	2999	12.76	<.001
R^2.level = 0	0.00	1	2999	12.39	<.001
R^2.full = R^2.level	0.06	16	2999	12.80	<.001
R^2.full = R^2.pattern	0.00	1	2999	13.04	<.001
High Achievers					
R^2.full = 0	0.05	17	2972	10.11	<.001
R^2.pattern = 0	0.05	16	2972	9.48	<.001
R^2.level = 0	0.00	1	2972	29.74	<.001
R^2.full = R^2.level	0.05	16	2972	8.80	<.001
R^2.full = R^2.pattern	0.00	1	2972	19.27	<.001
Advanced Achievers					
R^2.full = 0	0.09	17	1244	7.23	<.001
R^2.pattern = 0	0.07	16	1244	6.20	<.001
R^2.level = 0	0.01	1	1244	15.71	<.001
R^2.full = R^2.level	0.08	16	1244	6.61	<.001

R^2.full = R^2.pattern	0.02	1	1244	22.04	<.001

Note: R^2.full: the proportion of variance in the criterion variable explained by the full model. R^2.pattern: the proportion of variance explained by the pattern effect. R^2.level: the proportion of variance explained by the level effect.

3.4. The Effects of Item Difficulty and Position

To investigate the influence of item difficulty and position on the correlation between item response time and test scores, I constructed a table that presents the difficulty, position, and regression coefficient for each screen (refer to Table 7). These screen attributes are perceived as reflecting the attributes of the items. I then performed a Pearson correlation analysis. The results revealed that examinees tend to spend less time on screens located towards the end of the test (r = -0.20) and those containing easier items (r = 0.27). In addition, screen difficulty was strongly and positively correlated with the regression coefficient of screen response time (r = 0.54) and the corresponding criterion-related pattern (r = 0.59), implying that spending more time was linked with improved test performance, particularly for more difficult items. Furthermore, screen position showed a moderate association with the regression coefficient (r = 0.47) and criterion-related pattern (r = 0.49) of screen response time. This suggests that spending more time on items presented toward the end of the test tends to correlate with better test performance. These findings highlight the fact that the relationship between item response time and test performance is contingent upon item attributes such as difficulty and position.

Table 7

The Effect of Screen Difficulty and Position on Screen Response Time and the Relationship between Response Time and Test Performance

Screen	Difficulty	Position	b_v	b^*	Response Time	#Items
S1	1.23	1	-0.04	-0.09	68.79	1
S2	-0.15	2	-0.10	-0.16	123.12	3
S3	0.14	3	-0.19	-0.24	81.30	2
S4	0.57	4	0.21	0.15	148.90	2
S5	0.58	5	0.10	0.05	149.38	2
S6	0.56	6	0.18	0.12	137.13	2
S7	-0.70	7	-0.07	-0.13	91.66	2
S8	0.49	8	0.09	0.03	170.21	2
S9	0.71	9	0.04	-0.01	130.57	1
S10	0.94	10	-0.05	-0.01	191.39	1
S11	0.92	11	0.19	0.13	113.98	2
S12	0.26	12	-0.09	-0.15	34.83	1
S13	-0.73	13	0.03	-0.02	48.47	1
S14	1.06	14	0.12	0.07	115.44	1
S15	0.64	15	0.09	0.04	91.35	2
S16	1.39	16	0.23	0.17	103.27	2
S17	1.49	17	0.21	0.15	95.40	2

3.5. Summary of the Results

In conclusion, the results of this study provide answers to the RQs posed. For the first RQ, it was observed that examinees' standardized response time varied throughout the test. This means that while they spent more time on certain items compared to their peers, they spent less

time on other items. Regarding the RQ2, screen response time proved to be an effective predictor of test scores, with an R^2 value of 0.40. In addition, it was found that analyzing separate screen response times and within-person variability in response time yielded better predictions of test scores compared to using average response time alone. Furthermore, spending more time on certain items correlated with higher test scores, while spending more time on other items correlated with lower test scores. Thus, the relationship between response time and test scores varied across different items. Regarding the RQ3, it was observed that the examinees' response time was less successful in predicting their test scores within each achievement group. This suggests that the variation in test scores is primarily explained by differences among the groups. Moreover, differences in response time profiles were identified among the achievement groups. Advanced achievers and examinees who did not reach low achievement exhibited distinct criterion-related patterns, which differed from the patterns observed in the other groups. Finally, addressing the last RQ, it was found that criterion-related patterns were positively associated with item position and item difficulty, indicating that these factors influenced response time patterns.

4. Discussion

The goal of this study is to empirically explore the changes in examinees' response time within a test event and the value of using response time for predicting examinees' test performance. Based on the TIMSS 2019 PSI dataset, this study addressed these broad and progressive questions: (1) Do examinees' standardized response times fluctuate across screens or items? (2) To what extent can examinees' standardized response time predict their overall test performance? What is the form and contribution of separate response times in predicting the test performance, beyond the overall or average response time of all items? (3) To what degree can

the standardized response times of examinees serve as predictors for overall test performance within each achievement group? Are differences discernible in response time profiles and criterion-related patterns among different test achievement groups? and (4) How does the relationship between criterion-related patterns interact with item difficulty and item position? In this section, the findings corresponding to each question are discussed. Then, the limitations of this study are discussed. Finally, conclusions are drawn based on the findings and provide recommendations for future research on using response time to predict student test scores.

4.1. Discussions about the Results

In response to the first RQ, this study discovered that examinees' standardized response time varies across different screens. In comparison to other examinees' response times, an individual's response time may fluctuate, such as taking more time on one item than others but taking less time on the next. This is an anticipated finding, given that response time is influenced by several personal factors, including an individual's ability (Deribo, 2023) and time management skills (van der Linden, 2009). The finding also aligns with the literature as Thomas (2006) found that some examinees exhibited a tendency to spend more time on earlier items, while others preferred to devote more time to later ones. Moreover, as motivation levels and test engagement vary across different test items (Michaelides et al., 2020; Wise& Kingsbury, 2016), it is not surprising to observe fluctuations in an individual's standardized response time. The results of the first RQ lay the groundwork for the remaining RQs, as they highlight the existence of within-examinee variability in response time.

In investigating the effects of item attributes on response time and its relationship with test performance, the findings of this study indicate that, generally, examinees tend to spend less time on screens located toward the end of the test, which is aligned with existing literature

(Sweller et al., 1998; Vida et al., 2021). One possible reason for this pattern is that TIMSS PSI tasks imposed time constraints. As a result, examinees may feel compelled to accelerate their problem-solving pace, despite recognizing that the test scores may have a limited impact on their personal lives. Another potential explanation is that examinees' cognitive resources gradually become depleted, leading to disengagement in the later sections of the test. This disengagement can be observed as a decrease in response time. This study also revealed a negative correlation between response time and item difficulty. In other words, students spent less time on screens that contained easier items. This is expected as easier items are more likely to rely on automatic processes, which is a solution behavior involving faster response time (Schneider & Shiffrin, 1977).

This study used a multiple regression approach to predict examinees' test performance based on examinees' screen response times. Though previous studies have shown the value of response time in predicting examinees' performance (e.g., Castro et al., 2019), many have concentrated on item-level accuracy or correctness rather than overall test scores or ability estimates. This study's findings underscore that each response time differently predicts overall test performance in terms of strength and direction, as indicated by the regression coefficients and criterion-related patterns. These results align with my expectation, given that both item-specific factors and personal factors collectively influence response time, and such complex interplay can yield inconsistent relationships between response time and test performance across different items. For instance, more capable examinees tend to answer difficult questions more quickly and effectively than their counterparts (Rayner, 1998). In this case, a shorter response time might correspond to better overall test performances. Conversely, capable examinees often invest more time in later items in the test than other examinees (Thomas, 2006), implying a

negative correlation between response time and test performance. Moreover, some items in the test may be challenging or tricky, requiring a sufficient amount of consideration and thinking before arriving at the correct answer. Therefore, dedicating more time to these items could be linked to higher test performance. Moreover, for knowledge-based items where the response relies on information retrieval or for items that rely on routine solution processes, spending an excessive amount of time on them may indicate examinees' unfamiliarity with the relevant knowledge or solution processes, leading to lower performance. The finding corresponding to RQ4 of this study also supports these explanations and will be discussed in a subsequent section.

The current study illustrates that the predictive value of response time for test performance can be dissected into two effects using criterion-related profile analysis: the profile level effect and the pattern effect. Of these two, the pattern effect reveals the unique predictive value attributable to within-examinee variability in response time. Notably, the pattern effect exceeds the level effect. Therefore, this study substantiates that within-examinee variability explains a unique and larger portion of the variance in overall test performance, which is not accounted for by the average or total response time. These effects underwent cross-validation and were presented consistently in both random samples. The criterion-related pattern of response time in predicting test performance also highlights the value of using subscores to predict a criterion variable in the context of response time. It provides evidence that employing separate response times to predict test performance is more advantageous than using an averaged response time or the total time.

After determining the predictive value of individual response times for test performance using a full sample, I classified the examinees according to their levels of achievement. Then I found that high-achieving examinees generally spend less time on initial items but allocate more

time to later ones. In contrast, low achievers tend to dedicate more time to early items and less to those later in the test. One plausible explanation for this pattern, suggested by the dual-processing theory (Schneider & Shiffrin, 1977), is that proficient examinees may be better equipped with automatic processes or possess the skills and knowledge required to rapidly identify the correct solution. Conversely, examinees who rely more heavily on controlled processes may eventually deplete their cognitive resources, leaving them less motivated and less equipped to carefully evaluate later items (Sweller et al., 1998; Wise & Smith, 2011). The reduction in response time can thereby be attributed to the degree of disengagement. These findings are in line with those of Thoma's study, which discovered that examinees with the highest performance tend to maintain the quickest initial pace. Moreover, their response times show the most gradual decrease when compared to examinees with lesser abilities.

For the results of criterion-related profile analysis conducted for each performance group, it was observed that within each achievement group, predictions of test performance using response times were less accurate than when the full sample was employed. This result was anticipated, as grouping examinees into separate performance categories inherently accounts for considerable variation in test performance. In comparing the criterion-related patterns identified within each performance group, this study has found distinct correlations. For advanced achievers, a lesser amount of time spent on items is more likely to be associated with better performance. This may be explained by the fact that advanced achievers tend to use more automated processes, thus spending less time on items compared to other advanced achievers, reflecting their excellence in the tested knowledge materials and abilities. Conversely, for examinees who achieved the lowest performance, spending more time on items generally correlates with better performance. The reason may be that examinees who achieved the lowest

performance might have run out of cognitive resources and felt fatigued throughout the test event. Therefore, spending more time on items may indicate that they were still engaged in problem-solving behavior rather than engaging in rapid guessing. These differences in criterion-related patterns across varying performance groups indicate that the relationship between examinees' response time and test performance depends on the examinees' ability.

4.2. Implications

This study suggests that response times of different items are related to test performance in various ways, which presents some implications for the design and administration of educational assessments. In modern test administration, response time data have been integrated into adaptive testing systems to create a more personalized testing experience. For instance, computerized adaptive tests (CATs) can adjust the difficulty level of subsequent questions based on the examinee's response time in order to maintain their testing motivation and engagement (e.g., Choe et al., 2018). Typically, the item selection algorithm of CAT is based on IRT models (e.g., van der Linden, 2007). The finding of this study suggests that when applying IRT models and CATs, it is imperative to acknowledge that the relationship between response time and test performance varies across different items. Consequently, test developers should avoid presuming that longer response times always indicate examinees' difficulties or struggles in solving a question.

Response time, as a type of process data, has emerged as a valuable source of information for making psychological and educational inferences about examinees' response processes and solution strategies. Incorporating response time analysis can enhance the validity of test results (Molenaar, 2015). By recognizing and analyzing the within-examinee variability in response time, researchers can gain valuable insights into individual performance and engagement

throughout the test. For example, the identification of differential response time patterns between high-achieving and low-achieving examinees offers significant insights into their abilities and engagement. High-achievers tend to allocate less time to initial items but more time to later ones, suggesting the use of automated processes and better time-management skills. On the other hand, low-achievers devote more time to early items but less time to later items, potentially indicating cognitive resource depletion and guessing behaviors. These differential response time patterns serve as informative indicators of examinees' test-taking behaviors and underlying cognitive processes. By utilizing response time as a valuable indicator of examinee behavior, researchers can make inferences about examinees' problem-solving strategies, cognitive load, and engagement during test-taking. Therefore, the inclusion of response time analysis in educational assessments provides a valuable avenue for improving the validity of test outcomes.

Lastly, the findings of this study emphasize the value of utilizing separate response times of items to predict overall test performance because separate response times provide more granular information, enhancing the accuracy of predictions. Therefore, learning and assessment systems should collect more fine-grained response time data, such as item-level response time. This would allow for a more comprehensive understanding of the relationship between response time, item characteristics, and examinee performance. The use of these more fine-grained response time data will also guide the design and development of learning systems (e.g., Pelánek & Effenberger, 2020; Tseng et al., 2008), providing a more personalized learning and assessment experience for users.

4.3. Limitations

Several limitations should be considered before generalizing the findings of this study. The primary limitation is due to the TIMSS PSI dataset contains a large number of missing

response times, particularly for the later items and screens (Mullis et al., 2021). In order to conduct criterion-related profile analysis, missing data were omitted in this study; however, this decision may affect the generalizability of the results. The missing data could be associated with examinees struggling to meet the time constraints of the assessment. For example, some examinees might have spent excessive time on early items, resulting in insufficient time to answer the remaining questions within the allocated time frame. Furthermore, as the TIMSS PSI tasks are considered low-stakes assessments, it is possible that examinees did not approach the assessment with utmost seriousness, leading to a higher occurrence of missing responses. These missing data may indicate examinee disengagement due to fatigue or a recognition that the assessments have minimal impact on their lives. Future research may examine the distribution of missing data across achievement groups to evaluate its impact and gain a better understanding of the reasons behind its occurrence.

Another limitation is that the TIMSS PSI dataset only provides screen response time due to the assessment design. Consequently, all analyses in this study were conducted based on screen response time, while conclusions were drawn regarding item response time. Although screen response time was considered as a proxy for item response time, this approach may not accurately reflect the true relationship between response time and item difficulty. For example, this study examined the relationship between screen response time and the average difficulty of items on the screen. However, this averaged difficulty of items on the same screen may not truly represent the difficulty of individual items, thus their correlation with response time may differ. Future research should try to collect item-level response time in order to address this limitation.

Finally, it is important to note that in TIMSS PSI tasks, examinees answered the questions in a fixed order. This study investigated the influence of item position on examinees'

response time and the relationship between response time and test performance, revealing the significant influence of item position. Counterbalancing the item order could have enhanced the validity of the results. However, due to the use of secondary data, control over item sequencing was not possible. Therefore, it should be cautious when interpreting the effect of item position. For example, although this study found that examinees tended to spend less time on later items and attributed this effect to examinees' fatigue and exhaustion of testing motivation and cognitive resources, the reason might be that those items may be easier than earlier items. To address this limitation, future research may consider counterbalancing the presentation order of questions or randomly presenting items to examinees.

4.4. Conclusions

The present study investigated the relationship between examinees' response time and test performance using the TIMSS 2019 PSI data. The study revealed several key findings. Firstly, it highlighted the variability in standardized response time among examinees, indicating that examinees allocated time differently across items compared to others. Secondly, the study demonstrated that the relationship between response time and the overall test performance varied in terms of the strength and direction by items. While increased time on certain items correlated with higher total scores, the increased time on other items was linked to lower total scores. Furthermore, the study used criterion-related profile analysis to show that considering within-examinee variability in response time improved the accuracy of predicting test scores compared to relying solely on average or total response time. Therefore, to enhance prediction accuracy in test performance, it is beneficial to use the separate response time of individual items, as the predictive power of response time mostly comes from the within-examinee variability in response time. Moreover, the examination of different achievement groups suggests that the

predictive value and criterion-related patterns of response time are distinct across achievement groups. For example, response times show a stronger positive relationship with test performance in the low achievement group, while in the advanced achievement group, response times exhibit a more negative relationship with test performance. Finally, the study demonstrated that criterion-related patterns of response time were influenced by item position and item difficulty. It was found that item difficulty and position had positive relationships with the criterion-related patterns of response time, suggesting their impact on the relationship between response time and test performance. In conclusion, the study findings contribute to the understanding of the complex relationship between response time and test performance. The findings have practical implications for the design and administration of future educational assessments.

5. References

Becker, N., Schmitz, F., Göritz, A. S., & Spinath, F. M. (2016). Sometimes more is better, and sometimes less is better: Task complexity moderates the response time accuracy correlation. *Journal of Intelligence, 4*(3), 11. https://doi.org/10.3390/jintelligence4030011

Blommers, P., & Lindquist, E. F. (1944). Rate of comprehension of reading; its measurement and its relation to comprehension. *Journal of Educational Psychology, 35*(8), 449–473. https://doi.org/10.1037/h0054306

Bolsinova, M., de Boeck, P. & Tijmstra, J. (2017). Modelling Conditional Dependence Between Response Time and Accuracy. *Psychometrika 82*, 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bulut, O., Desjardins, C. D., & Desjardins, M. C. D. (2018). Package 'profileR'.

Castro, M. J., López, M., Cao, M. J., Fernández-Castro, M., García, S., Frutos, M., & Jiménez, J. M. (2019). Impact of educational games on academic outcomes of students in the Degree in Nursing. *PloS one, 14*(7), e0220388. https://doi.org/10.1371/journal.pone.0220388

Chen, H., De Boeck, P., Grady, M., Yang, C. L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence, 69*, 16-23. https://doi.org/10.1016/j.intell.2018.04.001

Choe, E. M., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 43*(2), 135-158.

Davison, M. L., & Davenport, E. C., Jr. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods, 7*(4), 468–484. https://doi.org/10.1037/1082-989X.7.4.468

Davison, M. L., Davenport Jr, E. C., Chang, Y. F., Vue, K., & Su, S. (2015). Criterion-related validity: Assessing the value of subscores. *Journal of Educational Measurement, 52*(3), 263-279. https://doi.org/10.1111/jedm.12081

Deribo, T., Goldhammer, F., & Kroehne, U. (2023). Changes in the Speed–Ability Relation Through Different Treatments of Rapid Guessing. *Educational and Psychological Measurement, 83*(3), 473-494. https://doi.org/10.1177/00131644221109490

Fishbein, B., & Foy, P. (2021). Scaling the TIMSS 2019 problem solving and inquiry data. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 17.1–17.51). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-17.html

Giordano, A. L., Prosek, E. A., Schmit, E. L., & Schmit, M. K. (2023). Examining coping and nonsuicidal self-injury among adolescents: A profile analysis. *Journal of Counseling & Development, 101*(2), 214-223. https://doi.org/10.1002/jcad.12459

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716

Hancock, C. R. (1994). Alternative Assessment and Second Language Study: What and Why? *ERIC Digest*.

Kobrin, J. L. (2000). An investigation of the cognitive equivalence of computerized and paper-and-pencil reading comprehension test items. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Krämer, R. J., Koch, M., Levacher, J., & Schmitz, F. (2023). Testing Replicability and Generalizability of the Time on Task Effect. *Journal of Intelligence, 11*(5), 82. https://doi.org/10.3390/jintelligence11050082

Martín, E. S., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183-203. https://doi.org/10.1177/0146621605282773

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. International Association for the Evaluation of Educational Achievement.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 187-205. https://doi.org/10.1080/15305058.2019.1706529

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences, 53*, 1-16. https://doi.org/10.1016/j.lindif.2016.10.002

Pelánek, R., & Effenberger, T. (2020). Beyond binary correctness: Classification of students' answers in learning systems. *User Model User-Adap Inter, 30*, 867–893. https://doi.org/10.1007/s11257-020-09265-5

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137*, 115-136. https://doi.org/10.1016/j.cognition.2014.12.004

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Schnipke, D. L. (1995). Assessing speededness in computer–based tests using item response times. *Dissertation Abstracts International, 57*(1), 759B. (University Microfilms No. 9617600).

Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Kreuger, S., & Norman, G. R. (2012). The relationship between response time and diagnostic accuracy. *Academic Medicine, 87*(6), 785-791. https://doi.org/10.1097/ACM.0b013e318253acbd

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Stanton, H. C., & Reynolds, C. R. (2000). Configural frequency analysis as a method of determining Wechsler profile types. *School Psychology Quarterly, 15*(4), 434–448. https://doi.org/10.1037/h0088799

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, 251-296. https://www.jstor.org/stable/23359412

Thomas, M. H. (2006). Modeling differential pacing trajectories in high-stakes computer adaptive testing using hierarchical linear modeling and structural equation modeling. PhD dissertation. University of North Carolina.

Tseng, J. C., Chu, H. C., Hwang, G. J., & Tsai, C. C. (2008). Development of an adaptive learning system with two sources of personalization information. *Computers & Education, 51*(2), 776-786. https://doi.org/10.1016/j.compedu.2007.08.002

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287.

Van Der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*(3), 247-272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Vida, L. J., Bolsinova, M., & Brinkhuis, M. J. (2021). Speeding up without Loss of Accuracy: Item Position Effects on Performance in University Exams. *International Educational Data Mining Society*.

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item Response Theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), Methods and Procedures: TIMSS 2019 Technical Report (pp. 11.1-11.25). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-11.html

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95-114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61. https://doi.org/10.1111/emip.12165

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, 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., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 139–153). American Psychological Association. https://doi.org/10.1037/12330-009

Yang, C. L., O Neill, T. R., & Kramer, G. A. (2002). Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement, 3*(3), 282-299. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/12147914/